Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Desambiguação Lexical Automática de Sentido: Um Panorama

Lucia Specia
Maria das Graças Volpe Nunes

NILC-TR-04-08

Agosto, 2004

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Neste relatório são discutidas diversas questões relacionadas à criação de um modelo para a tarefa de desambiguação lexical de sentido, incluindo a definição do conjunto de palavras a desambiguar, do repositório de sentidos, dos tipos de conhecimento a serem utilizados, do método de desambiguação, entre outras. Além disso, é apresentado um levantamento bibliográfico sobre os trabalhos de desambiguação lexical de sentido (monolingües ou multilingües) que vêm sendo desenvolvidos desde o surgimento dessa área. O estudo das questões relevantes e dos trabalhos existentes nessa área têm por objetivo subsidiar a proposta, em um trabalho posterior, de um modelo de desambiguação lexical de sentido para a tradução automática do inglês para o português do Brasil.

Índice

Figuras

Tabelas

1 Introdução

O desenvolvimento de sistemas de Processamento das Línguas Naturais (PLN) vem apresentando muitos avanços nos últimos anos, nas suas diversas áreas de aplicação. Contudo, há ainda muitos problemas não solucionados, tanto na interpretação quanto na geração das línguas. Grande parte desses problemas está relacionada à ambigüidade inerente às línguas naturais, nos seus diversos níveis, como o morfológico, lexical, sintático, semântico, contextual e pragmático.

Este trabalho focaliza o problema da ambigüidade lexical, ou seja, da necessidade de escolha por um dos possíveis sentidos (ou significados) de uma palavra quando da sua interpretação. Esse problema é comum em muitas aplicações, como a Recuperação de Informações, a Tradução Automática (TA), a Extração de Informações, a Análise de Conteúdo, etc. Na TA e em outras aplicações multilingües, em particular, os "sentidos" de uma palavra ambígua em um língua-fonte (LF) correspondem à sua tradução na língua-alvo (LA).

A ambigüidade lexical é causada, fundamentalmente, pela existência de algumas relações semânticas interlexicais, principalmente a polissemia e a homonímia. De acordo com a classificação adotada neste relatório (Lyons, 1977), na **polissemia** uma mesma palavra tem dois ou mais significados diferentes, mas relacionados entre si, sendo que, normalmente, somente um dos significados se ajusta a um determinado contexto. Na **homonímia** duas ou mais palavras com significados totalmente distintos, sem traços comuns, são idênticas quanto ao som (homofonia) e/ou à grafia (homografia)[1].

O problema da ambigüidade lexical pode, ainda, ser classificado como ambigüidade categorial ou ambigüidade de sentido (Ullmann, 1964). A **ambigüidade categorial** ocorre quando as duas ou mais opções de significados de uma dada palavra são de diferentes categorias gramaticais. Na tradução, um exemplo de ambigüidade categorial causada pela relação de homonímia é a palavra do inglês *field*, que pode ser traduzida para as palavras "campo" (substantivo) ou "interceptar" (verbo), no português. Já um exemplo de ambigüidade categorial derivada da relação de polissemia é a palavra do inglês *eats*, que pode ser traduzida no português como "mantimentos, víveres, gêneros alimentícios" (substantivos) ou "come" (verbo "comer" conjugado na terceira pessoa singular, presente do indicativo). A **ambigüidade de sentido**, por sua vez, ocorre quando as duas ou mais opções de sentido (ou tradução) de uma dada palavra têm a mesma categoria gramatical. Alguns exemplos são a palavra *know*, que pode ser traduzida como "saber" ou "conhecer", como um caso de polissemia, e a palavra *light*, que pode ser traduzida como "leve" ou "luz", como um caso de homonímia.

A ambigüidade categorial é, em geral, muito mais simples que a de sentido, uma vez que pode ser resolvida, na maioria das vezes, pela análise das características sintáticas das palavras, realizada por procedimentos de etiquetação gramatical ou análise sintática, por exemplo. Procedimentos dessa natureza alcançam, atualmente, resultados bastante satisfatórios. A resolução da ambigüidade de sentido, por sua vez, exige a análise da semântica das palavras e, eventualmente, a análise do uso de tais palavras (realizadas por procedimentos de análise semântica e pragmática, por exemplo). Portanto, o foco da maioria dos trabalhos voltados para o tratamento da ambigüidade lexical está no problema da ambigüidade de sentido, considerando tanto a relação de polissemia quanto a de homonímia. A área que se ocupa do tratamento desse problema é denominada **Desambiguação do Sentido das Palavras** (DLS), do inglês *Word Sense Disambiguation*.

[1] Como este trabalho aborda somente o processamento da língua escrita, apenas a homografia é considerada.

Para realizar a desambiguação de maneira automática, é necessário incorporar um módulo de DLS aos processos de interpretação e/ou geração da língua. Para a construção de um módulo de DLS, várias questões devem ser analisadas, as quais dão origem a muitas decisões de projeto. Essas questões incluem o conjunto de palavras a desambiguar, quais os possíveis sentidos dessas palavras, o método adotado para a desambiguação, como o módulo será avaliado, etc. Os trabalhos de DLS existentes apresentam, de fato, muitas variações, de acordo com as decisões tomadas com relação a essas e a muitas outras questões.

Este relatório procura justamente elencar e discutir essas questões, bem como apresentar os principais trabalhos de DLS já propostos desde o surgimento dessa área, os quais foram desenvolvidos de acordo com diferentes decisões com relação a tais questões. O levantamento apresentado é parte de um trabalho mais amplo, que visa à proposta de um modelo lingüístico-computacional de DLS a ser utilizado em um sistema de TA do inglês para o português do Brasil.

As questões consideradas relevantes para a proposta desse modelo são discutidas na Seção 2. Os trabalhos já desenvolvidos para a tarefa de DLS, em contextos monolingües e multilingües, são descritos na Seção 3. Algumas conclusões sobre o levantamento realizado para a proposta do modelo de DLS são apresentadas na Seção 4.

2 A tarefa de desambiguação lexical de sentido

Em termos gerais, a tarefa de DLS consiste em associar a uma palavra ambígua em uma sentença ou texto um sentido que é distinguível dos outros sentidos potencialmente atribuíveis a tal palavra, de acordo com o contexto dessa palavra. Para a criação de modelos para essa tarefa, são considerados os seguintes passos gerais (Ide & Véronis, 1998):

1) Determinação do conjunto de palavras a desambiguar: todas as palavras, todas as palavras de determinada(s) classe(s) gramatical(is), um subconjunto específico de palavras, etc.
2) Definição de todos os possíveis sentidos de cada palavra. No caso da DLS monolingüe, os sentidos são os possíveis significados da palavra, no caso da TA, as suas possíveis traduções.
3) Criação de um mecanismo para atribuir a cada ocorrência da palavra o sentido mais apropriado, incluindo decisões sobre a metodologia do mecanismo, os tipos de conhecimentos que serão utilizados e as suas fontes, etc., e a criação ou adaptação dos recursos necessários.
4) Avaliação desse mecanismo, independentemente da aplicação e/ou no contexto de alguma aplicação especifica, como a TA ou a Recuperação de Informações.

As decisões em cada uma dessas etapas permitem uma série de variações, que implicam diferentes propostas de DLS. Nesta seção, são abordadas diversas questões relacionadas a tais decisões, incluindo a definição do conjunto de palavras a desambiguar, a noção de "sentido" nos diferentes trabalhos, as metodologias que podem ser seguidas, os vários tipos de conhecimento idealmente e comumente empregados, os procedimentos usados para a criação de córpus de exemplos em trabalhos baseados em córpus, os processos de avaliação intrínseca e extrínseca da DLS, a inserção do módulo de DLS em sistemas de TA desenvolvidos sob diferentes métodos e alguns outros conceitos gerais e esclarecimentos sobre a terminologia importantes para a área de DLS.

Vale notar que os conceitos e questões discutidos são válidos tanto para a DLS monolingüe quanto para a DLS multilingüe. Os casos de especificidades entre as duas tarefas serão destacados. A maioria dos exemplos é apresentada na língua inglesa para que eles possam, sempre que possível, ser contextualizados no problema da TA dessa língua para o português.

2.1 Desambiguação, etiquetação ou discriminação de sentidos

A tarefa de DLS também é denominada por alguns autores de **etiquetação semântica** (*semantic tagging*) ou de **etiquetação de sentidos** (*sense tagging*), em uma analogia à tarefa de etiquetação morfossintática (*part-of-speech tagging*), ou, ainda, de **desambiguação semântica** (*semantic disambiguation*). Contudo, como esclarecem Wilks & Stevenson (1997a; 1997b), apesar de semelhantes, a DLS, a desambiguação semântica, a etiquetação semântica e a etiquetação de sentidos são tarefas distintas.

A etiquetação semântica consiste em anotar cada palavra de um texto com uma etiqueta que identifica sua categoria semântica. Essas categorias são relativamente genéricas, podendo ser constituídas, por exemplo, de etiquetas que denominam a área ou domínio de determinada palavra em um contexto ou de outras etiquetas ontológicas genéricas, como "humano", "animado", etc. Esse é o caso do trabalho de Segond et al. (1997), que utiliza as 45 etiquetas de classes semânticas da WordNet (Miller et al., 1990) em seu trabalho.

A etiquetação de sentidos é uma instância do problema de etiquetação semântica, no qual as etiquetas com que todas as palavras são anotadas são mais refinadas, correspondendo aos diferentes sentidos dessa palavra, de acordo com o conjunto de sentidos de um dicionário ou outro recurso lexical.

As tarefas de desambiguação, por sua vez, segundo os autores, constituem uma variação do problema de etiquetação semântica ou de sentidos, na qual não necessariamente todas as palavras de conteúdo precisam ser etiquetadas. Normalmente, os trabalhos de desambiguação são restritos a um subconjunto de palavras, por exemplo, as palavras consideradas mais ambíguas ou mais relevantes para a aplicação em questão.

Outra distinção terminológica importante deve ser feita entre os termos desambiguação e discriminação de sentidos. A discriminação de sentidos, diferentemente da desambiguação, consiste em identificar grupos de sentidos distintos e classificar as ocorrências da palavra como pertencentes a algum dos grupos. Não envolve rotular os sentidos ou associá-los a distinções de sentidos de alguma fonte de conhecimento externa, como um dicionário. Esse tipo de discriminação é geralmente realizado por trabalhos monolingües de DLS que agrupam ocorrências da palavra de acordo com sua similaridade. As distinções entre os grupos de sentidos gerados não correspondem, normalmente, às distinções de sentido tradicionais

Apesar dessas distinções, em geral, o termo “desambiguação” é utilizado de maneira genérica, para denominar qualquer uma das tarefas.

2.2 Desambiguação de todas as palavras ou de um subconjunto

A determinação do conjunto de palavras a serem desambiguadas, ou seja, a escolha entre a tarefa de etiquetação ou desambiguação, propriamente, depende essencialmente da aplicação do mecanismo de DLS. Em geral, consideram-se somente as palavras de conteúdo (palavras plenas). Alguns trabalhos monolingües independentes de aplicação realizam a desambiguação (ou etiquetação) dos sentidos de todas as palavras de conteúdo da língua. Esses trabalhos são válidos para aplicações como a categorização de textos de acordo com a sua área, por exemplo, nas quais a identificação do sentido de todas as palavras pode auxiliar na identificação dessa área.

Em outras aplicações, principalmente nas multilingües, no entanto, não é necessário identificar os sentidos de todas as palavras, mas só daquelas cuja ambigüidade implica algum problema. Esse é o caso da TA, uma vez que a tradução de determinadas palavras pode ser facilmente identificada por outros mecanismos do sistema, com base em informações de outra natureza, independentemente do mecanismo de DLS. Na TA, um mecanismo de DLS que atue na etiquetação de todas as palavras seria equivalente a um sistema completo de TA bastante simples, de tradução palavra a palavra.

Normalmente, nos trabalhos voltados para a desambiguação de todas as palavras, o mesmo modelo (ou tipo de regra) geral é utilizado, independentemente da palavra a desambiguar. Já nos trabalhos de desambiguação de um subconjunto de palavras, normalmente são criados modelos específicos para cada palavra, que podem ter características e estrutura totalmente distintas dos modelos para outra palavra.

Em geral, os trabalhos de etiquetação de todas as palavras são mais robustos e abrangentes, enquanto que os trabalhos de desambiguação de subconjuntos de palavras são mais precisos. Em exercícios de avaliação conjunta, como será descrito na Seção 2.8, há tarefas específicas para a avaliação de abordagens de ambos os tipos.

2.3 Desambiguação independente ou em cascata

Conforme mencionado, o processo de DLS considera o contexto da palavra ambígua para a desambiguação. Nos trabalhos em que apenas algumas palavras são desambiguadas, normalmente, há apenas uma ocorrência da palavra ambígua por contexto, por exemplo, em uma sentença. Assim, a desambiguação dessa palavra é realizada independentemente de desambiguações de quaisquer outras palavras e também as desambiguações anteriores dessa mesma palavra. Uma exceção é a proposta de Gale et al. (1992c), descrita na Seção 2.10, que considera que, em um discurso, uma palavra ambígua é sempre utilizada com o mesmo sentido. Neste caso, após a desambiguação da primeira ocorrência da palavra no discurso, às demais ocorrências seria atribuído o mesmo sentido.

Nos trabalhos que contemplam a desambiguação de todas as palavras de conteúdo ou, pelo menos, de várias palavras que podem ocorrer em um mesmo contexto, por outro lado, pode-se optar por: (a) desambiguar cada palavra independentemente das desambiguações eventualmente já realizadas para as outras palavras no mesmo contexto; ou (b) considerar as desambiguações já realizadas, em um processo "em cascata", em que a desambiguação de uma palavra depende não somente das informações sobre as palavras na sentença, mas também de informações sobre as escolhas de sentidos feitas para outras palavras.

Na desambiguação em cascata a desambiguação correta de uma palavra certamente pode auxiliar na desambiguação de outras palavras. Contudo, a desambiguação incorreta pode propagar os erros para várias outras palavras. Por essa razão, em geral, esse processo não é adotado pelos trabalhos de DLS.

Nos trabalhos multilingües, em especial, essa distinção representa uma opção importante: a desambiguação em cascata, considerando as palavras já traduzidas, torna o processo mais dependente da língua-alvo, já que podem ser consideradas características mais fortemente relacionadas à língua-alvo do que à língua-fonte. Por exemplo, podem ser consideradas co-ocorrências de palavras já traduzidas como combinações preferenciais. Uma observação importante é que podem ser consideradas palavras já traduzidas com ou sem o mecanismo de DLS. Assim, pode-se restringir a analise às traduções de palavras não ambíguas, realizadas independentemente do módulo de DLS. Contudo, mesmo essas traduções podem representar erros do sistema, decorrentes de outras escolhas, as quais podem, novamente, se propagar para outras palavras.

Em geral, a desambiguação em cascata também não é utilizada nos trabalhos de DLS para a TA. Uma exceção é o módulo de DLS de Dihn et al. (2003) (Seção 3.2.2). Há autores que consideram a desambiguação independente, mesmo para a TA, como desambiguação monolingüe, já que se baseia apenas em informações da língua-fonte. No entanto, neste trabalho, a desambiguação é considerada multilingüe, pois são analisadas apenas as ambigüidades que se manifestam na língua-alvo. Como mencionado por Lee (1997), em geral, não há relação direta entre o número de sentidos monolingües de uma palavra e o número de possíveis traduções para outra língua.

2.4 Os sentidos

Segundo Ide & Véronis (1998), a definição precisa do que é "sentido" é uma questão de considerável debate, sobre a qual não existe consenso. Uma definição normalmente aceita, e adotada neste projeto, é simplesmente aquela que descarta do "sentido" os problemas de ambigüidade categorial, concentrando-se na resolução da ambigüidade lexical de palavras da mesma categoria sintática. Contudo, mesmo seguindo essa definição, os diferentes trabalhos de DLS utilizam o termo "sentido" com diversas acepções, variando,

principalmente, no conjunto de sentidos usados e no nível de refinamento das distinções que caracterizam diferentes sentidos.

2.4.1 O repositório de sentidos

A definição do repositório de sentidos é um problema complexo, pois há diferentes visões sobre as variações que podem constituir os sentidos de uma palavra. Para contorná-lo, grande parte dos trabalhos de DLS utiliza os sentidos já estabelecidos em recursos lexicais, como dicionários. Nesses recursos, por exemplo, cada variação nas definições da palavra, normalmente associada a um código distinto, corresponderia a um sentido. Além disso, os sentidos poderiam ser hierarquizados de acordo com diferentes níveis de distinção. Assim, nos trabalhos monolingües, o sentido corresponderia a um código associado a cada uma das definições. Já nos trabalhos multilingües, poderiam ser utilizadas as diferentes traduções da palavra, diretamente, como sentidos.

Kilgarriff (1997a; 1997b), todavia, argumenta que não existe fundamentação teórica suficiente sobre o conceito de "sentido de palavra" nesses recursos. Ele afirma que não é possível definir claramente os sentidos das palavras, independentemente de aplicação, com base em quaisquer recursos lexicais. Segundo o autor, o uso de sentidos de dicionários não é adequado porque os sentidos só existem relativos a uma determinada tarefa. "Não á razão para esperar que um único conjunto de sentidos seja apropriado para diferentes aplicações de PLN" (Kilgarriff, 1997a, p. 2).

Ainda de acordo com Kilgarriff, a lista de significados especificados para cada palavra nos dicionários normalmente não é resultado da análise de como o significado das palavras opera, mas sim uma resposta a uma série de restrições impostas para a criação de tais dicionários, como o formato de impressão, a compactação, a simplificação e padronização do método de acesso, etc.

O autor acredita que tais sentidos devem ser abstrações criadas especificamente para cada tarefa de DLS, a partir de *clusters* de citações de córpus. A linguagem, segundo ele, funciona em seus usos, não cabendo, portanto, indagar sobre os significados das palavras, mas sobre suas funções práticas. Kilgarriff sugere dois critérios para determinar o conjunto de sentidos de uma palavra a partir da análise de córpus: freqüência e imprevisibilidade. Segundo o autor, deve-se considerar um sentido como um dos possíveis sentidos da palavra se sua freqüência no córpus for alta (*sufficiently frequent*) e se tal sentido não puder ser previsto ou derivado a partir do sentido básico (*unsufficiently predictable*).

Assim, para o autor, o conjunto de sentidos das palavras para uma língua é dependente da tarefa e é o córpus que dita tais sentidos das palavras, não os dicionários. Contudo, essa não é a visão da maioria dos pesquisadores em DLS, que procuram desenvolver soluções independentes de aplicação, utilizando recursos lexicais genéricos para a distinção dos sentidos.

Partindo da premissa definida por Gale et al. (1992a), de que os diferentes sentidos de uma palavra são traduzidos diferentemente para outra língua, Resnik & Yarowsky (1997a) sugerem que, para a DLS monolingüe, os diferentes sentidos de uma palavra sejam determinados considerando-se somente as distinções que são lexicalizadas em outras línguas. Eles propõem que um conjunto de línguas-alvo seja identificado e que as distinções de sentido aceitas sejam aquelas que são realizadas lexicalmente em um subconjunto mínimo dessas línguas-alvo. O conjunto de sentidos utilizado como base é o da WordNet. São utilizados dicionários bilíngües para identificar as possíveis traduções. O uso desse critério para a DLS na língua-fonte pode ser problemático porque algumas

ambigüidades são mantidas na língua-alvo. Contudo, a análise de diversas línguas-alvo pode minimizar esse problema.

Reforçando as declarações de Kilgarriff, alguns experimentos realizados por Véronis (1998) mostram que, usando os repositórios de sentidos com as distinções refinadas fornecidas por dicionários, seres humanos têm grandes dificuldades de escolha de um sentido na etiquetação de sentidos. Essa dificuldade é medida em função do nível de discordância entre diversos juízes humanos, especialistas, na tarefa de DLS. A concordância entre os juízes, nesses experimentos, não supera a concordância ao acaso para várias palavras. Isso implica, segundo o autor, que programas computacionais dificilmente conseguirão realizar uma desambiguação precisa utilizando esses sentidos. Assim, o autor também argumenta que, de modo geral, os sentidos de dicionários, *thesauri* ou, mesmo, da WordNet não são adequados para a tarefa de DLS.

Uma das principais razões para que os dicionários e outros recursos lexicais não sejam adequados para a DLS, segundo Véronis, é a falta de critérios distribucionais nas entradas dos dicionários em geral, incluindo a falta de informações mais úteis para a DLS. Os recursos preocupam-se apenas com a definição do significado das entradas, e não com informações mais superficiais (informações sintáticas, exemplos extraídos de textos reais, etc.) que são necessárias para identificar a correspondência dos sentidos dessas entradas com usos da palavra em textos reais.

O autor analisa, também, em seus experimentos, como o ser humano realiza a desambiguação e afirma que isso raramente é levado em consideração nos trabalhos de DLS automática. Diferentemente do que argumentam outros trabalhos que também defendem a inadequação do repositório de sentidos de recursos lexicais tradicionais para a DLS, o problema não reside apenas no alto nível de refinamento dos sentidos desses recursos. Segundo Véronis, os experimentos mostram que a maioria das discordâncias entre os anotadores ocorre em níveis mais abstratos de divisão das entradas. O que o autor sugere, então, é que são necessários recursos lexicais com outra estrutura e outras características para a tarefa de DLS, incluindo um critério distribucional mais bem definido e informações mais relevantes para a DLS, além da definição de cada entrada.

Com relação especificamente às distinções de sentido da WordNet, Palmer (1998) também argumenta que o nível de refinamento não é o único problema dessas distinções. Segundo a autora, em grande parte dos casos, as distinções nos sistemas de DLS são baseadas em restrições de seleção nos argumentos da palavra que podem ser codificadas e manipuladas em um léxico computacional. O problema é que, para tanto, essas distinções precisariam ser feitas com base em critérios concretos explícitos, e não com base em critérios abstratos envolvendo conhecimento de mundo, como normalmente é feito na WordNet.

Em outro trabalho (Palmer, 2000), a autora afirma que uma das áreas mais controversas na criação de dicionários e léxicos computacionais é a definição do que constitui uma separação clara em sentidos para uma palavra e como esses sentidos podem ser computacionalmente caracterizados e distinguidos. Certamente, essa definição se reflete nos sistemas de DLS que utilizam os dicionários ou léxicos computacionais como base para a distinção dos sentidos. Segundo Kilgarriff (1997b), é provável que léxicos computacionais desenvolvidos manualmente com base em córpus, e não extraídos a partir de dicionários tradicionais, sejam cada vez mais voltados para aplicações especificas. Assim, esses léxicos tendem a ser mais adequados como repositório de sentidos para a DLS que os dicionário tradicionais.

2.4.2 O nível de refinamento das distinções entre os sentidos

Uma das questões mais problemáticas na definição do repositório de sentidos é o nível de granularidade das distinções entre os sentidos, ou seja, o seu refinamento. Divisões com alto nível de granularidade, como aquelas encontradas em alguns dicionários, podem dificultar a DLS, pois introduzem efeitos combinatoriais significativos, requerem escolhas entre sentidos muito próximos e, conseqüentemente, exigem muito conhecimentos para a desambiguação. Os sentidos da WordNet, por exemplo, apesar de amplamente usados em muitos trabalhos de DLS (Seção 3.1.2), são considerados por diversos autores como muito refinados para a tarefa de desambiguação automática (Palmer, 1998; 2000; Ng, 1997b, etc.).

Por essa razão, alguns autores propõem divisões com baixo nível de granularidade, seja pela combinação de sentidos de distinção muito refinada em sentidos mais genéricos, seja pela utilização apenas das divisões de sentidos mais gerais de recursos com uma divisão hierárquica. É importante ressaltar que a decisão do nível de granularidade depende, novamente, da aplicação na qual a desambiguação de sentido será utilizada. A TA, por exemplo, exige sentidos mais refinados que a Recuperação de Informações.

A definição do repositório de sentidos tem implicação direta sobre o nível de refinamento para as distinções de sentidos: um dicionário pouco refinado agrupará vários sentidos similares em uma mesma entrada, com um único código, enquanto um dicionário mais refinado adotará entradas diferentes para cada variação de sentido. Entretanto, caso haja diferentes níveis de distinção nesses recursos, é possível selecionar apenas níveis menos refinados. Dependendo do recurso empregado, esses níveis são equivalentes às noções de homonímia e polissemia definidas anteriormente. Contudo, isso nem sempre é verdadeiro, pois os critérios utilizados pelos para a definição dos níveis de distinção na criação de tal recurso podem ser diversos dos usados neste trabalho para a distinção entre homonímia e polissemia.

Em vários trabalhos, Wilks & Stevenson (Wilks & Stevenson, 1998a; 1998b; Stevenson & Wilks, 2000; 2001), por exemplo, não usam os termos homonímia e polissemia como duas variantes da homografia. Em vez disso, eles diferenciam os dois fenômenos, a partir das distinções de um dicionário, mas por meio dos termos “homografia” e “polissemia”. Para os autores, os dicionários apresentam os vários significados de cada entrada em diversos homógrafos, que são conjuntos de sentidos com significado relacionado. Assim, entre homógrafos, os significados são totalmente distintos e o fenômeno da polissemia ocorre dentro de cada grupo homógrafo. Os diferentes homógrafos podem ser de categorias gramaticais distintas e cada grupo homógrafo pode apresentar também sentidos de classes gramaticais distintas. Um exemplo parcial, retirado de um dicionário para a palavra *bank* é ilustrado na Figura 1.

bank n
1 land along the side of a river, lake, etc.
2 earth which is heaped up in a fiel dor a garden, often making a border or division
3 a mass of snow, mud, clouds, etc.
4 a slope made at bends in a road ro race-track, so that they are safer for cars to go round
5 sandbank

bank v
1 (of a car or aircraft) to move with one side higher than the other, esp. when making a turn

bank n
1 a place where money is kept and paid out on demand, and where related activities go on
2 a place where something is held ready for use, ep. Organic product of human origin for medical use
3 (a person who keeps) a supply of money or pieces for payment or use in a game of chance
4 *break the bank* to win all the money of the bank in a game of chance

Figura 1. Exemplo de entrada em um dicionário (Wilks, & Stevenson, 1996, p.5)

Um exemplo de distinção em nível de homografia considerada pelos autores no caso de *bank* seria a distinção entre o sentidos de banco como "instituição financeira" (terceiro grupo de substantivos na Figura 1) e como "margem de um rio" (primeiro grupo de substantivos na Figura 1). Pode-se considerar, portanto, que o conceito de homógrafo desses autores corresponde, em princípio, ao conceito de homonímia (categorial e de sentido) deste trabalho. Ravin & Leacock (2000) também consideram a diferenciação entre polissemia e homonímia, tal como é feita neste trabalho, mas se referem, de maneira indistinta aos termos homonímia e homografia. Essa mesma distinção entre homografia e polissemia é também feita por Bräscher (2002), que afirma que a homografia ocorre por meio da colisão acidental entre as formas de dois signos lingüísticos distintos, enquanto a polissemia ocorre quando um único signo lingüístico envolve significados distintos.

Nas diversas publicações de Wilks & Stevenson, os resultados são sempre reportados em nível de homografia, considerada uma distinção pouco refinada, e em nível de sentidos, considerada uma distinção muito refinada. Outros trabalhos também relatam seus resultados considerando dois possíveis níveis de distinção. Em geral, esses trabalhos utilizam as distinções fornecidas por algum recurso lexical, como um dicionário, assim, não há a preocupação em identificar, automaticamente, se a ambigüidade constitui um caso de polissemia ou de homonímia. A desambiguação é realizada com base nas distinções fornecidas pelos recursos, ou por algum dos níveis de distinção desses recursos.

Contudo, a distinção de níveis baseada nas entradas de um dicionário nem sempre será possível, pois nem todos os recursos lexicais são organizados dessa forma: grupos de itens polissêmicos formam um homônimo. A WordNet, por exemplo, conforme será descrito na Seção 2.6.1.2, é organizada de maneira totalmente diferente.

Há também os trabalhos que utilizam córpus para a definição do seu repositório de sentidos, de modo que podem ocorrer tanto casos de homonímia quanto de polissemia. Ainda assim, nas avaliações, procura-se analisar se as distinções realizadas são mais ou menos refinadas.

Em geral, nas avaliações são diferenciados os trabalhos de DLS de resolução da ambigüidade em um nível pouco refinado (*homograph level* ou *coarse sense distinction*), dos trabalhos de resolução em um nível bastante refinado (*sense level*). Contudo, é importante mencionar que os diferentes níveis de refinamento nem sempre correspondem às distinções em nível de homonímia ou polissemia.

Certamente, quanto maior o nível re refinamento das distinções, maior a dificuldade na desambiguação. Por exemplo, a distinção entre os diversos sentidos de homônimos é mais simples que a distinção entre os sentidos de um item polissêmico, já que os contextos de ocorrência de itens homônimos são, naturalmente, muito distintos entre si, enquanto que nos contextos de um item polissêmico pode haver uma grande interseção entre as palavras utilizadas. Essa é uma característica importante a ser considerada na avaliação dos diferentes trabalhos.

É importante ressaltar que, para a TA, como reforçado por Mihalcea & Modovan (1999), é essencial a desambiguação considerando sentidos refinados.

2.4.3 O uso de córpus como base para a distinção entre os sentidos

Outra questão relacionada à definição dos sentidos é a base utilizada para essa distinção (Seção 2.4.1): o significado ou o uso das palavras. Conforme indicado por Kilgarriff, diversos trabalhos consideram que o sentido de uma palavra deve ser definido em função do seu uso ou papel em um córpus. Esses autores acreditam que o estudo do sentido das palavras não requer apenas um dicionário, mas também vários exemplos das palavras em uso. Para definir seus repositórios de sentido, os autores utilizam técnicas da lingüística de córpus.

A lingüística de córpus oferece a possibilidade de introduzir um alto grau de empirismo ao estudo de muitos aspectos das línguas. O problema, contudo, é a necessidade de um córpus representativo da(s) língua(s) em questão, ou seja, um córpus com variações suficientes sobre diversos aspectos relevantes da(s) língua(s), devidamente balanceadas ou distribuídas conforme necessário. Novamente, se analisado independentemente de aplicação, até mesmo o conceito de “representativo” é complexo. Em uma mesma língua, diversos aspectos podem ser considerados relevantes para um córpus, de acordo com a aplicação em foco, incluindo decisões entre: textos falados e/ou escritos, textos espontâneos e/ou preparados, variações de dialetos, variações de nível de formalidade, variações na idade/sexo/região do escritor, variações de domínio e de gênero, etc.

De modo geral, definidas as variações esperadas em um córpus, ele pode fornecer uma grande quantidade de padrões e estruturas que ocorrem com uma freqüência expressiva em uma língua. Um córpus também permite asserções quantitativas sobre a freqüência relativa de construções, ou a conclusão de que determinadas construções não ocorrem na língua.

Seguindo a sugestão de Resnik & Yarowsky (1997a), Ide (1999) analisa traduções de palavras do inglês em quatro línguas de famílias distintas (sem incluir a língua portuguesa) para verificar em que nível as traduções para os diferentes significados de um item polissêmico do inglês são lexicalizadas por itens diferentes nas quatro línguas. A autora procura determinar se essa informação pode ser usada para criar ou estruturar um conjunto de distinções de sentido monolingües (para a língua-fonte) úteis para aplicações de PLN. Para evitar o problema de que as ambigüidades podem ser preservadas entre diferentes línguas, Ide analisa línguas distintas entre si. Ainda, diferentemente de outros trabalhos, que utilizam repositórios de possíveis sentidos de recursos como dicionários bilíngües, Ide utiliza córpus paralelos alinhados, mapeando os sentidos que são traduzidos diferentemente em sentidos monolingües da WordNet.

De acordo com os resultados encontrados, a autora conclui que a lexicalização em diferentes línguas pode ser usada para definir e estruturas distinções de sentido. A maior motivação para o uso de córpus, segundo Ide, é que as traduções são atribuídas por tradutores experimentes, de modo que a chance de incorrer em erros é menor. O problema

com o uso desse tipo de córpus, segundo a autora, é a sua falta de textos paralelos alinhados para diversas línguas.

Ng et al. (2003) também empregam um procedimento similar para identificar um repositório de sentidos para a língua inglesa. Eles usam córpus paralelos inglês-chinês alinhados por palavras. O sentido das palavras nos textos em inglês é definido manualmente, com base nos sentidos da WordNet, mas revisado com base nas traduções para o chinês identificadas pelo alinhador de palavras. Assim, se duas palavras com dois sentidos diferentes da WordNet são traduzidos para um único termo do chinês, considera-se que essas duas palavras têm um único sentido, em inglês. Segundo os autores, com isso, elimina-se o problema da escolha pelo conjunto de sentidos possíveis, uma vez que esse conjunto, incluindo decisões sobre o nível de refinamento desses sentidos, é definido em função dos dados do córpus, em uma visão mais objetiva que a utilização de inventários de sentidos de recursos lexicais.

Na DLS para a TA, em especial, o uso de córpus para a definição dos sentidos isenta o desenvolvedor de uma série de decisões importantes, como a escolha entre o tratamento de casos de homonímia e/ou polissemia, a escolha do nível de refinamento dos sentidos, entre outras.

Na Seção 2.7, são descritos alguns trabalhos de criação de córpus de treinamento para trabalhos baseados em córpus que definem automaticamente o conjunto de sentidos (e o nível de distinção entre eles) a partir de córpus.

2.5 Tipos de conhecimento e fontes de informação

Uma questão importante na definição de um trabalho de DLS é a definição dos tipos de conhecimento que serão empregados e das fontes de informação que podem ser utilizadas para prover esses conhecimentos. Em geral, a desambiguação pode ser realizada com base na utilização de dois grandes grupos de conhecimento: o **conhecimento do contexto** da palavra-alvo (isto é, da palavra a ser desambiguada) e o **conhecimento externo** a respeito da palavra-alvo ou das palavras do contexto.

O contexto da palavra-alvo inclui informações contidas no texto do qual ela faz parte e também informações extralingüísticas sobre tal texto. Esse contexto pode ser definido de várias maneiras, como será explicado a seguir (*bag-of-words*, n-gramas, co-ocorrências, etc.). De modo geral, durante a desambiguação, as regras verificam se as palavras definidas como o contexto relevante da palavra ambígua para um determinado sentido coincidem com as palavras do contexto da palavra sendo desambiguada, considerando, em alguns casos, a posição e a ordem das palavras. Conforme mencionado, para a DLS na TA, em particular, o contexto normalmente consiste das palavras na língua-fonte.

O conhecimento externo inclui vários tipos de conhecimento que podem ser manualmente especificados ou provenientes de recursos lexicais, bem como de processos como etiquetadores morfossintáticos e analisadores sintáticos. Durante a desambiguação, as regras verificam se as palavras-alvo e/ou as palavras do contexto possuem determinadas características ou estão relacionadas de determinadas formas, por exemplo.

Alguns trabalhos consideram apenas conhecimento sobre o contexto, enquanto outros adicionam conhecimentos externos de diversas fontes. Analisando o trabalho de diversos autores, como McRoy, (1992), Hirst (1987) e Ng & Zelle (1997), foram levantados os seguintes tipos de conhecimento como os mais amplamente utilizados nos trabalhos de DLS:

1. *Bag-of-words*: conjunto de palavras que circundam a palavra a ser desambiguada em uma janela que pode variar desde uma quantidade pré-determinada de palavras na sentença ou no texto no qual a palavra está localizada à sentença inteira ou algumas sentenças do texto. Normalmente essa janela considera a palavra ambígua como centro e uma igual quantidade máxima de palavras em ambos os lados, sem considerar a sua ordem e independentemente das suas características.

Em geral, para tornar as propostas mais genéricas, as palavras são lematizadas e as palavras de classe fechada ou palavras irrelevantes são eliminadas por meio de uma lista de *stop words*. Esse tipo de contexto, especialmente quando contempla uma janela ampla, é mais voltado para capturar o tópico geral do texto. Por exemplo, um conjunto de palavras vizinhas incluindo *bank*, *loan* e *payment*, em quaisquer posições, podem levar à identificação do sentido de *interest* como "juros".

Em alguns trabalhos baseados em córpus, são selecionadas para o *bag-of-words* somente as palavras que ocorrem com uma determinada freqüência mínima no córpus, ou seja, palavras que carregam, potencialmente, informações mais relevantes para a desambiguação.

2. N-gramas e co-ocorrências: conjunto de duas ou mais palavras que ocorrem juntas em qualquer posição no contexto considerado. Podem incluir n-gramas, com n >1, ou co-ocorrências em geral. No caso dos n-gramas, são consideradas apenas as palavras que ocorrem adjacentes umas às outras, em uma ordem pré-definida. Por exemplo, um bigrama é constituído de duas palavras que ocorrem na seqüência dada, incluindo ou não a palavra ambígua. Novamente, podem ser representados apenas os n-gramas relevantes, ou seja, que ocorrem com freqüência. Nos trabalhos baseados em córpus, essa freqüência é dada por um limite estabelecido para a ocorrência do conjunto de palavras nos exemplos de treinamento, juntas e na ordem dada. Os n-gramas não excluem as palavras de classe fechada, porém, normalmente, pelo menos uma das palavras precisa ser de classe aberta.

As co-ocorrências são generalizações dos n-gramas, nas quais as palavras do conjunto não precisam ser adjacentes (um limite de distância pode ser estabelecido), mas a sua ordem mantém-se fixa. Em geral, palavras de classe fechada são ignoradas. Co-ocorrências relevantes também são determinadas pela sua freqüência. A distância estabelecida como limite entre as palavras do conjunto co-ocorrente define a informação mútua de desse conjunto, ou seja, mede o grau de associação entre as suas palavras.

3. *Collocations*: o termo *collocation* possui conotações muito variadas nos diversos trabalhos que empregam esse conhecimento. As *collocations* foram primeiramente utilizadas por Firth (1957) para caracterizar certos fenômenos lingüísticos de co-ocorrência de uma língua que se sustentam principalmente na competência lingüística dos seus falantes nativos. O autor enfatiza que esse fenômeno não se baseia somente na co-ocorrência da construção, mas no seu uso habitual ou comum.

Segundo Kilgarriff (1997a, p. 29), uma *collocation* é um grupo de duas ou mais palavras encontradas próximas umas às outras com uma freqüência significativamente maior do que o previsto, dada a freqüência de ocorrência de cada palavra individualmente. Elas podem ser ou não vizinhas imediatas, e o significado do todo pode ou não ser completamente determinado pelos significados das partes. A maioria das *collocations* é arbitrária. Por exemplo: *do the exercises* (em vez de *make the exercises*) e *make love* (em vez de *do love*), assim, o uso dessas colocações corretamente depende da fluência na língua.

Já de acordo com Ide & Véronis (1998), uma *collocation* pode ser definida como uma associação sintagmática entre itens lexicais *x* e *a*, *b*, *c*..., na qual a probabilidade do

item *x* co-ocorrer com os itens *a*, *b*, *c*... é maior que a probabilidade ao acaso. Yarowsky (1993), por sua vez, define uma *collocation* como a co-ocorrência de duas palavras em uma relação sintática pré-definida.

Em outros trabalhos, ainda, o termo *collocation* é utilizado em um sentido mais amplo, similar ao usado, neste trabalho, para denominar co-ocorrências: duas palavras fazem parte de uma *collocation* se elas são adjacentes ou ocorrem próximas uma da outra no córpus (na mesma sentença ou parágrafo, numa vizinhança pequena, de tamanho pré-definido), considerando-se ou não, além da distância da palavra com relação à palavra ambígua ou características específicas dessas palavras, como a sua categoria gramatical. Vale notar que, mesmo que a *collocation* seja constituída de apenas uma palavra, a análise do seu uso é realizada com relação à palavra a desambiguar. Assim, a *collocation* será formada, na verdade, por essa palavra e a palavra ambígua. Essa é a visão, por exemplo, de Martínez & Agirre (2000), que consideram, por exemplo, a primeira palavra de conteúdo à direita e à esquerda da palavra ambígua como duas *collocations*.

Neste trabalho, uma *collocation* consistirá de uma pequena seqüência de palavras próximas à palavra ambígua, considerando a sua ordem, a sua posição relativa à palavra-alvo na sentença e a sua categoria gramatical. Por exemplo, uma *collocation* pode ser formada pelo primeiro artigo e preposição à esquerda da palavra a desambiguar e a primeira preposição à sua direita. Essa seqüência pode representar um conjunto de palavras comumente usadas naquela ordem. Por exemplo, a seqüência "*in the interest of*" pode indicar o sentido correto de *interest* como "vantagem". Alguns outros exemplos de *collocations* são: primeiro substantivo à esquerda, primeiro verbo à direita, dois primeiros substantivos à direita, etc.

Vale notar que *collocations* são diferentes de expressões idiomáticas (Arnold et al., 1993). Ambas são unidades multi-palavras e dependem da competência lingüística do falante nativo. Contudo, o significado de expressões idiomáticas não pode ser completamente compreendido a partir do significado das suas partes, enquanto que isso é possível nas *collocations*. O que não é totalmente previsível, no caso de *collocations*, para falantes não nativos, são as palavras específicas, dentre vários sinônimos ou sentidos próximos, habitualmente utilizadas. Além disso, normalmente não é possível traduzir literalmente uma expressão idiomática, o que faz com que elas sejam geralmente tratadas como unidades lexicais individuais em sistemas de TA. No caso de *collocations*, isso não ocorre. A seguir, são ilustrados alguns exemplos de expressões idiomáticas, em (1), e de *collocations*, em (2).

(1) "*put the foot in the mouth*" (dizer coisas embaraçantes)
"*be like two peas in a pod*" (ser idênticos)
"*be head and shoulders above the rest*" (ser o melhor)

(2) "*rancid butter*" (em vez de "*sour butter*", "*rotten butter*"ou "*stale butter*")
"*take a walk*" (em vez de "*make a walk*")
"*do the homework*" (em vez de "*make the homework*")

No caso de *collocations* o conhecimento de que se deve usar *rancid* com a palavra *butter*, em vez de outras palavras que denominam a mesma propriedade (*sour*, *rotten* ou *stale*), não permite prever que se deve usar também *rancid* com a palavra *cream*, pelo fato dessa palavra ter significado próximo a *butter*. Na verdade, neste caso, o uso habitual é "*sour cream*".

Vale notar que nos diversos trabalhos apresentados na Seção 3, o termo *collocation* será mantido de acordo com sua utilização nesses trabalhos, mesmo com conotações distintas.

4. Relações sintáticas: relações superficiais entre as palavras da sentença, como sujeito-verbo, verbo-objeto e adjetivo-substantivo, usadas para a verificação de restrições de seleção. Por exemplo, na sentença "*He sold his interest in the joint venture*", a relação sintática verbo-objeto entre o verbo *sold* e o núcleo do sintagma nominal do objeto indica o sentido "parte na empresa" de *interest*.

5. Traços semânticos e restrições ou preferências de seleção: características semânticas das palavras, por exemplo, "humano", "animado", "concreto", etc., e restrições estabelecidas, com base nessas características e em relações sintáticas, geralmente para verbos, sobre as construções que são usadas como seus argumentos, e e adjetivos, sobre as construções que são modificadas por eles. Por exemplo, o sentido "conhecer" do verbo *know* exige como núcleo do seu objeto direto um substantivo com o traço "pessoa".

6. Categoria gramatical: categorias gramaticais da palavra ambígua e das palavras vizinhas. Pode-se analisar, por exemplo, as categorias das palavras definidas como n-gramas ou co-ocorrências, como a categoria da primeira palavra à esquerda, das duas palavras à direita, etc.

7. Traços morfológicos: traços morfológicos da palavra ambígua, como o seu número, para substantivos, e o tempo e modo, para verbos.

Alguns autores (Stevenson & Wilks, 1999; 2000; 2001, por exemplo), utilizam também outras informações disponíveis em dicionários e outros recursos lexicais, como as palavras nas definições de cada sentido nos dicionários, os exemplos associados às definições e os códigos de área atribuídos aos diferentes sentidos (medicina, direito, etc.), simulando a noção de domínio. Os códigos de área, segundo os autores, são úteis principalmente para a desambiguação de substantivos.

Em níveis mais abstratos de conhecimento, Kilgarriff (1997a) acrescenta a noção de situação como um conhecimento desejável para a DLS. De acordo com a semântica situacional, o mundo é formado por indivíduos, propriedades, relações e situações, sendo que as **situações** são partes limitadas do mundo que consistem de indivíduos tendo (ou não tendo) propriedades e participando (ou não participando) de relações. Para o autor, a situação é determinante para a interpretação de enunciados. Por exemplo, a situação, ainda que implícita, de uma sentença como "*I smoke*" envolve fogo e fumaça, características que podem ser essenciais para a desambiguação.

Outros autores (Arnold et al., 1993, por exemplo) mencionam a necessidade de conhecimento ontológico e conhecimento de mundo para a resolução de ambigüidades de sentido. Para esses autores, esse tipo de conhecimento é indispensável para entender como uma palavra contribui para o significado do enunciado no qual ela ocorre e, com isso, desambiguá-la.

Com base nos conhecimentos comumente utilizados e nas declarações de diversos autores, como os citados, sobre os tipos adicionais de conhecimentos para a DLS, Agirre & Martínez (2001) esquematizam um elenco mais completo de **tipos de conhecimento** apontados pela literatura como úteis ou desejáveis para a DLS monolingüe:

1. Etiquetas gramaticais;
2. Morfologia;
3. *Bag-of-words*, co-ocorrências e *collocations*;
4. Associações semânticas entre as palavras, incluindo:

   i. Organização em uma taxonomia;
   ii. Situação;
   iii. Tópico;
   iv. Relação núcleo-argumentos;
5. Informações sintáticas, como a estrutura de subcategorização da palavra;
6. Papéis semânticos;
7. Preferências de seleção;
8. Domínio;
9. Freqüência dos sentidos;
10. Pragmática.

Agirre & Martínez (2001) procuram fazer também um paralelo com as **fontes de informação** disponíveis, em termos de recursos computacionais, para a aquisição desses conhecimentos automática ou semi-automaticamente. Segundo os autores, normalmente, as bases de conhecimento lexicais contendo vários dos conhecimentos citados, principalmente os mais profundos, são construídas manualmente e organizadas de modo a integrar os diferentes conhecimentos. McRoy (1992), por exemplo, organiza o conhecimento relativo a 10.000 lemas em quatro componentes inter-relacionados:

- Léxico: constituído de um léxico central, capturando os tipos de conhecimento 1, 2, 5 e 9, e um léxico dinâmico, capturando o tipo de conhecimento 8.
- Hierarquia conceitual, incluindo os tipos de conhecimento 4i, 6 e 7.
- Padrões de *collocations*, capturando o tipo de conhecimento 3.
- *Clusters* de definições relacionadas, capturando os tipos de conhecimento 4ii e 4iii.

Contudo, atualmente, há várias fontes de informação que podem ser usadas para a extração automática de alguns desses tipos de conhecimento. Alem das fontes citadas por Agirre & Martínez, a lista a seguir considera alguns processos de PLN que permitem a geração automática de vários dos conhecimentos mencionados.

- Etiquetadores morfossintáticos (1): etiquetas gramaticais são previamente atribuídas às palavras.
- Analisadores morfológicos (2): da mesma forma, em uma etapa prévia ao processo de DLS, analisadores morfológicos reduzem as palavras aos seus lemas ou raízes, anotando as informações morfológicas referentes à sua derivação ou flexão na sentença.
- Analisadores sintáticos (4iv e 5): também em uma etapa anterior ao processo de DLS, é realizada a análise sintática (geralmente, superficial) da sentença para extrair as relações entre seus elementos;
- Dicionários eletrônicos (4, 5, 6, 7, 8, 9): em geral, o primeiro sentido apresentado nos dicionários pode ser usado como o sentido mais freqüente (9). Os demais conhecimentos são disponibilizados apenas por alguns dicionários mais detalhados e/ou específicos e, normalmente, são divididos entre diversos dicionários.
- Ontologias (4i): a organização taxonômica como a ontologia da WordNet pode ser empregada em trabalhos de DLS, em geral.
- Córpus (3, 4ii, 4iii, 5, 7, 8, 9): um córpus etiquetado com sentidos e outras informações importantes pode ser usado para fornecer, além da freqüência de uso de cada sentido da palavra, diversas informações referentes ao seu contexto.

De modo geral, os diferentes trabalhos de DLS usam um ou mais dos tipos de conhecimento mencionados. Trabalhos que combinam vários conhecimentos podem, em teoria, apresentar melhores resultados. Propostas como as de Hearst (1991), McRoy (1992), Ng & Lee (1996) e Stevenson & Wilks (1999; 2000; 2001), por exemplo, procuram combinar uma grande quantidade de conhecimentos e, com isso, obtêm uma precisão maior que a das outros trabalhos que consideram apenas um ou mais dos conhecimentos, isoladamente.

Como se pode perceber, grande parte dos conhecimentos pode ser extraída de córpus e, outra parte, de dicionários eletrônicos. Com base nisso, Agirre & Martínez (2001) sugerem algumas combinações entre fontes de informação que podem suprir deficiências por tipos de conhecimento não disponíveis nessas fontes, se tomadas isoladamente. As combinações consideradas úteis e viáveis são: dicionário-córpus, ontologia-córpus, dicionário-ontologia. Certamente, deve haver uma interação entre esses conhecimentos

Alguns dos tipos de conhecimento citados não possuem uma fonte de informação correspondente e, por essa razão, não são (ou raramente são) utilizados na DLS. Por exemplo, praticamente não há trabalhos que utilizam papéis semânticos.

Com relação à importância de cada tipo de conhecimento, Agirre & Martínez analisam os resultados de alguns algoritmos para a DLS que utilizam as fontes de informação citadas (incluindo as combinações), de modo a verificar quais são as mais importantes para a desambiguação. Dadas as diferenças entre os algoritmos em questão, foram realizados testes sob configurações específicas: diferentes línguas, inventários de sentidos e conjuntos de treinamento e teste. Contudo, manteve-se a padronização quanto ao repositório de sentidos (da WordNet) e o conjunto de testes (todas as ocorrências de oito substantivos do córpus SEMCOR (Seção 2.7.1) ou todos os substantivos polissêmicos em quatro arquivos randomicamente selecionados do SEMCOR).

De acordo com os testes, os autores afirmam que os trabalhos que combinam informações provenientes de córpus e de outras fontes apresentam resultados muito melhores que os de trabalhos que só usam informações provenientes de uma fonte. O uso isolado de cada tipo de conhecimento, em particular, leva a resultados bastante insatisfatórios. De modo geral, os autores indicam os principais tipos de conhecimento como determinantes para a DLS:

- co-ocorrências e *collocations* adquiridas de um córpus;
- associações semânticas de palavras sobre um determinado tópico ou situação, adquiridas de um córpus;
- informações sintáticas;
- freqüência do uso dos sentidos, calculada com base em um córpus.

Esses resultados confirmam os apresentados por McRoy, de que *collocations* e associações semânticas de palavras são os tipos e conhecimento mais importantes para DLS. A conclusão dos autores é que um córpus etiquetado com sentidos é a melhor fonte de informação para a aquisição automática dos tipos de conhecimento considerados. As *collocations* e co-ocorrências são, de fato, muito relevantes para a DLS. Como será apresentado na próxima seção, a maioria dos trabalhos utiliza algum tipo de *collocation* ou co-ocorrência, principalmente porque podem ser automaticamente extraídas a partir de um córpus.

Pedersen (2002a), em particular, relatam resultados razoavelmente bons utilizando apenas características lexicais que podem ser automaticamente extraídas de córpus, como bigramas e co-ocorrências. Mas, como o próprio autor relata, essa abordagem simples atinge um patamar de acurácia (menos de 70%) que não pode ser melhorado, mesmo com a

sofisticação das técnicas de extração de conhecimento e o aumento do córpus. Assim, somente com outros tipos de conhecimento pode-se melhorar essa acurácia.

Martínez et al. (2002) discorrem especificamente sobre a contribuição de uma série de características sintáticas para auxiliar a DLS em trabalhos baseados em córpus. Informações provenientes da estrutura sintática são sugeridas como base para a identificação de alternâncias no significado decorrentes de diferentes usos da língua por diversos autores (por exemplo, Levin, 1993). No entanto, segundo Martínez et al., a maioria dos trabalhos em DLS não explora, efetivamente, informações dessa natureza. Eles realizam experimentos considerando características básicas, utilizadas por grande parte dos trabalhos de DLS, como os lemas das palavras vizinhas, *collocations* e n-gramas. Adicionalmente a essas características, são consideradas, então, informações sintáticas. Essas informações são extraídas dos exemplos a partir de um *parser* que gera estruturas sintáticas completas para as sentenças (e não apenas relações superficiais, como as utilizadas em outros trabalhos de DLS). Elas incluem relações sintáticas diretas entre as palavras (ou seja, entre palavras contíguas na sentença, como verbo-objeto) e relações sintáticas indiretas (como núcleo de um sintagma preposicional modificador).

Para verificar a importância dessas características, os autores configuram e executam dois experimentos, utilizando dois algoritmos de aprendizado diferentes e os dados do exercício de avaliação SENSEVAL-2 (Seção 2.8.2). Os resultados são reportados com relação à adição das informações sintáticas ao conjunto de características básicas citado. Segundo os autores, pode-se perceber que, independentemente das outras variações nos testes, o uso das relações sintáticas contribui significativamente para o aumento tanto na precisão quanto na cobertura dos trabalhos.

Por fim, Lee & Ng (2002) (Seção 3.2.2), em uma avaliação comparativa com quatro tipos de conhecimento e diferentes algoritmos de aprendizado para trabalhos baseados em córpus, concluem que a contribuição relativa de cada um dos tipos de conhecimento para a acurácia do modelo depende do algoritmo utilizado. Por exemplo, *collocations* representam a maior contribuição para um algoritmo, mas as categorias gramaticais são mais relevantes quando outro algoritmo é considerado. Certamente, outras características também influenciam na acurácia com o uso de cada tipo de conhecimento, como o método de desambiguação, o tipo de palavra-alvo, etc.

2.5.1 Contexto local x contexto global

Como mencionado, a utilização do contexto de ocorrência de uma palavra é imprescindível para determinar o seu sentido. Portanto, todos os trabalhos em DLS utilizam, de alguma forma, o contexto da palavra-alvo para prover informações para sua desambiguação. Duas classes de informações contextuais podem contribuir para a seleção do sentido mais adequado: informações do **contexto local**, ou **micro-contexto**, e informações do **contexto local**, ou **contexto de tópico**. Atualmente, há vários trabalhos que integram ambos os tipos de contextos, no entanto, não há um consenso, entre os diversos trabalhos de DLS, sobre o papel e a importância de cada uma dessas informações, bem como do seu inter-relacionamento.

O **micro-contexto** geralmente consiste de uma pequena janela de palavras próximas à ocorrência da palavra alvo, considerando ou não as relações entre essas palavras e a palavra alvo. A maioria dos trabalhos de DLS utiliza esse tipo de contexto. Alguns trabalhos consideram apenas o conjunto de palavras, sem relações com a palavra-alvo (n-gramas e *bag-of-words*). Contudo, como mostra Yarowsky (1992), trabalhos que consideram as relações entre as palavras são mais efetivos, principalmente na desambiguação de verbos. Os principais tipos de relações consideradas são as de distância

da palavra em relação à palavra ambígua (co-ocorrências), as *collocations* e as relações sintáticas. Nesse caso, considera-se como contexto apenas as palavras que aparecem a uma determinada relação de distância, que possuem determinadas características que indicam collocations ou que ocorrem em determinadas relações sintáticas com a palavra-alvo.

A distância da palavra depende do tamanho da janela de contexto considerada. Para esse tamanho, não há um valor ideal, válido para todos os trabalhos. Esse valor depende, entre outras coisas, do tipo de ambigüidade. Yarowsky (1993; 1994), por exemplo, sugere uma janela de 3 ou 4 palavras à direita e à esquerda da palavra-alvo.

O **contexto global** inclui uma quantidade maior de palavras que co-ocorrem com um dado sentido de uma palavra. Não há um consenso sobre quando um contexto deixa de ser local para se tornar global. Contudo, considera-se, normalmente, uma janela de várias sentenças, seguindo a estratégia *bag-of-words*. O tamanho dessa janela também pode variar. Por exemplo, Yarowsky (1993; 1994) sugere uma janela de 20 a 50 palavras à direita e à esquerda da palavra-alvo.

Os trabalhos que usam um contexto mais amplo exploram a redundância nos textos, ou seja, o uso repetido das palavras que são semanticamente relacionadas em todo o texto sobre um determinado tópico. Por exemplo, a palavra *base* é ambígua, mas seu uso em um documento contendo palavras como *pitcher*, *ball*, etc., mesmo a uma grande distância de *base*, pode determinar o seu sentido adequado (neste caso, uma base de arremesso no beisebol), bem como o sentido dessas outras palavras, igualmente ambíguas.

Apesar da potencial eficiência, o uso desse tipo de contexto é relativamente recente e mais comum apenas nos trabalhos baseados em córpus. Um dos motivos pode ser o custo de processamento de uma grande quantidade de palavras e, possivelmente, das informações relativas a essas palavras. Outra restrição é que nem sempre o contexto da palavra a ser desambiguado inclui um número suficiente de palavras. Por exemplo, a unidade de entrada do sistema pode ser constituída de sentenças isoladas.

Com relação à distinção entre os dois tipos de contexto, vários autores (Yarowsky, 1992, por exemplo), com base em experimentos com seus trabalhos, afirmam que o contexto global é mais indicado para a desambiguação de substantivos, pois eles requerem informações que podem estar mais distantes no texto. Já o contexto local é mais apropriado para desambiguação de verbos ou adjetivos, pois eles necessitam de informações que geralmente estão próximas, como os argumentos de um verbo ou os elementos modificados por um adjetivo.

Em um experimento para determinar a influência dos dois tipos de contexto na desambiguação, Leacock et al. (1996) analisam diferentes trabalhos baseados em córpus que se baseiam unicamente em contexto global ou local, sendo que o contexto global é dado por todas as palavras da sentença a ser desambiguada e da sentença anterior no texto, quando existir, e contexto local é dado por *collocations*. Ao avaliar o desempenho dos trabalhos para a desambiguação dos seis sentidos de ocorrências da palavra *line*, os autores concluem que o contexto global fornece uma boa indicação do sentido (70%, em média), mas que não é suficiente para a DLS. Segundo os autores, o contexto local é superior ao contexto global como indicar do sentido adequado de maneira mais precisa. Eles também realizam um experimento com humanos na desambiguação, no qual percebem que o contexto local é mais efetivo. Por outro lado, nos trabalhos que usam apenas contexto local a cobertura obtida é muito baixa, pois dificilmente o contexto do novo caso a desambiguar coincide com o contexto aprendido. Assim, sugerem que os dois tipos de contexto devem ser empregados para obter resultados mais abrangentes e precisos de DLS.

Dagan & Itai (1994) apontam, também, que informações sobre o contexto global são menos sensíveis a distinções lexicais e semânticas refinadas e, por isso, não são muito úteis para desambiguar diferentes sentidos de uma palavra que aparecem em contextos

similares. Por outro lado, como o contexto global contém mais palavras, pode prover mais informações para a desambiguação, principalmente nos casos em que a distinção é baseada no tópico do discurso. Assim, os autores sugerem igualmente que as duas fontes de informação são complementares e podem ser combinadas para obter um conhecimento maior sobre o contexto de ocorrência da palavra ambígua. Isso se mostra importante principalmente para a desambiguação em larga escala, de palavras de várias categorias gramaticais já que, conforme mencionado, cada contexto fornece informações mais adequadas para determinadas categorias.

2.5.2 A importância do domínio

A noção de domínio, ainda que implicitamente, é utilizada vários trabalhos de DLS, que desambiguam as palavras de acordo com seu uso em um dado domínio. Trabalhos que implementam mais fortemente a noção de domínio são aquelas restritas a domínios específicos, por exemplo, os sistemas de TA KMBT (Goodman & Nirenburg, 1991) e Mikrokosmos (Beale, 1997), que serão descritos no Capitulo 4. Outros sistemas de TA, como o Systran, procuram, por meio do contexto da sentença, identificar o domínio para utilizar dicionários específicos daquele domínio.

Alguns trabalhos de DLS empregam a noção explícita de domínio, conferindo-lhe mais ou menos importância. Gale et al. (1992c), por exemplo, defende sua teoria de que somente um sentido é utilizado somente para cada palavra ambígua em cada discurso (de um dado domínio) (Seção 2.10).

Wilks & Stevenson, em seus diversos trabalhos (Wilks & Stevenson, 1997b; 1998; Stevenson & Wilks, 1999; 2000; 2001), empregam os códigos de área do LDOCE para tentar identificar o domínio do contexto da palavra ambígua para então verificar qual o sentido que pertence a tal domínio. Contudo, no LDOCE, só uma pequena parte das entradas possuem uma identificação de área. Assim, eles empregam outras fontes de informação complementares para a desambiguação.

Magnini et el. (2002) usam uma identificação mais estruturada de domínio, fornecida por uma adaptação da WordNet, denominada WordNet Domain, na qual são inseridos códigos de domínio para cada sentido (*medicine, sportes,* etc.). A partir desse recurso, é escolhido o sentido de uma palavra cujo domínio é o mais próximo do domínio identificado pelo contexto dessa palavra. Nesse trabalho, a estrutura hierárquica da WordNet permite explorar mais profundamente esse conhecimento. Todos os sentidos possuem um código de domínio, entretanto, muitos sentidos são de domínio genérico (por exemplo, o verbo *to be*) e, portanto, o seu código não auxilia na desambiguação. Os autores estimam que apenas cerca de 20% dos sentidos possuem códigos que podem, efetivamente, auxiliar na desambiguação. Assim, se a sentença for pequena ou composta por palavras de domínio genérico, esse método falha.

Vários outros trabalhos, como será descrito na seção seguinte, procuram identificar, automaticamente, o domínio do contexto da entrada para a DLS. Certamente, a informação sobre o domínio é importante, mas em trabalhos que pretendem ser aplicáveis a vários domínios, esse conhecimento deve ser coordenado com outros, pois nem sempre é possível identificar automaticamente, com um nível de confiança satisfatório, qual é o domínio. Mesmo com a delimitação da proposta a um dado domínio ou a identificação automática do domínio do texto, somente esse conhecimento, isoladamente, não elimina todas as ambigüidades. O nível de influência do domínio para a tarefa de DLS depende de vários fatores, como o gênero de texto (técnico, jornalístico, etc.), o nível de refinamento entre os sentidos da palavra a ser desambiguada.

2.6 Métodos de DLS

Os trabalhos de DLS podem seguir os diferentes métodos de PLN: 1) método profundo, baseado em conhecimento lingüístico e/ou extralingüístico explicitamente especificado; 2) método empírico, baseado em córpus de exemplos e em algoritmos de aprendizado de máquina para adquirir conhecimento automaticamente a partir dos exemplos; ou 3) método híbrido, que combina características dos métodos profundos e empíricos. Essa classificação para a DLS é recomendada pelo grupo de estudos lexicais do EAGLES (EAGLES, 1998).

Além dessa classificação principal, neste trabalho, são consideradas subdivisões desses métodos, de acordo com algumas variações relevantes, como a maneira com que o conhecimento é codificado, os tipos de recursos lexicais que podem ser utilizados, a representação dos modelos aprendidos a partir de córpus, etc. A seguir, são apresentadas as principais características de cada método e suas subdivisões, bem como suas vantagens e desvantagens no contexto da DLS. A mesma classificação será mantida para descrição dos trabalhos de DLS na Seção 3.

2.6.1 Método baseado em conhecimento

Nos trabalhos baseados em conhecimento, a desambiguação é realizada com o uso de informações explicitamente especificadas, manualmente ou a partir de recursos lexicais já disponíveis. Vale notar que não são considerados, aqui, trabalhos baseados em conhecimento aqueles que utilizam informações provenientes de processos lingüísticos, como um *parser* ou um lematizador, uma vez que essas informações, ainda que explícitas no momento da desambiguação, são geradas por sistemas independentes, e não pelo mecanismo de DLS. Esses processos podem, inclusive, ter sido criados empiricamente, com base em córpus. Além disso, as informações geradas podem conter erros, derivados de problemas próprios desses processos.

2.6.1.1 Conhecimento manualmente codificado

Os primeiros trabalhos baseados em conhecimento para DLS utilizavam técnicas da Inteligência Artificial para representar o conhecimento, que era manualmente codificado. Esses trabalhos começaram a ser definidos na década de 1960, normalmente para resolver o problema da compreensão da língua natural e, como parte desse problema, a DLS. A maioria deles é baseada em alguma teoria de compreensão da língua humana, envolvendo o uso de conhecimento detalhado sobre a sintaxe e semântica da língua para realizar tal tarefa. Os trabalhos podem ser divididos em duas categorias, que variam quanto ao tipo de técnica empregada para a representação do conhecimento: **abordagens simbólicas**, que utilizam técnicas simbólicas, e **abordagens conexionistas**, que utilizam técnicas numéricas.

A grande vantagem dos trabalhos baseados em conhecimento manualmente codificado é que o seu nível de especialização e consistência pode levar a resultados bastante precisos, principalmente na desambiguação de palavras altamente ambíguas e com distinções de sentido bastante refinadas. Em contrapartida, o problema desses trabalhos é que a tarefa de codificação manual é lenta e custosa. Isso pode ser facilmente verificado pela observação de Small (1980) (Seção 3.1.1), de que a descrição de um modelo para uma única palavra ocupou seis páginas e ainda poderia ocupar um espaço dez vezes maior, caso fosse feita minuciosamente.

Assim, de modo geral, as dificuldades para criar manualmente as fontes de conhecimento requeridas pelos métodos baseados em IA[2] acabam restringindo a implementação dos sistemas que utilizam esses métodos, com poucas exceções (McRoy, 1992, por exemplo), a protótipos bastante limitados, restritos a pequenos subconjuntos da língua natural, em domínios e gêneros bem delimitados. Como conseqüência, esses sistemas de DLS não são facilmente generalizáveis e, portanto, não podem ser utilizados em aplicações reais, de larga escala. À medida que a quantidade de conhecimento para a desambiguação cresce, a manutenção manual e as expansões do sistema tornam-se cada vez mais complexas.

Um problema mais pontual é que a maioria dos trabalhos explora apenas o conhecimento disponível no nível sentencial. Além disso, essas sentenças normalmente são artificialmente construídas, de modo a apresentarem casos de ambigüidades complexos, que exigem distinções bastante refinadas entre os sentidos. Segundo Ide & Véronis (1998), a maioria das sentenças analisadas raramente são usadas no mundo real, o que reforça a dúvida sobre a possibilidade de aplicação desses trabalhos.

Segundo Kilgarriff (1992), os principais problemas dos trabalhos que seguem esse método são que, além da quantidade muito pequena de palavras estudada, não há justificativa de como esse subconjunto de palavras ambíguas foi escolhido, com base em quais critérios. Os possíveis sentidos para cada palavra ambígua também são, em geral, definidos pelo pesquisador, como um recorte dos sentidos encontrados em dicionários.

2.6.1.2 Conhecimento pré-codificado

A partir da década de 1980, quando se tornaram disponíveis recursos lexicais eletrônicos em larga escala, como dicionários eletrônicos, começaram a surgir trabalhos em DLS baseados em conhecimento pré-codificado, isto é, extraído desses recursos, os quais podem ter sido criados com várias finalidades. Esse conhecimento inclui os sentidos possíveis de cada palavra, informações associadas a esses sentidos, como a sua categoria gramatical, marcadores de tópico e área, restrições de seleção, definições textuais, etc., bem como relações entre os sentidos ou entre grupos de sentido, como a sinonímia e a antonímia.

Alguns trabalhos utilizam diretamente o conhecimento disponível em recursos lexicais (em dicionários eletrônicos, por exemplo), enquanto outros utilizam bases de conhecimento criadas manualmente, especificamente para serem manipuladas por sistemas de PLN em geral. Na descrição dos diferentes trabalhos, os recursos de conhecimento pré-codificado são divididos, de acordo com a classificação de Ides & Véronis (1998), em **léxicos computacionais**, **dicionários eletrônicos** e ***thesauri***, com base, entre outras características, no seu método de organização dos dados: em um dicionário, a entrada principal é no nível da palavra, sendo que cada entrada representa os vários sentidos dessa palavra. Em um *thesaurus*, a entrada principal é um grupo de palavras relacionadas. Em um léxico, a entrada principal é o sentido, que pode corresponder a várias palavras.

Os trabalhos que extraem conhecimento pré-codificado de diferentes fontes normalmente empregam alguma técnica estatística para a utilização desse conhecimento, que pode variar desde uma freqüência relativa a cálculos mais complexos. Por essa razão, alguns autores (Manning & Schütze, 2001, por exemplo) classificam esses trabalhos como "estatísticos". Contudo, vale notar que neste trabalho são consideradas abordagens estatísticas aquelas que adquirem conhecimento a partir de exemplos em um córpus e representam esse conhecimento por meio de modelos estatísticos (Seção 2.6.2).

[2] Esse problema ficou conhecido como "o gargalo da aquisição de conhecimento" (Gale et al., 1993).

Léxicos computacionais

Os **léxicos computacionais** são recursos lexicais criados (em geral, manualmente) especificamente para o tratamento computacional. São também chamadas de **bases de dados lexicais** (*Lexical Databases*). Exemplos de léxicos computacionais incluem a WordNet (Miller et al., 1990)[3], para o inglês, o ACQUILEX (Briscoe, 1991), para o inglês, o italiano, o espanhol e o holandês, o SIMPLE (Lenci et al., 1999), para o italiano, e a DIADORIM (Greghi et al., 2001), para o português do Brasil.

Os léxicos computacionais visam permitir a representação e manipulação de diversos tipos de informação sobre cada item lexical. Sua estrutura permite que se estabeleçam conexões tanto entre itens lexicais distintos quanto entre características que pertençam a itens distintos. Assim, segundo Correia (1996), um léxico computacional pode ser visto como uma complexa rede de relações (morfológicas, sintagmáticas, semânticas e paradigmáticas), na qual o conhecimento sobre um item lexical é composto por vários níveis ou camadas. O objetivo principal de um léxico computacional, conforme mencionado, é servir a sistemas de PLN, contudo, ele pode facilmente ser compreendido e manipulado por humanos, como é o caso da WordNet, amplamente utilizada tanto em tarefas computacionais como em consultas por humanos.

A construção desses léxicos segue, fundamentalmente, duas propostas: a abordagem **enumerativa**, na qual todos os sentidos de uma palavra são explicitamente fornecidos como diferentes entradas, e a abordagem **gerativa**, na qual as informações semânticas são subespecificadas e as informações de sentido são derivadas a partir de regras geração. Assim, na abordagem gerativa, normalmente há apenas uma entrada para cada palavra e os seus vários sentidos são derivados dessa entrada por meio das operações gerativas.

Entre os léxicos enumerativos, a WordNet é o recurso mais utilizado para DLS do inglês. Adicionalmente às definições e exemplos de uso para os diversos sentidos de cada palavra, esse recurso especifica relações semânticas entre tais palavras e entre grupos de palavras, incluindo a sinonímia, a antonímia, a meronímia e a hiponímia, entre outras. A sinonímia é a principal relação entre palavras da WordNet: grupos de palavras sinônimas são organizadas em conjuntos denominados *synsets*. Cada *synset* representa um conceito cujo sentido é válido para todas as palavras do conjunto. Por exemplo, o *synset* {*plant*, *flora*, *plant life*} representa o conceito que é compartilhado pelas três palavras. Esse conceito é diferenciado dos demais por meio de um código.

Um *synset* contém, então, além de todas as formas de palavras que podem se referir a um dado conceito, uma glosa e, na maioria dos casos, sentenças de exemplo. Palavras e *synsets* são inter-relacionados por meio de ligações lexicais e semântico-conceituais, respectivamente. Por exemplo, a sinonímia e antonímia ligam palavras individuais, enquanto a hiponímia e a hiperonimia ligam *synsets*. Com isso, a WordNet define também uma organização conceitual hierárquica, como um *thesaurus*.

Como pode ser verificado, a WordNet provê uma série de informações necessárias para a DLS em um único recurso. Nos trabalhos que utilizam esse recurso, em geral, os *synsets* definem o repositório de sentidos: cada *synset* é um sentido. Contudo, alguns trabalhos utilizam grupos de *synsets* mais genéricos como sentido, por exemplo, sub-hierarquias de *synsets*, definidas de acordo com algum critério, ou os grupos de *synsets* definidos em diferentes categorias semânticas genéricas.

[3] Vale notar que alguns autores, como Palmer (1998), não consideram a WordNet como um léxico computacional, mas como um "recurso lexical *on-line*", já que sua criação, segundo a autora, não foi primordialmente voltada para o seu uso por programas computacionais, mas para o uso por humanos.

A exploração de léxicos gerativos para a DLS, por outro lado, é pouco comum. Segundo Pedersen (1997), o uso de léxicos gerativos como o de Pustejovsky não é operacional para a DLS, por conta da dificuldade em se definir adequadamente as estruturas lexicais, principalmente a sua estrutura *qualia*, e da possibilidade de geração de sentidos desnecessários. Para Kilgarriff (1997a), pouco acréscimo será feito à cobertura do sistema se ele explorar princípios gerativos para derivar novos sentidos para uma palavra. Os usos não padrão da palavra tendem a ter uma história particular, específica, com um uso não-padrão derivando outros usos não-padrões, com conexões muito específicas à palavra ou a um campo lexical, altamente complexas e desnecessárias para a tarefa de DLS. O autor afirma que os métodos práticos para estender a cobertura de aplicações de PLN para usos não comuns não geram novos significados, mas sim uma lista deles.

Uma distinção importante, ressaltada por Kilgarriff (1995), das bases de dados lexicais com relação aos dicionários eletrônicos, é que apesar de ambos compartilharem do objetivo de descrever o recurso lexical de uma determinada língua (ou línguas), os dicionários são criados para utilização por humanos e, por isso, podem apresentar algumas inconsistências e, principalmente, omissões sistemáticas de sentidos, ou seja, omissão daqueles sentidos que podem ser facilmente inferidas pelo ser humano a partir do sentido básico de uma palavra. Nas bases de dados lexicais, por outro lado, essas omissões devem ser representadas ou inferidas por mecanismos gerativos.

Dicionários Eletrônicos

Os **dicionários eletrônicos** constituem fontes de informação para a desambiguação extraídas automaticamente a partir de recursos não eletrônicos já existentes, criados com outras finalidades, normalmente para uso por seres humanos. Normalmente, esses recursos são as versões eletrônicas de dicionários em papel.

Esses dicionários, também chamados de **dicionários de máquina**, normalmente são divididos em Dicionários Legíveis por Máquina (MRDs – *Machine-Readable Dictionaries*) e Dicionários Tratáveis por Máquina (MTDs – *Machine-Tractable Dictionaries*) (cf. Correia, 1996). Os MRDs são os dicionários que possuem uma versão publicada em papel e outra eletrônica, são elaborados por lexicógrafos e concebidos para uso por humanos. Normalmente, não são passíveis de serem manipulados por sistemas de PLN. Já os MTDs, conforme definido por Wilks et al. (1990), são aqueles que possuem as informações de um MRD, mas transformadas para um formato apropriado para utilização em sistemas de PLN. Para tanto, a descrição das informações é feita por meio de um formalismo reconhecível pelo computador, ou seja, uma linguagem que traduz o conhecimento dos dicionários criados para uso por humanos em uma descrição formal para uso por máquinas. Contudo, nem sempre essas variações são consideradas. Normalmente, quando se trata de sistemas de DLS, usa-se o termo MRD, de modo geral, para designar dicionários que são utilizados como fonte de informações. Neste trabalho, igualmente é utilizado um termo genérico para esse tipo de recurso, denominado **dicionário eletrônico**.

Desde o surgimento de dicionários eletrônicos, diversas técnicas e ferramentas têm sido desenvolvidas para a extração automática de informações dessas fontes. Os dicionários eletrônicos provêem informações de grande importância sobre os sentidos das palavras. Por essa razão, diversos trabalhos utilizam esse recurso para a tarefa de DLS.

Thesauri

Assim como os dicionários eletrônicos, os ***thesauri*** usados na DLS também consistem, geralmente, de versões eletrônicas de algum *thesaurus* em papel. Os ***thesauri*** contêm

informações sobre as categorias semânticas das palavras ou conceitos, bem como sobre as relações semânticas entre as palavras, como a sinonímia, a antonímia e a hiponímia, entre outras. A suposição dos autores que empregam esse recurso como fonte de conhecimento é de que as categorias semânticas de uma palavra em um contexto determinam a categoria semântica do contexto como um todo e essa categoria, por sua vez, determina quais sentidos das palavras são usados.

O primeiro thesaurus eletrônico é o *Roget International Thesaurus*, criado na década de 1950 e utilizado, a partir de então, em diversas aplicações, incluindo a DLS. Esse *thesaurus* também fornece uma hierarquia conceitual explícita, consistindo de até oito níveis de refinamento. Normalmente, em um thesaurus, cada ocorrência de uma palavra sob diferentes categorias representa um sentido diferente para aquela palavra. Assim, as categorias correspondem, aproximadamente, aos sentidos, e as palavras da mesma categoria são semanticamente relacionadas.

A principal vantagem da utilização de bases de conhecimento pré-codificado (léxicos computacionais, dicionários e *thesauri*) para a DLS é o fato de que não é necessário codificar todo o conhecimento manualmente. Contudo, essas bases também apresentam alguns problemas.

Um dos maiores problemas, principalmente no caso das bases extraídas automaticamente a partir de fontes de informação não eletrônica, é que pode haver inconsistências nos dados eletrônicos, uma vez que os recursos são criados para uso por humanos, e não por computadores.

Outro problema é que a maioria desses recursos não dispõe de informações pragmáticas, uma vez que se espera, novamente, que seres humanos possam realizar interpretações nesse nível do conhecimento. Em se tratando de *thesauri*, especificamente, segundo Ide & Véronis (1998), apesar da riqueza de informações que podem ser fornecidas, essa fonte não tem se mostrado adequada e, conseqüentemente, não tem sido usada extensivamente na DLS. Um dos motivos é que mesmo a definição dos níveis mais altos das hierarquias conceituais em *thesauri* gera discordâncias entre as várias propostas, um fato que tende a piorar à medida que os sentidos são mais refinados.

Uma característica dos recursos lexicais, principalmente dos dicionários eletrônicos, que pode representar um problema é o fato de eles serem normalmente de domínio genérico e, portanto, de pouca validade para a DLS em domínios específicos. Recursos específicos normalmente são criados por cada trabalho de DLS.

Uma limitação da utilização de quaisquer bases de conhecimento pré-codificado que diz respeito somente às aplicações multilingües é o fato de que existem poucas bases multilingües com informações suficientes para a DLS, em especial, contemplando a língua portuguesa.

2.6.2 Método baseado em córpus

Nos últimos anos, com os avanços na área de Aprendizado de Máquina (AM), tem crescido no PLN a utilização de métodos que permitem extrair conhecimento automaticamente a partir de córpus, visando minimizar o problema do gargalo da aquisição de conhecimento. Um córpus provê um conjunto de exemplos que, quando submetidos a algoritmos de AM, permitem o desenvolvimento de modelos capazes de descrever esses exemplos e de predizer o comportamento de novos exemplos. Os trabalhos baseados em córpus, também chamadas de **empíricos,** realizam a desambiguação, portanto, com o uso de informações obtidas automaticamente a partir de um córpus. Nesses trabalhos, normalmente há uma etapa de treinamento, que resulta no aprendizado do modelo de desambiguação, com base no córpus. Após o treinamento, o modelo de desambiguação é gerado (com exceção dos

trabalhos baseados em instâncias) e pode ser usado para desambiguar novos casos de ambigüidades.

Alguns trabalhos de DLS tidos como baseados em conhecimento (manualmente codificado ou pré-codificado) utilizam córpus, mais especificamente, técnicas da lingüística de córpus para extrair automaticamente algum conhecimento útil a ser utilizado no modelo de DLS, como é o caso de McRoy (1992) (Seção 3.1.1), que extrai de um córpus um léxico de *collocations*. Contudo, neste trabalho, são considerados trabalhos que seguem o método baseado em córpus apenas aqueles cujo processo de desambiguação é automaticamente "aprendido" a partir de córpus, por meio de técnicas de aprendizado de máquina. Excluem-se dessa classificação, portanto, os trabalhos que apenas aplicam estratégias baseadas em córpus para adquirir conhecimento como suporte ao processo de desambiguação. Excluem-se, ainda, trabalhos que mapeiam o conhecimento do córpus em regras de desambiguação, mas sem utilizar técnicas de aprendizado de máquina (por exemplo, Brun, 2000).

De modo geral, as principais vantagens dos trabalhos baseados em córpus são: (1) não é necessário codificar todo o conhecimento manualmente; (2) podem ser utilizados algoritmos tradicionais de AM, já implementados e disponíveis, para a aquisição automática ou semi-automática desse conhecimento; (3) os modelos criados podem ser mais facilmente generalizáveis para outros gêneros e/ou domínios de textos e, com isso, aplicáveis em larga escala; e (4) os modelos gerados podem expressar algum conhecimento novo sobre o uso da palavra ambígua ou sobre as informações utilizadas para a distinção entre os seus sentidos.

Por outro lado, esses trabalhos também apresentam problemas: (1) os córpus utilizados para a criação do modelo precisam ser representativos da língua natural para o gênero e/ou domínio em questão. Apesar da atual disponibilidade de vários córpus de tamanho considerável, nem sempre eles apresentam as informações necessárias (como as etiquetas gramaticais das palavras); (2) no caso de trabalhos supervisionados, nas quais são necessários exemplos com as etiquetas de sentido das palavras ambíguas, o esforço para a criação do córpus é ainda maior, pois normalmente é preciso identificar essas etiquetas manualmente; e (3) não há garantia de que os resultados serão adequados, em função de várias características do processo automático de aprendizado, como a possibilidade de inconsistências nos exemplos (ruídos), a dificuldade de avaliação do modelo gerado.

Em se tratando de DLS para aplicações multilingües, como a TA, o maior problema é que praticamente não há córpus adequados disponíveis, uma vez que os córpus suficientemente abrangentes e com as informações necessárias (etiquetas de sentido, no caso dos trabalhos supervisionados, ou anotações de outras naturezas, como a gramatical, no caso de trabalhos não-supervisionados) são monolingües.

2.6.2.1 Modos de aprendizado

O córpus de exemplos para o aprendizado do modelo de DLS pode ser anotado (rotulado, etiquetado) ou não-anotado. Em um córpus **anotado**, os exemplos possuem etiquetas de sentido, normalmente atribuídas com base no conjunto de sentidos de um dicionário ou outro recurso lexical. Esse córpus já é, portanto, desambiguado, e os trabalhos baseados nele seguem o modo de aprendizado supervisionado (**DLS supervisionada**). Em um córpus **não-anotado**, os exemplos não possuem nenhuma forma de anotação de sentido, ou seja, o córpus não é desambiguado. Os trabalhos baseados nesse tipo de córpus seguem o modo de aprendizado não-supervisionado (**DLS não-supervisionada**).

Todos os algoritmos já empregados para a DLS requerem, em geral, que as características (ou atributos) dos exemplos sejam previamente extraídas e convertidas para

o formato atributo-valor e estruturadas em vetores de características, como é comum na maioria das aplicações de aprendizado de máquina. Esses vetores possuem o formato (([característica$_1$][valor$_1$])...([característica$_n$][valor$_n$])), com a lista de características ([característica$_i$]) e seus valores ([valor$_i$]). As características correspondem à representação dos diferentes tipos de conhecimento citados na Seção 2.5. Conforme mencionado, podem-se considerar as mesmas características para todas as palavras ambíguas (incluindo a palavra ambígua) ou definir características específicas para cada palavra.

Principalmente nos modelos genéricos, válidos para todas as palavras ambíguas, nem todas as características são relevantes para todas as palavras. Assim, podem ser empregados métodos da área de AM para a seleção automática das características mais importantes, a partir de um conjunto de características dadas. Esses métodos podem, por exemplo, verificar a freqüência de ocorrência das características nos exemplos de treinamento para determinar sua relevância.

Conforme observam Manning & Schütze (2001), a DLS supervisionada, no contexto das possíveis tarefas supervisionadas de AM (Figura 2), corresponde à tarefa de **classificação**, ou seja, de criação de um modelo que possa identificar (prever) a classe (sentido) mais adequada para novos casos de ambigüidade com base em exemplos já classificados. Já no caso da DLS não-supervisionada, a tarefa de DLS, dentre as tarefas não-supervisionadas de AM, pode ser considerada uma tarefa de ***clustering***, ou seja, de descrição dos exemplos de acordo com a similaridade (ou dissimilaridade) entre eles. Contudo, como ressalta Ng & Zelle (1997), trabalhos não-supervisionados de DLS não necessariamente se referem à tarefa de *clustering*.

Vale lembrar, aqui, a distinção já mencionada entre **desambiguação de sentidos** e **discriminação de sentidos**, ressaltada por Manning & Schütze (2001). A tarefa supervisionada provê a desambiguação de sentidos, uma vez que um sentido é escolhido e realmente atribuído à ocorrência de uma palavra. Já a tarefa de *clustering* provê apenas a discriminação ou indução de sentidos. Como enfatiza Yarowsky (1995), a tarefa de *clustering* consiste do uso de medidas de similaridade para particionar os exemplos de palavras em grupos, os quais não necessariamente recebem alguma atribuição de sentido e, ainda, podem não ter relação alguma com distinções de sentido padrão.

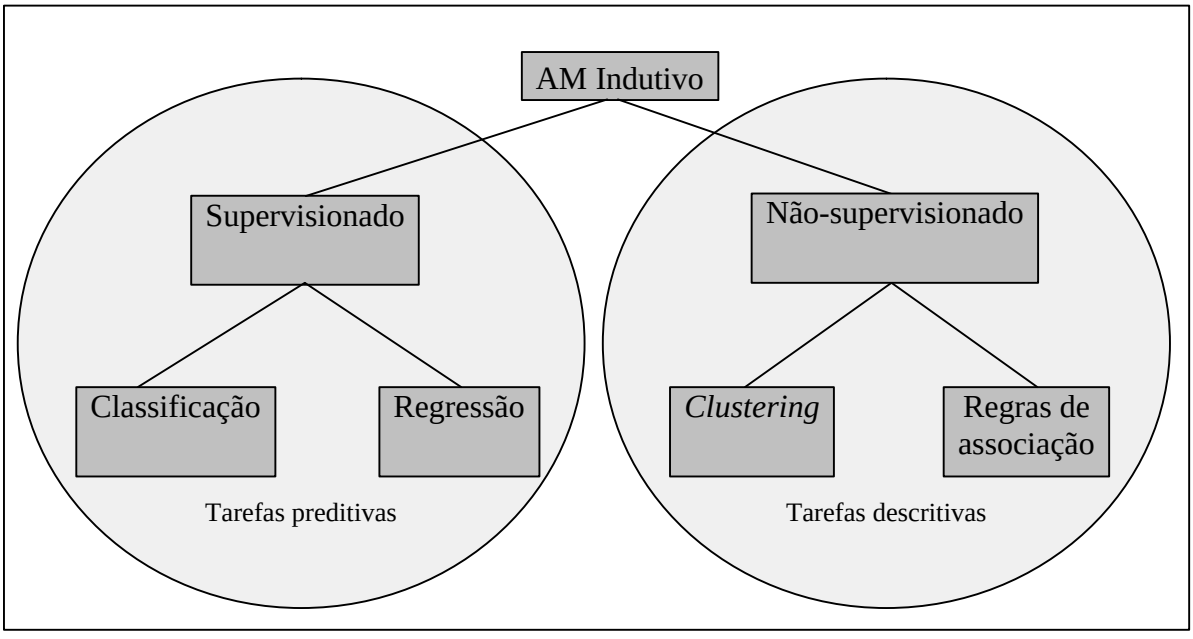

Figura 2. Modos e tarefas de aprendizado em AM (Monard & Baranauskas, 2003, p. 91)

DLS Supervisionada

Nos trabalhos baseados no modo supervisionado, um conjunto pré-definido de sentidos é especificado e cada exemplo do córpus é etiquetado com um desses sentidos. Normalmente, os exemplos desse córpus são constituídos das características de sentenças (ou textos) cujas palavras são semanticamente ambíguas e das etiquetas de sentido correspondentes a cada palavra ambígua como uma das características, podendo apresentar também características de outras naturezas. Os algoritmos de aprendizado supervisionado devem, a partir desse córpus, gerar um modelo para tentar prever o sentido de novos casos de ambigüidade, não anotados, dentre os possíveis sentidos pré-definidos.

A principal vantagem dos trabalhos supervisionados é o fato de que os sentidos podem ser especificados previamente, provendo uma etiquetação mais adequada e refinada. Com isso, é possível criar modelos mais eficientes para a DLS, principalmente para aplicações multilingües. As duas edições do exercício de avaliação específico para a área de DLS (SENSEVAL), que será descrito na Seção 2.8.2, comprovam que as propostas supervisionadas são as que apresentam o melhor desempenho tanto para a tarefa de desambiguação de um pequeno grupo de palavras, quanto para a tarefa de desambiguação de todas as palavras da sentença (etiquetação de sentidos).

O problema com esses trabalhos é a necessidade de etiquetação de um córpus de treinamento, normalmente feita manualmente, por humanos. Esse problema acaba por restringir a abrangência de muitos trabalhos a poucas palavras, pois não há, ainda, córpus representativos com etiquetas de sentido, visando a uma ampla utilização para a DLS. Isso ocorre, muito provavelmente, porque o mesmo córpus não pode ser usado em diferentes aplicações que envolvam a DLS, uma vez que o nível de refinamento da etiquetação, bem como as etiquetas de sentido, propriamente ditas, dependem da aplicação e, em alguns casos, do domínio dessa aplicação.

Com relação ao domínio, Wilks & Stevenson (1997a) mencionam que um córpus etiquetado criado para um domínio específico (por exemplo, notícias de esporte) dificilmente poderá ser utilizado para a geração de modelos em domínios distintos (por exemplo, produtos eletrônicos). A solução, neste caso, parece ser a criação de córpus mais genéricos, cobrindo satisfatoriamente (ainda que não perfeitamente) diferentes domínios.

O compartilhamento de córpus entre aplicações é ainda mais complexo. Em aplicações monolingües, como a Recuperação de Informações, a etiqueta corresponde ao sentido, na língua do texto de origem, que distingue a ocorrência de cada palavra ambígua. Em aplicações multilingües, como a TA, por outro lado, a etiqueta corresponde à tradução, em outra língua, que distingue cada tradução da palavra ambígua. Em especial, em se tratando de tradução inglês-português, foco deste trabalho, não há córpus com etiquetas de sentidos disponíveis.

Com isso, cada aplicação acaba criando seu próprio córpus. Uma exceção significativa é o córpus DSO, descrito na Seção 2.7.1, que vem sendo usado em várias aplicações de DLS monolingüe. Nessa seção, serão descritos também vários esforços que têm sido empregados para a etiquetação automática de sentidos em córpus.

DLS Não-supervisionada

Nos trabalhos que utilizam métodos não-supervisionados, os exemplos do córpus consistem de características de sentenças (ou textos) cujas palavras são semanticamente ambíguas, podendo apresentar anotação de outras naturezas que não a de sentido, por exemplo, etiquetas gramaticais. Dessa maneira, não há sentidos específicos pré-definidos

para uma palavra. Os algoritmos procuram identificar grupos de sentidos similares e, diante de uma nova palavra, verificar a qual dos grupos ela mais se assemelha.

De modo geral, a grande vantagem dos trabalhos não-supervisionados é o fato de que não há necessidade de etiquetação do córpus de exemplos. Por outro lado, uma vez que a desambiguação não é realizada com respeito a um conjunto pré-definido de sentidos, essa abordagem pode não ser apropriada para uma série de aplicações, principalmente para aquelas que necessitam de uma explicitação dos sentidos. O conjunto de *clusters* e, com isso, o grau de generalidade/especialização dos sentidos, depende muito do algoritmo usado e dos seus parâmetros. Os grupos obtidos podem ou não corresponder a diferentes níveis de uma hierarquia lexical padrão. Além disso, nem sempre os *clusters* de sentidos gerados são bem definidos.

Segundo Wilks & Stevenson (1997a), é difícil visualizar o que pode ser feito com os grupos de sentidos identificados pelos trabalhos não-supervisionados. Para a aplicação na TA, em especial, esses trabalhos dificilmente poderiam ser utilizados, pois como os sentidos correspondem às traduções, eles precisam ser explicitados. Certamente, isso poderia ser feito por meio de pós-edição humana, como ocorre no trabalho de Schütze (1992). Contudo, ainda assim não haveria garantia alguma de que os grupos formados corresponderiam a distinções de sentido válidas.

Manning & Schütze (2001) enfatizam o problema de que não há garantia de que os sentidos identificados distinguem adequadamente as ocorrências das palavras ambíguas, uma vez que eles podem ser muito genéricos ou muito refinados, ou, ainda, não corresponder às distinções padrão. Segundo Wilks (1997), nesses trabalhos sempre haverá a atribuição das palavras ambíguas a algum grupo, de modo que a DLS sempre ocorre. O problema é como interpretar os resultados.

DLS Semi-supervisionada

Alguns autores mencionam um terceiro modo de aprendizado, denominado **semi-supervisionado** ou **fracamente supervisionado**. Neste trabalho, são consideradas semi-supervisionadas diversas variações desse modo de aprendizado, incluindo o co-treinamento (*co-training*), o autotreinamento (*self-training*) e o aprendizado ativo (*active learning*), tanto para a tarefa de classificação quanto para a tarefa de *clustering*. Nos trabalhos que seguem esse modo de aprendizado, o processo é iterativo: parte-se de um córpus com apenas alguns exemplos manualmente rotulados, que são usados para treinar o sistema para etiquetar novos casos. O objetivo é aumentar o número de exemplos de treinamento, adicionando os casos etiquetados a esse conjunto de exemplos, contudo, respeitando-se um limite mínimo de confiabilidade para a classe (ou *cluster*) atribuída. A verificação desse limite pode ser feita de diferentes maneiras. Nos trabalhos de co-treinamento para a classificação, por exemplo, são usados dois ou mais classificadores, treinados com diferentes visões (conjuntos disjuntos de características, por exemplo), sendo que todos eles (ou a maioria deles) precisam concordar com a classe atribuída para que o exemplo seja inserido no conjunto de treinamento. Esse processo, denominado muitas vezes de *bootstrapping*, pode se repetir várias vezes, de modo a aumentar cada vez mais o número de exemplos de treinamento. Em alguns trabalhos, o objetivo é etiquetar a maior quantidade possível de exemplos. Em outros, o mais importante é garantir a confiabilidade na etiquetação dos novos casos.

Nos algoritmos de *clustering*, o aprendizado semi-supervisionado também consiste em partir de um córpus de treinamento pequeno, com os exemplos já agrupados nos possíveis *clusters*, para atribuir outros exemplos, não rotulados, aos *clusters*. São usadas medidas de similaridade para identificar a proximidade entre os exemplos já anotados e os

casos não anotados. Novamente, novos casos só são atribuídos a algum *cluster* se essa proximidade atinge um limite mínimo pré-estabelecido.

Analisando-se o processo de aprendizado, tanto na tarefa de classificação quanto na de *clustering*, contudo, pode-se perceber que o treinamento ocorre sempre a partir de exemplos etiquetados, seja nas etapas intermediárias ou na etapa final, quando a etiquetação de todo o córpus de exemplos estiver concluída. Assim, os trabalhos que utilizam essas técnicas são consideradas, aqui, como supervisionados.

DLS supervisionada com córpus artificial

EAGLES (1998) fazem uma distinção mais refinada sobre os córpus anotados no contexto de DLS: um **córpus desambiguado** designa um córpus cujos exemplos possuem etiquetas de sentido manualmente atribuídas, enquanto um **córpus artificial** designa um córpus cujos exemplos também possuem etiquetas de sentido, mas que é criado artificialmente para a DLS. O grupo cita dois casos de córpus com essas características: (a) córpus cujas etiquetas de sentido são automaticamente identificadas a partir de córpus multilingües paralelos; e (b) córpus cujos casos de ambigüidade a serem manipulados são propositadamente inseridos por meio de pseudopalavras (*pseudo-words*). Pseudopalavras são duas ou mais palavras, normalmente sem relação alguma de ambigüidade, que são agrupadas e consideradas com uma única palavra ambígua. Todas as ocorrências de cada uma dessas palavras no texto são substituídas pela pseudopalavra, de modo que o córpus se torna ambíguo. Por exemplo, pode ser criada a pseudopalavra *author-baby*. Todas as ocorrências de *author* e *baby* no córpus são então substituídas por *author-baby* e a meta é desambiguar as ocorrências de *author-baby* como *author* ou *baby*. A resposta correta para avaliar o sistema é o texto original, antes da substituição das pseudopalavras.

Sanderson (1994), por exemplo, utiliza pseudopalavras de diferentes tamanhos (2, 3, 4, 5 e 10) para analisar a influência da ambigüidade lexical na precisão da recuperação de informações, bem como avaliar os efeitos da resolução dessa ambigüidade, usando um mecanismo de DLS, na precisão da recuperação. Trabalhos que usam pseudopalavras são válidos apenas em contextos como esse, ou seja, para testar métodos ou técnicas de DLS, já que eliminam a necessidade de etiquetação manual do córpus. Contudo, não podem ser aplicados a sistemas reais. Não há consenso sobre como os resultados da desambiguação devem ser interpretados, ou seja, se eles podem ser generalizados para os casos de ambigüidades reais.

Os córpus criados a partir de textos paralelos, por outro lado, possuem, em principio, as mesmas características que os córpus anotados manualmente. O que diferencia esses córpus, para o grupo EAGLES, portanto, é o seu processo de produção. Neste trabalho, tal distinção não é relevante. Assim, um córpus anotado poderá ter sido criado manualmente ou a partir de córpus paralelos. O termo córpus artificial será utilizado apenas para os córpus com pseudopalavras.

2.6.2.2 Paradigmas de aprendizado

Além da distinção com relação ao modo de aprendizado, decorrente do tipo de córpus utilizado, há variações nos trabalhos baseados em córpus no que diz respeito ao paradigma de aprendizado empregado. Os principais paradigmas são: **simbólico**, **estatístico**, **conexionista** ou **baseado em instâncias** (*instance-based* ou também chamados, no inglês, *case-based*, *memory-based* ou *exemplar-based*). A utilização de determinado paradigma diz respeito à linguagem de descrição do modelo gerado (em paradigmas que geram modelos).

Na tarefa de classificação, por exemplo, os trabalhos simbólicos dão origem a modelos simbólicos de DLS, como árvores de decisão, regras de decisão ou listas de decisão. Os trabalhos sob o paradigma estatístico, por sua vez, dão origem a modelos matemáticos, normalmente baseados em probabilidades, cujos parâmetros são estimados a partir dos exemplos de treinamento. Por fim, os trabalhos conexionistas dão origem a modelos matemáticos de redes neurais, cujos parâmetros também são estimados partir dos exemplos de treinamento. Trabalhos baseados em instâncias, por outro lado, não geram modelos de classificação, em vez disso, analisam toda a base de exemplos, sempre que um novo caso precisa ser classificado, para verificar qual o exemplo mais próximo e atribuir a classe sua ao novo exemplo.

No aprendizado não-supervisionado, normalmente, o modelo consiste de *clusters* de exemplos, agrupados de acordo com sua similaridade, com um dos exemplos (ou uma abstração dos vários exemplos) representando as características mais determinantes daquele grupo (por exemplo, o centróide de um *cluster*).

Algoritmos tradicionais de AM dos diferentes paradigmas podem ser empregados nos trabalhos de DLS baseados em córpus, por exemplo, C4.5 e C4.5rules (Quinlan, 1988), variações do KNN (*K-Nearest Neighbor*), Naive Bayes (Duda & Hart, 1973), etc.

Uma vantagem dos trabalhos simbólicos com relação aos desenvolvidos sob os demais paradigmas, relevante para este trabalho, é o fato de que o conhecimento adquirido automaticamente é facilmente compreensível por seres humanos, mesmo que eles não sejam especialistas no domínio em questão (a desambiguação, neste caso). Assim, o modelo gerado pode ser ajustado, seja pela inclusão/exclusão de exemplos de treinamento, seja pela inclusão/exclusão de conhecimento explícito (por exemplo, a inclusão/exclusão de regras de decisão). Certamente, como ocorre nos vários domínios de aplicação do AM, o tamanho do modelo gerado pode ser demasiadamente grande, dificultando sua compreensão e avaliação.

2.6.3 Método híbrido

Trabalhos híbridos para a DLS combinam características dos métodos baseados em conhecimento (codificado manualmente ou pré-codificado) e em córpus, podendo seguir modos supervisionados e não-supervisionados e empregar diferentes paradigmas de aprendizado. Os tipos de conhecimento utilizados, bem como a interação entre esse conhecimento e o AM podem variar.

As vantagens dos trabalhos que seguem o método híbrido, em teoria, correspondem às vantagens dos trabalhos baseados em conhecimento e em córpus. Contudo, essa relação pode não ser tão direta: é preciso encontrar uma maneira adequada para combinar as características de ambos os métodos, de modo a minimizar seus problemas.

2.7 Criação de córpus para trabalhos supervisionados

Para a criação de propostas baseadas em córpus, é preciso produzir um córpus de treinamento substancial. Conforme mencionado, os trabalhos dessa natureza que obtêm os melhores resultados são as de aprendizado supervisionado, que utilizam córpus anotados com sentidos, o que dificulta ainda mais a tarefa de criação do córpus. Para a desambiguação monolingüe do inglês, já há alguns córpus com esse tipo de anotação. Já para a desambiguação em outras línguas e, principalmente, para a desambiguação multilingüe, apenas recentemente começaram a ser desenvolvidos trabalhos de etiquetação de córpus.

Os córpus podem ser criados manual ou automaticamente. Aqui, é importante lembrar a distinção feita anteriormente entre as tarefas de DLS e a etiquetação de sentidos (Seção 2.1). As propostas de criação automática de córpus para a DLS realizam a etiquetação de sentidos. Apesar de, novamente, alguns autores considerarem essa tarefa como a DLS, propriamente dita, neste trabalho, elas são consideradas separadamente, já que o seu objetivo é o de criação de córpus para o uso posterior em algum trabalho de DLS supervisionada

A seguir, são brevemente apresentados alguns córpus criados manualmente comumente utilizados pelos trabalhos descritos na Seção 3. Na seqüência, são apresentadas algumas propostas de criação automática de córpus e um problema que deve ser considerado quando da criação de córpus: a falta ou insuficiência de exemplos para determinados sentidos.

2.7.1 Córpus etiquetados manualmente

Os principais exemplos de córpus já disponíveis e que são comumente utilizados para o treinamento e avaliação de trabalhos de DLS são os córpus DSO (Ng & Lee, 1996) e SEMCOR (Miller et al., 1994). Ambos os córpus foram criados manualmente, para a desambiguação monolingüe, utilizando os sentidos da WordNet.

O maior e mais significativo desses córpus é o DSO. Ele consiste de 192.800 sentenças de exemplo contendo 192.874 ocorrências dos 121 substantivos e 70 verbos mais freqüentes da língua inglesa, extraídas do córpus Brown (Francis & Kucera, 1979) e de um córpus de artigos do *Wall Street Journal*. Em média, cada verbo considerado possui 12 sentidos, enquanto cada substantivo possui 7.8 sentidos. Para cada palavra, foram extraídos até 1.500 exemplos. O processo de etiquetação manual do córpus consumiu um ano. Apesar da marcação manual, os autores estimam que o córpus apresenta de 10 a 20% de etiquetas que podem ser consideradas, por outros anotadores, como erros. Segundo os autores, esse córpus é bastante representativo, pois 191 palavras correspondem a cerca de 20% de todas as ocorrências de palavras em qualquer texto do inglês.

O córpus SEMCOR (Miller et al., 1994) consiste de um subconjunto do córpus Brown com cerca de 200.000 palavras, das quais as palavras de conteúdo foram manualmente etiquetadas com os sentidos da WordNet. Outros córpus menores, mas também utilizados principalmente para a avaliação de sistemas, incluem os córpus criados para determinados trabalhos de DLS e disponibilizados para uso. Por exemplo, Leacock et al. (1993) e Bruce & Wiebe (1994), cada um com pouco mais de 2.000 sentenças de exemplos com seis diferentes sentidos da palavra *line* e *interest*, respectivamente. Outros exemplos são os córpus usados nas três edições do exercício de avaliação SENSEVAL (Seção 2.8.2). Com exceção da primeira edição, os demais córpus são baseados nos sentidos da WordNet.

Contudo, como afirma Ng (1997b), esses córpus, incluindo o DSO, são ainda muito pequenos para serem utilizados para a criação de trabalhos irrestritos de DLS. Com base no DSO, o autor examina o efeito do tamanho do córpus de treinamento, em termos do número de exemplos. Para tanto, ele propõe um trabalho baseado em instâncias e realiza testes com vários subconjuntos do córpus, de modo a obter as curvas de aprendizado nesse córpus. Os resultados do experimento mostram que a precisão aumenta à medida que o número de exemplos do córpus cresce e que todos os exemplos do córpus são efetivamente utilizados pelo algoritmo empregado.

Como conclusão desses experimentos, o autor estima que um córpus de 3.200 palavras diferentes etiquetadas com seus sentidos é suficiente para construir um sistema de DLS de ampla cobertura e alta precisão, considerando-se qualquer palavra de conteúdo, em

textos irrestritos da língua inglesa. Assumindo uma média de 1.000 ocorrências etiquetadas por sentido por palavra, isso significa um córpus de 3.2 milhões de palavras etiquetadas. Com base na sua experiência com a criação do DSO, segundo o autor, a criação manual desse córpus demandaria um tempo de 16 anos, considerando-se o esforço de um etiquetador humano. O autor sugere, como alternativa para minimizar o esforço de criação de córpus, o uso de técnicas de seleção de exemplos informativos, evitando a anotação redundante.

Focalizando a importância da seleção de exemplos relevantes na construção de córpus para a DLS, Fujii et al. (1998) empregam um método de amostragem seletiva de exemplos, de acordo com sua utilidade para o treinamento de um sistema de DLS. Eles desenvolvem um trabalho baseado em instâncias para a construção semi-automatizada do córpus. Apenas a desambiguação de um conjunto de verbos é contemplada.

O método apresentado pelos autores é de construção iterativa da base de exemplos de treinamento simula, em parte, o aprendizado semi-supervisionado. É necessário um número inicial mínimo de exemplos manualmente desambiguados e um conjunto de exemplos não desambiguados de qualquer tamanho. Cada novo exemplo é submetido ao sistema, que atribui ao seu verbo uma etiqueta de sentido. Esse exemplo é então analisado pelo método de amostragem seletiva, para determinar a sua utilidade para o treinamento, com base (a) no conjunto de exemplos não desambiguados, analisando-se a quantidade desses exemplos que se assemelha a ele, de modo que ele possa cobrir um grande número de novos casos; e (b) no conjunto de exemplos já desambiguados e pertencentes à base de exemplos de treinamento do sistema, analisando-se a sua diferença com relação a esses exemplos, de modo a evitar exemplos redundantes. Os exemplos selecionados por sua utilidade são então submetidos à revisão e/ou correção humana da etiquetação de sentido realizada pelo sistema e, em seguida, acrescentados à base de exemplos de treinamento. Em uma nova iteração, esse exemplo já é utilizado pelo sistema. Os exemplos não selecionados retornam para a base de exemplos não desambiguados.

Outra alternativa para o problema da etiquetação manual que tem sido investigada ultimamente se mostra viável principalmente na TA é a etiquetação completamente automática dos sentidos dos exemplos.

2.7.2 Córpus etiquetados automaticamente

Segundo Agirre & Martínez (2004), a criação automática de corpus é uma das estratégias mais indicadas para minimizar o problema do gargalo da aquisição do conhecimento, contudo, é ainda muito pouco explorada. Segundo Dagan & Itai (1994), além de permitir a aquisição de córpus mais representativos, a etiquetação automática permite capturar distinções diferentes das que seriam atribuídas por um anotador humano, por exemplo, distinções específicas de algum domínio ou pouco comuns. Alguns dos trabalhos recentes de criação automática de córpus etiquetados são descritos a seguir.

Continuando o trabalho de exploração de córpus paralelos para a identificação dos sentidos a serem utilizadas na DLS monolingüe iniciado anteriormente (Ide, 1999, Seção 2.4.3), Ide et al. (2001; 2002) realizam experimentos mais significativos, estendendo o número de línguas (para sete), aumentando o número de palavras ambíguas e o tamanho dos córpus paralelos. Um algoritmo de *clustering* é utilizado para criar grupos de sentidos de acordo com as diferentes traduções de cada palavra do inglês, nas diferentes línguas. As distinções de sentido são, então, adquiridas a partir do córpus.

Para avaliar seu método, os grupos resultantes são comparados a grupos formados, sobre os mesmos dados, a partir da atribuição de sentido por juízes humanos. Essa

comparação mostra que o algoritmo de *clustering* usando córpus paralelos provê distinções de sentido bastante refinadas, próximas das distinções feitas pelos juízes humanos. O único problema ressaltado pelos autores é o da falta de córpus paralelos substanciais entre várias línguas.

Dyvik (2002) também destaca a relevância de córpus paralelos para a extração de informações semânticas. Segundo o autor, córpus paralelos são importantes fontes de informações semânticas. As traduções são, em muitos sentidos, fontes mais confiáveis que descrições de significado providas por um lexicógrafo ou semanticista.

Diab & Resnik (2002) propõem uma abordagem para a criação de um corpus de exemplos etiquetados com sentidos para ser usado, posteriormente, em aplicações de DLS supervisionada monolingüe. Nessa abordagem, são utilizados córpus paralelos bilíngües e um inventário de sentidos pré-definido para uma das duas línguas. O principal objetivo é a criação de um córpus substancial anotado com sentidos para a língua da qual se dispõe do inventário de sentidos. Como conseqüência da etiquetação desse córpus, é possível etiquetar também o córpus da segunda língua, utilizando o mesmo inventário de sentidos. Isso seria realizado por meio da identificação das traduções das palavras da primeira língua (as quais já possuem seu sentido atribuído) nessa segunda língua.

Os autores utilizam, como exemplo, a tradução entre uma língua-fonte e uma língua-alvo, sendo que o inventário de sentidos é válido para a língua-alvo. A língua cujos textos devem ser inicialmente etiquetados é, portanto, a língua-alvo.

Para a criação do córpus, é utilizado um sistema de TA para gerar córpus paralelos entre as duas línguas. Em seguida, os textos paralelos são automaticamente alinhados por sentenças e por palavras. Esse alinhamento permite identificar, nos textos da língua-alvo, quais as traduções correspondentes a palavras da língua-fonte. As palavras da língua-alvo que são traduções de uma mesma forma na língua-fonte são, então, agrupadas. Para cada um dos conjuntos gerados, são considerados todos os possíveis sentidos para cada palavra. A etiqueta de sentido adequada para cada palavra é atribuída de acordo com a sua similaridade semântica com as outras palavras no grupo.

Apesar da facilidade na geração do córpus paralelo alinhado, os autores ressaltam que esses córpus podem apresentar diversos erros decorrentes de traduções automáticas ou alinhamentos automáticos inadequados. Esses erros podem se propagar pelo processo de criação do córpus etiquetado e, certamente, influenciarão no desempenho dos trabalhos supervisionados criados utilizando tais córpus como base.

Agirre & Martínez (2004) descrevem o processo de criação automática de um córpus de exemplos etiquetados, focalizando a análise do desempenho desse córpus em um trabalho supervisionado de DLS e, principalmente, a análise do papel do *bias* de distribuição dos sentidos nesse córpus, ou seja, do número de exemplos para cada sentido de cada palavra.

O método de criação do córpus de exemplos empregado é o proposto por Leacock et al. (1998), que se baseia nos “parentes” não-polissêmicos dos itens ambíguos para obter exemplos etiquetados com sentidos para esses itens.

Para os testes, foram considerados como itens ambíguos 29 substantivos e os seus parentes não-polissêmicos indicados pela WordNet. Os parentes, nesse experimento, são os sinônimos desses itens ambíguos. São então realizadas buscas na *web*, considerando sentenças de busca com os sinônimos não-polissêmicos para recuperar exemplos contendo esses sinônimos. A suposição do método é de que para um determinado sentido da palavra ambígua, se for possível encontrar um sinônimo não-ambíguo desse sentido, então os

exemplos que contém esse sinônimo devem ser muito similares ao sentido da palavra ambígua e podem, portanto, ser usados para treinar um modelo supervisionado para aquele sentido da palavra.

Segundo os autores, uma característica que pode ser determinante na precisão e, principalmente, na cobertura de um trabalho supervisionado é o *bias* de distribuição dos exemplos para cada sentido no córpus. Para verificar o impacto desse *bias*, eles realizam diversos experimentos, incluindo: (a) nenhum *bias*, ou seja, considerando a mesma quantidade de exemplos para cada sentido; (b) o *bias* dos exemplos adquiridos automaticamente da *web*; (c) o *bias* do córpus SEMCOR, ou seja, considerando a mesma distribuição de sentidos desse córpus. Para os testes, é considerado o córpus de teste disponibilizado na segunda edição do exercício de avaliação SENSEVAL e um algoritmo supervisionado simbólico. Os resultados mostram que diferentes distribuições implicam diferentes resultados, sendo que os melhores resultados, principalmente em termos de cobertura, são obtidos a partir do *bias* do SEMCOR. Os resultados utilizando o *bias* automático de distribuição dos exemplos adquiridos da *web*, contudo, não são muito inferiores. Com relação à precisão do método de criação do córpus, considerando também o *bias* da *web*, os autores concluem que ela é maior que a de outros trabalhos da mesma categoria avaliadas no SENSEVAL.

Assim como Agirre & Martínez, Fernández et al. (2004) também apresentam uma estratégia para a criação automática de córpus baseada na formação de sentenças de busca a partir das definições e relações da WordNet e na busca de exemplos com essas sentenças em córpus ou na *web*. Cada *synset* a que pertence uma palavra na WordNet é caracterizado, por meio de suas relações com outros *synsets* ou palavras, como uma potencial sentença de busca. Contudo, os critérios para a construção das sentenças de busca são mais elaborados e flexíveis.

No trabalho de Agirre & Martínez, bem como em outras similares (Leacock et al., 1998), a estrutura das sentenças de busca é definida previamente, por exemplo, ela é constituída sempre do contexto da palavra-alvo e de mais um sinônimo não ambíguo dessa palavra. Fernández et al., por outro lado, definem uma linguagem para especificação de padrões de sentenças de busca, de modo que várias estratégias de busca podem ser previamente definidas para formar diferentes sentenças para a busca nos córpus. Com isso, a proposta torna-se mais flexível e as buscas podem retornar um número muito maior de exemplos.

A linguagem criada inclui operadores lógicos, funções para indicar que parte da WordNet deve ser usada para extrair as palavras da sentença de busca (glosas, relações, etc.) e palavras, sentidos ou relações específicas da WordNet. Assim, como a definição de uma única estratégia, podem ser geradas diversas sentenças de busca.

Os autores realizam um experimento inicial considerando apenas o córpus do SEMCOR, localmente armazenado, mas afirmam que o método pode ser usado em qualquer buscador *web*. São criadas seis estratégias de busca, sendo que algumas são baseadas nas estratégias usadas em outros trabalhos similares, como os citados. Essas estratégias são aplicadas às 73 palavras ambíguas usadas na segundo edição do exercício de avaliação SENSEVAL. As sentenças de busca geradas são então utilizadas para recuperar exemplos no SEMCOR. Cada estratégia envolve um possível sentido da palavra ambígua e as sentenças de busca mantêm esse sentido. Assim, os exemplos recuperados já possuem, automaticamente, uma etiqueta de sentido, neste caso, um sentido da WordNet. O córpus SEMCOR foi utilizado justamente porque possui etiquetas de sentido, também da WordNet, assim, os sentidos atribuídos pelo sistema podem ser comparados com os sentidos originais.

A precisão e a cobertura média de todas as estratégias são baixas, entretanto, os resultados também são apresentados considerando-se cada uma das estratégias, mostrando que algumas estratégias apresentam valores bem mais altos. Para todas as palavras, as sentenças de busca de todas as estratégias recuperam, em conjunto, 48.980 exemplos (não necessariamente todos corretos de acordo com sentido buscado). Esse pode ser considerado um número alto, já que o SEMCOR é um córpus relativamente pequeno. Se as buscas forem feitas em textos da *web*, certamente, esse número pode aumentar.

Vale notar que alguns dos trabalhos descritos na Seção 3, apesar de voltados para o processo de desambiguação, podem ser vistos, alternativamente, como propostas para a criação automática ou semi-automática de córpus de treinamento. Podem ser citados, por exemplo, Hearst (1991), Schütze (1992), Dagan et al. (1991) e Dagan & Itai (1994). Contudo, nesses trabalhos, os exemplos são etiquetados durante o processo de DLS. O seu objetivo não é, portanto, a criação de córpus de exemplos, mas uma alternativa de DLS parcialmente supervisionada.

Nos trabalhos voltados para a TA, uma estratégia simples para facilitar a criação de córpus de exemplos é a utilização de textos paralelos entre as línguas, já alinhados no nível das palavras, ou submetidos a alinhadores de textos. Contudo, os alinhamentos precisam estar corretos, ou uma revisão manual posterior é necessária. Além disso, é necessário um conjunto substancial de textos para extrair um número representativo de exemplos. Para várias línguas, entretanto, não há grandes conjuntos de textos paralelos disponíveis. Por essas razões, essa estratégia é pouco explorada, ainda.

2.7.3 O problema dos dados esparsos

O problema de dados escassos ou esparsos (*sparseness data*) é comum nos trabalhos baseados em córpus, principalmente nos supervisionados. Ele ocorre quando não há exemplos para todos os sentidos no córpus ou, ainda, há uma quantidade muito pequena de exemplos para alguns sentidos, que acabam se tornando estatisticamente insignificantes. De modo geral, a escassez de dados indica que, em um grande espaço de interpretações alternativas produzidas por palavras ambíguas, somente uma pequena parte é utilizada.

Esse problema, que é comum em todas as aplicações que utilizam córpus, é especialmente grave para a DLS. Primeiramente, quantidades muito grandes de textos são necessárias para tentar garantir que todos os sentidos de todas as apalavras ambíguas consideradas sejam representados. Além disso, como os trabalhos de DLS são geralmente baseados em co-ocorrências da palavra ambígua com outras palavras, muitas das possíveis co-ocorrências são improváveis ou pouco freqüentes mesmo em córpus muito grandes.

Algumas estratégias têm sido empregadas para a minimização desse problema. Em geral, elas procuram estimar a probabilidade de co-ocorrência de sentidos que não ocorrem nos dados de treinamento, de modo que essa probabilidade não seja assumida como nula. Ide & Véronis (1998) dividem essas estratégias em: (a) técnicas de suavização (*smoothing*) (Gale & Church, 1991); (b) modelos baseados em classe (Pereira et al., 1993); e (c) métodos baseados em similaridade (Dagan et al., 1993). Ide & Véronis apontam para os métodos baseados em similaridade, especialmente, para o método de Dagan et al., como os mais elaborados, que apresentam os melhores resultados. Dagan et al. procuram estimar a probabilidade de co-ocorrência de sentidos inexistentes no córpus por meio da analogia entre cada co-ocorrência específica não observada e outras co-ocorrências que contêm palavras semelhantes, de acordo com uma medida de similaridade entre as palavras.

Os autores apresentam o exemplo da desambiguação da palavra *chapter*, que é sucedida pela palavra *describes* em uma sentença. No córpus de exemplos, não consta o par *(chapter, describes)*, mas constam os pares *(book, describes)* e *(section, describes)*. São

utilizadas, então, métricas de similaridade para indicar que *chapter* é similar a *book* e *section* e que, portanto, *chapter* deve ser utilizado no mesmo sentido que essas palavras, se co-ocorrer com a palavra *describes*.

O método é avaliado em dois cenários, o primeiro considerando a desambiguação para a TA, em um sistema desenvolvido pelos autores (Dagan etc al., 1991), e o segundo considerando a Recuperação de Informações, comparativamente a outros métodos de estimação. No primeiro cenário, o método aumentou em 15% a cobertura e possibilitou uma melhoria na precisão do mecanismo de escolha da tradução mais adequada. No segundo cenário, o método proporcionou uma estimativa 27% mais precisa que a estimação baseada em freqüência (suavização).

Outros autores usam alternativas mais simples para o problema dos dados esparsos. Towell & Voorhess (1998), por exemplo, permitem que o classificador gerado não atribua nenhum sentido a palavras para as quais ele não possui evidências suficientes para a classificação. Assim, uma das classes do sistema é denominada "*do not know*". Na sua avaliação, os autores relatam uma melhoria na acurácia do sistema com essa simplificação. Certamente, em alguns casos, uma resposta como essa pode ser mais indicada que uma classificação incorreta. Contudo, na maioria das aplicações, principalmente na TA, essa resposta é de pouca validade.

2.8 A avaliação dos trabalhos de DLS

Haja vista que DLS é uma tarefa intermediária, que pode ser utilizada em outras tarefas maiores, há duas possibilidades de avaliação dos sistemas nessa área (Sparck-Jones & Galliers, 1996): (a) a **avaliação intrínseca**, na qual os sistemas são testados considerando o objetivo específico para o qual foram desenvolvido, neste caso, para a desambiguação lexical, independentemente da tarefa em que serão aplicados; e (b) a **avaliação extrínseca** (ou **validação**), na qual os resultados dos sistemas são avaliados em termos da sua contribuição para o desempenho global de um sistema criado para determinada aplicação, como a TA. Nas avaliações extrínsecas do módulo de DLS, apenas o resultado final da tarefa maior considerado, sujeito à avaliação apropriada nessa tarefa. Conforme a nomenclatura utilizada por Ide & Véronis (1998), as avaliações intrínseca e extrínseca são denominadas, respectivamente, avaliação *in vitro* e avaliação *in vivo*.

A avaliação intrínseca normalmente consiste em comparar a saída do sistema para determinadas entradas com os resultados esperados (corretos), obtidos por meio da atribuição manual de sentidos a um *corpus* de referência. Os resultados dessas comparações são reportados de acordo com diferentes medidas, em geral, calculadas individualmente para cada palavra ambígua.

Alguns autores apresentam seus resultados em termos de precisão (*precision*) e cobertura (*recall*) (e, ainda, a medida combinada *f-measure*), outros, apenas, de acurácia. Como mencionam Lee & Ng (2002), precisão e cobertura são, na verdade, duas maneiras de expressar a acurácia do sistema, sendo que a **precisão** indica o percentual de palavras ambíguas corretamente classificadas pelo sistema com relação a todas as palavras ambíguas do conjunto de teste para as quais alguma etiqueta é atribuída por esse sistema. Já a **cobertura** indica o percentual de palavras ambíguas corretamente classificadas pelo sistema com relação a todas as palavras ambíguas do conjunto de teste.

Precisão e cobertura, de acordo com essa definição, são as medidas usadas no exercício de avaliação SENSEVAL, descrito logo seguir. Alguns autores ainda expressam uma medida de **abrangência** (*coverage*) geral do sistema, que identifica o percentual de palavras ambíguas etiquetados, com relação ao total de palavras ambíguas do conjunto de teste, independentemente de a etiquetação ser correta ou incorreta.

Apesar do padrão estabelecido pelo SENSEVAL, a maioria autores, principalmente nos trabalhos anteriores a esse exercício, expressa o desempenho do sistema em termos da sua **acurácia**, definindo-a como o percentual de palavras ambíguas corretamente etiquetadas em relação ao total de palavras ambíguas. Nesse sentido, a medida de acurácia corresponde à medida de cobertura descrita. Outros autores, entretanto, expressam os resultados utilizando o termo "acurácia" mas, na verdade, referem-se à precisão. Assim, não é possível realizar uma comparação direta entre os resultados de alguns trabalhos.

Mesmo entre os trabalhos nos quais é explicitamente mencionado como é calculada a medida usada, realizar comparações apenas com base nos resultados nem sempre é possível ou indicado, em função do grande número de diferenças entre os trabalhos. A seguir, são apresentados alguns critérios adotados para avaliações individuais, os principais problemas de avaliações comparativas e alguns esforços voltados para a realização de avaliações dessa natureza.

2.8.1 Avaliações (intrínsecas) individuais

Os resultados de avaliações individuais, realizadas em cada trabalho, são normalmente comparadas aos resultados de algum limite inferior pré-definido como o mínimo que se espera que seja superado por um sistema de DLS. Um critério básico, proposto por Gale et al. (1992b) e normalmente utilizado nas avaliações individuais de diferentes trabalhos, é a definição do sentido mais freqüente, independentemente do contexto, como *baseline* (ou limite inferior) para a comparação dos trabalhos. De acordo com esse critério, atribui-se o sentido mais freqüente a todas as ocorrências da palavra ambígua nas sentenças de teste e avalia-se se o trabalho em questão alcança um desempenho melhor no processo de DLS.

Outros autores, contudo, comparam seus trabalhos com a *baseline* de escolha aleatória por um sentido, principalmente nos trabalhos não-supervisionados, nos quais não se dispõe de um córpus anotado com sentidos para extrair o sentido mais freqüente. Para esses casos, Gale et al. sugerem um mecanismo para estimar o sentido mais freqüente por meio do sentido mais provável, por meio das distribuições de freqüências em um córpus. McCarthy et al. (2004) também propõem um método para extrair os sentidos predominantes, mais freqüentes, a partir de um córpus não anotado e de informações de um *thesaurus*. Segundo os autores, esses sentidos podem ser utilizados como *baseline* nas avaliações.

Gale et al. também sugerem um limite superior para a comparação dos trabalhos, considerado o "ideal". Esse valor ideal corresponderia ao desempenho atingido por humanos na tarefa de DLS, ou à estimativa desse desempenho. Dessa maneira, o limite superior seria dependente da configuração do experimento e, principalmente, do nível de concordância entre os humanos na atribuição de sentidos. A partir de alguns experimentos de desambiguação com humanos, os autores apontam para uma precisão entre 97% e 99%. Contudo, eles só atingem essa precisão em seus experimentos porque consideram a desambiguação entre sentidos muito distantes e, mais importante, não analisam a tarefa de desambiguação, propriamente dita, mas a capacidade dos juízes identificarem se duas ocorrências da mesma palavra em uma sentença são ou não exemplos do mesmo sentido (visando testar a sua hipótese de um sentido por discurso - Seção 2.10). Considerando sentidos mais refinados e a tarefa de desambiguação, Véronis (1998) aponta uma grande discordância entre os juízes, de modo que a utilização desse critério como limite superior nas comparações não é tão simples e direta.

Em um experimento para analisar a concordância entre seis anotadores humanos na marcação de sentidos de itens polissêmicos do córpus usado no exercício de avaliação conjunta de línguas românicas ROMANSEVAL (Seção 2.8.2), considerando a língua

francesa, Véronis relata resultados que indicam uma discordância consideravelmente grande entre os juízes: para as 60 palavras ambíguas analisadas (20 substantivos, 20 verbos e 20 adjetivos), com cerca de 60 exemplos cada uma, o índice Kappa (Carletta, 1996) é, em média, menor que 50%. No caso de algumas palavras, o índice indicou uma concordância menor que a concordância ao acaso. Segundo o autor, um dos motivos dessa discordância é o alto nível de refinamento dos sentidos, extraídos de um dicionário. Ainda segundo Véronis, a alta discordância é preocupante no sentido de que tem implicações diretas no processo de avaliação de sistemas de DLS. De fato, neste caso, o experimento foi realizado com um córpus usado em um exercício de avaliação conjunta.

No caso de distinções de sentido muito refinadas, a comparação direta com a etiquetação humana se torna, de fato, inadequada, pois o sentido escolhido pelo sistema pode ser distinto, mas muito próximo do escolhido pelo juiz humano. Pode ser necessária, então, uma avaliação menos restritiva. Nesse sentido, Resnik & Yarowsky (1997b) sugerem medidas mais elaboradas de avaliação, em vez da simples precisão e cobertura. Uma das medidas propostas atribui pesos diferentes para cada "erro" do sistema de acordo com a distância entre o sentido indicado por ele e o sentido considerado correto: erros em um nível de distinção de sentidos altamente refinado devem ser menos punidos que erros em níveis menos refinados de distinção, já que, nesse último caso, pode-se atribuir um sentido totalmente diferente do sentido adequado.

2.8.2 Avaliações (intrínsecas) comparativas

Conforme ressaltado por Ide & Véronis (1998), comparar os resultados de avaliações individuais de trabalhos de DLS é uma tarefa muito difícil, uma vez que há diferenças substanciais nas configurações dos testes realizados nesses trabalhos. Por exemplo, diferentes tipos de texto são utilizados, incluindo textos especializados a um determinado domínio, nos quais o uso dos sentidos é limitado, e textos gerais, nos quais o uso dos sentidos é mais variável.

As palavras que são testadas nesses trabalhos também costumam variar, incluindo palavras com graus de ambigüidade muito diferentes, palavras com diferentes categorias gramaticais e palavras com usos figurados (metafóricos, metonímicos, etc.), etc. Segundo Ng & Zelle (1997), por exemplo, verbos são mais difíceis de desambiguar do que substantivos e quanto mais abrangente com relação a gênero e domínio for o córpus, maior a dificuldade. Além disso, os critérios para a avaliação da correção da atribuição de sentido também variam e a forma como a atribuição de sentido é julgada nesses trabalhos geralmente não é muito clara, uma vez que os julgamentos, na maioria das vezes, são feitos por humanos e a falta de concordância entre eles não é claramente documentada.

Como destacam Ng & Zelle (1997), não há córpus de teste representativos para a avaliação dos diferentes trabalhos. Para a DLS monolingüe, alguns córpus em menor escala foram manualmente criados em diferentes trabalhos e disponibilizados para teste, conforme descrito na Seção 2.7.1. Os criadores desses córpus e de outros trabalhos de DLS apresentam alguns resultados do processo de desambiguação que têm sido usados como referência em outros trabalhos que utilizam os mesmos córpus.

A avaliação comparativa da precisão de trabalhos baseados em córpus não-supervisionados é ainda mais complexa, em função das grandes variações nos córpus de teste empregados e dos diferentes grupos de sentidos que podem ser gerados e, com isso, da falta de resultados de referência adequados. Conforme mencionado, trabalhos não-supervisionados normalmente apresentam desempenho inferior, quando comparados aos trabalhos supervisionados. Em alguns casos, o desempenho é inferior à *baseline* do sentido mais freqüente. Avaliações com distinções menos refinadas, principalmente nos casos em

que apenas dois sentidos possíveis são considerados, podem levar a valores bem mais altos.

Para analisar a adequação de diferentes métodos, modos de aprendizado, tipos de conhecimento, algoritmos de aprendizado, entre outras possíveis variáveis, algumas avaliações comparativas entre diferentes trabalhos, considerando os mesmos conjuntos de palavras e sentidos para os testes, vem sendo realizadas (Mooney, 1996; Paliouras, 2000; Escudero et al., 2000a; Escudero et al., 2000c; Pedersen, 2002a; Lee & Ng, 2002, por exemplo). Contudo, os resultados não são conclusivos, pois principalmente nos trabalhos baseados em córpus, a variação de um único parâmetro pode mudar completamente o seu desempenho.

Na tentativa permitir avaliações comparativas, a partir de discussões do grupo SIGLEX (*ACL Special Interest Group on the Lexicon*) e das sugestões apresentadas por Resnik & Yarowsky (1997a), foi criado o Projeto SENSEVAL[4]. Esse Projeto tem por objetivo prover padrões para a avaliação e comparação de diversos trabalhos de DLS.

Para tanto, são promovidos exercícios de avaliação, para os quais são definidos e disponibilizados córpus de treinamento e teste, com um determinado repositório de sentidos, bem como as tarefas a serem realizadas. As duas principais tarefas do evento são: (a) a desambiguação (ou etiquetação) de todas as palavras de conteúdo em um córpus (*all-words task*); e (b) a desambiguação de todas as ocorrências de determinadas palavras em um córpus (*lexical sample task*).

Já foram realizadas duas edições do evento de avaliação conjunta, SENSEVAL-1, em 1998, e SENSEVAL-2, em 2001, e uma terceira, SENSEVAL-3, está em andamento. Além das tarefas de avaliação, propriamente ditas, o Projeto promove workshops específicos para apresentação, discussão e divulgação dos resultados e de perspectivas para a área. Na primeira edição do evento, as avaliações foram monolingües, considerando-se apenas a língua inglesa.

Paralelamente a essa primeira edição do evento, como uma das suas ramificações, foi realizado um evento de avaliações correspondentes voltado especificamente para a desambiguação em algumas línguas românicas (especialmente, o francês e o italiano), denominado ROMANSEVAL[5]. Essa divisão foi realizada em decorrência das diferenças entre essas línguas e a língua inglesa, as quais, segundo os organizadores do evento, demandam padrões de avaliação distintos. Os resultados foram apresentados em conjunto com o workshop do SENSEVAL-1. Edições posteriores a esse evento foram embutidas nas edições do SENSEVAL.

Na segunda edição do SENSEVAL foram incluídas várias outras línguas (por exemplo, tcheco, alemão, estoniano, chinês, italiano e espanhol) no mesmo procedimento de avaliação, com córpus específicos, e uma tarefa de desambiguação tradução para o japonês. O repositório de sentidos para a DLS monolingüe passou a ser constituído, desde então, dos *synsets* da WordNet.

Para a terceira edição, além das línguas já incluídas, está prevista uma tarefa de desambiguação multilingüe de cerca de 50 palavras, com grupos inscritos para a DLS do inglês para o francês e do inglês para o híndi. A língua portuguesa, contudo, ainda não foi incluída. Estão previstas também novas tarefas monolingües, como a etiquetação de papéis semânticos e a desambiguação das definições da WordNet.

De modo geral, nas duas primeiras edições, os trabalhos baseados em córpus supervisionados apresentaram o melhor desempenho (Kilgarriff & Palmer, 2000; Edmonds & Cotton, 2001). Em especial, o primeiro colocado na primeira edição foi um sistema usando uma abordagem simbólica de listas de decisão hierárquicas e, na segunda, um

[4] http://www.senseval.org/
[5] http://aune.lpl.univ-aix.fr:16080/projects/romanseval/

sistema combinando diversos algoritmos de aprendizado (listas de decisão, Naive Bayes, etc.). Os resultados referentes à terceira edição do evento ainda não foram divulgados.

Em geral, na avaliação extrínseca da DLS são realizadas análises contrastivas, comparando o desempenho do sistema em que o módulo de DLS é inserido com e sem esse módulo. Como a maioria dos trabalhos de DLS, conforme mencionado, não é voltada para nenhuma aplicação específica, em geral, só são relatadas as avaliações intrínsecas desses trabalhos. Para a avaliação extrínseca na TA, em particular, é necessário considerar o uso do módulo de DLS em algum sistema de TA. Certamente, as possibilidades de inserção do módulo de DLS na TA dependem fortemente do método de TA empregado no desenvolvimento do sistema, conforme descrito a seguir.

2.9 O módulo de DLS em um sistema de TA

Considerando-se a utilização de um módulo de DLS em um sistema de TA, pode-se identificar, basicamente, dois modos de atuação: posterior e interativo. No modo posterior, a desambiguação é feita depois da tradução. Neste caso, o sistema de TA deve gerar todas as possíveis traduções das palavras ambíguas e, a partir dessas possibilidades, o módulo de DLS deve realizar sua escolha. Com isso, intuitivamente, o contexto a ser observado poderia ser o da sentença já traduzida, ou seja, as demais palavras da sentença, já na LA. Esse modo de atuação é bastante simples e potencialmente mais independente que o interativo. Contudo, pode ser pouco eficaz, pois pode haver ambigüidade também nas demais palavras da sentença, que seriam utilizadas como contexto para desambiguar determinada palavra. Assim, seria necessário, de qualquer maneira, retomar a sentença original.

No modo de atuação interativo, a desambiguação é realizada durante a tradução, em cada ponto em que é detectada ambigüidade lexical de sentido. Com isso, o contexto de desambiguação, normalmente, são as palavras da sentença original, podendo ou não ser utilizadas, adicionalmente, as traduções de palavras já desambiguadas. Esse modo de atuação torna a construção do módulo de DLS mais dependente do sistema de TA, contudo, é potencialmente mais eficaz que o modo posterior.

Considerando-se o modo interativo, a localização do módulo de DLS e o momento da sua atuação em um sistema de TA ainda dependem, entre outras coisas, do método de TA utilizado para construir o sistema e do nível de profundidade no tratamento lingüístico desse sistema.

Os sistemas de TA podem se basear em diferentes métodos, que se referem à organização global do processamento do sistema. Esses métodos podem ser classificados como **diretos** ou **indiretos**, sendo esses últimos divididos, ainda, em métodos **por transferência** e **por interlíngua** (Dorr et al. 2000).

A TA direta procura transformar as sentenças da língua-fonte em sentenças da língua-alvo sem utilizar representações intermediárias, realizando o mínimo de processamento lingüístico possível. Esse processamento pode variar, incluindo a simples substituição das palavras de uma sentença na língua-fonte por sua(s) correspondente(s) na língua-alvo (tradução palavra-por-palavra) ou a realização de tarefas mais complexas, como uma análise sintática para a reordenação das palavras na sentença da língua-alvo e a inclusão de preposições. O número de estágios de processamento desses sistemas depende, portanto, da proximidade das línguas envolvidas e do nível de profundidade da tradução pretendido. Geralmente, o processo de tradução segue o esquema ilustrado na Figura 3.

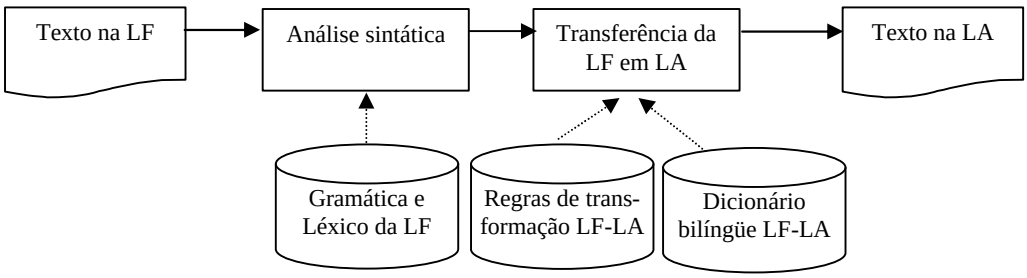

Figura 3. TA pelo método direto

Em sistemas que seguem o modo direto de tradução, o módulo de DLS deve atuar no componente de transferência da língua-fonte em língua-alvo, no momento da escolha da palavra, na língua-alvo, para traduzir a palavra da língua-fonte. Contudo, como os sistemas diretos não realizam análise semântica, a DLS dificilmente pode ser realizada de maneira efetiva, apenas com os conhecimentos sintáticos superficiais explorados.

Nos sistemas de tradução indireta, a análise da língua-fonte e a geração da língua-alvo constituem processos independentes. As sentenças na língua-fonte são primeiramente transformadas numa representação intermediária e, a partir dela, são geradas as sentenças na língua-alvo. Essa representação intermediária pode ser única e independente de língua (tradução por interlíngua), ou diferente para cada língua e dependente dessa língua (tradução por transferência).

Na TA por transferência, a tradução consiste nos seguintes passos ilustrados na Figura 4: (1) alteração da estrutura e/ou palavras da sentença de entrada para produzir uma representação intermediária da língua-fonte, podendo envolver processos complexos como a análise semântica, mas, em geral, limitando-se à análise sintática (fase de análise); (2) transformação da representação (árvore sintática, em geral) gerada na etapa anterior em uma estrutura intermediária da língua-alvo, por meio de regras de mapeamento que indicam as correspondências lexicais e sintáticas entre tais estruturas (fase de transferência); e (3) geração da sentença na língua-alvo a partir dessa estrutura (fase de geração).

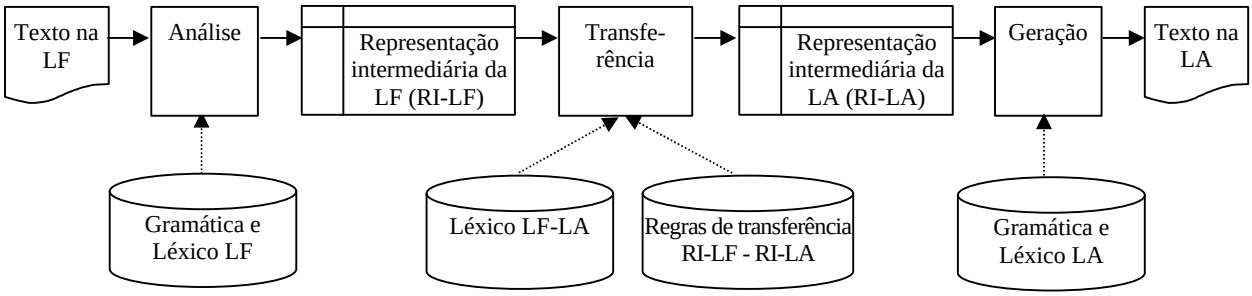

Figura 4. TA pelo método indireto por transferência

No método indireto por transferência, a localização do módulo de DLS depende do nível de profundidade considerado na transferência. Os sistemas que realizam apenas a análise sintática da língua-fonte, mantendo as palavras fonte na representação intermediária da língua-fonte (RI-LF) devem incorporar o módulo de DLS no componente de transferência, para que possa auxiliar nas escolhas do léxico da língua-fonte-língua-alvo e, também, nas regras de transferência. Por exemplo, na TA inglês-português, o verbo *know* não seria

considerado ambíguo na análise da língua inglesa, pois não haveria necessidade de identificar o seu sentido mais adequado. Esse verbo só precisaria ser desambiguado no componente de transferência. Já os sistemas por transferência que realizam a análise semântica ou análises mais profundas podem incorporar o módulo de DLS ao processo de análise da língua-fonte e, dependendo das diferenças entre as representações intermediárias das duas línguas, no módulo de geração da língua-alvo.

Na tradução por interlíngua, a representação intermediária (isto é, a interlíngua) é única, capturando as características comuns às duas línguas envolvidas, ou seja, as características independentes de língua, que representam o significado dos enunciados a serem traduzidos. Nesse caso, a saída da análise da língua-fonte corresponde à entrada do componente de geração na língua-alvo e a tradução é realizada segundo as etapas da Figura 5: (1) análise completa do texto na língua-fonte, extraindo seu significado e representando-o na interlíngua; e (2) geração do texto na língua-alvo, partindo da representação interlingual e expressando o mesmo significado. Nesse contexto, o processo de geração do texto na língua-alvo a partir da representação de significado caracteriza-se mais como uma paráfrase que como uma tradução.

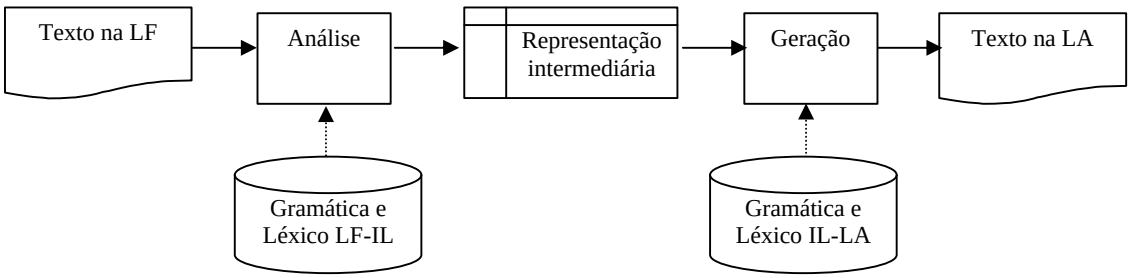

Figura 5. TA pelo método indireto por interlíngua

Considerando-se que a interlíngua deve ser suficientemente expressiva para representar os conceitos de todas as línguas envolvidas no sistema, no método indireto por interlíngua, o módulo de DLS deve atuar no componente de análise da língua-fonte, de modo a permitir a escolha pelo conceito da interlíngua mais adequado para as palavras da língua-fonte. Assim, a representação intermediária já estará desambiguada e, no léxico que faz a correspondência entre a interlíngua e a língua-alvo, a cada conceito deverá corresponder apenas uma palavra da língua-alvo. Segundo Pedersen (1997), o problema está justamente em estabelecer uma representação neutra e suficientemente expressiva para diferenciar ambigüidades interlinguais. Ainda segundo autora, apesar de o método por interlíngua representar o caminho ideal para a TA, sistemas de TA por transferência permitem abordagens mais práticas para o tratamento da DLS, já que o mapeamento língua-fonte-língua-alvo é mais controlado.

Esses exemplos de localização e atuação do módulo de DLS não constituem, certamente, as únicas possibilidades, pois além do método de TA, outras questões influenciam nessa atuação. Por exemplo, a DLS pode guiar as escolhas já durante o processamento sintático, somente no processamento semântico ou, ainda, depois que esses processos forem completamente realizados, para a escolha de uma das representações sintáticas/semânticas (no método por transferência), representações conceituais (no método por interlíngua) ou traduções finais (no método direto).

2.10 Um sentido por discurso e um sentido por *collocation*

Nesta seção são discutidos dois conceitos muito influentes em muitos trabalhos de DLS: as noções de **um sentido por discurso** (*one sense per discourse*) (Gale et al., 1992c) e **um sentido por *collocation*** (*one sense per collocation*) (Yarowsky, 1993). Ambas as noções foram desenvolvidas a partir da observação dos autores durante a realização de experimentos e não constituem, portanto, propostas de DLS.

Primeiramente, Gale et al., em experimentos de DLS, observaram um forte relacionamento entre o discurso e o significado. Com base nisso, propuseram a hipótese que define que quando uma palavra ocorre mais de uma vez em um discurso, todas as ocorrências dessa palavra compartilham o mesmo significado. A noção de discurso, contudo, não fica clara no trabalho. Em princípio, poderia consistir de um documento, um artigo de jornal, um parágrafo, entre outros. O "sentido" considerado corresponde à noção de homonímia, ou seja, significados distintos, não relacionados entre si.

Para testar sua hipótese, os autores conduziram um experimento com cinco juízes humanos. Cada juiz recebeu um conjunto de definições para nove palavras ambíguas e 82 pares de linhas de com duas ocorrências de cada palavra. Eles deveriam determinar, em cada par de ocorrências, se o sentido da palavra ambígua era o mesmo ou não. Como resultado, os autores reportaram a probabilidade de 94% das duas ocorrências da palavra ambígua no mesmo discurso terem o mesmo sentido. Vale que praticamente não houve problemas de discordância entre os juízes no experimento porque eles não precisavam indicar o sentido e as distinções para cada palavra ambígua, quando existentes, eram no nível de homonímia.

Com o experimento, os autores concluem que sua hipótese pode ser usada para melhorar o desempenho dos algoritmos de DLS, bem como para auxiliar na avaliação dessa tarefa. Somente uma ocorrência da palavra ambígua em um discurso precisaria ser desambiguada; as demais ocorrências poderiam ser automaticamente etiquetadas com o mesmo sentido.

Apesar de ser usada em alguns trabalhos de DLS, essa hipótese recebeu também várias críticas. Wilks (1997), por exemplo, afirma que o problema não é tão simples como colocado pelos autores, ou seja, a identificação do sentido do discurso não resolve o problema da DLS. Além disso, mesmo que haja apenas um sentido para cada palavra ambígua em um discurso, a sua identificação não deixa de ser complexa. O experimento realizado não mostrou essa complexidade, pois o seu objetivo não era a desambiguação, propriamente dita, mas apenas verificar se duas palavras em um discurso tinham ou não o mesmo sentido.

Segundo Wilks, raramente as apalavras em um texto aparecem com uma quantidade pequena de sentidos, a não ser em textos técnicos ou muito específicos. Assim, a noção de um sentido por discurso só se aplicaria para a DLS desse tipo de textos. Além disso, o autor afirma que o experimento realizado, com apenas dois sentidos distintos para cada palavra ambígua e baseado no julgamento humano para identificar se pares de palavras pertenciam ou não ao mesmo sentido, é muito limitado.

A hipótese de Gale et al. também foi criticada por Krovetz (1998). O autor procura repetir o experimento dos autores, mas considerando sentidos da WordNet, nos córpus DSO e SEMCOR. O discurso, no seu experimento, corresponde a um arquivo do SEMCOR ou um documento do DSO. Como resultado, ele mostra que essa hipótese não se aplica para distinções de sentidos mais refinadas. Em vez dos 6% de ocorrências de múltiplos sentidos por discurso para uma palavra ambígua relatados por Gale et al, Krovetz encontra 33%, em média, dessas ocorrências.

O autor apresenta também algumas motivações para considerar, no experimento, não apenas distinções de sentido em nível de homonímia, mas também em nível de polissemia. Segundo ele, de acordo com os julgamentos dos juízes seria possível determinar, além da validade ou não da hipótese de um sentido por discurso, a distinção empírica entre casos de homonímia e polissemia. Os casos de julgamentos com alto nível de discordância corresponderiam a ocorrências de sentidos diversos de uma palavra polissêmica, enquanto os casos de julgamento com uma concordância mínima corresponderiam a ocorrências de homonímia.

Em um trabalho posterior, Yarowsky (1993) restringe a noção de um sentido por discurso de acordo com a hipótese de um sentido por *collocation*. Segundo essa hipótese, duas palavras dificilmente aparecem em uma mesma *collocation* com sentidos diferentes em um mesmo córpus. Wilks (1997) afirma que se trata de um trabalho estatisticamente expressivo, mas que esse não pode ser um critério utilizado isoladamente para a DLS em larga escala, uma vez que não há um consenso sobre o que pode constituir uma *collocation* suficiente para a DLS, independentemente do contexto.

Segundo Martínez & Agirre (2000), em geral, os trabalhos atuais de DLS são voltados para o uso de distinções mais refinadas de sentido. Assim, eles analisam também a hipótese de Yarowsky, considerando as distinções da WordNet no córpus DSO, a exemplo do trabalho de Krovetz, e considerando, também, as mesmas *collocations* em dois córpus de gênero e domínio distintos, para verificar se tal hipótese se aplica nessas duas condições. Em sua proposta, Yarowsky considerou a distinção entre apenas dois sentidos muito distintos, no nível de homonímia, para cada palavra ambígua e não analisou o efeito das mesmas *collocations* em vários córpus.

Diferentemente dos resultados de Yarowsky de 99% de precisão nos experimentos para verificar sua hipótese, Martinez & Agirre reportam uma precisão menor que 80% considerando-se as distinções de sentido mais refinadas nos experimentos com o córpus para treinamento e teste. Nos experimentos considerando várias porções do córpus (casos de treinamento e teste de córpus de diferentes gêneros e domínios), a precisão e a cobertura foram ainda menores: cerca de 70%. Com isso, os autores concluem que a hipótese de um sentido por *collocation* se mantém de um córpus para outro, contudo, as informações sobre *collocations* adquiridas de um córpus não são válidas para desambiguar outros córpus porque eles apresentam poucas *collocations* em comum.

Os autores afirmam que variações de gênero e domínio afetam o desempenho da desambiguação: textos que têm gêneros e domínio similares têm maior chance de ter *collocations* em comum e, conseqüentemente, a desambiguação de um texto com um modelo treinado em outros textos do mesmo gênero e domínio tende a apresentar um desempenho melhor.

3 Trabalhos de desambiguação lexical de sentido

Desde a década de 60, quando a ambigüidade lexical passou a ser vista como um problema prático e a DLS tornou-se uma área de pesquisa, vários trabalhos vêm sendo propostas. Nesta seção, são apresentadas os trabalhos considerados mais importantes, seja por seu papel histórico ou inovador na área de DLS, seja pelas suas características distintivas com relação às demais abordagens ou, ainda, pela sua relevância para o modelo de DLS a ser proposta.

Em sua maioria, os trabalhos são monolingües e foram desenvolvidos para língua inglesa, independentemente de aplicação. Na descrição a seguir, eles são apresentados em ordem cronológica crescente, divididos de acordo com a classificação da Seção 2.6. Somente nos trabalhos voltados para aplicações específicas ou para outras línguas essas informações (aplicação e língua(s)) são destacadas. Os trabalhos cujas características serão diretamente exploradas na proposta do modelo de DLS já mencionada são descritos com mais detalhes.

Apesar da proposta referir-se a um modelo multilingüe, a descrição não é limitada apenas aos trabalhos multilingües pois, de modo geral, os principais conceitos e procedimentos de DLS se aplicam tanto a aplicações monolingües quanto multilingües. As principais características das aplicações monolingües podem, em parte, ser adaptadas para a TA. Além disso, conforme mencionado, são poucos os trabalhos voltados para aplicações multilingües, em particular, para a TA. Assim, a exclusão dos trabalhos monolingües implicaria na desconsideração de importantes experiências já relatadas.

Antes de apresentar os trabalhos, propriamente ditos, na Tabela 1, todos eles são listados, em ordem cronológica, de acordo com sua aplicação, as línguas envolvidas e o método de DLS empregado.

Tabela 1. Lista geral dos trabalhos de DLS descritos

Trabalho	Aplicação	Língua(s)	Método
(Masterman, 1961)	Tradução Automática	inglês - ?	baseado em conhecimento manualmente codificado
(Quillian , 1961)	independente (monolingüe)	inglês	baseado em conhecimento manualmente codificado
(Hayes, 1976)	independente (monolingüe)	inglês	baseado em conhecimento manualmente codificado
(Wilks, 1975)	Compreensão da Língua Natural	inglês	baseado em conhecimento manualmente codificado
(Boguraev,1979)	independente (monolingüe)	inglês	baseado em conhecimento manualmente codificado
(Small, 1980)	Compreensão da Língua Natural	inglês	baseado em conhecimento manualmente codificado
(Hirst, 1987)	Compreensão da Língua Natural	inglês	baseado em conhecimento manualmente codificado
(Beale, 1997)	Tradução Automática	inglês - ?	baseado em conhecimento manualmente codificado
(Goodman & Nirenburg, 1991)	Tradução Automática	inglês – japonês	baseado em conhecimento manualmente codificado
(Copeland, 1991)	Tradução Automática	várias	baseado em conhecimento manualmente codificado
(Gajek, 1991)	Tradução Automática	várias	baseado em conhecimento manualmente codificado
Systran	Tradução Automática	várias	baseado em conhecimento manualmente codificado
(McRoy, 1992)	Compreensão da Língua Natural	inglês	baseado em conhecimento manualmente codificado

(Dorr, 1993)	Tradução Automática	inglês – espanhol e francês	baseado em conhecimento manualmente codificado
(Egedi et al., 1994)	Tradução Automática	coreano – inglês	baseado em conhecimento manualmente codificado
(Pedersen, 1997)	Tradução Automática	dinamarquês – inglês	baseado em conhecimento manualmente codificado
(Dorr & Katsova, 1998)	Tradução Automática	inglês – espanhol	baseado em conhecimento manualmente codificado
(Leffa, 1998)	Tradução Automática	inglês – português	baseado em conhecimento manualmente codificado
(Bräscher, 2002)	Recuperação de Informações	português	baseado em conhecimento manualmente codificado
(Cottrell & Small, 1983)	independente (monolingüe)	inglês	baseado em conhecimento manualmente codificado
(Waltz & Pollack, 1985)	independente (monolingüe)	inglês	baseado em conhecimento manualmente codificado
(Cottrell, 1989)	Compreensão da Língua Natural	inglês	baseado em conhecimento manualmente codificado
(Voorhees, 1993)	Recuperação de Informações	inglês	baseado em conhecimento pré-codificado
(Sussna, 1993)	Recuperação de Informações	inglês	baseado em conhecimento pré-codificado
(Resnik, 1995a)	independente (monolingüe)	inglês	baseado em conhecimento pré-codificado
(Agirre & Rigau, 1996)	independente (monolingüe)	inglês	baseado em conhecimento pré-codificado
(Mihalcea & Moldovan, 1999)	independente (monolingüe)	inglês	baseado em conhecimento pré-codificado
(Montoyo et al., 2002)	Recuperação de Informações multilingüe	espanhol e inglês – catalão e basco	baseado em conhecimento pré-codificado
(Lesk, 1986)	Recuperação de Informações	inglês	baseado em conhecimento pré-codificado
(Wilks et al., 1990)	independente (monolingüe)	inglês	baseado em conhecimento pré-codificado
(Véronis & Ide, 1990)	independente (monolingüe)	inglês	baseado em conhecimento pré-codificado
(Guthrie et al., 1991)	Recuperação de Informações	inglês	baseado em conhecimento pré-codificado
(Cowie et al., 1992)	independente (monolingüe)	inglês	baseado em conhecimento pré-codificado
(Wilks & Stevenson, 1996)	independente (monolingüe)	inglês	baseado em conhecimento pré-codificado
(Wilks & Stevenson, 1997a; 1997b)	independente (monolingüe)	inglês	baseado em conhecimento pré-codificado
(Brun, 2000)	independente (monolingüe)	inglês	baseado em conhecimento pré-codificado
(Masterman, 1957)	Tradução Automática	latim – inglês	baseado em conhecimento pré-codificado
Patrick (1985)	independente (monolingüe)	inglês	baseado em conhecimento pré-codificado
Brown et al. (1991)	Tradução Automática	francês – inglês	baseado em córpus
(Schütze, 1992; 1998)	Recuperação de Informações	inglês	baseado em córpus
(Schütze & Pedersen, 1995)	Recuperação de Informações	inglês	baseado em córpus
(Stokoe et al., 2003)	Recuperação de Informações	inglês	baseado em córpus
(Pedersen & Bruce,	independente (monolingüe)	inglês	baseado em córpus

1997)			
(Pedersen & Bruce, 1998)	independente (monolingüe)	inglês	baseado em córpus
(Dini et al., 1998)	Recuperação de Informações	inglês	baseado em córpus
(Pantel & Lin, 2002)	Recuperação de Informações	inglês	baseado em córpus
(Rapp, 2004)	independente (monolingüe)	inglês	baseado em córpus
(Black, 1988)	independente (monolingüe)	inglês	baseado em córpus
(Hearst, 1991)	Compreensão da Língua Natural	inglês	baseado em córpus
(Yarowsky, 1995)	independente (monolingüe)	inglês	baseado em córpus
(Yarowsky, 2000)	independente (monolingüe)	inglês	baseado em córpus
(Mooney, 1996)	independente (monolingüe)	inglês	baseado em córpus
(Ng & Lee, 1996)	independente (monolingüe)	inglês	baseado em córpus
(Ng, 1997a)	independente (monolingüe)	inglês	baseado em córpus
(Towell & Voorhess, 1998)	Recuperação de Informações	inglês	baseado em córpus
(Pedersen, 2000)	independente (monolingüe)	inglês	baseado em córpus
(Zinovjeva, 2000)	Tradução Automática	inglês - sueco	baseado em córpus
(Agirre & Martinez, 2000)	independente (monolingüe)	inglês	baseado em córpus
(Escudero et al., 2000a)	independente (monolingüe)	inglês	baseado em córpus
(Escudero et al., 2000b)	independente (monolingüe)	inglês	baseado em córpus
(Escudero et al., 2000c)	independente (monolingüe)	inglês	baseado em córpus
(Escudero et al., 2001)	independente (monolingüe)	inglês	baseado em córpus
(Pedersen, 2002a)	independente (monolingüe)	inglês e espanhol	baseado em córpus
(Pedersen, 2002b)	independente (monolingüe)	inglês e espanhol	baseado em córpus
(Lee & Ng, 2002)	independente (monolingüe)	inglês	baseado em córpus
(Florian et al., 2002)	independente (monolingüe)	inglês	baseado em córpus
(Park et al., 2003)	independente (monolingüe)	coreano	baseado em córpus
(Dihn et al., 2003)	Tradução Automática	inglês – vietnamita	baseado em córpus
(Mihalcea, 2004)	independente (monolingüe)	inglês	baseado em córpus
(Dagan et al., 1991) e (Dagan & Itai, 1994)	Tradução Automática	inglês - hebreu e alemão	híbrido
(Yarowsky, 1992)	independente (monolingüe)	inglês	híbrido
(Resnik, 1997)	independente (monolingüe)	inglês	híbrido
(Lee, 1997)	Tradução Automática	inglês - coreano	híbrido
(Karov & Edelman, 1998)	independente (monolingüe)	inglês	híbrido
(Wilks & Stevenson, 1998) e (Stevenson & Wilks, 1999; 2000; 2001)	independente (monolingüe)	inglês	híbrido
(Paliouras et al., 1999)	Extração de Informações	inglês	híbrido
(Paliouras et al., 2000)	Extração de Informações	inglês	híbrido

3.1 Método baseado em conhecimento

3.1.1 Conhecimento manualmente codificado

Abordagens simbólicas

Os trabalhos de DLS baseados em conhecimento, em especial, os mais antigas, utilizam técnicas simbólicas de representação e manipulação do conhecimento da Inteligência Artificial, como redes semânticas e *frames*. Dentre os trabalhos voltados para a TA, alguns constituem módulos independentes de DLS, enquanto outros consistem de módulos ou procedimentos de DLS embutidos nos sistemas de TA, considerando tanto sistemas de TA comerciais quanto sistemas acadêmicos.

O trabalho de Masterman (1961) é voltada para a TA por interlíngua, utilizando redes semânticas como formalismo de a representação do conhecimento. A rede semântica é usada para derivar representações das sentenças na interlíngua, que é constituída dos conceitos fundamentais da língua. São codificados 15.000 conceitos em um dicionário, classificados de acordo com 100 tipos de conceitos primitivos (*thing*, *do*, etc.), sendo que esses tipos são organizados em um reticulado com o mecanismo de herança de conceitos mais genéricos para os mais específicos. As distinções de sentido são realizadas implicitamente pela seleção das representações que refletem grupos de nós fortemente relacionados na rede.

Quillian (1961) também utiliza uma rede semântica que inclui relacionamentos entre palavras e conceitos, as quais são rotuladas com nomes de relações semânticas ou indicam associações entre as palavras. A rede é inicialmente criada manualmente a partir de definições de um dicionário, mas é estendida com outros conhecimentos, também manualmente codificados. Para cada duas palavras representadas na rede, a ativação gradual dos nós de conceitos ao longo do caminho dos relacionamentos originados a partir de cada palavra de entrada é simulada por meio da passagem de marcadores. A desambiguação ocorre porque apenas um nó de conceito associado com uma determinada palavra de entrada pode estar envolvido no caminho mais direto encontrado entre as duas palavras de entrada.

Várias características desses dois trabalhos, com o uso de tipos primitivos e estratégias como a de passagem de marcadores, foram reproduzidas em trabalhos subseqüentes, como os descritos a seguir.

O trabalho de Hayes (1976) explora o uso de *frames* e de uma rede semântica para uma representação mista do conhecimento. Ele utiliza vários tipos de conhecimento na desambiguação lexical, enfatizando a busca por associações semânticas entre as palavras. Os sentidos de substantivos são representados como nós na rede, enquanto que os verbos são representados como *frames*. Essa representação, que pode ser vista como uma rede semântica na qual cada frame é uma parte da rede, facilita a busca por associações semânticas entre as palavras. Outras associações são identificadas para nós e *frames* antecedentes ou descendentes, de acordo com as relações hierárquicas padrão de hiponímia e meronímia das redes semânticas, levando a uma cadeia de conexões na base de conhecimento.

Seu sistema usa também estruturas de caso e restrições de seleção (absolutas e preferenciais). Na desambiguação, os métodos mais restritivos são testados primeiro, na seguinte ordem: estruturas de casos, associações e restrições de seleção. É considerada

apenas a desambiguação dos substantivos, que é realizada a partir da estrutura sintática produzida para a sentença de entrada. Nesse processo de desambiguação, todos os substantivos ambíguos são considerados em conjunto, com todas as possibilidades analisadas em paralelo, até que o sistema possa convergir para uma solução.

Segundo Hayes, as associações semânticas são úteis principalmente para tratar casos de homonímia, dada a grande diferença entre os seus sentidos, mas não são muito indicadas para o tratamento de polissemia. Isso ocorre porque na polissemia, normalmente, encontra-se uma associação para mais de um sentido, já que eles são semanticamente relacionados. Segundo Hirst (1987), esse deve ser um dos motivos pelos quais Hayes não usou seu trabalho para a desambiguação de verbos, uma vez que verbos tendem a ser polissêmicos, enquanto os substantivos tendem a ser homônimos.

Trabalhos baseados em semântica de preferência como o de Wilks (1975) e o de Boguraev (1979) especificam restrições de seleção semânticas para a combinação dos itens lexicais da sentença que podem ser "amenizadas" quando não são encontras palavras com as restrições desejadas. Esses trabalhos são baseados no mecanismo de traços semânticos (ou primitivas semânticas) de Katz & Fodor (1963), também usado por Masterman, no qual os itens lexicais de substantivos são marcados com traços semânticos, como "humano", "animado" e "abstrato", os itens lexicais representando verbos são marcados com restrições de seleção, especificadas em termos dos traços semânticos exigidos por seus argumentos e os itens lexicais de adjetivos e preposições são marcados com restrições de seleção, especificadas em termos dos traços semânticos exigidos pelo substantivo ou frase nominal seguinte ao item ou que ele modifica.

Para cada sentido de uma mesma palavra, um novo item lexical é criado, assumindo-se que tal item deve ser descrito por diferentes restrições ou traços semânticos. De acordo com a semântica de preferência, construções nas quais as restrições de seleção são infringidas não são simplesmente consideradas incorretas e descartadas. Wilks acredita que nenhuma construção fornecida a um sistema de interpretação da língua natural deve ser considerada incorreta, já que faz sentido para o seu autor. O sistema, portanto, deve ser capaz de produzir alguma interpretação válida para tal construção. Nos casos em que o sistema consegue chegar a duas possíveis interpretações da construção, sendo que apenas uma delas infringe as restrições de seleção, a interpretação que satisfaz essas restrições é "preferida". Por outro lado, se todas as possíveis interpretações infringirem as restrições, será escolhida aquela que menos compromete as restrições. Dessa maneira, os sistemas podem manipular alguns usos metafóricos, como o da sentença "*My car drinks gasoline*". Nela, apesar de *drink* esperar um sujeito com o traço "animado", ele aceita um sujeito como o da sentença.

O sistema de Wilks usa cerca de 60 traços semânticos e suas entradas lexicais são modelos para a descrição de palavras, de acordo com sua categoria gramatical. Todo processo de interpretação da língua é guiado pelas restrições semânticas, por meio de um analisador semântico e de mecanismos para desambiguação e resolução anafórica. Uma linguagem específica é utilizada para a geração das interpretações produzidas pelo sistema.

Segundo Kilgarriff (1992), essa linguagem e o sistema de interpretação, como um todo, são severamente limitados pelo uso das restrições de seleção semânticas como o único mecanismo para a DLS. O trabalho de Wilks não considera informações mais expressivas provenientes da sintaxe ou das relações entre palavras vizinhas nas sentenças. Além disso, não considera construções que não apresentam relações do tipo núcleo-modificadores.

De fato, segundo Hirst (1987), o problema da abordagem de semântica de preferência é que as preferências ou, mesmo, restrições de seleção são inadequadas para

tratar todos os casos de ambigüidades lexical, como aquelas que necessitam de um contexto global, principalmente os casos de polissemia.

Apesar dessas limitações, o trabalho de Wilks, com seu conceito de preferências de seleção baseadas em traços semânticos, influenciou muitos dos trabalhos de DLS e de interpretação da LN, as quais procuram integrar esses conceitos a outras informações.

Boguraev (1979), por exemplo, utiliza o formalismo de Wilks para representação semântica, mas procura eliminar os seus problemas, utilizando não só as preferências, mas o que chama de julgamentos semânticos. Ele considera que somente a semântica de preferência não é suficiente para o tratamento da ambigüidade lexical de sentido, principalmente no caso de verbos polissêmicos. Seu trabalho explora o uso de características sintáticas e semânticas de modo integrado para a DLS. Enquanto Wilks trata apenas a desambiguação entre dois diferentes sentidos de uma palavra, ambos com a mesma categoria gramatical, Boguraev realiza a desambiguação entre interpretações da palavra com diferentes categorias gramaticais ou entre sentidos alternativos com a mesma categoria gramatical.

O sistema criado por Boguraev consiste de um *parser* ao qual são adicionados procedimentos semânticos como ações a serem realizadas para a conclusão sobre cláusulas e sintagmas nominais. Esses procedimentos são responsáveis pela desambiguação semântica e estrutural e constroem a representação semântica da sentença, de acordo com o formalismo de Wilks. Primeiramente, os procedimentos aplicam restrições e preferências de seleção e estruturas de caso, entre outros recursos, para tentar desambiguar as palavras. Desambiguações lexicais posteriores são integradas a um processo de desambiguação estrutural. O sistema analisa a coerência semântica das possíveis interpretações para os diferentes sentidos das palavras ambíguas. As interpretações consideradas improváveis são descartadas. Caso todas as interpretações sejam consideradas improváveis, o *parser* procura um novo caminho de análise. Os julgamentos semânticos são feitos apenas com base no conhecimento lexical, pois o sistema não utiliza nenhum conhecimento de mundo. Assim, o autor mostra que é possível desambiguar entre os sentidos considerados usando um conjunto limitado de informações sintáticas e semânticas disponíveis no léxico. O léxico criado possui 400 palavras, o que representa um número relativamente grande para o padrão dos sistemas baseados em conhecimento. Além disso, esse número mostra que tais palavras dificilmente representam casos especiais e, assim, que o trabalho tem grandes possibilidades de extensão.

No sistema de Boguraev, assim como em vários outros sistemas baseados em restrições de seleção, a desambiguação de sentido só ocorre depois da análise sintática e, em alguns casos, somente depois que o domínio da sentença é identificado. O trabalho de Small (1980), por outro lado, baseia-se na teoria de que o conhecimento humano sobre a língua é primariamente organizado na forma de conhecimento sobre as palavras, e não na forma de regras. Assim, no léxico devem estar armazenadas as informações relativas aos papéis sintáticos, as informações semânticas e sobre o domínio do discurso das palavras da sentença a ser desambiguada, de modo que é por meio do processo de desambiguação que são determinados o sentido, a estrutura e o significado da sentença.

No seu sistema de interpretação da língua natural, denominado *Word Expert Parser*, cada palavra é representada no léxico por um *word expert*, um mecanismo que contém informações e procedimentos para a discriminação de todos os seus sentidos. Diversos desses mecanismos especialistas (um para cada palavra) operam em conjunto, diretamente sobre a sentença de entrada, coordenando a escolha do sentido adequado e, ao mesmo tempo, realizando a análise sintática (a mínima necessária) e semântica, produzindo

a representação semântica da sentença como resultado. Assim, não há estrutura sintática intermediária. Os vários *experts* usados para a análise de uma sentença são disparados e controlados em um ambiente distribuído.

Segundo Small, cada palavra apresenta idiossincrasias suficientes para que os *experts* sejam criados caso a caso, um por palavra. Assim, são mais importantes as diferenças entre as palavras do que as suas similaridades. O significado de uma palavra é determinado por um procedimento e uma rede de discriminação que apresenta todos os seus possíveis significados. A partir do contexto de palavras vizinhas das entradas lexicais, o procedimento acessa as informações necessárias dessas palavras e as usa para determinar o caminho a seguir na rede, de modo a encontrar um único significado. A cada resultado parcial, ou seja, cada vez que é determinado o sentido correto de uma palavra, o *expert* adiciona o resultado intermediário à estrutura conceitual que está sendo construída para representar a sentença. Palavras são, então, processos ativos na compreensão da língua e, segundo o autor, a interpretação por regras uniformes não condiz com o processo compreensão da língua realizado pelos humanos.

No trabalho de Small, os *word experts* são, portanto, repositórios de uma grande quantidade de informações que em outros trabalhos normalmente estão distribuídas entre os módulos de análise sintática, léxico, gramática, entre outros. O autor afirma que as informações de controle sobre a desambiguação devem ser armazenadas juntamente com o conhecimento declarativo sobre a palavra. Segundo ele, trabalhos nos quais as informações de controle estão distribuídas em diferentes módulos não apresentam ganhos em termos de funcionalidade e são difíceis de entender ou modificar. Contudo, o autor ressalta que um *expert* não é uma simples cópia de outro, com pequenas alterações, e que por isso cada um deles requer um processo elaborado para a sua criação. Por exemplo, a descrição do *expert* para o verbo *throw* possui seis páginas e, segundo o autor, poderia ser ainda mais detalhadamente especificada, com um tamanho dez vezes maior.

A falta de mecanismos de generalização, em função do princípio de definição individualizada e altamente refinada de cada *expert*, certamente limita a abrangência do sistema. Contudo, o trabalho de Small é relevante ao mostrar que regras e conhecimento específicos sobre cada palavra tornam a desambiguação mais precisa.

Hirst (1987) emprega um modelo de representação do conhecimento baseado no formalismo de *frames*. Seu trabalho consiste de um sistema completo de PLN, que incorpora um léxico, uma gramática, módulos de análise sintática e semântica e módulos específicos de desambiguação estrutural e lexical. As sentenças de entradas e as estruturas produzidas pelo sistema também são representadas usando o mesmo formalismo de *frames*. Uma vantagem desse sistema é a possibilidade de utilizar várias fontes de informação diferentes para a seleção lexical, as quais podem atuar em conjunto.

As palavras de entrada para o sistema são restritas a um pequeno conjunto da língua inglesa, em um domínio específico. O autor procura modelar computacionalmente o processo de desambiguação lexical realizado pelos humanos, de acordo com os princípios da Ciência Cognitiva. Para tanto, ele utiliza a técnica de *semantic priming*. Essa técnica procura reproduzir um processo inerente aos seres humanos, por meio do qual o processamento mental de alguns conceitos influencia e facilita o processamento de conceitos semanticamente relacionados, introduzidos subseqüentemente. Assim, a representação de uma palavra reduziria o tempo de resposta para o reconhecimento de outra palavra semanticamente relacionada (por exemplo, após ouvir/ver a palavra “fruta”, o ser humano reconhece mais rapidamente o conjunto de caracteres (escrito ou falado) que forma “maçã” como uma palavra).

Os modelos de *semantic priming* são geralmente baseados em conceitos de ativação propagada. Neles, a representação mental de conceitos é uma rede, na qual conceitos semanticamente relacionados estão próximos uns dos outros. O uso de um conceito, nessa rede, faz com que ele seja ativado. Por exemplo, processar uma palavra fará com que seu significado seja ativado. Essa ativação não se limita ao conceito acessado, mas se propaga a partir da origem para os nós vizinhos, fazendo com que eles também sejam ativados. No entanto, a ativação se torna mais fraca à medida que os conceitos se distanciam da origem, de modo que o grau de ativação de um conceito será dado em função da sua proximidade semântica com a origem. Além disso, os conceitos não permanecem ativados por muito tempo: sua ativação diminui com o tempo, até que eles voltem ao estado normal.

Hirst emprega a técnica de *semantic priming* por meio de uma estratégia de passagem de marcadores de Quillian, a qual consiste de um modelo discreto de ativação propagada, que passa marcadores pela base de conhecimento para estabelecer quais sentidos de quais palavras são mais fortemente associados a quais sentidos de quais outras palavras na entrada. O objetivo e encontrar o menor caminho de associação entre *frames* de sentido de palavras em um contexto, para escolher o sentido mais apropriado. Esse processo envolve marcar o nó na base de conhecimento para cada sentido da palavra e, então, em um processo iterativo, marcar todos os nós para os sentidos da palavra que aparecem nas entradas lexicais contendo um nó já marcado. Todos os sentidos de todas as palavras na sentença são marcados simultaneamente. Há um conjunto de regras para determinar por quantos passos o processo deve continuar, que ação realizar quando um "sucesso direto" (quando um sentido para uma das palavras sendo marcadas é identificado no primeiro passo, pelo sentido de outra), que ação realizar no caso de "sucesso indireto" (quando um nó é marcado a partir de duas diferentes fontes), entre outros procedimentos.

Hirst denomina seu mecanismo mais especializado de desambiguação de "palavras polaroid". Nele, os sentidos inadequados são progressivamente eliminados com base em evidências sintáticas fornecidas por um analisador sintático, juntamente com as restrições de seleção na rede de *frames*. Ele desenvolve processos de palavras polaroid para substantivos, preposições e verbos. As informações sobre as palavras são mantidas em dois lugares: no pacote de conhecimento manipulado pelo processo da palavra polaroid estão os sentidos alternativos da palavra e na base de conhecimento está o conhecimento de mundo sobre o que cada sentido da palavra denota. O mecanismo usado pelas palavras polaroid é baseado em restrições de seleção, mas como a base de conhecimento já contém uma taxonomia com a hierarquia de tipos de objetos, não há necessidade de definir traços semânticos especialmente para o uso das restrições. Exceto nos casos de uso das palavras em sentidos metafóricos ou metonímicos, Hirst afirma que somente um sentido permanece após a aplicação do mecanismo de palavras polaroid.

De modo geral, o trabalho de Hirst é de grande importância, pois mostra que é possível integrar técnicas de diferentes áreas da representação do conhecimento em Inteligência Artificial e PLN. Reconhecidamente, a DLS requer informações provenientes de várias fontes, como da sintaxe local, das restrições de seleção, das relações entre as palavras, etc. O trabalho de Hirst permite que essas informações co-operem, visando à interpretação da sentença de entrada, e, conseqüentemente, como parte desse processo, a DLS.

Beale (1997) descreve o sistema de Mikrokosmos, cuja tradução é baseada em conhecimento lingüístico profundo. Esse sistema segue o método por interlíngua, em uma representação denominada TMR (*Text Meaning Representation*). O sistema focaliza um domínio específico, de aquisição e fusão entre corporações. Para esse domínio, é definida uma ontologia e um léxico semântico cujas entradas estão relacionadas aos conceitos da

ontologia. No léxico, o significado dos termos de entrada é definido por informações lingüísticas de diversas naturezas (morfológicas, sintáticas, semânticas, etc.), divididas em zonas, de acordo com a sua natureza. São também codificadas algumas informações extralingüísticas, por meio das chamadas micro-teorias.

Nesse sistema, não há procedimentos específicos de DLS, mas as ambigüidades lexicais de sentido são resolvidas pelo analisador semântico, que combina o conhecimento disponível na ontologia e no léxico, aplicando-o à sentença de entrada, para produzir as TMRs. No léxico, há entradas para todos os diferentes sentidos de cada palavra. O significado dos itens lexicais só é definido em termos do seu mapeamento com os conceitos da ontologia e da sua contribuição para a estrutura TMR sendo gerada para a sentença à qual pertence. Para tanto, são estabelecidas restrições de seleção para esse mapeamento, na ontologia, e essas restrições são mutuamente verificadas para a composição da estrutura TMR para toda a sentença de entrada.

Basicamente, o processo de DLS consiste em, a partir de todas as entradas do léxico para uma palavra ambígua, isto é, todos os seus sentidos, que satisfazem as restrições sintáticas da sentença atual, verificar se as características semânticas da entrada de cada sentido satisfazem uma série de restrições para o seu mapeamento em conceitos da ontologia, considerando também os possíveis mapeamentos das palavras vizinhas na sentença também em conceitos da ontologia. Todas as possíveis combinações entre todos os sentidos das palavras da sentença de entrada são verificadas e um escore é calculado de acordo com a satisfação mútua das restrições em cada combinação. A TMR cuja combinação apresenta o maior escore é escolhida como representação semântica da sentença.

O autor avalia o processo de DLS do seu sistema considerando textos com 68 sentenças, com 119 palavras com ambigüidade de sentido. 91% dessas palavras ambíguas foram corretamente desambiguadas pelo sistema. Contudo, o autor menciona que esses mesmos textos foram usados como base para a construção do sistema. Para minimizar a tendenciosidade da avaliação, ele realiza uma outra avaliação, com 17 palavras com ambigüidade de sentido. 14 dessas palavras (88.5%) foram corretamente desambiguadas.

Goodman & Nirenburg (1991) descrevem a criação de um sistema semelhante ao Mikrokosmos, de TA por interlíngua para a tradução de manuais técnicos (sobre computadores) entre o inglês e o japonês. Esse sistema, também baseado em conhecimento lingüístico profundo, é denominado KBMT (*Knowledge-Based Machine Translation*). A sua interlíngua consiste de uma hierarquia conceitual que foi manualmente construída, especificamente para esse sistema. Os itens lexicais são representados em dicionários monolingües e mapeados nos conceitos dessa ontologia, que são independentes de língua e, em princípio, não ambíguos.

Nesse sistema, não há um módulo específico de DLS, mas as ambigüidades na língua-fonte são resolvidas durante o processo de mapeamento dos itens lexicais da língua-fonte em conceitos não ambíguos da interlíngua, por meio de restrições de seleção. Certamente, isso é possível porque a ontologia é delimitada a um único domínio. Para sistemas independentes de domínio, trabalhos de DLS fundamentados principalmente em uma ontologia são pouco viáveis, dada a complexidade para a construção de uma ontologia dessa natureza e a quantidade limitada de conhecimento que ela poderá prover.

Outros sistemas de TA, como o EUROTRA (Copeland, 1991) e METAL (Gajek, 1991), empregam procedimentos mais simples de DLS. Eles procuram tratar a ambigüidade lexical por meio da definição de estruturas argumentais e de restrições ou preferências de seleção sobre essas estruturas. No sistema EUROTRA, em particular, uma

hierarquia simples de tipos semânticos (entidade, humano, não-humano, etc.) é utilizada para tratar os casos de desambiguação mais refinada, ainda com base em preferências de seleção. O sistema aplica a noção de distância semântica entre os nós dessa hierarquia. Para a desambiguação de um substantivo que complementa o verbo em uma sentença, por exemplo, a hipótese é de que quanto menor a distância entre o nó que representa o sentido de um substantivo e os nós que representam as restrições impostas na estrutura argumental do verbo em questão, maior a indicação de que esse é o sentido do substantivo. Contudo, de modo geral, como afirma Pedersen (1997), essas restrições são simples e limitadas, de modo que resolvem apenas alguns casos mais simples de ambigüidade. A autora sugere que é necessária uma proposta mais elaborada, no entanto, não tão complexa quanto as interlinguais.

Sistemas comerciais de TA que oferecem algum tratamento à DLS empregam métodos mais simples, em função da necessidade de abrangência a qualquer gênero e domínio de textos. O Systran®[6], considerado por muitos como o melhor sistema de TA disponível atualmente, adota uma visão bastante prática do processo de DLS (cf. Flanagan & McClure, 2002): procura identificar o domínio do texto sendo traduzido para acessar dicionários específicos de cada domínio. Isso é feito com base na análise de traços sintático-semânticos (objeto concreto, sujeito humano, etc.) e das categorias semânticas (dispositivo, propriedade, etc.) das palavras do contexto, armazenadas nos dicionários do sistema. Contudo, nem todas as entradas possuem essas informações e o seu uso não é efetivo, na maior parte dos casos. Além disso, dependendo do tamanho do texto a ser traduzido e da sua natureza, a identificação do domínio não é possível. Para os casos mais simples, o Systran também possui entradas específicas para algumas expressões idiomáticas, locuções comuns e termos técnicos de diversas áreas.

Os sistemas comerciais não disponibilizam muitas informações sobre o seu processo de DLS. Contudo, segundo Mowatt (1999), de modo geral, esses sistemas não incorporaram conhecimentos semânticos de várias fontes, como os mencionados na Seção 2.5, necessários para resolver o problema da ambigüidade lexical de sentido. Em um experimento com os dois principais sistemas de TA, Globalink Power Translator® Pro[7] e Systran, Mowatt compara a tradução de 3.000 palavras do inglês para o francês realizada por esses sistemas à tradução humana. Segundo o autor, o Systran, em especial, incorpora alguns conhecimentos semânticos, mas em quantidade muito pequena, insuficiente para manipular muitos problemas, como a ambigüidade.

Seguindo a proposta de Hirst, McRoy (1992) define e implementa um trabalho no qual a desambiguação é realizada a partir de informações de várias fontes. Ela utiliza esse modelo em um sistema de interpretação da língua natural. Além das informações sugeridas por Hirst, McRoy acrescenta outras: traços morfológicos, categorias gramaticais, *collocations* e relações sintáticas com restrições de seleção.

Uma característica marcante no trabalho de McRoy é a abrangência do seu sistema, considerando-se que é um sistema baseado em conhecimento manualmente codificado. Enquanto a maioria dos sistemas dessa natureza se limita a manipular poucas dezenas ou, no máximo, algumas centenas de palavras, seu sistema possui um léxico com 13.000 sentidos. Em função dessa grande quantidade de sentidos, o sistema não realiza a análise das sentenças antes da DLS, uma vez que o número de estruturas possíveis geradas seria muito grande. Assim, a desambiguação ocorre em conjunto com o processo de análise sintática.

[6] http://www.systransoft.com
[7] http://www.bmsoftware.com/powertranslator.htm

O léxico citado é denominado “central” e é independente de gênero e domínio. Nele, são armazenadas apenas as distinções de sentido principais (mais genéricas), válidas em todos as situações, além de informações sintáticas e associações com uma hierarquia conceitual. Distinções de sentidos mais refinadas são armazenadas em outro léxico, denominado léxico “dinâmico”. Esse léxico contém sentidos válidos apenas em contextos particulares (domínios e *collocations* específicos, por exemplo), os quais são ortogonais com relação aos sentidos do léxico central.

McRoy utiliza lingüística de córpus para extrair automaticamente padrões de *collocations* a partir de um córpus do mesmo gênero e domínio dos textos a serem desambiguados. As *collocations* são armazenadas no léxico dinâmico. A suposição da autora é de que se duas ou mais palavras em uma *collocation* ocorrem juntamente em um texto de entrada, isso provê uma forte evidência de que essas palavras estão sendo usadas como *collocation*.

McRoy cria também uma hierarquia conceitual, com cada sentido sendo relacionado a um conceito “pai” no léxico. Com esse recurso, pode-se simular, de maneira otimizada, o processo de passagem de marcadores de Hirst: para testar se dois conceitos são semanticamente relacionados, o sistema busca o ponto mais baixo na taxonomia em que ainda é possível identificar um antecedente comum para ambos; quanto mais baixo na taxonomia esse antecedente estiver, maior a relação semântica entre os conceitos.

O uso da hierarquia conceitual permite, ainda, que as restrições ou preferências de seleção, bem como outras informações em comum para determinados *clusters*, sejam estabelecias apenas para alguns itens lexicais (os itens mais altos na hierarquia), já que os itens derivados podem automaticamente herdar essas informações. Além disso, a hierarquia conceitual permite determinar um contexto semântico. Esse contexto é modelado por meio de *clusters* conceituais, extraídos de um córpus, que agrupam sentidos que compartilham o mesmo conceito, por exemplo, *c-published-document*, para indicar o grupo de sentidos relacionados a “documentos publicados”. Segundo a autora, *clusters* dessa natureza podem ser considerados *collocations* para as quais não há um padrão sintático previsível. Os *clusters* são definidos no léxico dinâmico do sistema e têm essencialmente o mesmo propósito que a rede semântica de Hirst: um *cluster* contendo o sentido de uma palavra será ativado se ele contiver qualquer um dos sentidos sob consideração para outras palavras no contexto atual.

No processo de desambiguação do sistema de interpretação da língua natural de McRoy, os sentidos candidatos de todas as palavras em uma sentença são pré-selecionados depois de uma fase de pré-processamento que inclui indicações morfológicas, sintáticas (etiquetação gramatical), das *collocations* e dos *clusters* conceituais. Esses sentidos são então fornecidos ao analisador sintático e semântico. Esse processo calcula o escore de cada sentido, com base em diferentes pesos atribuídos para as diferentes características consideradas. O sentido com o escore mais alto é eleito pelo sistema como o mais adequado. Por fim, a representação semântica considerando tal sentido é gerada pelo sistema.

A autora não avalia o desempenho do módulo de desambiguação, individualmente, alegando a dificuldade para criar material de referência para contrastar com os resultados do sistema. Segundo Kilgarriff (1992), um problema desse trabalho é que não há uma justificativa clara para a suposição de que distinções mais refinadas entre sentidos sejam sempre dependentes de domínio. Essa característica limita o processo de desambiguação em determinados domínios. Por exemplo, a palavra *engage* só pode ser interpretada com o sentido de “ataque” no domínio militar.

Além disso, os sentidos do léxico central são considerados os mais freqüentes e, por isso, preferidos no processo de DLS. Os sentidos dos léxicos específicos de cada domínio

são considerados pouco freqüentes e somente são escolhidos quando a escolha falha para todos os sentidos do léxico central, por exemplo, porque todos infringem as restrições de seleção. Com isso, a escolha pode tornar-se altamente tendenciosa a privilegiar somente os sentidos mais freqüentes. Outro problema é que com o uso do léxico dinâmico o sistema torna-se dependente de domínio e precisaria então ser personalizado, com um novo léxico dinâmico para abranger para cada domínio.

Apesar dessas limitações, o trabalho de McRoy é bastante importante para a área de DLS, pois integra uma grande variedade de informações, além das já utilizadas em outros trabalhos, e, adicionalmente, ressalta a importância das *collocations* para a desambiguação. Vale notar que, apesar de extrair informações de um corpus, esse trabalho não é considerado baseado em corpus, pois o mecanismo de desambiguação é especificado manualmente. Além disso, apesar de utilizar um grande léxico, também não é considerado baseado em conhecimento pré-codificado, pois esse léxico foi criado manualmente, especificamente para o seu sistema.

O sistema UNITRAN de tradução automática por interlíngua entre o inglês, o espanhol e o francês (Dorr, 1993) é um exemplo representativo dos trabalhos tradicionalmente empregados para o tratamento da ambigüidade lexical nos sistemas de TA não comerciais. Como a maioria dos sistemas, o UNITRAN não dispõe de um mecanismo específico para esse problema. Em vez disso, o seu tratamento é embutido em outros módulos do sistema.

O sistema utiliza uma adaptação das estruturas conceituais lexicais (*Lexical Conceptual Structures*) de Jackendoff (1990), tanto para a representação dos itens lexicais quanto das estruturas conceituais compostas por vários itens. A interlíngua do sistema corresponde à composição de várias dessas estruturas para a representação de sentenças específicas, de acordo com as palavras da sentença.

No UNITRAN, todos os problemas (ou “divergências”) de tradução, em diversos níveis, são tratados de acordo com uma estratégia similar. A ambigüidade lexical, em especial, é considerada uma das variações do problema de divergência lexical. Por ser um sistema por interlíngua, a ambigüidade pode ocorrer tanto na geração das estruturas conceitual (na interlíngua) a partir da língua-fonte quanto na realização lexical das estruturas da interlíngua para a língua-alvo. Contudo, autora não menciona o problema da ambigüidade lexical na geração das estruturas conceituais, supostamente porque as entradas lexicais são descritas na forma de estruturas conceituais parciais, com diversas informações que permitem identificar, univocamente, qual a estrutura conceitual parcial correspondente a cada palavra na sentença.

Assim, a ambigüidade lexical é tratada apenas como um problema de seleção lexical, na realização das estruturas conceituais compostas para a língua-alvo. A necessidade de escolha ocorre quando uma parte da estrutura conceitual composta (que representa um conceito) da pode combinar com mais de uma estrutura conceitual lexical da língua-alvo, ou seja, quando um conceito pode ser realizado por mais de uma palavra. Essa escolha é feita por meio da verificação das estruturas lexicais que satisfazem as restrições de seleção sintáticas e semânticas presentes na estrutura conceitual composta, por meio de um processo similar ao de unificação. Tais restrições incluem vários níveis da teoria de Jackendoff: primitivas, tipos, campos e traços de cada estrutura conceitual lexical devem combinar com os mesmos itens da estrutura conceitual composta. Várias estruturas podem combinar em todos esses itens, de modo que, em alguns casos, o sistema retorna mais de uma realização lexical. Contudo, segundo a autora, a idéia não é, de fato, encontrar a melhor combinação, mas simplesmente encontrar combinações satisfatórias.

A autora afirma que essas restrições não capturam distinções que não sejam caracterizadas por propriedades puramente sintáticas, por exemplo, por distinções que dependem de conhecimento do discurso, de domínio ou de mundo, conhecimento de usos idiomáticos, etc. Essa limitação nos tipos de conhecimento considerados no sistema, bem como o fato de o sistema poder retornar várias realizações lexicais, implica, certamente, que muitos casos de ambigüidade não são resolvidos. De fato, segundo a autora, um tratamento mais explícito a esse problema, considerando-se, por exemplo, restrições contextuais e *collocations*, poderia melhorar o processo de escolha lexical do sistema.

Egedi et al. (1994) apresentam um sistema de TA por transferência entre coreano e o inglês, implementado de acordo com o formalismo de representação *Synchronous Tree Adjoining Grammar* (STAG) (Shieber & Schabes, 1990). O sistema possui um módulo de DLS para tratar da ambigüidade de alguns verbos, com base na unificação de restrições de seleção semânticas definidas na estrutura argumental desses verbos com os traços semânticos definidos para os substantivos que podem ser utilizados como seus argumentos. As regras de transferência, incluindo as restrições de seleção e os traços semânticos, são manualmente especificadas.

A DLS ocorre no processo de transferência lexical, com base nas possíveis traduções especificadas em um dicionário bilíngües e nas restrições de seleção e traços semânticos especificados na língua-alvo. Os autores justificam a especificação desse conhecimento na língua-alvo porque, segundo eles, a seleção lexical normalmente depende da existência de traços semânticos nos elementos da língua-alvo que são completamente irrelevantes para a língua-fonte. Eles citam, como exemplo, a tradução do verbo *wear*, do inglês para o coreano. No coreano, a tradução depende do complemento do verbo: "*wear clothes*" e "*wear socks*" são traduzidos por verbos completamente diferentes. No entanto, no inglês, não há distinção.

Um problema com esse trabalho é que a desambiguação de um verbo depende da tradução correta dos seus argumentos, já que estes precisam ser primeiramente traduzidos para a língua-alvo para que possam ser identificados seus traços semânticos. Certamente, poderão surgir problemas se os argumentos do verbo também forem ambíguos.

Pedersen (1997) descreve um trabalho baseado em teorias da semântica lexical para a desambiguação de um subconjunto de verbos de movimento polissêmicos na TA do dinamarquês para o inglês. A autora considera apenas o fenômeno da polissemia sistemática desse subconjunto. O seu objetivo é identificar padrões para o tratamento de polissemia sistemática, ou seja, que possam ser aplicados a diversos verbos com significado relacionado, dentre os verbos de movimento, formalizar e implementar esses padrões na forma de regras lexicais que possam ser usadas para a DLS na TA.

Para tanto, primeiramente, foi realizada uma análise das ocorrências de 100 verbos de movimento em diferentes córpus do dinamarquês para verificar propriedades estatísticas (freqüência, co-ocorrências, etc.) e outras características do uso desses verbos, bem como os tipos de conhecimento que são necessários para diferenciar os seus sentidos. Para essa análise, foram selecionados de 100 a 300 exemplos de ocorrência de cada um dos verbos. A partir da análise, os exemplos foram manualmente categorizados de acordo com suas propriedades sintáticas e semânticas. Nessa etapa, foram estabelecidas várias delimitações, por exemplo, foram descartados exemplos do uso do verbo em expressões idiomáticas e metafóricas. No processo de categorização foram agrupados os exemplos de acordo com os padrões de valência do verbo de movimento, separados os exemplos com verbos que possuíam elementos modificadores de direção dos que não possuíam, etc. Como resultado dessa etapa, foram formados grupos de exemplos com propriedades similares, por

exemplo, exemplos de verbos de movimento que têm uma direção específica, cujo agente é animado e que implica o movimento de partes do corpo ou de uma máquina.

A partir dessa análise, os verbos foram classificados em uma taxonomia para os verbos de movimento, de acordo com suas propriedades sintáticas e semânticas e, principalmente, com as regularidades nos desvios do significado básico para os demais sentidos. Nesses verbos, segundo a autora, a polissemia deve se manifestar de maneira sistemática, de modo que todos os verbos do grupo podem receber o mesmo tratamento na DLS.

Para representar os verbos dos grupos, a autora definiu um modelo lexical, baseado na teoria *Frame Semantics* (Atkins & Fillmore, 1994)[8] e na Estrutura de Eventos do léxico gerativo de Pustejovsky. Também foram definidas uma hierarquia conceitual parcial e restrições de seleção para substantivos distribuídos nessa hierarquia. Os verbos foram então especificados de acordo com o modelo definido, utilizando uma grande quantidade de informações lingüísticas na língua-fonte, em diversos níveis, que indicam os desvios de significado e, portanto, podem auxiliar na desambiguação. Os esquemas especificados foram implementados na forma de regras lexicais e incorporados a um sistema de interpretação do dinamarquês. A autora realizou um teste com 42 sentenças com os verbos ambíguos. Desses verbos, 39 foram corretamente desambiguados.

Apesar da aplicação voltada para a TA, o foco desse trabalho de DLS é na especificação das regras lexicais com informações suficientes para permitir capturar os padrões sistemáticos entre os diferentes sentidos de um verbo, de modo a evitar descrições ambíguas. Assim, a desambiguação ocorre, basicamente, na língua-fonte.

Dorr & Katsova (1998) definem um mecanismo de seleção lexical para verbos e substantivos deverbais na TA (entre o inglês e o espanhol) que se baseia na estrutura argumental desses elementos, representada por meio de estruturas conceituais lexicais (LCSs), e nos sentidos da WordNet. A hipótese é de que a tradução de um elemento da língua-fonte pode ser desambiguada se forem escolhidos, na língua-alvo, elementos que tenham a mesma LCS e que pertençam ao mesmo *synset* da WordNet, ou seja, que sejam sinônimos do elemento na língua-fonte.

Para testar sua hipótese, as autoras implementam um algoritmo de seleção lexical que utiliza um sistema já existente para codificar sentenças em suas representações LCSs. Esse sistema também possui um léxico do inglês e outro do espanhol, cujas entradas estão codificadas como LCSs, com um código correspondente ao *synset* da WordNet ao qual pertencem (anotado manualmente). Com base na estrutura gerada pelo sistema para uma sentença, o algoritmo extrai a estrutura LCS genérica do verbo a ser desambiguado, sem as constantes que representam as palavras da sentença, e recupera do léxico do espanhol todas as entradas correspondentes a verbos que têm a LCS com as mesmas propriedades estruturais. Por exemplo, para o verbo *sap*, são recuperados 358 verbos do espanhol com a mesma estrutura de LCS. Desse conjunto de verbos, o algoritmo seleciona apenas aqueles que apresentam o mesmo código do *synset* que o verbo sendo desambiguado. Se o verbo puder pertencer a vários *synsets*, são selecionados todos os verbos em todos os seus *synsets*. Para o *verbo sap*, apenas um verbo (*escurir*) pertence ao mesmo *synset*. Esse verbo é então escolhido como a tradução mais adequada para o verbo do inglês.

Caso haja mais de um verbo com a mesma estrutura e o mesmo código de *synset*, o algoritmo retorna todos eles. Por outro lado, caso não seja encontrado nenhum verbo com a

[8] Atkins, S.; Fillmore, C. (1994). Starting Where the Dictionaries Stop: The Challenge of Corpus Lexicography. In B. Atkins & A. Zampolli (eds), *Conceptual approaches to the lexicon*. Oxford University Press, Oxford. Apud Pedersen (1997).

LCS equivalente no mesmo *synset*, o algoritmo estende a busca aos *synsets* hiperônimos em um nível (mais genéricos) de todos os *synsets* aos quais o verbo pertence.

O algoritmo pode operar também na DLS monolingüe. Nesse caso, as buscas por LCSs equivalentes são feitas no léxico da própria língua. As autoras realizam experimentos para a DLS de três verbos do inglês, monolingüe e multilingüe. Na DLS monolingüe, dois dos verbos possuem exatamente um equivalente em estrutura e *synset*, enquanto para o outro verbo só é encontrado um equivalente quando são analisados os *synsets* hiperôminos. Na DLS multilingüe, um verbo possui exatamente uma tradução, enquanto os outros dois possuem duas e quatro traduções. Nenhum outro tipo de conhecimento é empregado para filtrar essas possíveis traduções.

As autoras afirmam que o seu método é mais efetivo para a DLS monolingüe. Mencionam também que se a DLS monolingüe for realizada como um pré-processamento para a multilingüe, ela pode reduzir o número de ambigüidades, melhorando a precisão na tradução.

Não são realizados experimentos de avaliação mais abrangentes, mostrando se a proposta é realmente viável. Um problema desse trabalho é que como são recuperadas todas as LCSs estruturalmente equivalentes, podem ser recuperados verbos que, apesar de estarem no mesmo *synset*, não são válidos como tradução do verbo na língua-fonte. Um possível filtro, bastante simples, seria buscar apenas as estruturas dos verbos que podem ser traduções do verbo na língua-fonte, a partir da consulta a um dicionário bilíngüe. Com relação à abrangência, o sistema é limitado às LCSs que já estão codificadas no léxico, às quais já foi atribuído um código de *synset*. Além disso, o fato de serem retornadas todas as traduções possíveis indica que o sistema não elimina todas as ambigüidades.

Apesar de utilizar informações da WordNet, esse trabalho não é considerado baseado em conhecimento pré-codificado, pois a maior parte do conhecimento é especificada manualmente, incluindo os *synsets* correspondentes às entradas lexicais.

O único trabalho multilingüe voltado explicitamente para a DLS e envolvendo o português é o de Leffa (1998), que focaliza a importância do uso do contexto local da palavra ambígua, isto é, das palavras vizinhas a ela na sentença, na forma de *collocations*, para a desambiguação na TA. Ele afirma que *collocations* são mais efetivas para a DLS que outras características mais profundas, como conhecimento de mundo, devido à dificuldade em se representar e utilizar esse conhecimento e à natureza dinâmica do uso das palavras.

Leffa também defende a análise do uso das palavras em córpus para definir o seu conjunto de possíveis sentidos. Segundo ele, em um contexto multilingüe, é possível estabelecer uma metodologia bastante objetiva para a definição desse conjunto de sentidos, a partir de exemplos de tradução. Para investigar sua hipótese, o autor realiza um experimento para desambiguar 20 substantivos ambíguos do inglês para o português, contextualizados em exemplos de tradução extraídos de um córpus de 20.000.000 de palavras

Para cada palavra, foram aleatoriamente selecionados 200 exemplos, sendo que cada exemplo consiste de um segmento com 20 palavras, em média. O autor não menciona como as regras de desambiguação são construídas, apenas que são incorporadas às regras de um sistema de TA inglês-português em fase inicial de construção. Ao que tudo indica, as regras são manualmente codificadas, com base em um conjunto de *collocations* pré-definidas, que também não são explicitadas no trabalho. Na sua avaliação, o autor relata uma acurácia média de 94% para as 20 palavras. No entanto, como esse modelo se baseia apenas nas palavras da sentença, na forma de *collocations*, sua abrangência deve ser

bastante limitada. Tanto o módulo de DLS quanto o sistema de TA mencionados não foram concluídos[9].

Em um outro trabalho para a língua portuguesa que menciona o tratamento explícito ao problema da ambigüidade lexical monolingüe, Bräscher (2002) emprega um sistema de PLN para a resolução de vários tipos de ambigüidades na Recuperação de Informações monolingüe, incluindo ambigüidades morfológicas, lexicais, sintáticas, predicativas e semânticas. Esse sistema se baseia fundamentalmente em informações relativas à valência sintático-semântica das unidades lexicais que compõem um enunciado, de acordo com a Teoria de Valências de Borba (1996), representadas por Gráficos Conceituais (Sowa, 1984)[10]. Essas informações correspondem à estrutura de argumentos (ou relação núcleo-argumentos), traços semânticos e restrições de seleção.

O sistema de PLN adotado foi desenvolvido para outros fins, podendo ser personalizado (manualmente) em diversos aspectos, incluindo principalmente aqueles relacionados à língua envolvida no processo de compreensão. A autora não cita detalhes da utilização efetiva desse sistema como auxiliar na Recuperação de Informações; realiza apenas alguns testes, procurando ilustrar casos em que as informações da valência sintático-semântica das unidades lexicais provêem a resolução de ambigüidades, nos diversos níveis, e casos em que somente essas informações não são suficientes. No que se refere às ambigüidades lexicais, especificamente, a autora mostra diversos casos, tanto de polissemia quanto de homografia, que não podem ser resolvidos apenas com essas informações. Com isso, apesar do sistema de PLN resolver outros casos de ambigüidade, a conclusão da autora é de que quanto mais conhecimento lingüístico/cognitivo for incorporado ao sistema, maior será a precisão na recuperação.

Abordagens conexionistas

Alguns trabalhos das décadas de 60 e 70, fundamentados da Ciência Cognitiva, procuravam modelar a técnica de *semantic priming* citada por meio de redes neurais artificiais. Além de permitirem "simular" o processamento do cérebro humano, em teoria, as redes neurais se mostram mecanismos bastante apropriados para a tarefa de DLS, uma vez que provêem um ambiente adequado: muitos processos exigindo que várias restrições sejam simultaneamente satisfeitas. O trabalho de Quillian (1961), já citado, é considerado precursor dos modelos de ativação propagada para a DLS. Contudo, esse trabalho é simbólico, enquanto os trabalhos usando redes neurais são numéricos. Assim, diferentemente desse trabalho, na qual a propagação é implementada por meio de estratégias como a passagem de marcadores, nos trabalhos conexionistas, a propagação é numérica, baseada em modelos de redes neurais.

Várias abordagens de redes neurais foram desenvolvidas com base em modelos de ativação propagada para implementar a técnica de *semantic priming*, por exemplo, Cottrell & Small (1983). Nesse trabalho, assim como nos simbólicos, cada nó representa uma palavra ou um conceito em uma rede semântica. Os conceitos são ativados durante o uso e a ativação se espalha para os nós conectados. Os nós que recebem ativações de diversos outros nós são progressivamente reforçados e têm, portanto, mais chance de serem escolhidos como o sentido mais adequado. Por outro lado, a ativação se torna mais fraca à medida que se espalha.

[9] De acordo com comunicações pessoais com o autor.
[10] Sowa, J.F. (1984). *Conceptual Structres: Information Processing in Mind & Machine*. Addison-Wsley, Massachusetts. Apud Bräscher (2002).

Outros trabalhos implementaram modelos similares a esse para a DLS, acrescentando (manualmente) informações aos nós que representam os conceitos. Waltz & Pollack (1985), por exemplo, em vez de considerar simplesmente as palavras como nós da rede, adicionam a cada nó conjuntos de micro-traços semânticos. A motivação para a inclusão dessas informações é que as palavras codificadas nas redes podem não estar presentes no contexto da palavra a ser desambiguada. Assim, as características poderiam ser usadas para permitir a correspondência com palavras semelhantes. Esses micro-traços correspondem a distinções semânticas básicas (animado/inanimado, comestível/não-comestível, etc.), características de duração de eventos (segundo, minuto, hora dia, etc.), localizações (cidade, país, etc.), entre outras.

Em seu trabalho, precisam ser codificados, segundo os autores, milhares de micro-traços. Cada nó da rede é ligado, por meio de relações de inibição ou de ativação bidirecionais, a apenas um subconjunto do conjunto total de micro-traços. Uma palavra pode compartilhar micro-traços com outras palavras, com as quais está semanticamente relacionada. Essa palavra irá, portanto, ativar os nós correspondentes a essas palavras relacionadas quando ela for ativada. Os conjuntos de micro-traços precisam ser manualmente inicializados para ativar um contexto para desambiguar uma palavra de entrada subseqüente.

Segundo Véronis & Ide (1990), a idéia de micro-traços é problemática devido à dificuldade em se definir um conjunto apropriado de traços, que, em essência, correspondem às primitivas semânticas de outros trabalhos de PLN.

Um trabalho bastante significativo desenvolvido com base em redes neurais simulando a técnica de *semantic priming* é a de Cottrell (1989). O foco do trabalho do autor é na Ciência Cognitiva, não no PLN. Ele concentra-se na análise da técnica de *semantic priming*, usando a DLS apenas para testar a aplicação dessa técnica.

Cottrell criou um sistema de compreensão da LN, com um módulo lexical, um módulo sintático, para a determinação da categoria gramatical das palavras ambíguas, e um módulo de interpretação de casos, para a utilização de restrições ou preferências de seleção. Todos os módulos cooperam para atribuir um sentido para cada item lexical ambíguo, com base em um repositório de sentidos.

O módulo lexical é o mecanismo de entrada para a rede neural. À medida que cada palavra da sentença é lida, sua representação é disparada. As palavras ativam seus sentidos, e estes ativam os papéis sintáticos e os casos que eles esperam preencher ou ter preenchidos, bem como todos os conceitos relacionados, como na passagem de marcadores. Uma interpretação para a sentença é encontrada quando, para cada palavra, somente uma unidade de sentido, uma unidade de caso e uma unidade de papel sintático é disparada.

O modelo de Cottrell é bastante expressivo, porém, muito complexo. Seu trabalho aplica, sem distinção, um estilo de representação e processamento para informações sintáticas, semânticas, lexicais, restrições de seleção e associações entre as palavras. Seus argumentos para essa decisão são de que: (a) no cérebro humano, todas essas informações fluem em um meio comum, por meio de impulsos ao longo dos neurônios, e (b) há evidencias psicolingüísticas de que o processamento de todas essas informações ocorre em paralelo nos seres humanos. Apesar de válidos, esses argumentos, segundo Kilgarriff (1992), não indicam que uma abordagem conexionista é a melhor alternativa para modelar ou reproduzir o comportamento humano. Uma abordagem mais modular, em que as diferentes informações são representadas e processadas por componentes distintos, certamente seria mais fácil de compreender, manipular e estender.

Em uma rede neural como a de Cottrell, voltada para a compreensão da LN, não é possível isolar a parte do processamento responsável apenas pela desambiguação lexical e, com isso, não é possível compreender como a desambiguação esta sendo realizada, avaliar esse processo ou, ainda, modificá-lo para tentar aprimorá-lo. Como ressalta Kilgarriff, a possibilidade de considerar diversas informações, em conjunto, é importante para a DLS, mas seu valor depende da sua correção em análises especializadas de cada tipo de informação, o que só é possível se elas puderem ser isoladas.

Os trabalhos baseados em conhecimento manualmente codificado descritos acima são listados, de maneira resumida e em ordem cronológica, na Tabela 2, de acordo com a abordagem empregada, o conjunto de palavras para as quais forma criados e/ou testados, o nível de refinamento das distinções entre os sentidos e a acurácia apresentada. Vale notar que, em alguns casos, determinadas informações não são explicitadas nas publicações referentes a tais trabalhos (representadas por "?").

Tabela 2. Lista dos trabalhos de DLS baseados em conhecimento manualmente codificado

Trabalho	Abordagem	Conj. palavras	Nível de refinamento dos sentidos	Acurácia
(Masterman, 1961)	simbólica	?	?	?
(Quillian, 1961)	simbólica	?	?	?
(Hayes, 1976)	simbólica	alguns substantivos	?	?
(Wilks, 1975)	simbólica	?	?	?
(Boguraev,1979)	simbólica	?	?	?
(Small, 1980)	simbólica	?	alto	?
(Hirst, 1987)	simbólica	?	?	?
(Beale, 1997)	simbólica	17 palavras	?	88.5%
(Goodman & Nirenburg, 1991)	simbólica	?	alto	?
(Copeland, 1991)	simbólica	?	?	?
(Gajek, 1991)	simbólica	?	?	?
Systran	simbólica	irrestrito	?	?
(McRoy, 1992)	simbólica	13.000 palavras	alto	?
(Dorr, 1993)	simbólica	?	baixo	?
(Egedi et al., 1994)	simbólica	alguns verbos	?	?
(Pedersen, 1997)	simbólica	100 verbos de movimento	alto	92%
(Dorr & Katsova, 1998)	simbólica	verbos e substantivos deverbais	alto	?
(Leffa, 1998)	simbólica	20 substantivos	?	94%
(Bräscher, 2002)	simbólica	?	?	?
(Cottrell & Small, 1983)	conexionista	?	?	?
(Waltz & Pollack, 1985)	conexionista	?	?	?
(Cottrell, 1989)	conexionista	?	?	?

3.1.2 Conhecimento pré-codificado

Léxicos computacionais

Conforme será descrito a seguir, a grande maioria dos trabalhos baseados em léxicos computacionais emprega a WordNet ou versões multilingües desse recurso, como a EuroWordNet, como repositório de sentidos e/ou fonte de informações variadas. Nenhum dos trabalhos é voltado para a TA.

Voorhees (1993) propõe um trabalho voltado para a Recuperação de Informações do inglês com base na WordNet, com grupos de sentidos menos refinados que os *synsets*, porém, não tão genéricos quanto as suas classes semânticas (sub-hierarquias pricipais). Para tanto, a autora define uma construção, denominada "toldo", para representar categorias de sentidos agrupando vários *synsets*, com base nas relações de hiponímia entre *synsets* de substantivos da WordNet. Um toldo de um determinado *synset s* é o maior subgrafo conexo que contém *s* e contém os *synsets* descendentes de um hipônimo de *s*, mas não contém os *synsets* que têm um descendente que inclui outra instância de um membro de *s* como membro. Os sentidos em seu trabalho são, portanto, os toldos, sendo que o identificador de um toldo é o *synset* que está na sua raiz.

Para desambiguar uma palavra em um texto, as palavras desse texto que ocorrem em cada um dos diferentes toldos são contadas. Isso é feito tanto para a sentença de busca (soma local) quanto para os documentos encontrados (soma global). O sentido correspondente à raiz do toldo para o qual a diferença entre as somas global e local é a maior é escolhido para aquela palavra.

O trabalho de Voorhees é uma tentativa de utilização de distinções menos refinadas que os *synsets* da WordNet, já que eles são, reconhecidamente, muito refinados, principalmente para a Recuperação de Informações. Experimentos relatados pela autora, contudo, indicam que seu trabalho não é mais adequado para a recuperação de informações que os procedimentos de busca tradicionais, na maioria das vezes. Contudo, isso não se deve ao uso dos toldos, mas sim ao fato de que não é possível identificar o toldo adequado em função do tamanho reduzido das sentenças de busca, que não provêem um contexto suficiente para a desambiguação.

Sussna (1993) apresenta uma trabalho também voltado para a Recuperação de Informações que se baseia nas distâncias entre os sentidos na hierarquia conceitual da WordNet. A hipótese do autor é de que, para um dado conjunto de termos ocorrendo próximos uns dos outros em um texto, cada um deles podendo ter vários sentidos, o sentido que minimiza a distância entre eles na hierarquia representa o sentido mais adequado. Essa distância é denominada "distância semântica".

O trabalho de Sussna considera a desambiguação de todos os substantivos de um texto, previamente identificados por um etiquetador morfossintático. Para calcular a distância semântica para cada substantivo do texto a ser desambiguado, o sistema analisa a distribuição na hierarquia conceitual da WordNet de todos os possíveis sentidos desse substantivo, bem com de todos os possíveis sentidos das palavras vizinhas na sentença em uma janela de contexto pré-definida, visando identificar o grau de relação entre os pares de sentidos das palavras.

Esse grau é computado por uma medida que atribui pesos às diferentes relações da WordNet de acordo com o seu tipo (holonímia, sinonímia, hiponímia, etc.) e calcula a distância entre dois sentidos na hierarquia, com base no número de arcos do mesmo tipo partindo de um nó (um sentido) e na profundidade de uma determinada aresta na hierarquia global. Como resultado, um escore global do grau de relacionamento é computado entre cada possível sentido e os sentidos das palavras do contexto. O sentido com o maior escore, que corresponde ao caminho de menor distância na hierarquia, é então escolhido.

Sussna realiza diversos experimentos de avaliação considerando cinco documentos jornalísticos e variações no tamanho e no tipo da janela de contexto, bem como diferentes esquemas de pesos das relações hierárquicas para o cálculo da distância semântica. Para a comparação, é utilizada como *baseline* a escolha ao acaso e como valores de referência, as etiquetas atribuídas manualmente. A partir da etiquetação manual, o autor seleciona para os

testes somente os substantivos cuja desambiguação é considerada possível pelo humano e cujos sentidos são bastante distintos entre si ("não-triviais").

A avaliação mostra que os resultados do sistema são significativamente melhores que a *baseline* de escolhas ao acaso (39%, em média). Nas melhores configurações, o sistema atinge uma acurácia de 53% a 55% na desambiguação. A acurácia da desambiguação humana é de 78%. Segundo o autor, esses resultados são relevantes, considerando-se que nenhum conhecimento lingüístico adicional ao disponível na WordNet é necessário.

Um dos problemas desse trabalho, apontado por Resnik (1995a), é que são testadas todas as combinações possíveis entre todos os sentidos do substantivo ambíguo e de todos os demais substantivos no seu contexto. Assim, se forem considerados contextos maiores, o custo computacional dessa abordagem pode torná-la inviável. Além disso, com o filtro aplicado para os testes, o autor considera apenas casos de ambigüidade de resolução mais simples.

Resnik (1995a) desenvolve um trabalho para identificar o sentido de agrupamentos de substantivos, de acordo com os sentidos providos pela WordNet. O foco é na desambiguação de grupos de palavras já estabelecidos e, portanto, com alguma relação implícita, e não na desambiguação de palavras dado seu contexto em textos livres. Essa característica diferencia o trabalho de Resnik da maioria dos outros em DLS. O autor assume que os grupos são previamente gerados por um processo independente do seu sistema, por exemplo, a partir das classes de um *thesaurus* ou de um algoritmo de *clustering*.

A atribuição do sentido é realizada por meio de uma medida de similaridade entre os substantivos do grupo, com base na hierarquia da WordNet, como no trabalho de Sussna. A suposição de ambos os autores é a de que o sentido de um grupo é o que maximiza a relação entre os sentidos de todos os elementos do grupo. Contudo, diferentemente de Sussna, Resnik não define essa relação em termos do tamanho do caminho entre os sentidos, mas sim em função de informações de conteúdo. A medida, definida em Resnik (1995b), procura identificar qual o conceito (ou seja, o sentido) mais específico na hierarquia, considerando apenas a relação de hiponímia, que subsume todos os conceitos das palavras do grupo. A hipótese dessa medida é de que quanto mais específico o conceito que subsume duas ou mais palavras, mais semanticamente relacionadas são essas palavras. Esse sentido mais específico é então escolhido como o sentido para o grupo. Para tanto, simplesmente são listadas todas as alternativas de sentido apresentadas na WordNet para cada substantivo do grupo e é escolhido o sentido compartilhado por todos os substantivos que ocorrem em um nível mais baixo na hierarquia.

O uso dessa medida representa outro diferencial no trabalho de Resnik com relação a outros trabalhos que medem similaridade entre sentidos: dependendo do grupo, podem ser atribuídos sentidos mais refinados ou mais genéricos, de categorias superiores na hierarquia da WordNet.

Para avaliar seu trabalho, o autor realiza um teste considerando 125 grupos provenientes de categorias do *thesaurus Roget* e compara os resultados com a identificação de sentidos (da WordNet) realizada manualmente por dois juízes, considerada como o limite superior da desambiguação. Para evitar o problema de discordância entre juízes humanos, somente os sentidos atribuídos com um alto grau de confiança por esses juízes foram avaliados. Dos sentidos desambiguados com uma alta confiabilidade pelo primeiro e segundo juizes, 65.7% e 68.6%, respectivamente, estavam corretos. Para os mesmos grupos de palavras, o sistema acertou em 58.6% e 60.5% das vezes.

Segundo o autor, os resultados são significativos, considerando-se que o sistema permite a desambiguação entre os sentidos refinados da WordNet, mas também entre sentidos menos refinados, quando a desambiguação refinada não é possível. Contudo, como discutido em Kilgarriff (1992; 1997), não há garantias de que o nível de distinção adequado para diferentes casos pode ser obtido automaticamente a partir da hierarquia da WordNet.

O trabalho de Agirre & Rigau (1996) usa as classes semânticas e a hierarquia de classes da WordNet para a DLS, com uma medida já definida e denominada por eles de "densidade conceitual". Essa medida calcula a distância entre conceitos, ou seja, a proximidade do significado entre eles. Em termos de uma hierarquia semântica, a distância conceitual pode ser definida como o comprimento do menor caminho que conecta dois conceitos nessa hierarquia. Essa medida é similar à "distância semântica" utilizada por Sussna e a "similaridade semântica" empregada por Resnik.

A medida é aplicada à desambiguação apenas de substantivos. Para tanto, são analisados todos os substantivos em contextos parametrizáveis de ocorrência do substantivo a desambiguar, os possíveis sentidos desse substantivo na WordNet e a sua distribuição, juntamente com a distribuição dos sentidos dos substantivos no contexto, na hierarquia da WordNet. Mais especificamente, verifica-se, na sub-hierarquia da WordNet para cada um dos possíveis sentidos do substantivo a ser desambiguado, a quantidade de sentidos dos demais substantivos do contexto. A sub-hierarquia que contiver mais sentidos, relativamente ao total de sentidos da hierarquia, leva a uma densidade mais alta na medida de densidade conceitual. O sentido do substantivo ambíguo nessa hierarquia é, então, escolhido para desambiguar tal substantivo. Assim, o sentido escolhido como mais adequado para o substantivo ambíguo é o que maximiza a densidade conceitual entre esse sentido e os dos substantivos vizinhos. Em alguns casos, o procedimento falha e retorna vários ou nenhum sentido.

Os autores avaliam seu trabalho em um conjunto de quatro textos aleatoriamente selecionados, de domínios distintos, do SEMCOR (Seção 2.7). Um dos objetivos da avaliação é verificar as configurações mais adequadas dos parâmetros na medida de densidade conceitual, incluindo o tamanho da janela de contexto, o nível de refinamento dos sentidos (*synsets* ou classes semânticas da WordNet), etc. Uma constatação importante sobre o tamanho da janela de contexto é que o tamanho mais adequado das janelas varia de acordo com o texto. Isso pode ser devido à não delimitação dos textos em unidades como tópicos ou parágrafos. Assim, em textos com parágrafos pequenos, como pequenos diálogos, o trabalho apresenta uma precisão maior com janelas pequenas, enquanto que em textos com unidades maiores, apresenta maior precisão com janelas maiores.

Os resultados relatados pelos autores, proporcionalmente ao número de substantivos ambíguos avaliados em cada texto, considerando-se o melhor tamanho de janela em cada texto, indicam uma precisão média de 43%, para distinção entre sentidos, e de 53.9% para a distinção entre as classes semânticas da WordNet. Esses resultados, apesar de mais baixos do que os de outros trabalhos, são significativos, segundo os autores, pois consideram a desambiguação de todos os substantivos, enquanto que outros trabalhos visam à desambiguação de apenas um subconjunto de palavras.

A medida de densidade conceitual também é utilizada por Mihalcea & Moldovan (1999), contudo, os autores utilizam um filtro para selecionar somente os sentidos com probabilidade maior de co-ocorrência (mais freqüentes), antes de aplicar essa medida. As informações de freqüência de co-ocorrências são obtidas a partir de buscas em textos da *web* e a densidade conceitual é medida com base na hierarquia da WordNet.

O objetivo dos autores é a desambiguação de substantivos, verbos, advérbios e adjetivos em textos irrestritos, usando os sentidos da WordNet. Para tanto, são extraídos contextos de co-ocorrência de duas palavras para cada palavra ambígua (bigrama), incluindo essa palavra, de modo que o contexto de cada palavra a ser desambiguada é constituído por apenas uma palavra. Para desambiguar a palavra ambígua em um bigrama, são realizadas buscas na *web* (usando o buscador AltaVista®), nas quais cada sentença de busca consiste de um dos possíveis sentidos da palavra ambígua e da outra palavra do bigrama, que se mantém fixa. Os sentidos são então ordenados de acordo com o número de documentos resultantes nas respectivas buscas. Os sentidos com mais documentos são considerados mais freqüentes. Isso é feito com todas as palavras da sentença a desambiguar.

Considerando apenas os sentidos mais freqüentes para cada palavra, de acordo com um limite inferior pré-definido, a próxima etapa do processo de DLS é refinar a ordenação dos sentidos utilizando a medida de densidade conceitual. Essa medida, neste caso, considera o número de palavras comuns que estão a uma distância semântica, dada pela hierarquia da WordNet, de duas ou mais palavras. Conforme mencionado, quanto maior o número de palavras em comum, maior o relacionamento semântico entre as palavras e, portanto, maior a medida de densidade conceitual. Vale notar que essa segunda etapa é realizada apenas para os verbos e substantivos, pois adjetivos e advérbios não estão incluídos na hierarquia da WordNet. Para as palavras dessas duas classes gramaticais, somente as informações de freqüência de co-ocorrência são utilizadas.

Diferentemente da maioria dos trabalhos de DLS, em vez de um único sentido, o sistema retorna os vários possíveis sentidos classificados de acordo com sua densidade conceitual (e freqüência). Segundo os autores, essa característica permite a escolha por outras opções de sentido, quando a primeira não for aplicável.

Em uma avaliação realizada considerando 384 pares de palavras do SEMCOR e a escolha do primeiro sentido da classificação produzida, os autores reportam uma acurácia média de 80%.

Montoyo et al. (2002) apresentam uma interface para a desambiguação de substantivos e verbos desenvolvida para ser acoplada a sistemas multilingües de Recuperação de Informações. Nessa interface, são considerados o espanhol e o inglês como língua-fonte para a definição das sentenças de busca e a geração de sentenças equivalentes em catalão e basco.

Os autores utilizam a taxonomia da EuroWordNet (Vossen, 1998), uma versão da WordNet que inclui outras línguas além do inglês, para realizar o mapeamento entre as palavras dessas duas línguas para o catalão e o basco. Além das definições das palavras, dos exemplos e das relações entre as palavras já existentes na WordNet, a EuroWordNet possui também ligações entre as palavras das diversas línguas, por meio de um índice interlingual[11]. Assim, a identificação do sentido de uma palavra em uma língua provê, automaticamente, o conjunto de sentidos equivalentes em outras línguas.

O módulo de DLS é incluído em uma arquitetura de um sistema multilingüe que realiza a análise sintática dos textos (em inglês ou espanhol), retornando a estrutura sintática completa de cada sentença para o módulo de DLS. Esse módulo consulta a EuroWordNet para buscar todos os possíveis sentidos da palavra a ser desambiguada e seleciona o método específico para a DLS, dependendo da classe gramatical das palavras da estrutura (verbo ou substantivo). Para a desambiguação de substantivos é empregado um método de marcas de especificação e para a desambiguação de verbos, um método de

[11] Esse recurso não inclui a língua portuguesa.

similaridade semântica.

O método de marcas de especificação se baseia na EuroWordNet para identificar quantas palavras do contexto da palavra ambígua na sentença estão relacionadas a cada marca de especificação, que corresponde a uma classe semântica da WordNet. O sentido da palavra na sub-hierarquia que contiver o maior número de palavras com a marca de especificação correspondente é escolhido para desambiguar um substantivo.

O método de similaridade semântica se baseia na relação verbo-objeto (com o verbo ambíguo) retornada pelo *parser*. Todos os possíveis sentidos do verbo e do substantivo que representa o núcleo do seu complemento são extraídos da EuroWordNet. Para cada sentido do verbo, são extraídos os substantivos das definições do verbo na sub-hierarquia correspondente da EuroWordNet. Em seguida, é aplicada uma função para determinar a similaridade entre cada sentido desses substantivos e os sentidos do substantivo na relação verbo-objeto na sentença. A combinação de sentidos mais similar, dada pelo sentido do verbo que possui, nas definições da sua sub-hierarquia, mais substantivos similares ao substantivo da sentença indica o sentido do verbo.

Os autores mencionam que, na tarefa de desambiguação parcial monolingüe do SENSEVAL-2, seu sistema obtém resultados superiores aos de sistemas similares.

Como pode ser observado, a desambiguação realizada consiste em identificar qual é o sentido correspondente à palavra a ser desambiguada, ainda na língua-fonte e, em seguida, encontrar a(s) palavra(s) na língua-alvo com o mesmo código da EuroWordNet. Assim, embora seja voltada para aplicações multilingües, a desambiguação é feita de maneira monolingüe. Além disso, para aplicações multilingües, esse trabalho só é possível para línguas previstas no projeto EuroWordNet, para as quais já existem códigos correspondentes aos itens lexicais.

Dicionários Eletrônicos

Diversos trabalhos utilizam informações de diferentes dicionários eletrônicos diretamente para a DLS. Os tipos de informação utilizados também variam, de acordo com o dicionário empregado. Nenhum dos trabalhos é voltado para a TA.

O trabalho de Lesk (1986), voltado para a Recuperação de Informações, é a primeiro a utilizar esse recurso. Ele baseia-se na premissa de que cada sentido de uma palavra ambígua é distinguível dos demais a partir dos sentidos de um contexto limitado a uma pequena quantidade de palavras e que um dicionário eletrônico consistente deve permitir identificar esse sentido por meio das palavras disponíveis na definição de cada sentido. Dessa maneira, quaisquer palavras ambíguas em textos irrestritos poderiam ser desambiguados com a utilização de um dicionário eletrônico.

Para verificar sua hipótese, o autor associa uma “assinatura” a cada sentido de uma palavra ambígua e a cada sentido das suas palavras vizinhas em uma sentença. Essa assinatura consiste da lista de palavras de conteúdo que aparecem na definição daquele sentido em um dicionário, sem considerar os exemplos apresentados nesse dicionário. A desambiguação é realizada pela seleção do sentido da palavra ambígua cuja assinatura apresenta o maior número de sobreposições com alguma assinatura das palavras vizinhas na sentença.

O autor realiza alguns experimentos para analisar possíveis variantes no seu trabalho, como o uso de dicionários com definições maiores ou menores, a atribuição de pesos para diferentes tipos de sobreposições e diferentes tamanhos de janelas de contexto. Com relação ao tamanho das definições nos dicionários, Lesk afirma que uma pequena definição já é suficiente, e que não é necessário considerar os exemplos de uso dos

sentidos, pois eles podem acrescentar ruídos à assinatura. O autor também não considera viável a atribuição de pesos às sobreposições. Quanto aos diferentes tamanhos de janela, ele afirma que há pouca diferença nos resultados ao considerar janelas de 4, 6, 8 ou 10 palavras. Como padrão, seu sistema utiliza 10 palavras.

Lesk não realiza uma avaliação sistemática do seu trabalho, mas apenas alguns testes com pequenos exemplos. Nesses testes, ele obtém uma acurácia de 50% a 70%. Segundo ele, esses resultados são significativos, considerando-se que são empregadas distinções de sentido relativamente refinado (dadas pelo dicionário), que não há dependência de contexto global e que apenas as informações disponíveis no dicionário são utilizadas. Ele sugere que essa abordagem seja utilizada como complemento a algum mecanismo que utilize outras informações, como as relações sintáticas entre as palavras.

Um problema com esse trabalho é que ele é dependente das definições de um dicionário específico: a presença ou ausência de uma palavra na definição daquele dicionário pode mudar completamente os resultados. Além disso, a simples contagem das palavras nas assinaturas pode privilegiar a escolha de sentidos com definições mais extensas. Contudo, tal trabalho é de grande relevância para a área de DLS, pois serviu de base para vários outros trabalhos estatísticos. De fato, as técnicas usadas pelo autor são refinadas em vários outros trabalhos, que acrescentam diferentes campos de informação dos dicionários ou de outras fontes para aprimorar os resultados da DLS.

Assim como Lesk, Wilks et al. (1990) analisam a sobreposição de definições em um dicionário eletrônico. O dicionário utilizado é o LDOCE (*Longman Dictionary of Contemporary English*). Os autores afirmam que a análise somente das definições diretas no dicionário é problemática quando as definições são muito curtas, muitas vezes insuficientes para que ocorra alguma sobreposição.

Para minimizar esse problema, Wilks et al. consideram um contexto maior para cada definição no dicionário, procurando estender o conhecimento associado aos sentidos de cada palavra. Para tanto, são computadas informações sobre a vizinhança lexical de todas as palavras do dicionário inteiro por meio da freqüência de co-ocorrência das palavras.

A partir dessa freqüência, são aplicadas técnicas de *clustering* para particionar as palavras de acordo com os sentidos a que elas correspondem e essas partições são usadas para classificar os sentidos. Assim, são derivadas métricas sobre o grau de relacionamento entre as palavras, as quais são usadas em um vetor de relacionamentos entre cada palavra e seu contexto. A classificação é iterativa: primeiramente são comparados os vizinhos mais locais, depois, os vizinhos dos vizinhos, e assim sucessivamente.

Em experimentos com a palavra ambígua *bank*, os autores reportam uma acurácia máxima de 53% na identificação dos 13 possíveis sentidos do LDOCE. Em testes para uma classificação considerando sentidos menos refinados (homógrafos) para essa mesma palavra, os autores relatam entre 85% e 90% de acurácia.

Segundo Véronis & Ide (1990), tanto o trabalho de Wilks et al. quanto o de Lesk perdem informações sobre como as palavras das definições se inter-relacionam. Em seu trabalho, os autores também utilizam informações de um dicionário, mas procuram manter as relações semânticas entre as palavras. Para tanto, eles definem um mecanismo para automaticamente converter as definições do dicionário do CED (*Collins English Dictionary*) e os relacionamentos entre suas palavras em uma rede neural.

Nessa rede neural, há nós para cada entrada do dicionário e para cada um dos seus possíveis sentidos. Os nós das entradas são ligados aos nós dos seus sentidos por relações de ativação, os quais são ligados às palavras em suas definições, que também estão ligadas

aos seus sentidos, e assim sucessivamente, criando uma rede complexa. O contexto é dado, portanto, pela conectividade na rede.

Durante a desambiguação, os nós correspondendo às palavras da sentença a ser desambiguada são ativados. Cada nó ativa, então, os nós de seus sentidos, os quais ativam os nós das palavras com as quais estão conectados, e assim sucessivamente, por vários ciclos. A cada ciclo, os nós de palavras e sentidos recebem um retorno de ativação dos nós conectados. Os nós que competem entre si enviam inibições uns para os outros. Depois de vários ciclos, a rede estabiliza em um estado no qual um sentido para cada palavra da sentença de entrada estiver mais ativo que os outros. Esses sentidos são, então, escolhidos para as palavras ambíguas.

O trabalho de Véronis & Ide é similar aos trabalhos conexionistas baseados em conhecimento manualmente codificado, no entanto, diferentemente desses trabalhos, a rede é criada automaticamente, sem a necessidade de codificação manual, a partir do dicionário. Com isso, esse trabalho pode ser aplicado em larga escala, para a desambiguação de todas as palavras de uma sentença.

Na avaliação de seu trabalho, os autores mencionam apenas que ele encontra o sentido correto de todos os exemplos citados por Lesk. Eles concluem que seu trabalho é mais robusto que outros similares, pois permite o uso de um contexto maior, de modo a tornar o sistema menos dependente das palavras da sentença e das suas definições no dicionário, bem como menos sensível a ruídos. Além disso, ele não exige a representação de outras formas de conhecimento, como os micro-traços de Waltz & Pollack (1985).

Diversos outros autores procuram refinar o trabalho de Lesk incorporando campos de informação adicionais fornecidas pelo dicionário eletrônico para os diferentes sentidos, além da definição das palavras. Grande parte dos trabalhos emprega o dicionário LDOCE, utilizando principalmente suas informações de freqüência dos sentidos, os códigos de área (economia, engenharia, etc.), os traços semânticos (abstrato, humano, etc.) de substantivos e as restrições de seleção sobre argumentos de verbos.

Guthrie et al. (1991), por exemplo, descrevem um trabalho voltado para a Recuperação de Informações que se baseia nos códigos de área do LDOCE. Esse trabalho é similar ao de Lesk, mas em vez de considerar a sobreposição entre as definições de todos os possíveis sentidos da palavra ambígua e das suas vizinhas, considera apenas a sobreposição das definições em que o código de área também coincide com o código das palavras vizinhas. Com isso, procuram estabelecer relações de co-ocorrência dependentes de área entre as palavras.

Para identificar essas relações, a partir de uma palavra ambígua, são coletadas algumas palavras (por exemplo, 10) que aparecem nas definições de todos os sentidos de cada uma das suas possíveis áreas. Nem todos os sentidos possuem códigos de área, uma vez que alguns são considerados genéricos, independentes de área. As palavras que aparecem nas definições desses sentidos são então agrupadas em uma área “geral”. Apesar da hierarquia provida pelo LDOCE disponibilizar códigos de área em dois níveis, somente o nível principal (cerca de 100 códigos) é utilizado.

A verificação das correspondências de sentido entre os códigos de área é realizada em duas etapas. Na primeira etapa, é verificada a intersecção das palavras coletadas para cada área da palavra ambígua com as suas palavras vizinhas na sentença. A área que apresenta a maior intersecção (mais palavras coincidentes), respeitando-se um limite inferior pré-estabelecido, é escolhida como a área da palavra ambígua.

Segundo os autores, os possíveis sentidos de cada área (exceto os da área geral) apresentam uma relação de significado implícita, que os diferencia dos demais sentidos em outras áreas, mas que não é suficiente para a desambiguação. Assim, em uma segunda

etapa, eles procuram identificar qual o sentido na área selecionada. Para tanto, é verificada a intersecção das definições dos possíveis sentidos da área selecionada com as palavras da sentença. O sentido com a maior intersecção é escolhido como o sentido da palavra ambígua.

Na primeira etapa, se o limite inferior não for atingido, um processo é disparado para aumentar a vizinhança a ser analisada, considerando também as definições de sentidos das palavras coletadas para cada área. Esse processo é iterativo e continua a considerar mais definições de sentidos relacionadas, até que o número mínimo de palavras coincidentes na definição da área (e das áreas relacionadas) e na sentença seja alcançado.

Os autores não avaliam seu trabalho, apenas mencionam o problema da quantidade limitada de informações disponíveis no LDOCE para cada código de área e sugerem, como trabalho futuro, que sejam utilizados, em vez das definições do dicionário, córpus de diferentes áreas para a identificação de palavras significativas em cada área.

Um trabalho similar ao de Guthrie et al. é o de Cowie et al. (1992). Nele, os autores adicionam a técnica para otimizar as verificações entre as possíveis combinações de definições. Segundo os autores, na maioria dos demais trabalhos é analisada a desambiguação de uma única palavra, em diferentes contextos. Assim, não é possível visualizar se (e como) a desambiguação de uma palavra pode influenciar na desambiguação das demais na sentença ou ser influenciada por essa desambiguação (por exemplo, se a escolha de um sentido pode ser desfeita). Os autores sugerem o uso da técnica de otimização computacional *simulated annealing* nesses trabalhos, uma vez que ela torna possível considerar toda a sentença como contexto e procurar a melhor combinação entre todos os sentidos de todas as palavras, simultaneamente.

Essa técnica é voltada para a resolução de problemas de minimização combinatorial em larga escala. Para verificar sua utilidade na DLS, os autores adaptam seu trabalho anterior (Guthrie et al, 1991), incorporando essa técnica a ele.

O trabalho é avaliado em um experimento considerando a desambiguação 50 sentenças extraídas do LDOCE, com uma média de 5.5 palavras ambíguas por sentença. O resultado do sistema foi comparado com a etiquetação manual das 50 sentenças. Os autores reportam uma acurácia de 72% na identificação de homógrafos do LDOCE e de 47% para distinções entre sentidos mais refinados. Apesar de baixos, vale notar que esses valores correspondem à desambiguação de todas as palavras de conteúdo em uma sentença.

Um problema que se mantém nesse trabalho é que definições mais longas no dicionário tendem a ser mais privilegiadas que definições curtas, em função do número maior de combinações possíveis nas definições longas.

Wilks & Stevenson (1996) defendem a utilização da categoria gramatical das palavras ambíguas como fonte de informação essencial para a desambiguação. Em seu trabalho, os autores consideram também a desambiguação entre palavras de categorias gramaticais distintas. Em um experimento realizado com 1.700 palavras, eles procuram analisar se somente essa característica (isto é, a categoria gramatical), associada à informação de freqüência dos sentidos do LDOCE, é suficiente para a desambiguação em nível de homografia.

Um etiquetador morfossintático com uma precisão de 95% é utilizado para identificar as categorias de todo o texto. Em seguida, é realizado, manualmente, o mapeamento das etiquetas identificadas para as etiquetas gramaticais mais simples do LDOCE. Nesse mapeamento, para cada palavra ambígua, se existir mais de um homógrafo no LDOCE com a mesma categoria gramatical, o primeiro deles é escolhido, com base no fato de que o primeiro homógrafo listado é sempre o mais freqüente.

Considerando-se apenas as palavras ambíguas, os autores relatam uma acurácia de 87.4% para o seu trabalho na atribuição do homógrafo correto. Com isso, concluem que a grande maioria das distinções de sentido pouco refinadas pode ser resolvida identificando-se apenas a sua categoria gramatical.

Wilks & Stevenson (1997a; 1997b) discutem vários trabalhos de DLS que podem ser empregados para a tarefa de etiquetação de sentido e propõem uma abordagem especificamente voltada para essa tarefa. Essa abordagem faz uso de diferentes tipos de informação, disponíveis no dicionário LDOCE, bem como de procedimentos auxiliares. A suposição dos autores é a de que a combinação de vários métodos "fracos", que apresentam baixo desempenho isoladamente, pode levar a um método "forte".

Os tipos de informação citados pelos autores como úteis para a etiquetação de sentidos são: etiquetas gramaticais, códigos de área (ou categorias de um *thesaurus*), *collocations*, restrições ou preferências de seleção e definições de dicionário. Com exceção das etiquetas gramaticais e das *collocations*, os demais tipos de informação podem ser extraídos do LDOCE.

Na primeira implementação da sua proposta (Wilks & Stevenson, 1997a), os autores utilizam apenas um etiquetador morfossintático como filtro, para eliminar os sentidos que pertencem a categorias distintas da indicada para a palavra ambígua, e a busca das definições no dicionário para verificar as sobreposições entre as definições dos sentidos das palavras vizinhas da palavra ambígua na sentença e as definições de cada sentido da palavra ambígua nesse dicionário, como no trabalho de Lesk. Além disso, utilizam o algoritmo de *simulated annealing* para otimizar o processo de escolha entre as muitas combinações possíveis.

Em um experimento com 10 sentenças com palavras ambíguas, os autores obtêm 86% de acurácia em nível de homografia e 57% em nível de polissemia. Apesar de pouco significativo, em função do tamanho reduzido do conjunto de teste, esses resultados já se mostram melhores que os obtidos por Cowie et al. utilizando o mesmo conjunto.

Em uma segunda implementação da sua abordagem (Wilks & Stevenson, 1997b), os autores incluem um identificador de nomes de entidades para isolar nomes próprios, que não precisam ser desambiguados, e informações do LDOCE para verificar a sobreposição entre os códigos de área dos sentidos das palavras vizinhas à palavra a ser desambiguada e dos diversos sentidos da palavra no dicionário, como no trabalho de Guthrie et al. Um mecanismo simples é empregado para combinar os resultados dos diversos processos, o qual escolhe pelo sentido indicado pelo maior número desses processos.

A avaliação dessa versão do trabalho, em um córpus com 14 palavras ambíguas, indica um desempenho superior ao da versão anterior: 88% de acurácia em nível de homografia e 60% em nível de polissemia.

Por fazer uso de diversas fontes de conhecimento, essa abordagem é de especial interesse para o modelo de DLS a ser proposto. Ela vem sendo aperfeiçoado em trabalhos posteriores dos autores (Wilks & Stevenson, 1998; Stevenson & Wilks, 1999; 2000; 2001), nos quais eles incorporam, além do uso de outras características do LDOCE, informações provenientes de córpus, por meio de algoritmos de aprendizado de máquina. Essas versões mais recentes são consideradas híbridas e serão descritas na Seção 3.3.

Brun (2000) apresenta uma proposta na qual um dicionário eletrônico é utilizado de maneira diferenciada dos demais trabalhos. O modelo de DLS proposto é incorporado a um ambiente integrado para o processamento de textos, que dispõem de um dicionário bilíngüe bidirecional inglês/francês e de mecanismos de acesso a esse dicionário, de um *parser* superficial e de um extrator de regras de DLS.

O extrator de regras utiliza as definições monolingües do dicionário para automaticamente extrair regras de DLS para o inglês. Para tanto, para cada entrada do dicionário que possui mais de um sentido, os exemplos fornecidos para cada um dos sentidos são analisados pelo *parser*, que gera as relações de dependência funcional nesses exemplos (basicamente, sujeito-verbo, verbo-objeto e modificadores de vários tipos). Quando a relação de dependência envolve o lema da palavra de entrada do dicionário, é produzida uma regra lexical que indica o sentido daquela entrada, dado o seu uso naquela relação sintática, com a(s) palavra(s) em questão.

As regras lexicais produzidas para os sentidos de cada palavra, considerando cada palavra que co-ocorre com ela na relação sintática, são então generalizadas para classes de palavras co-ocorrentes. Para tanto, para cada regra construída, a hierarquia conceitual da WordNet é consultada e a palavra que co-ocorre com a palavra ambígua na regra é substituída por todas as suas classes semânticas, ou seja, pelo código de todas as classes da WordNet que contêm essa palavra. Essa generalização aumenta a abrangência do sistema, que pode cobrir casos de ambigüidade com outras palavras vizinhas, além daquelas apresentadas nos exemplos do dicionário.

Todas as regras são armazenadas em uma base, que é consultada para resolver novos casos de ambigüidade. Nos casos em que nenhuma regra se aplica, é automaticamente atribuído o primeiro sentido do dicionário, considerado o mais freqüente, à palavra ambígua. No processo de aplicação das regras, regras para diferentes partes de uma sentença podem cooperar entre si. Por outro lado, também pode haver conflito entre as regras, quando várias regras podem ser aplicadas para a desambiguação da mesma palavra. Neste caso, a autora sugere uma medida que analisa o contexto das dependências funcionais.

O trabalho é avaliado considerando-se duas configurações: (a) as 34 palavras e o córpus de 8.500 sentenças usados no SENSEVAL-1; e (b) todas as palavras de 400 sentenças de um jornal. Em ambos os casos, a precisão média fica em torno de 79% e a cobertura, em torno de 36%.

Pelos valores da cobertura, pode-se perceber que a generalização usada não é suficiente para garantir regras abrangentes. Contudo, deve-se considerar que o trabalho é simples, pois não requer processamento lingüístico profundo, e permite a desambiguação de todas as palavras, de acordo com um conjunto de sentidos refinado. Apesar de a autora não mencionar o uso dessa proposta para a TA, isso seria possível, pois cada sentido de uma palavra do inglês no dicionário usado, além da definição e dos exemplos em inglês, possui a sua tradução para o francês.

Thesauri

São poucos os trabalhos que utilizam apenas um *thesaurus* como fonte de informações, geralmente, trabalhos mais antigos. Trabalhos mais recentes normalmente utilizam *thesauri* juntamente com informações de outras fontes, como é o caso de Yarowsky (1992), que emprega também informações extraídas de um córpus (Seção 3.3).

O primeiro trabalho a usar informações fornecidas por um *thesaurus* é o de Masterman (1957), que utiliza, também, um dicionário bilíngüe. Neste trabalho, é utilizado o *thesaurus Roget* para a DLS em um sistema de TA do latim para o inglês. Primeiramente, a tradução da raiz de cada palavra em latim é recuperada de um dicionário latim-inglês. O sistema busca, então, para cada tradução, seu índice de classes correspondente no *Roget*. Dessa maneira, cada raiz de palavra do latim é associada a uma lista de índices do *Roget* relacionados a seus equivalentes em inglês. Os índices para todas as palavras na mesma

sentença são então examinados para verificar as sobreposições entre eles. As palavras em inglês correspondentes aos índices com maior sobreposição são escolhidas para a tradução.

Outro trabalho de DLS que utiliza o *thesaurus Roget* é o sistema de Patrick (1985). Patrick usa esse *thesaurus* para discriminar entre os sentidos de verbos, examinando agrupamentos semânticos derivados desse *thesaurus*, em função das relações de sinonímia mais fortes. As distinções de sentido são bastante refinadas, uma vez que são baseadas apenas nas palavras mais fortemente semanticamente relacionadas no *thesaurus*. Patrick afirma que seu sistema é capaz de desambiguar entre vários sentidos de verbos como *inspire* (causar inspiração, inalar, respirar, etc.) e *question* (duvidar, fazer uma pergunta) com um alto nível de confiabilidade. Contudo, seu trabalho limitou-se a um pequeno conjunto de palavras e não foi continuado ou estendido.

Vale notar que os trabalhos que utilizam a hierarquia conceitual da WordNet, descritos anteriormente, classificados como baseadas em léxicos computacionais, poderiam ser também considerados baseados em *thesaurus*, já que a hierarquia da WordNet, apesar de mais complexa, inclui as informações de um *thesaurus*.

Os trabalhos baseados em conhecimento pré-codificado descritos acima são resumidamente elencados na Tabela 3, de acordo com a abordagem empregada, o(s) recurso(s) lingüístico(s) usado(s), o conjunto de palavras para as quais forma criados e/ou testados, o nível de refinamento das distinções entre os sentidos e a acurácia apresentada. Novamente, em alguns casos, determinadas informações não são explicitadas nas publicações referentes a tais trabalhos (representadas por "?"). Além disso, a medida de desempenho disponibilizada pode não ser na forma de acurácia, mas sim de precisão e cobertura, conforme indicado na tabela.

Tabela 3. Lista dos trabalhos de DLS baseados em conhecimento pré-codificado

Trabalho	**Abordagem**	**Recurso**	**Conj. palavras**	**Nível de refinamento dos sentidos**	**Acurácia**
(Voorhees, 1993)	simbólica	WordNet	?	médio	?
(Sussna, 1993)	simbólica	WordNet	substantivos	alto	53 a 55%
(Resnik, 1995a)	simbólica	WordNet	125 grupos de substantivos	alto e baixo	58.6 a 60.5%
(Agirre & Rigau, 1996)	simbólica	WordNet	substantivos de 4 textos do SEMCOR	alto e baixo	43%: distinções refinadas, 53.9%: distinções pouco refinadas (precisão)
(Mihalcea & Moldovan, 1999)	simbólica	WordNet	alguns substantivos, verbos, advérbios e adjetivos	alto	80%
(Montoyo et al., 2002)	simbólica	EuroWord Net	verbos e substantivos	alto	?
(Lesk, 1986)	simbólica	MRD	?	alto	50 a 70%
(Wilks et al., 1990)	simbólica	MRD LDOCE	1 palavra	alto e baixo	53%: distinções refinadas, 85 a 90%: distinções pouco refinadas
(Véronis & Ide, 1990)	conexionista	MRD CED	?	?	?
(Guthrie et al., 1991)	simbólica	MRD LDOCE	?	?	?
(Cowie et al.,	simbólica	MRD	palavras de 50	alto e baixo	47%: distinções

1992)		LDOCE	sentenças do LDOCE		refinadas, 72%: distinções pouco refinadas
(Wilks & Stevenson, 1996)	simbólica	MRD LDOCE	1700 palavras	baixo	87.4%
(Wilks & Stevenson, 1997a; 1997b)	simbólica	MRD LDOCE	palavras de 10 sentenças	alto e baixo	57 a 60%: distinções refinadas, 86 a 88%: distinções pouco refinadas
(Brun, 2000)	simbólica	MRD e WordNet	palavras do Senseval-1 e de 400 sentenças	alto	79% (precisão), 36% (cobertura)
(Masterman, 1957)	simbólica	*thesaurus* Roget e MRD bilíngue	?	baixo	?
Patrick (1985)	simbólica	*thesaurus* Roget	alguns verbos	alto	?

3.2 Método baseado em córpus

Os trabalhos baseados em córpus descritos nesta seção contemplam exemplos de abordagens supervisionadas e não-supervisionadas de DLS, bem como de abordagens desenvolvidas sob os diferentes paradigmas de AM. Alguns dos trabalhos são voltados para a TA.

Além dos trabalhos específicos para a TA que são apresentados, uma observação importante é que existem, atualmente, algumas propostas de sistemas de TA completos baseados em córpus, que não são discutidas aqui. Esses sistemas, normalmente supervisionados e estatísticos, geram regras de tradução completas a partir de exemplos de treinamento, normalmente na forma de regras de transferência direta da língua-fonte para a língua-alvo, com informações de diferentes níveis (lexical e sintático, em geral).

Nesses sistemas, a DLS é realizada implicitamente, como parte do modelo gerado, e se baseia em informações simples, como a freqüência das ocorrências de cada tradução, considerando um determinado contexto. Exemplos de sistemas dessa natureza são o módulo Stattrans do sistema VERMOBIL (Vogel et al., 2000), de tradução entre o alemão e o inglês, o sistema PALGLOSS-LITE (Frederking & Brown, 1996), de tradução entre o inglês e o espanhol e entre o inglês e o sérvio-croata, e o sistema apresentado por Brown et al. (1990), de tradução do francês para o inglês.

3.2.1 DLS não-supervisionada

Nos diversos trabalhos não-supervisionados descritos a seguir são empregados diferentes algoritmos, em especial, algoritmos baseados em diferentes técnicas de *clustering*, para realizar a discriminação de sentidos. Apesar de existirem muitos trabalhos não-supervisionados baseados em *clustering*, somente alguns são discutidos, pois, conforme mencionado (Seção 2.6.2.1), esse tipo de abordagem dificilmente poderá ser aplicado para a TA. Também são apresentados trabalhos que empregam outros algoritmos, em geral, associados ao uso de córpus paralelos, para a desambiguação na TA.

Brown et al. (1991) usam um modelo estatístico da Teoria da Informação baseado em informação mútua para a seleção lexical de itens ambíguos na TA do francês para o

inglês. Para tanto, são extraídas de um córpus paralelo entre as duas línguas as possíveis traduções das palavras do francês para o inglês que apresentam um alinhamento direto (um para um). Em seguida, é definido um conjunto de possíveis características úteis para distinguir, com base no contexto local das sentenças na língua-fonte (francês) ou na língua-alvo (inglês, considerando a sentença parcialmente traduzida por um sistema de TA), qual é a tradução adequada para uma palavra na língua-alvo. As características incluem diferentes palavras do contexto da palavra ambígua na língua-fonte, por exemplo, o primeiro substantivo à direita, o primeiro verbo à direita, a primeira palavra à esquerda, etc.

O modelo estatístico empregado, denominado algoritmo Flip-Flop, tenta encontrar, para cada palavra ambígua, uma única característica (dentre as pré-definidas) que indica, com um alto nível de confiabilidade, qual a sua tradução. Nesse sentido, o modelo é bastante diferente do bayesiano, que considera todas as características. Esse algoritmo considera uma desambiguação binária, ou seja, a escolha entre apenas duas possíveis traduções de uma palavra ambígua. O processo é iterativo e, a cada interação, o algoritmo procura aumentar a informação mútua obtida com o emprego da característica para a desambiguação da palavra em questão. O critério de parada é indicado pela estabilização da informação mútua, ou seja, na iteração em que essa medida não pode mais ser aumentada.

Definida a característica que melhor divide o conjunto de treinamento, os exemplos (em francês) podem ser divididos em dois grupos, para as duas traduções possíveis, de acordo com o valor que apresentam para essa característica. Na verdade, cada um dos grupos pode ter várias traduções, mas elas são ranqueadas de acordo com uma estimativa da probabilidade de cada uma das traduções no córpus do inglês. A tradução com a maior probabilidade de ocorrência em cada grupo é então escolhida para etiquetar a palavra ambígua em todos os exemplos do francês selecionados.

Para a avaliação do módulo de DLS, o modelo foi treinado para a desambiguação das 500 palavras mais comuns do inglês e as 200 mais comuns do francês e o módulo resultante foi incorporado a um sistema de TA por transferência, também estatístico, na fase de análise. Na tradução de 100 sentenças aleatoriamente selecionadas com essas palavras, os autores relatam uma diminuição de 13% na taxa de erro das traduções resultantes do sistema com o uso módulo.

Nos trabalhos de Schütze (1992; 1998), voltados para a Recuperação de Informações, é utilizado um córpus não etiquetado, com uma grande quantidade de exemplos e apenas informações de contexto para a são consideradas discriminação de sentidos. Essa discriminação é realizada por um processo de *clustering* contextual.

O autor propõe um esquema de representação no qual as palavras, os sentidos e os contextos são representados por vetores de alta dimensionalidade em um espaço vetorial, sendo que a similaridade entre os vetores corresponde à proximidade semântica entre as palavras. Essa similaridade é baseada na co-ocorrência de segunda ordem entre as palavras do contexto: duas palavras (dois contextos) na vizinhança da palavra ambígua são atribuídas ao mesmo *cluster* se elas co-ocorrem com alguma palavra que, por sua vez, co-ocorre com palavras do córpus de treinamento. Segundo o autor, esse tipo de co-ocorrência é mais robusto e menos esparso que a co-ocorrência simples, de primeira ordem. A hipótese do autor é a de que duas ocorrências de uma palavra ambígua pertencem ao mesmo sentido se suas representações contextuais são similares.

O contexto de cada ocorrência de uma palavra ambígua no córpus de treinamento é representado como um vetor formado pelas co-ocorrências de segunda ordem, basicamente, esse vetor possui a soma das co-ocorrências das palavras vizinhas à palavra

ambígua em uma janela (por exemplo, de 1000 caracteres). Como a dimensionalidade dos vetores pode ser muito alta, uma técnica de redução de dimensionalidade é aplicada.

A partir dos vetores, o sistema procura identificar as suas relações de significado, agrupando-os por um processo de *clustering* hierárquico. Os vetores de contexto são agrupados de modo que ocorrências similares, de acordo com a co-ocorrência definida, sejam atribuídas ao mesmo *cluster*. É utilizado um algoritmo tradicional que identifica, automaticamente, o número de *clusters* mais apropriado para dividir o conjunto. Os *clusters* resultantes representam, implicitamente, os vários sentidos da palavra. Assim, um "sentido" é simplesmente um grupo de palavras com contextos similares. Na discriminação da ocorrência de uma palavra ambígua é computado o seu vetor de contexto de segunda ordem e ela é atribuída ao *cluster* cujo centróide é mais próximo dessa representação.

Para a avaliação do seu trabalho, o autor realiza experimentos com 10 palavras ambíguas e relata uma acurácia média de 92%. Contudo, é importante notar que ele considera apenas a desambiguação binária (entre dois sentidos) e entre sentidos totalmente distintos, inclusive, de categorias gramaticais distintas, em alguns casos.

O autor focaliza apenas a etapa de discriminação de sentidos, ou seja, de formação dos *clusters*, e não a etapa posterior de atribuição de algum sentido aos *clusters*. Para uma tarefa de desambiguação de sentidos, os *clusters* precisam ser manualmente rotulados com tais sentidos. Assim, esse trabalho elimina a necessidade de atribuir um sentido para cada ocorrência da palavra, entretanto, ainda há a tarefa manual de análise dos *clusters* para a determinação do sentido predominante em cada *cluster*. Além disso, cada sentido pode ser atribuído a diversos *clusters*. Segundo o autor, a atribuição manual requer, em cada *cluster*, a análise de 10 a 20 membros. Essa atribuição poderia ser feita de maneira automática utilizando o sistema de Resnik (1995a), descrito na Seção 3.1.2.

O autor afirma que seu trabalho, que considera apenas o contexto da palavra ambígua, é válido para a discriminação de sentidos, mas não com a desambiguação. Com isso, não é adequado para algumas aplicações, como a desambiguação na TA.

Segundo Ide & Véronis (1998), o problema mais grave desse trabalho é que não há garantia de que os sentidos derivados dos *clusters* correspondam a distinções reais de sentido e que, por isso, essas distinções dificilmente poderão ser aplicadas em outras tarefas, uma vez que elas são totalmente dependentes do córpus usado. Outro problema desse trabalho, segundo Yarowsky (1995), é que ele trata os exemplos como *bag-of-words* e, por isso, não analisa muitas características importantes presentes nos exemplos e que poderiam ser facilmente extraídas, como *collocations* e outras relações de distância entre as palavras.

Em um trabalho posterior, também voltado para a Recuperação de Informações, Schütze & Pedersen (1995) usam praticamente a mesma proposta de Schütze nos trabalhos anteriores citados. Eles comparam sua proposta de recuperação baseada em vetores de sentidos com as propostas tradicionais, baseadas em vetores de palavras. Como resultado, os autores relatam um aumento de 7.4% na precisão de um sistema de recuperação com o uso da recuperação baseada em sentidos, em vez de baseada em palavras, e um aumento de 14.4% na precisão considerando a combinação das duas formas de recuperação.

É importante ressaltar que esse trabalho foi o único, durante muito tempo, a reportar resultados positivos sobre o uso de mecanismos de DLS na recuperação de informações (Sanderson, 2000), em termos de um aumento significativo na precisão dessa área.

Mais recentemente, Stokoe et al. (2003), com um trabalho baseado em córpus supervisionado simples, usando estatísticas de co-ocorrência e *collocations* do córpus SEMCOR e as freqüências dos sentidos na WordNet, também relatam resultados positivos

para a recuperação de informações. Os autores reportam um aumento significativo na precisão da recuperação de informações com a aplicação do mecanismo de DLS (62.1%), comparada à precisão da recuperação usando a medida tradicional TF-IDF (*Term Frequence - Inverse Document Frequence*) (45.9%).

Pedersen & Bruce (1998) apresentam um trabalho não-supervisionado em que todo o conhecimento é extraído de um córpus. Os autores empregam técnicas para estimar os parâmetros de um modelo que descreve a distribuição condicional dos grupos de sentido, dadas as características contextuais de cada ocorrência da palavra ambígua. Essas características são descritas por vetores de características que armazenam as propriedades selecionadas do contexto (sentença) no qual cada palavra ambígua ocorre. O objetivo é dividir todos os exemplos do córpus, cada um representado por um vetor, em um número de grupos. Para tanto, são consideradas duas técnicas para estimar os parâmetros: o algoritmo *Expectation Maximization* e um método baseado em Cadeias de Markov. Os autores pretendem também avaliar qual das duas técnicas é mais adequada para a tarefa.

Com os grupos formados, para atribuir cada ocorrência de ambigüidade a um grupo de sentidos, o contexto da ocorrência da palavra ambígua é observado e o sentido mais provável é atribuído, de acordo com um modelo Naive Bayes cujos parâmetros são estimados pelas técnicas citadas.

Os autores realizam testes com diversas ocorrências de 13 palavras ambíguas, incluindo verbos, substantivos e adjetivos, considerando dois ou três sentidos de cada uma delas. Além de variar a técnica de estimação dos parâmetros, eles definem três tipos de conjuntos de características para os vetores de características, de modo a avaliar quais características são mais relevantes. As características incluem informações morfológicas, categoria gramatical, co-ocorrências e *collocations*. O objetivo dos testes é avaliar quão bem os grupos de sentido automaticamente definidos podem ser mapeados para grupos de sentidos definidos por humanos.

Com relação às técnicas de estimação de parâmetros empregadas, os autores concluem que não há diferenças significativas entre elas. Já as características consideradas mais determinante para a atribuição do sentido dos verbos são as *collocations*.

Os resultados de diversos testes com o trabalho apresentaram variações de acordo com a técnica de estimação de parâmetros e o conjunto de características empregado. A maior acurácia média para substantivos foi de 64%, para adjetivos, de 72% e para verbos, de 70%.

Em um trabalho anterior (Pedersen & Bruce, 1997), Pedersen & Bruce já haviam comparado três algoritmos estatísticos de aprendizado não-supervisionado baseados em *clustering* para a DLS, incluindo o *Expectation Maximization*, utilizando as mesmas 13 palavras e as mesmas variações no conjunto de características. Os resultados das comparações foram similares: 65% a 66% de acurácia, dependendo do algoritmo de aprendizado usado. Esses resultados são ruins, se considerada que, no corpus usado, a *baseline* do sentido mais freqüente permite desambiguar corretamente 73% dos exemplos.

Uma constatação importante nesses testes foi que todos os algoritmos analisados, por serem não-supervisionados, são mais aplicáveis a substantivos do que para verbos ou palavras de outras categorias gramaticais.

O trabalho de Dini et al. (1998) para a DLS utiliza relações funcionais extraídas a partir das estruturas sintáticas de um córpus não-supervisionado, gerado por um analisador sintático superficial. Com base nessas relações, um algoritmo de aprendizado, derivado de

um algoritmo usado na tarefa de etiquetação morfossintática, produz regras de desambiguação, considerando o grupo de 45 etiquetas das classes semânticas da WordNet.

O algoritmo se baseia, fundamentalmente, em restrições de seleção e em traços semânticos. Ao analisar as relações funcionais para uma sentença, se pelo menos uma palavra tiver uma etiqueta semântica atribuída de maneira não ambígua, o processo de desambiguação pode partir dessa palavra. As regras geradas são simbólicas, de modo que podem ser posteriormente editadas por humanos para melhorar o desempenho do sistema.

Os autores realizam um experimento de avaliação do desempenho do seu trabalho e reportam resultados superiores, quando comparados aos de outros trabalhos de DLS para a Recuperação de Informações que não utilizam técnicas de PLN (por exemplo, Resnik, 1997).

Também no contexto da Recuperação de Informações, Pantel & Lin (2002) propõem um algoritmo de *clustering* distribucional para agrupar as palavras de textos em grupos semanticamente similares, que correspondem aos sentidos das palavras. Para tanto, é analisado o contexto de ocorrência da palavra em um córpus. A hipótese das abordagens distribucionais é de que palavras que ocorrem nos mesmos contextos tendem a ser similares. Assim, o algoritmo proposto identifica os sentidos das palavras agrupando-as, de acordo com a sua similaridade distribucional, em *clusters* com identificações numéricas que representam os sentidos.

No seu algoritmo, cada palavra ambígua é representada por um vetor de características, sendo que cada característica corresponde a uma palavra com a qual a palavra ambígua co-ocorre. Para a formação inicial dos *clusters*, a similaridade entre duas palavras é calculada usando uma medida que analisa a informação mútua entre os seus vetores. O centróide de cada *cluster* é construído a partir da média dos vetores de um subconjunto de todos os vetores do *cluster*. Esse subconjunto é considerado determinante para eleger quais os novos membros do *cluster*. O algoritmo concentra-se na escolha desse subconjunto, de modo que o centróide represente o vetor com as características mais comum do sentido representado pelo *cluster*. Novas palavras ambíguas são atribuídas a um *cluster* de acordo com a sua similaridade ao centróide do *cluster*. Diferentemente dos algoritmos tradicionais, os novos elementos adicionados a um *cluster* não são utilizados para recalcular o seu centróide, que permanece o mesmo, calculado com base nos exemplos de treinamento.

Os autores avaliam seu trabalho com base nos sentidos da WordNet, fazendo, para tanto, um mapeamento entre os sentidos dessa base e os indicados pelos *clusters*. Os textos empregados são do gênero jornalístico e comumente utilizados em avaliações de trabalhos para a área de Recuperação de Informações. Foram selecionadas 13.403 palavras, com uma média de 740.8 características por palavra. Os resultados são reportados em termos de precisão (60.8%), cobertura (50.8%) e *f-measure* (55.4%).

Os autores comparam um subconjunto de 1% desse conjunto de teste com a atribuição de sentidos manual, reportando 88% de concordância entre a etiquetação manual e automática. Segundo os autores, esses resultados mostram que seu algoritmo supera outros algoritmos de *clustering*, como os hierárquicos ou híbridos, na tarefa de DLS.

Rapp (2004) apresenta uma proposta alternativa aos algoritmos de *clustering* tradicionais para a identificação e discriminação de sentidos em textos não etiquetados e a sua utilização para a DLS. Seu trabalho emprega o método *Singular Value Decomposition* (SVD) para a representação dos exemplos em vetores de dimensionalidade reduzida.

Segundo Rapp, a maioria dos trabalhos de *clustering* para discriminação de sentidos representa os exemplos por meio de vetores de co-ocorrência global, ou seja,

considerando todas as palavras de todos os exemplos. No trabalho proposto, o autor sugere que os exemplos sejam representados por vetores de co-ocorrência local, considerando apenas o contexto local das palavras ambíguas. Essa postura, segundo o autor, está relacionada à idéia de um sentido por discurso, de Gale et al. (1992c) (Seção 2.10).

O *clustering* nesse trabalho é realizado sobre vetores de co-ocorrência locais, baseados no contexto de uma única palavra. Assim, as matrizes de co-ocorrência são construídas considerando-se a co-ocorrência entre palavras e contextos, e não entre todas as palavras do córpus. Para tanto, para cada palavra é criada uma matriz cujos vetores de contexto, quando somados, formam um vetor global. Quando são construídas matrizes para todas as palavras, juntas elas formam um vetor tri-dimensional, com as duas primeiras dimensões sendo as palavras e a terceira, todos os contextos.

Um algoritmo de *clustering* poderia ser aplicado aos vetores de contexto gerados para identificar a similaridade entre eles. Contudo, essa representação é extremamente esparsa e, assim, dificilmente levaria a um resultado ótimo global. Por essa razão, os autores empregam o método SVD de redução de dimensionalidade. Esse método calcula a similaridade entre os valores da matriz, truncando os menores valores, levando a uma redução significativa no número de colunas e, alternativamente, no número de linhas. Assim, não é necessário aplicar um algoritmo de *clustering*, pois o método SVD, implicitamente, realiza o agrupamento dos exemplos. Contudo, é preciso interpretar esses agrupamentos e analisar se eles são úteis como discriminadores de sentidos.

Na tentativa de interpretar a semântica dos agrupamentos gerados, a hipótese dos autores é de que as colunas resultantes estão ordenadas por sua importância, de acordo com a similaridade dos exemplos. Essa interpretação é explicada por meio de um experimento considerando uma única palavra, utilizando todas as ocorrências do córpus BNC (*British National Corpus*) (Burnard, 2000) para essa palavra (2.054) e uma janela de contexto de 20 palavras de conteúdo à esquerda e à direita da palavra ambígua. Assim, foram gerados 2.054 contextos para a palavra ambígua. A matriz resultante consiste de 2.054 linhas x 10.610 colunas. Aplicando o método SVD a essa matriz, ela foi reduzida a duas colunas.

A interpretação dos autores para essas duas colunas é de que a primeira delas apresenta os níveis das associações entre as palavras, de modo que palavras mais próximas à palavra ambígua são apresentadas primeiramente. A segunda dimensão, mais importante, apresenta uma divisão das ocorrências em dois sentidos: a primeira parte das palavras é relacionada a um sentido da palavra ambígua, enquanto a segunda parte, a outro sentido. Assim, essa segunda dimensão poderia ser usada para a identificação dos possíveis sentidos da palavra e, implicitamente, para a discriminação dos sentidos.

Os autores não apresentam uma avaliação mais abrangente do seu trabalho, de modo que não é possível saber se sua interpretação dos agrupamentos gerados pode se estender para outras palavras, com um número maior de sentidos. Todavia, esse trabalho é importante porque representa uma alternativa ao processo de *clustering* tradicional, em um processo relativamente simples e computacionalmente barato.

3.2.2 DLS supervisionada

Os trabalhos descritos a seguir são desenvolvidos de acordo com diferentes paradigmas de aprendizado (simbólico, estatístico, conexionista e baseado em instâncias), utilizando diversos algoritmos. Contudo, eles não são separados, aqui, de acordo com o seu paradigma, pois alguns trabalhos envolvem a combinação ou comparação entre diferentes paradigmas, enquanto outros realizam aperfeiçoamentos de abordagens anteriores, considerando também o uso de outros paradigmas. Assim, para facilitar sua compreensão, eles são descritos mantendo-se a ordem cronológica, sendo que, em alguns casos,

abordagens similares dos mesmos autores são descritas na seqüência, abolindo-se a ordem cronológica. Alguns desses trabalhos são voltados para a TA.

Um dos primeiros trabalhos de DLS baseada em córpus é o de Black (1988), que desenvolveu um modelo utilizando árvores de decisão, criadas manualmente, mas simulando um processamento automático similar ao do algoritmo C4.5. A partir de um córpus de 22 milhões de palavras, foram selecionados cinco substantivos (*interest*, *point*, *power*, *state* e *terms*) e 2.000 sentenças de exemplo para cada um deles. Todos os exemplos foram manualmente etiquetados com o sentido do substantivo ambíguo. Nesses exemplos, quatro substantivos possuem quatro sentidos, enquanto um deles possui apenas três. Os exemplos foram então divididos em 1.500 exemplos de treinamento e 500 exemplos de teste.

O principal objetivo do autor era testar três configurações de características, denominadas por ele de "métodos" de desambiguação. Cada método consiste de 81 características, chamadas "categorias contextuais" e determinadas de maneira diferente e individualmente para cada uma das cinco palavras de teste. O contexto de uma palavra ambígua consiste, então, da presença ou ausência das categorias contextuais na sua sentença.

O primeiro método, denominado DG, é de domínio geral. As 81 categorias contextuais são obtidas para cada uma das cinco palavras, com base nas 500 palavras mais freqüentes em todos os exemplos daquela palavra, organizadas de acordo com o seu código de área no LDOCE. Assim, cada categoria contextual consiste de uma lista de palavras de um determinado código de área.

Os dois outros métodos são de domínio específico, denominados DS1 e DS2, respectivamente. O método DS1 é baseado nas freqüências dos diferentes itens lexicais dos 1.500 exemplos de treinamento de cada palavra. As categorias contextuais são constituídas das 41 palavras mais freqüentes em uma janela de duas palavras à esquerda e à direita do substantivo ambíguo e das 40 palavras mais freqüentes nos exemplos, em qualquer posição, excluindo-se as palavras de classe fechada. Assim, cada categoria contextual consiste de uma palavra.

No método DS2, 20 categorias contextuais são também baseadas nas freqüências dos itens lexicais dos exemplos de treinamento de cada palavra, como no DS1. As demais 61 categorias contextuais são derivadas de outras 100 sentenças do córpus inicial, de 22 milhões de palavras, que não incluem as cinco palavras de teste. O conteúdo dessas sentenças foi manualmente analisado, resultando em possíveis categorias temáticas presentes no córpus (por exemplo, *document*, *energy*, *powerful_people*). Assim, 20 categorias contextuais consistem de uma palavra, enquanto as outras 61 consistem de palavras representando as categorias temáticas.

O autor construiu uma árvore de decisão para cada um dos métodos, em cada uma das cinco palavras, num total de 15 árvores. Para verificar a acurácia de cada método, as três árvores para cada palavra foram testadas no conjunto de 500 exemplos. A acurácia média (para todas as palavras) do método DG foi de 47%, do método DS1, de 72%, e do método DS2, de 75%. Em todos os casos, a acurácia é maior que a *baseline* de escolhas ao acaso (37%). Segundo o autor, esses resultados mostram que as categorias contextuais que determinam as principais decisões nas árvores são aquelas relacionadas à estrutura ou ao conteúdo temático da palavra ambígua e, portanto, essas categorias são mais informativas.

É importante ressaltar que todo o processo de DLS relatado no trabalho é realizado manualmente, desde a extração das janelas de contextos à identificação dos valores referentes às categorias temáticas do método DS2, tanto para os exemplos de treinamento quanto para os exemplos de teste. Para que esse processamento fosse realizado

automaticamente, seria necessário um recurso como um *thesaurus*. Neste caso, as categorias também precisariam ser definidas de acordo com as categorias presentes no *thesaurus*. Certamente, vários problemas, em diferentes níveis de dificuldade, estariam envolvidos nesse processo.

O trabalho de Hearst (1991), que também visa à desambiguação de substantivos, pode ser considerado de aprendizado semi-supervisionado, de acordo com a nomenclatura adotada neste trabalho. Esse trabalho parte de um conjunto de exemplos com sentidos manualmente etiquetados para várias ocorrências dos substantivos a desambiguar e emprega técnicas de *bootstrapping* para aumentar o número de exemplos etiquetados com base no conjunto inicial. Isso é feito considerando-se os exemplos já etiquetados como o conjunto de treinamento para um algoritmo de aprendizado supervisionado. A partir dos exemplos de treinamento, o algoritmo classifica parte dos exemplos não anotados e, caso essa classificação atinja um nível de confiabilidade pré-definido, os exemplos são adicionados ao conjunto de treinamento.

Antes do treinamento do sistema, as sentenças de exemplo são pré-processadas para determinar a categoria gramatical das suas palavras e gerar sua estrutura sintática superficial. No treinamento, o algoritmo implementado analisa o contexto do substantivo a desambiguar, extraindo uma série de características ortográficas, sintáticas e lexicais do substantivo e da sua vizinhança. Essas características incluem: se a palavra ambígua ou seus modificadores são capitalizados, se a palavra ambígua modifica ou é modificada por outra, se a palavra ambígua faz parte de uma locução preposicionada, etc. Essas características são representadas em vetores, sobre os quais são usadas técnicas estatísticas para gerar o modelo de desambiguação, com base na importância das características para a distinção dos sentidos e na sua freqüência.

Depois de estabelecida a importância relativa de cada uma das características, os exemplos não rotulados são pré-processados e submetidos ao algoritmo. Os resultados da desambiguação que atingirem um determinado limite de confiança são adicionados ao conjunto de treinamento, sendo tratados como exemplos manualmente etiquetados e, portanto, usados como evidência adicional para a desambiguação de novos exemplos. O processo continua até que o número máximo possível de exemplos seja etiquetado. A partir de então, o algoritmo pode ser utilizado para desambiguar novos casos.

A autora determina que são necessárias pelo menos 10 ocorrências iniciais já desambiguadas, mas que um número bem maior (20 ou 30) é necessário para uma desambiguação mais precisa. Contudo, como não são necessários mecanismos complexos para uma análise lingüística profunda, a desambiguação é relativamente “barata”. Hearst realiza um teste com quatro substantivos, considerando de três a quatro conjuntos de exemplos de treinamento inicialmente etiquetados, com 20 a 70 exemplos cada um, por sentido, de cada palavra. Ela relata que a acurácia aumenta de acordo com o número de exemplos etiquetados fornecidos e que a acurácia média é de 80% (considerando apenas os exemplos etiquetados manualmente, ou seja, sem o *bootstrapping*). Os mesmos testes são realizados considerando mais uma iteração do algoritmo, incorporando também os exemplos etiquetados pelo sistema ao conjunto de treinamento. A acurácia média aumenta para 83%.

Segundo a autora, esses resultados são comparáveis ou melhores que os de trabalhos anteriores, como os de Lesk (1986) e Guthrie et al. (1991), e mostram que a técnica de *bootstrapping*, além da vantagem de permitir um número muito pequeno de exemplos manualmente etiquetados, não causa nenhuma degradação nos resultados. Vale notar que seu trabalho visa apenas à desambiguação entre dois sentidos totalmente distintos.

Yarowsky (1994) propõe a adaptação da técnica de Listas de Decisão de Rivest (1987) para a resolução de ambigüidades lexicais de vários tipos e aplica essa técnica em diversos trabalhos posteriores (Yarowsky, 1995; 2000) especificamente para a ambigüidade lexical de sentido.

A técnica de listas de decisão consiste em inicialmente processar os exemplos de treinamento para extrair as características consideradas, que recebem pesos de acordo com uma medida de verossimilhança. Em princípio, essa medida considera distinções entre apenas dois sentidos para cada palavra, mas pode ser facilmente adaptada para distinções entre qualquer quantidade de sentidos (isso é feito, por exemplo, em (Agirre & Martínez, 2000), que será descrito posteriormente). A lista de todas as características ordenadas (em ordem decrescente) de acordo com seus pesos constitui a lista de decisão. Para desambiguar novos exemplos, a lista de decisão é percorrida em ordem e a característica com o peso mais alto no exemplo de teste seleciona o sentido mais apropriado.

O primeiro trabalho para a DLS (Yarowsky, 1995), apesar de denominado pelo autor de não-supervisionado, se enquadra, na classificação adotada deste trabalho, como uma proposta semi-supervisionada. O trabalho é voltado para a identificação de sentidos em larga escala e se baseia em duas hipóteses já estabelecidas pelo autor em trabalhos anteriores: a hipótese de um sentido por *collocation*, aceita como válida partindo-se da suposição de que a hipótese de um sentido por discurso é verdadeira (Seção 2.10).

Esse trabalho provê mecanismos para alimentar o sistema com "sementes" (*seeds*) para o processo de desambiguação supervisionada. As sementes correspondem a um pequeno conjunto de *collocations*, definidas pelo autor como ocorrências de cada palavra a ser desambiguada, juntamente com o seu contexto e a indicação do seu sentido, naquele contexto. As *collocations* podem ser geradas de três diferentes maneiras, com ou sem a interferência do usuário: (a) são usadas definições de cada sentido das palavras em um recurso lexical qualquer, como um dicionário eletrônico; (b) o usuário identifica uma *collocation* para cada sentido da palavra ambígua; (c) o usuário escolhe o sentido adequado para *collocations* fornecidas pelo sistema, identificadas a partir de informações sobre a co-ocorrência da palavra ambígua com outras palavras no córpus. O sistema determina, portanto, entre quais sentidos a desambiguação deve ocorrer a partir dos sentidos do dicionário eletrônico ou dos sentidos indicados pelo usuário nas sementes.

A partir dessas sementes, para cada palavra ambígua, em um processo iterativo, primeiramente, são classificados os casos "óbvios" do córpus, ou seja, as sentenças do córpus que contêm uma das sementes. Em seguida, utilizando listas de decisão, o conjunto de sentenças do córpus classificadas é examinado para obter mais indicadores sobre os sentidos da palavra para classificar os demais exemplos, denominados "resíduos". Esses indicadores são ordenados de acordo com a quantidade de evidência que eles provêem para cada sentido. Com isso, é possível classificar uma nova quantidade de sentenças do córpus, diminuindo a quantidade de resíduos e produzindo mais indicadores para cada sentido.

Esse processo continua até que a lista de decisão torne-se estável e que todas (ou uma proporção acima de um limite pré-estabelecido) as sentenças do córpus com a palavra ambígua sejam classificadas. A lista ordenada de indicadores de sentido resultante representa um conhecimento generalizado sobre o córpus, ou seja, o modelo de desambiguação, e pode ser utilizada para desambiguar novos casos. Segundo Stevenson & Wilks (1998), essa lista pode ser considerada um conjunto de regras simbólicas.

Nesse processo iterativo, a hipótese de um sentido por *collocation* pode ser verificada na generalização dos sentidos identificados nas *collocations* para outros exemplos do córpus com as mesmas *collocations*. Já a hipótese de um sentido por discurso pode ser verificada em duas situações: (a) na utilização de um filtro na generalização dos

sentidos, por exemplo, corrigindo a etiqueta de uma palavra em um discurso com um sentido contrário ao da maioria das demais ocorrências da palavra no mesmo discurso; e (b) no processo de generalização dos sentidos, atribuindo a uma palavra o mesmo sentido já empregado a várias ocorrências dessa palavra no mesmo discurso.

Yarowsky avalia seu trabalho em um conjunto de 12 palavras ambíguas e obtém um desempenho de cerca de 96%, superior ao de diversos trabalhos supervisionados. Contudo, os testes realizados consideram a desambiguação apenas entre dois sentidos muito distintos de cada palavra ambígua. Como foi discutido na Seção 2.10, essa configuração é a única na qual as hipóteses de um sentido por discurso e por *collocation* se mostram completamente válidas. Vale notar que o trabalho de Yarowsky pode ser considerado híbrido, se for utilizado um recurso lexical (como um dicionário eletrônico) para a identificação das *collocations* semente. Como o autor exemplifica apenas o procedimento considerando a identificação feita por um usuário, o sistema é considerado, aqui, apenas baseado em córpus.

Em um trabalho posterior (Yarowsky, 2000), o autor apresenta uma proposta de listas de decisão hierárquicas como melhoria dos trabalhos anteriores, que utilizavam listas de decisão planas.

A diferença das listas hierárquicas, com relação às planas, é que elas permitem ramificações condicionais, de modo a dividir o fluxo de controle do procedimento de decisão em caminhos especializados relativamente independentes para modelar as necessidades de cada parte da divisão. Isso já é feito nas tradicionais árvores e regras de decisão, contudo, nas árvores ou regras de decisão a ramificação ocorre em todas as características. Segundo o autor, esse excesso de ramificações leva a uma fragmentação desnecessária nos exemplos de treinamento. Por essa razão, o seu mecanismo de listas de decisão hierárquicas prevê o pré-estabelecimento das características nas quais deve ocorrer um particionamento dos exemplos, de acordo com os seus valores para tais características.

Para a lista de decisão hierárquica apresentada e avaliada no seu trabalho é utilizado um conjunto de cinco características: a categoria gramatical da palavra ambígua, as flexões morfológicas da palavra ambígua, expressões idiomáticas mais importantes, traços sintáticos e, por fim, subsentidos, quando um sentido apresenta uma estrutura de sentidos mais gerais e mais específicos.

O trabalho é avaliado na competição SENSEVAL-2, apresentando a melhor precisão, dentre todos os candidatos (média de 78.9%, considerando todas as categorias gramaticais). Uma observação importante é que a precisão é significativamente superior para substantivos (87%) do que para outras classes, por exemplo, para verbos (74.3%). O autor também compara esse desempenho com o obtido por meio das listas de decisão planas, usadas anteriormente. A precisão diminui cerca de 8% com as listas planas. Assim, a conclusão do autor é de que seu trabalho, considerando o conjunto de características empregado e a técnica de listas de decisão hierárquicas, melhora a qualidade da desambiguação, mantendo os benefícios do fluxo de dados geral das listas de decisão planas.

Mooney (1996) realiza experimentos de avaliação comparativa com 7 algoritmos de aprendizado supervisionado, de modo a analisar o seu desempenho na DLS. Foram analisados algoritmos de diferentes paradigmas de aprendizado: estatístico (Naive Bayes), conexionista (perceptrons), baseado em instâncias (KNN) e simbólico (C4.5, listas de decisão, DNF - *Disjunctive Normal Form* e CNF - *Conjuctive Normal Form*).

Todos os algoritmos foram testados considerando uma única palavra, *line*, e os seus 6 sentidos na WordNet. Para tanto, foram coletados 1.200 exemplos de treinamento e 894

casos de teste. As características empregadas são todas as palavras da sentença ambígua e as palavras da sentença anterior no córpus, sem considerar sua ordenação.

Os resultados mostram que os algoritmos Naive Bayes e de redes neurais apresentam os melhores resultados, seguidos dos algoritmos de listas de decisão e C4.5. O bom desempenho do classificador Naive Bayes sob as condições simples de teste é confirmado em trabalhos posteriores, por exemplo, Ng & Zelle (1997). Contudo, como apontado por Ng (1997a), as condições de teste usadas por Mooney são limitadas: apenas uma palavra ambígua, poucos exemplos e somente características referentes às palavras, representadas como *bag-of-words*. Além disso, não foram exploradas configurações mais adequadas nos parâmetros dos algoritmos. Por exemplo, o algoritmo KNN foi testado para analisar a semelhança com apenas três vizinhos (k = 3). Como mostram Escudero et al. (2000b), valores maiores desse parâmetro tendem a apresentar resultados melhores.

Ng & Lee (1996) propõem um trabalho para DLS utilizando um algoritmo baseado em instâncias e integrando diversos tipos de conhecimento: categoria gramatical das palavras vizinhas, traços morfológicos (número para substantivos e modo/tempo/pessoa para verbos), co-ocorrência de palavras vizinhas não ordenadas, *collocations* e relações sintáticas verbo-objeto (apenas para substantivos).

O sistema implementado provê a etiquetação de sentidos para todas as palavras de conteúdo de sentenças irrestritas. Ele assume, como entrada, as palavras etiquetadas com a sua categoria gramatical e seus traços morfológicos. Os sentidos utilizados são os fornecidos pela WordNet.

O algoritmo baseado em instâncias empregado é denominado PEBLS (Cost & Salzberg, 1993), um algoritmo KNN que permite a atribuição de pesos para os exemplos e as características e que permite características simbólicas. São considerados como características todos os tipos de conhecimento citados. A medida de similaridade adotada considera que a distância entre dois exemplos é a soma da distância entre todas as características desses exemplos.

Os autores inicialmente avaliam seu trabalho considerando o conjunto de teste definido por Bruce & Wiebe (1994), que contém 2.369 sentenças com uma ocorrência da palavra *interest*, etiquetada com os sentidos do LDOCE. Como resultado, relatam uma precisão média de 87.4%, maior do que a alcançada por Bruce & Wiebe (78%). Considerando o mesmo conjunto de teste, os autores realizam outros experimentos para analisar a contribuição de cada tipo de conhecimento empregado. Em cada experimento, consideram apenas um tipo de conhecimento. A maior precisão (80.2%) foi obtida com o uso apenas de *collocations*. Segundo os autores, esse resultado corrobora observações de outros autores (Choueka & Lusignan, 1985)[12], de que o ser humano usa uma janela de contexto de algumas poucas palavras para realizar a DLS. Esses experimentos também representam um diferencial com relação aos outros trabalhos envolvendo vários tipos de conhecimento (por exemplo, McRoy, 1992), que não avaliam a contribuição individual de cada um dos tipos.

Para uma avaliação mais substancial, Ng & Lee constroem, manualmente, o DSO, um córpus com 192.800 exemplos dos 121 substantivos e 70 verbos mais freqüentes da língua inglesa, etiquetado com os sentidos da WordNet (descrito na Seção 2.7.1). Dois testes com subconjuntos do córpus de tamanhos diferentes foram realizados. Os resultados de ambos os testes indicam que o córpus com uma quantidade maior de exemplos apresenta uma acurácia maior (68.6% contra 54%), o que indica que o algoritmo baseado em instâncias tende a apresentar um desempenho significativamente maior com o aumento

[12] Choueka, Y.; Lusignan, S. (1985). Disambiguation by Short Contexts. Computers & the Humanity, 19, pp. 147-157. Apud Ng & Lee (1996).

do número de exemplos. Além disso, esses resultados superam a *baseline* de escolha pelo sentido mais freqüente em ambos os testes (63% e 47.1%).

Posteriormente, Ng (1997a) procura aperfeiçoar seu trabalho baseado em exemplos utilizando um procedimento para a validação cruzada sobre os exemplos de treinamento para determinar, automaticamente, qual é o melhor valor para os *k* vizinhos mais próximos como parâmetro no algoritmo PEBLS para o treinamento desse algoritmo. Esse procedimento aponta para um valor maior de *k* do que o freqüentemente usado (*k* = 1, que é o valor padrão do algoritmo). Por exemplo, Mooney (1996), na sua avaliação comparativa descrita anteriormente, utiliza *k* = 3 para o algoritmo KNN e obtém um desempenho inferior ao do algoritmo probabilístico Naive Bayes. Segundo Ng, o parâmetro *k* é determinante no desempenho do algoritmo PEBLS.

Ng sugere uma avaliação mais substancial que a de Mooney para comparar seu trabalho aos probabilísticos, bem como ao apresentado na sua proposta anterior (Ng & Lee, 1996). Para tanto, utiliza, novamente, os dois subconjuntos do córpus DSO, mas considera apenas *collocations* como características na representação dos exemplos, já que elas se mostraram as características individuais mais importantes na avaliação da proposta anterior.

Os resultados da avaliação mostram que a determinação automática do parâmetro *k* melhora a acurácia da desambiguação, mesmo considerando apenas *collocations* (75.2% e 58.7% nos dois conjuntos, contra os 68.6% e 54% apresentados na proposta anterior). Além disso, com essa configuração o algoritmo PEBLS apresenta uma acurácia similar à do Naive Bayes (74.5% e 58.2%) nos mesmos conjuntos de teste.

Towell & Voorhess (1998) apresentam um trabalho supervisionado, voltado para a Recuperação de Informações, que utiliza redes neurais para aprender um modelo de classificação. São empregadas duas redes que consideram características distintas, uma com o contexto local e outra com o contexto global. O classificador final combina as saídas das duas redes.

O contexto global é constituído de substantivos que têm a probabilidade de co-ocorrer com determinados sentidos da palavra ambígua. O contexto local inclui informações sobre a ordem das palavras, suas distâncias e algumas informações sobre a estrutura sintática, considerando todas as palavras vizinhas à palavra ambígua na sentença. A categoria gramatical das palavras é determinada em uma etapa prévia, mas essa informação não é utilizada como característica, apenas para separar os modelos para diferentes categorias.

Diferentemente da rede neural de Véronis & Ide (1990, Seção 3.1.2), nesse trabalho, a rede precisa ser inicialmente alimentada com exemplos etiquetados. Para tanto, os autores utilizam um córpus do qual extraem exatamente o mesmo número de exemplos para cada sentido de uma palavra ambígua, visando eliminar os efeitos da freqüência de cada sentido.

O trabalho é testado em três palavras ambíguas de três classes gramaticais distintas (verbo, substantivo e adjetivo), para avaliar o desempenho individual de cada rede, bem como o desempenho da combinação de ambas, com base em diferentes números de exemplos. A maior acurácia é atingida com a combinação das redes, com o maior número de exemplos possível: 87%, 90% e 81%, respectivamente, para as três palavras.

Um problema decorrente do uso de redes neurais para o aprendizado é o grande número de exemplos etiquetados necessários e, como conseqüência, o tempo elevado para o treinamento da rede e também para a classificação de novos casos. Esse tempo elevado

pode causar prejuízos para a sua utilização, por exemplo, em um sistema de recuperação de informações *on-line*.

Com relação à quantidade de exemplos necessários, para tentar minimizar a necessidade de etiquetação manual, os autores propõem um algoritmo semi-supervisionado, em que a rede neural procura gerar classificadores precisos com um pequeno número de exemplos de treinamento etiquetados e um grande número de exemplos não etiquetados. Esse algoritmo se baseia na similaridade dos exemplos etiquetados com aqueles que não possuem uma etiqueta para criar "exemplos sintéticos".

Para criar um exemplo sintético, um exemplo etiquetado é utilizado como semente. A partir dele, são coletados exemplos próximos. O exemplo que representa o centróide desse conjunto é computado. Não são incluídos, nesse conjunto, exemplos que já possuem outras etiquetas ou exemplos que são mais próximos de outros exemplos etiquetados. Desse modo, formam-se grupos de exemplos que recebem o sentido do exemplo mais similar.

Os autores comparam essa variação do trabalho com a anterior, que usa apenas exemplos etiquetados, considerando o mesmo número de exemplos, contudo, divididos em etiquetados e não etiquetados. Eles relatam uma melhoria na acurácia da rede considerando exemplos não rotulados, contudo, não discutem os possíveis motivos desses resultados.

Pedersen (2000) apresenta um trabalho em que vários classificadores estatísticos (Naive Bayes) são combinados, em um conjunto (*ensemble*) de classificadores, de modo que a escolha do sentido é determinada pelo voto da maioria dos classificadores. Esse trabalho é motivado pela observação do autor de que o aumento do conjunto de características ou a escolha de um algoritmo de aprendizado mais complexo geralmente não implica um desempenho na desambiguação melhor do que o obtido com o uso de características lexicais superficiais (como co-ocorrências) e um algoritmo de aprendizado supervisionado simples. De acordo com avaliações mencionadas em vários trabalhos (por exemplo, Mooney, 1996; Ng & Lee, 1996), os classificadores estatísticos são os que apresentam melhores resultados se considerados isoladamente, dentre os trabalhos baseados em córpus. Assim, sua suposição é de que a combinação de vários classificadores estatísticos simples, empregando diferentes conjuntos de características, pode melhorar significativamente o desempenho da DLS.

A combinação de Pedersen inclui 81 classificadores simples, que consideram diferentes janelas de contexto como vetores de características. As variações incluem o tamanho total da janela (0, 1, 2, 3, 4, 5, 10, 25 ou 50 palavras) e diferentes configurações desse tamanho para contextos dos dois lados da palavra ambígua (direito e esquerdo).

Esse trabalho foi avaliado em dois grupos de palavras ambíguas, *line* e *interest*, considerando-se os seis possíveis sentidos distintos da WordNet para cada uma delas. Cada um dos possíveis classificadores foi testado individualmente, para cada palavra. A combinação de classificadores foi testada para cada palavra. Comparado ao desempenho do melhor classificador individual, o desempenho da combinação de classificadores aumentou 4% para a palavra *line* (de 84% para 88%) e 3% para a palavra *interest* (de 86% para 89%).

Considerando-se que são utilizadas apenas características lexicais superficiais, sem outras fontes de informação e sem nenhum processamento lingüístico mais profundo (somente a lematização das palavras), o desempenho relatado é bastante significativo. De fato, esses resultados são superiores aos de outros trabalhos similares, baseados em técnicas estatísticas ou em outros paradigmas de aprendizado, mas que utilizam apenas um classificador.

O trabalho de Zinovjeva (2000) emprega o método de aprendizado por transformações (TBL - *Transformation Based Learning*) (Brill, 1995) com o objetivo de aprender automaticamente regras (simbólicas) para traduzir corretamente palavras ambíguas do inglês para o sueco, em textos irrestritos, de qualquer gênero e domínio.

Um conjunto de exemplos de treinamento é criado a partir de sentenças do córpus BNC manualmente etiquetadas com a tradução dos verbos e substantivos ambíguos. A partir desses exemplos, são realizados alguns experimentos de aprendizado, cada um considerando determinados tipos de conhecimento. Com esses experimentos, a autora pretende verificar quais conhecimentos são mais adequados para e, assim, empregá-los na construção do seu modelo de DLS. Os experimentos consideram cada palavra ambígua, individualmente, e assumem que as palavras da sentença já possuem etiquetas gramaticais, corretamente atribuídas em uma etapa de pré-processamento.

No primeiro experimento, são consideradas apenas as palavras vizinhas à palavra ambígua na sentença. São criados modelos para três palavras, dois substantivos e um verbo. Os exemplos incluem 4.800 ocorrências de cada substantivo e 780 ocorrências do verbo. Cerca de 10% desses exemplos são usados para teste e o restante, para o treinamento do modelo. A acurácia obtida para cada palavra foi de 92.1%, 95.2% e 73.1%.

O segundo experimento considera, em vez das palavras vizinhas, as suas categorias gramaticais. Com base na mesma configuração que a do primeiro experimento, a acurácia obtida aumentou para 93.6%, 95.4% e 80.8%.

O terceiro experimento considera as relações sintáticas das palavras do contexto da palavra ambígua, produzidas por um *parser*. Apenas o modelo para o verbo é gerado, a partir de 78 das suas ocorrências. A acurácia obtida foi de 83.3%. Segundo a autora, essa acurácia relativamente baixa deve-se ao tamanho reduzido do conjunto de treinamento.

O quarto experimento considera a combinação das etiquetas gramaticais com as relações sintáticas. Novamente, apenas o modelo para o verbo, com 78 das suas ocorrências, é gerado. A acurácia obtida foi de 84.6%, pouco maior que a do experimento anterior.

A cada experimento, uma etapa subseqüente de alteração manual das regras foi realizada, visando aperfeiçoar regras muito genéricas ou muito específicas. A avaliação dos modelos considerando essas alterações levou a uma acurácia superior, em todos os casos. Vale notar que essa alteração só foi possível porque as regras geradas são simbólicas.

Zinovjeva não menciona a cobertura do seu trabalho nos diversos experimentos, sua medida de acurácia corresponde, na verdade, à precisão das regras, já que são considerados apenas os exemplos classificados.

A autora destaca que os tipos de conhecimento empregados em cada regra podem variar, pois os experimentos não mostraram, em geral, um único tipo ou uma combinação de tipos de conhecimento mais adequado para todas as palavras ambíguas. Isso pode depender de uma série de fatores característicos de cada palavra e do córpus em uso. Com isso, Zinovjeva reforça a necessidade da geração de modelos específicos para cada palavra ou grupos de palavras similares.

As regras que apresentaram a maior acurácia são incorporadas a um sistema de TA já existente, também baseado em transformações. Sem o módulo de DLS, o sistema necessita da interação com o usuário para que ele escolha entre todas as possíveis traduções de palavras ambíguas. São ilustrados apenas alguns exemplos de traduções de sentenças do inglês para o sueco com e sem a utilização das regras. Uma avaliação mais substancial do uso dos modelos no sistema de TA não é apresentada.

O trabalho de Zinovjeva é bastante próximo do modelo de DLS que se pretende propor, uma vez que é voltado especificamente para a TA, considera o aprendizado de máquina para a geração das regras, cria um modelo específico para cada palavra e analisa a

influência de diversos tipos de conhecimento no processo de DLS. Além disso, a autora realiza a desambiguação com base em restrições estabelecidas para as palavras na língua-fonte, não na língua-alvo. Assim, a DLS pode ser considerada monolingüe, contudo, são realizadas apenas as distinções de sentido que se manifestam na língua-alvo. Segundo a autora, não é necessário preocupar-se com a representação de todos os possíveis sentidos das palavras ambíguas na língua-fonte, já que vários sentidos podem possuir a mesma tradução na língua-alvo e, inversamente, uma palavra que possui um único sentido na língua-fonte pode ser traduzida por diferentes palavras na língua-alvo. Outras semelhanças do seu trabalho com o que se pretende propor dizem respeito ao uso de uma abordagem simbólica para a geração de regras e à análise das regras geradas para uma possível edição manual.

Motivados pelos resultados encorajadores do trabalho baseado em listas de decisão na primeira edição da competição SENSEVAL, Agirre & Martínez (2000) empregam as listas de decisão de Yarowsky em sua abordagem de DLS. Entretanto, diferentemente de Yarowsky, eles se propõem a analisar o uso dessa técnica para a distinção dos sentidos refinados da WordNet.

Para o treinamento e teste do trabalho são utilizados os córpus SEMCOR e DSO. Em ambos os córpus, os resultados da avaliação indicam uma precisão de cerca de 70%, um valor significativamente mais alto que o das duas *baselines* consideradas: escolha randômica (cerca de 13%) e escolha pelo sentido mais freqüente (cerca de 54%). Segundo os autores, esse resultado pode ser considerado bom, dado o nível de refinamento dos sentidos e o fato de que são desambiguadas todas as palavras das sentenças. Em um teste considerando sentidos menos refinados (as classes semânticas da WordNet), o trabalho obteve 83% de precisão nos dois córpus. A cobertura do trabalho também é significativa: maior que 90%.

Os autores procuram avaliar também a concordância na etiquetação manual dos dois córpus, treinando o modelo com um desses córpus e testando-o com o outro. A precisão obtida foi, como esperado, bastante baixa, mostrando a grande discordância entre os etiquetadores humanos na distinção entre sentidos muito refinados, já discutida anteriormente (Seção 2.4.1).

Escudero et al. (2000a) apresentam os resultados de uma comparação sistemática entre um algoritmo do paradigma estatístico (Naive Bayes) e um algoritmo do paradigma baseando em instâncias (KNN). O algoritmo KNN foi testado com diferentes configurações, considerando a atribuição de pesos aos exemplos e/ou aos atributos e diferentes métricas de similaridade. O objetivo, com isso, é verificar qual dos algoritmos leva resultados mais precisos na DLS.

O córpus usado para treinamento e teste dos algoritmos é um subconjunto do DSO para 15 verbos e substantivos ambíguos. Para cada palavra foram coletados todos os exemplos do córpus (373 a 1.500 exemplos por palavra). Foram usadas sete características referentes ao contexto local na desambiguação, consistindo de algumas palavras vizinhas à palavra ambígua (duas palavras à esquerda e duas à direita, por exemplo). Além das variações nos algoritmos, os autores criaram dois conjuntos com diferentes números de características, de modo a avaliar o comportamento dos algoritmos mediante o aumento do número de características.

Os resultados mostram que o algoritmo baseado em instâncias (com a atribuição de pesos) apresenta uma acurácia maior que a do Naive Bayes, em ambos os conjuntos de características. Além disso, somente o algoritmo baseado em instâncias obteve uma melhora na sua acurácia com o conjunto maior e mais significativo de características, o que

indica que o algoritmo Naive Bayes não é muito sensível à inclusão de novas características, ainda que elas sejam relevantes.

Os autores também avaliam a acurácia dos algoritmos considerando uma variação na qual são analisadas somente as características "positivas" em cada exemplo, ou seja, somente as informações sobre as palavras que o exemplo contém, excluindo-se as informações sobre palavras que ele não contém. Essa variação é importante, segundo os autores, para a manipulação grandes quantidades de atributos. Os resultados da avaliação mostram que essa variação melhora ou pelo menos mantém a acurácia dos métodos sem essa restrição.

Escudero et al. (2000b) empregam o algoritmo de *boosting* AdaBoost (Freund & Schapire, 1996), já utilizado em outras tarefas de aprendizado de máquina (incluindo no PLN), para a DLS. Algoritmos de *boosting*, de modo geral, procuram combinar um número grande de hipóteses simples e com acurácia moderada (geradas por classificadores fracos) em uma única hipótese com uma acurácia alta. Os classificadores fracos são treinados seqüencialmente e, assim, a cada iteração a hipótese fraca aprendida (na forma de regra) tende a cobrir exemplos mais difíceis de classificar pelas hipóteses anteriores. As hipóteses fracas aprendidas são combinadas linearmente para formar a hipótese final.

O algoritmo é avaliado na mesma configuração definida em Escudero et al. (2000a), descrito acima. Além da versão padrão do algoritmo criado, é considerada uma variação na qual o critério de parada é otimizado para evitar o super-ajuste das hipóteses aos exemplos de treinamento (*overfitting*). Os resultados da avaliação, comparados aos dois trabalhos dos autores já citados e à *baseline* do sentido mais freqüente, mostram que as duas variações do algoritmo de *boosting* apresentam uma acurácia melhor na desambiguação de 14 das 15 palavras (ambas com 68% de acurácia, contra 66%, em média, dos outros dois trabalhos, e 53% da *baseline*).

Conceitualmente, o algoritmo AdaBoost, segundo os autores, é bastante adequado para a tarefa de DLS, contudo, ele apresenta a desvantagem de ser computacionalmente caro em termos de complexidade e tempo e, portanto, não poderia ser aplicado para a DLS em larga escala. Isso ocorre em função da alta dimensionalidade do espaço de atributos que é explorado pelos classificadores fracos: quanto maior o número de exemplos e de sentidos possíveis para cada palavra ambígua, maior a dimensionalidade. Por essa razão, os autores propõem uma adaptação do método, utilizando uma técnica denominada *LazyBoosting*, que reduz o número de atributos que são examinados em cada iteração do algoritmo. Essa variação do algoritmo foi avaliada na mesma configuração. A acurácia média obtida foi de 69.5%.

Em um trabalho posterior (Escudero et al., 2000c), os autores descrevem uma avaliação comparativa do seu algoritmo *LazyBoosting* com outros quatro algoritmos supervisionados: Naive Bayes, KNN, listas de decisão e SNoW (*Sparse Network of Winnows*) (Carlson et al., 1999), um algoritmo que gera uma rede de funções lineares com regras de atualização. Além de verificar o desempenho desses algoritmos, os autores visavam analisar a sua dependência do domínio dos textos, explorando, para tanto, conjuntos de treinamento e teste distintos.

Os algoritmos foram testados em um subconjunto de 21 palavras altamente ambíguas do córpus DSO, incluindo 13 substantivos e 8 verbos, com uma média de 10 sentidos e 1.000 exemplos para cada palavra. Foram utilizadas 15 características locais, consistindo de palavras de conteúdo vizinhas em determinadas posições e das etiquetas gramaticais de algumas dessas palavras, e características globais, consistindo de todas as palavras da sentença.

Em sete testes com diferentes combinações dos conjuntos de treinamento e teste, os resultados da avaliação mostram que o algoritmo *LazyBoosting* supera todos os demais, com uma acurácia média de 62%, enquanto que o algoritmo Naive Bayes, na última colocação, apresenta uma acurácia média de 56.6%. Nessa acurácia média estão computados os valores dos testes considerando tanto conjuntos de treinamento e teste de mesmo domínio quanto de domínios distintos.

Os autores realizam também outros experimentos, considerando apenas conjuntos de treinamento e teste distintos e, gradualmente, inserindo exemplos ao conjunto de treinamento do mesmo domínio do conjunto de teste. A acurácia de todos os classificadores tende a aumentar à medida que mais exemplos do domínio sendo testado são inseridos ao conjunto de treinamento. Com isso, todos os algoritmos mostram uma forte dependência do domínio, de modo que seria necessária alguma forma de adaptação para o uso dos modelos em córpus de domínios diferentes daqueles para os quais foram inicialmente criados.

Em outro trabalho posterior (Escudero et al., 2001), os autores apresentam os resultados da avaliação de uma nova versão do seu algoritmo *LazyBoosting* na competição SENSEVAL-2. Nessa versão são incluídas outras características: um número maior de características sobre o contexto local e uma característica para identificar o domínio. O algoritmo obteve 59.4% e 67.1% de acurácia, considerando, respectivamente, sentidos mais refinados e mais genéricos. Segundo os autores, o uso de informações sobre o domínio da palavra ambígua foi determinante para a obtenção de resultados precisos.

Pedersen (2002a) também realiza alguns experimentos com trabalhos baseados em córpus supervisionados, utilizando classificadores estatísticos (Naive Bayes) e simbólicos (árvores de decisão - C4.5), para verificar a utilidade do uso apenas de características lexicais simples, facilmente extraídas dos exemplos de treinamento, para a desambiguação. Como características, são consideradas todas as palavras (não ordenadas) do contexto (*bag-of-words*), todas as possíveis bigramas e co-ocorrências. Cada exemplo é descrito simplesmente pela indicação binária de possuir ou não cada uma das características.

As variações nos experimentos realizados incluem classificadores Naive Bayes individuais, cada um com um dos três tipos de características, o *bagging* de uma árvore de decisão considerando somente bigramas, e uma combinação (*ensemble*) do *bagging* de três árvores de decisão, cada uma utilizando um dos três tipos de características. A técnica de *bagging* simula uma combinação de classificadores: o conjunto de exemplos de treinamento é dividido em 10 amostras, sendo que uma árvore de decisão é aprendida para cada uma das amostras e o sentido atribuído é aquele indicado pela maioria das 10 árvores geradas. O uso dessa técnica permite para minimizar a instabilidade e alta variância dos algoritmos baseados em árvores de decisão

Os dados utilizados são os da competição SENSEVAL-2, para o inglês e o espanhol. Os melhores resultados de todos os experimentos são os apresentados pelas árvores de decisão, em particular, pela combinação do *bagging* de três árvores de decisão: acurácia de 54% para o espanhol e 60% para o inglês. Esses resultados, segundo o autor, são entre 7% e 10% menores que os dos sistemas que atingiram os melhores resultados na competição.

Segundo o autor, os resultados mostram o potencial das características lexicais simples. De fato, partindo de um conjunto simples de características, que não exige nenhum processamento lingüístico mais aprofundado, pode-se realizar testes diversos, com vários paradigmas, técnicas, algoritmos e parâmetros de aprendizado.

As vantagens de uso de características como as testadas, além da facilidade para sua extração, incluem a independência de língua. Pelo menos nas duas línguas analisadas, que são relativamente próximas, são usadas as mesmas características e é empregado o mesmo procedimento para a extração dessas características dos exemplos de treinamento.

Em um trabalho posterior (Pedersen, 2002b), o autor apresenta os resultados da avaliação de um trabalho de conjuntos (*ensembles*) de árvores de decisão para a desambiguação. Ele combina três classificadores simbólicos simples, baseados em árvore de decisão, utilizando o algoritmo C4.5. Cada classificador utiliza apenas uma característica: *bag-of-words*, bigramas ou co-ocorrências. O objetivo principal é verificar se essas três características são complementares (cada modelo classifica determinados casos de ambigüidade) ou redundantes (todos os modelos levam a resultados similares). Para tanto, é analisado o desempenho dos três classificadores em conjunto, individualmente e em diversas combinações de dois a dois.

Os testes são feitos com base nos conjuntos de teste do SENSEVAL (edições 1 e 2) e os resultados são comparados aos dos sistemas avaliados na competição. As combinações com três e dois classificadores ficam abaixo apenas de dois sistemas avaliados na competição, apresentando uma acurácia de 71.3% na primeira edição, 57.3% na segunda, ambas para a língua inglesa, e 61.2% para a segunda edição, considerando a língua espanhola.

Na avaliação dos classificadores individuais e das combinações de dois a dois, contudo, o autor concluiu que a acurácia do conjunto não é significativamente melhor que a acurácia dos classificadores individuais ou combinados de dois a dois. Considerando-se as avaliações individuais, o classificador baseado em co-ocorrências é o que apresenta a melhor acurácia (bastante próxima da acurácia do conjunto).

Os resultados da avaliação mostram, de modo geral que os três classificadores não são complementares e sim, na maioria das vezes, redundantes. Certamente, essa redundância ocorre principalmente entre os classificadores considerando bigramas e co-ocorrências, já essas características são muito similares. A combinação de classificadores com características mais distintas poderia indicar uma complementação maior entre eles.

Lee & Ng (2002) apresentam uma avaliação comparativa de quatro algoritmos de aprendizado para DLS e quatro tipos de conhecimento. O objetivo dos autores á analisar a contribuição de cada tipo de conhecimento nos diferentes algoritmos e a viabilidade do uso de um método de seleção automática de características para o treinamento.

Os tipos de conhecimento utilizados são: (a) características para a categoria gramatical da palavra ambígua e de três palavras vizinhas à esquerda e à direita; (b) características binárias para todas as palavras vizinhas à palavra ambígua na sentença, lematizadas, excluindo-se palavras de uma lista de *stop-words*; (c) características binárias para 11 *collocations*, por exemplo, a primeira palavra à direita e à esquerda da palavra ambígua; e (d) características para relações e informações sintáticas da sentença, geradas por um *parser* baseado em dependências. O número e a natureza das relações sintáticas depende da categoria gramatical da palavra ambígua. Por exemplo, para o substantivo são representados o seu núcleo na relação de dependência, a categoria gramatical do núcleo, a voz do núcleo, caso ele seja um verbo, e a posição relativa do núcleo.

Os algoritmos analisados são: SVM (*Support Vector Machines*), AdaBoost, Naive Bayes e C4.5. Todos foram usados nas versões implementadas no ambiente Weka[13], com seus parâmetros padrões.

[13] http://www.cs.waikato.ac.nz/~ml/weka/

Para todas as características binárias, em que o algoritmo simplesmente verifica a sua existência ou não no exemplo para gerar o modelo, os algoritmos foram testados com e sem o uso de um método para a seleção de características. Esse método consiste de um único parâmetro, configurado com um valor (três, neste caso), que determina a quantidade mínima de vezes que a característica ocorre nos exemplos de treinamento. Assim, na geração do modelo para cada palavra ambígua, para cada característica, o método verifica se ela ocorre para algum sentido da palavra pelo menos três vezes. Se isso acontecer, tal característica é utilizada.

Para avaliar os diferentes trabalhos, os autores utilizam os exemplos de treinamento e teste das duas primeiras edições do SENSEVAL para a tarefa de classificação de um conjunto de palavras. Na primeira edição, foram consideradas 36 palavras, na segunda, 73 (incluindo substantivos, verbos e adjetivos). Todos os algoritmos foram testados utilizando cada um dos tipos de conhecimento individualmente, com e sem o método de seleção de características e, por fim, com todos os tipos de conhecimento combinados. O uso do método de seleção de características permitiu resultados melhores, em todos os casos. A melhor acurácia para todos os classificadores foi obtida com a combinação de todos os tipos de conhecimento, com o método de seleção de características.

Dentre os classificadores, o SVM apresentou a maior acurácia (79.2% e 65.4%, nos dois exercícios). O autor também comparou seus resultados com os obtidos pelos três sistemas mais bem colocados nas duas edições do SENSEVAL. Em ambos os casos, a acurácia média do classificador SVM é maior que a dos três sistemas avaliados no SENSEVAL.

Por fim, a contribuição relativa de cada um dos tipos de conhecimento para a acurácia do trabalho depende do algoritmo utilizado. Por exemplo, as características de *collocations* são as que mais contribuem para o SVM, enquanto que as características das categorias gramaticais são as mais relevantes para o Naive Bayes.

Florian et al. (2002) descrevem várias possibilidades de combinação dos resultados de diferentes classificadores supervisionados em um único trabalho de DLS. Sua hipótese é de que classificadores com diferentes *bias* de aprendizado, diferentes métodos para a seleção de características e diferentes fontes de conhecimento levam a resultados distintos que, se combinados, podem melhorar o desempenho dos classificadores individuais.

Para verificar essa hipótese, primeiramente é analisado o desempenho individual de seis algoritmos de classificação: Naive Bayes e Cosine estendidos (com a atribuição de pesos às características), Bayes Ratio (Gale et al., 1992d), MMVC (*Mixture Maximum Variance Correction*) (Cucerzan & Yarowsky, 2002)[14], Listas de Decisão e TBL. Os quatro primeiros algoritmos são estatísticos, enquanto os dois últimos são simbólicos. Todos os algoritmos são utilizados considerando um conjunto variado de características locais e globais, incluindo lemas e etiquetas gramaticais de n-gramas (com duas e três palavras), *collocations*, relações sintáticas e *bag-of-words*.

Os classificadores foram individualmente treinados e testados com os exemplos de treinamento da tarefa de classificação de um conjunto de palavras do SENSEVAL-2, para quatro diferentes línguas. O foco dos autores era a análise de vários métodos para a combinação dos resultados de cada classificador individual para formar um único classificador. Foram discutidos vários métodos de combinação, variando desde a simples contagem dos votos de cada classificador até a contagem dos votos com pesos para cada classificador de acordo com seu desempenho. Esse último método levou ao melhor desempenho dos classificadores em conjunto e foi então escolhido para a combinação.

[14] Cucerzan, S.; Yarowsky, D. (2002). Augnented Mixture Models for Lexical Disambiguation. In *Proceedings of the EMNLP-2002*, pp. 33-40. Apud Florian et al. (2002).

A acurácia de cada classificador individual, em cada uma das línguas, foi comparada com a acurácia obtida com a combinação de todos os classificadores. Em todos os casos, a acurácia da combinação (71.82%) superou a acurácia individual (68.9% para o melhor classificador). Essa acurácia também superou a obtida pelo sistema de melhor desempenho no SENSEVAL-2 (70.3%), considerando o mesmo conjunto de teste.

Para analisar a contribuição de cada classificador para a combinação, cada um deles foi excluído da combinação, e a acurácia foi então verificada. Os classificadores que mais influenciaram positivamente na combinação foram os simbólicos, primeiramente, o TBL e, na seqüência, o baseado em listas de decisão.

Park et al. (2003) descrevem um trabalho de DLS da língua coreana, baseado na noção de amostragem seletiva dos exemplos, similar ao aprendizado semi-supervisionado (de árvores de decisão), para aumentar o número de exemplos etiquetados. As características usadas incluem a categoria gramatical e a função sintática da palavra ambígua, algumas palavras vizinhas e suas relações sintáticas com a palavra ambígua e a existência ou não de algumas relações sintáticas na sentença.

São empregados 15 classificadores, sendo que cada classificador utiliza as mesmas características, contudo, com subconjuntos distintos de exemplos de treinamento, gerados por uma técnica de amostragem randômica com reposição e repetição. O sentido aprendido é o indicado por pelo menos 10 dos classificadores. Sempre que um exemplo for rotulado, ele é adicionado ao conjunto de treinamento. Esse processo é repetido, iterativamente, até que todos os exemplos sejam etiquetados.

Os autores avaliaram seu trabalho considerando quatro substantivos ambíguos, com exemplos selecionados em um córpus de um milhão de palavras. Para cada palavra foram selecionados entre 350 e 876 exemplos, os quais foram então anotados manualmente com o sentido da palavra ambígua. A acurácia relatada para o classificador final, com todos os exemplos rotulados, é de 87%.

Dihn et al. (2003) descrevem um sistema de TA do inglês para o vietnamita, desenvolvido de acordo com um método híbrido: parte do sistema é constituída de regras manualmente criadas e outra parte, de regras aprendidas a partir de córpus, com base no aprendizado baseado em transformações. Esse sistema possui módulos específicos para cada tipo de ambigüidade, incluindo um módulo para a DLS. As regras desse módulo são geradas por um modelo baseado em córpus.

O córpus de exemplos é criado a partir de textos paralelos entre as duas línguas, de diversos gêneros e domínios, por meio do alinhamento automático das palavras, revisado manualmente. As características para o aprendizado consistem de n-gramas (de uma a quatro palavras), etiquetas gramaticais e funções sintáticas. Além disso, o algoritmo considera as etiquetas já atribuídas às palavras vizinhas na sentença, ou seja, as palavras já traduzidas. Isso é possível porque no córpus de exemplos, todas as palavras estão etiquetadas com a tradução correspondente. Os autores não avaliam os módulos individuais do sistema, tampouco a influência desses módulos no desempenho geral do sistema de TA.

Mihalcea (2004) analisa as vantagens da utilização de córpus semi-supervisionado para o aprendizado supervisionado estatístico. Como nos trabalhos de Hearst (1991) e Yarowsky (1995), por exemplo, esse trabalho permite a construção de classificadores utilizando apenas alguns exemplos etiquetados e uma grande quantidade de exemplos não etiquetados.

Duas técnicas de aprendizado são analisadas, co-treinamento (*co-training*) e autotreinamento (*self-training*), ambas visando aumentar o número de exemplos

etiquetados por meio de exemplos não etiquetados para melhorar o desempenho do aprendizado supervisionado. No co-treinamento, são gerados vários classificadores treinados com base em diferentes "visões" que, neste caso, consistem de diferentes conjuntos de características dos exemplos etiquetados. No autotreinamento, é gerado apenas um classificador, considerando todas as características. Em ambos os métodos, a cada iteração, os exemplos que são etiquetados com um alto grau de confiabilidade são adicionados aos exemplos de treinamento.

Para avaliar o co-treinamento, são usados dois subconjuntos de características, um para cada um dos classificadores gerados. O primeiro conjunto é constituído de características locais, incluindo as palavras vizinhas na sentença, a categoria gramatical da palavra ambígua das palavras vizinhas, relações sintáticas e quatro *collocations*. O segundo conjunto é constituído de características globais: 10 palavras co-ocorrendo pelo menos três vezes com a palavra ambígua. Para avaliar o autotreinamento, são usadas todas as características. Os classificadores, em ambos os casos, são construídos utilizando o modelo probabilístico Naive Bayes.

O autor avalia algumas variações de ambas as técnicas de aprendizado considerando os conjuntos de treinamento e teste do SENSESAL-2 e um subconjunto de exemplos não etiquetados extraídos do córpus BNC. Apenas a desambiguação de substantivos é testada. Para tanto, são definidas duas configurações de teste: a configuração ideal dos diferentes parâmetros do modelo e uma definição empírica dessa configuração. Esses parâmetros consistem do número de repetições do processo de etiquetação iterativa, do número de exemplos não rotulados selecionados a cada iteração e da quantidade dos exemplos cuja etiquetação atingiu um alto nível de confiabilidade para ser inserida no conjunto de exemplos de treinamento. Com os parâmetros ideais, obtidos a partir de medições nos exemplos de teste, tanto para a técnica de co-treinamento como para a de autotreinamento, o autor reporta uma diminuição de 25.5% no erro da desambiguação. Já com os parâmetros empiricamente definidos, a redução de erro máxima é de 9.8%, alcançada por uma abordagem de *bagging* de treinamento cooperativo.

Como a configuração ideal dificilmente pode ser identificada em aplicações reais, o autor sugere que o trabalho que utiliza um *bagging* de classificadores empregando co-treinamento para suavizar as curvas de aprendizado é o mais indicado.

Os trabalhos baseados em córpus descritos acima são listados na Tabela 4, de acordo com o modo e o paradigma de aprendizado, o conjunto de palavras para as quais forma criados e/ou testados, o nível de refinamento das distinções entre os sentidos e a acurácia apresentada. As informações não explicitadas nas publicações referentes a tais trabalhos são representadas por "?". Novamente, a medida de desempenho pode ser na forma de precisão e cobertura.

Tabela 4. Lista dos trabalhos de DLS baseados em córpus

Trabalho	**Modo**	**Paradigma**	**Conj. palavras**	**Nível de refinamento dos sentidos**	**Acurácia**
Brown et al. (1991)	não-supervisionado	-	500 palavras do inglês e 200 do francês	?	diminuição de 13% no erro do sistema de TA
(Schütze, 1992; 1998)	não-supervisionado	-	10 palavras	baixo	92%
(Schütze & Pedersen, 1995)	não-supervisionado	-	?	baixo	aumento de 14.4% na precisão do

					sistema de Recuperação de Informações
(Stokoe et al., 2003)	não-supervisionado	-	?	?	aumento de 16.2% na precisão do sistema de Recuperação de Informações
(Pedersen & Bruce, 1997)	não-supervisionado	-	13 palavras	?	65 a 66%
(Pedersen & Bruce, 1998)	não-supervisionado	-	13 palavras	?	64 a 72%
(Pantel & Lin, 2002)	não-supervisionado	-	13.403 palavras	alto	60.8% (precisão), 50.8% (cobertura)
(Dini et al., 1998)	não-supervisionado	-	?	baixo	?
(Rapp, 2004)	não-supervisionado	-	1 palavra	?	?
(Black, 1988)	supervisionado	simbólico	5 palavras	?	75%
(Hearst, 1991)	semi-supervisionado	simbólico	4 substantivos	baixo	83%
(Yarowsky, 1995)	semi-supervisionado	simbólico	12 palavras	baixo	96%
(Yarowsky, 2000)	supervisionado	simbólico	palavras do Senseval-2	alto	78.9% (precisão)
(Mooney, 1996)	supervisionado	comparação entre simbólico, estatístico, conexionista e baseado em instâncias	1 palavra	alto	algoritmo estatístico - melhor desempenho
(Ng & Lee, 1996)	supervisionado	baseado em instâncias	1 palavra (*interest*) e palavras do DSO	alto	87.4% (*interest*) (precisão), 68.6% (DSO)
(Ng, 1997a)	supervisionado	baseado em instâncias	palavras do DSO	alto	75.2%
(Towell & Voorhess, 1998)	supervisionado	conexionista	3 palavras	?	86%
(Pedersen, 2000)	supervisionado	estatístico	2 palavras	alto	88 a 89%
(Zinovjeva, 2000)	supervisionado	simbólico	3 palavras	alto	73.1 a 95.4%
(Agirre & Martinez, 2000)	supervisionado	simbólico	palavras do SEMCOR e DSO	alto e baixo	70%: distinções refinadas, 83%: distinções pouco refinadas (precisão), 90% (cobertura)
(Escudero et al., 2000a)	supervisionado	estatístico e baseado em	15 verbos e substantivos	alto	algoritmo baseado em

		instâncias			instâncias - melhor desempenho: 66%
(Escudero et al., 2000b)	supervisionado	simbólico	15 verbos e substantivos	alto	68 a 69.5%
(Escudero et al., 2000c)	supervisionado	comparação entre simbólicos, estatísticos, baseados em instâncias, etc.	21 verbos e substantivos do DSO	alto	algoritmo simbólico - melhor desempenho: 62%
(Escudero et al., 2001)	supervisionado	simbólico	palavras do Senseval-2	alto e baixo	59.4%: distinções refinadas, 67.1%: distinções pouco refinadas
(Pedersen, 2002a)	supervisionado	estatístico e simbólico	palavras do Senseval-2	alto	54% para o espanhol e 60% para o inglês
(Pedersen, 2002b)	supervisionado	simbólico	palavras do Senseval-1 e 2	alto	inglês: 71.3% no Senseval-1, 57.3% no Senseval-2. espanhol: 61.2%
(Lee & Ng, 2002)	supervisionado	comparação entre estatístico, simbólico, etc.	palavras do Senseval-1 e 2	alto	algoritmo SVM - melhor desempenho: 79.2% no Senseval-1 e 65.4% no Senseval-2
(Florian et al., 2002)	supervisionado	combinação de estatísticos, simbólicos, etc.	palavras do Senseval-2	alto	71.82% - combinação
(Park et al., 2003)	semi-supervisionado	simbólico	4 substantivos	?	87%
(Dihn et al., 2003)	supervisionado	simbólico	?	?	?
(Mihalcea, 2004)	semi-supervisionado	estatístico	substantivos do Senseval-2	?	diminuição de 9.8% no erro de DLS

3.3 Método híbrido

Os trabalhos descritos a seguir incluem diferentes combinações consistindo do uso de conhecimento codificado manualmente ou proveniente de qualquer recurso lingüístico já existente e do uso de córpus em alguma das variações no modo de aprendizado (supervisionado, não-supervisionado ou semi-supervisionado), nos paradigmas e nos algoritmos.

Dagan et al. (1991) e Dagan & Itai (1994) propõem um mecanismo para a desambiguação monolingüe em uma língua (L1), baseada em córpus comparáveis entre L1 e uma outra língua (L2) e nas correspondências lexicais entre essas duas línguas indicadas por um dicionário bilíngüe. A suposição dos autores é de que os diferentes sentidos de uma palavra ambígua em uma língua se manifestam por diferentes itens lexicais em outras línguas. A idéia, portanto, é desambiguar os sentidos das palavras de L1, usando as suas traduções em L2. Para tanto, inicialmente, o mecanismo procura identificar a tradução correta das palavras de conteúdo de uma sentença em L1 na L2. Na nomenclatura adotada pelos autores, esse processo é denominado "seleção da palavra-alvo".

Para a seleção da palavra-alvo, o mecanismo se divide em duas fases: fase lingüística e fase estatística. Na fase lingüística, primeiramente são identificadas todas as possíveis traduções dessa palavra na L2 disponíveis em um dicionário bilíngüe L1-L2. Em seguida, é realizada a análise sintática superficial da sentença na qual a palavra ambígua ocorre, de modo a identificar relações sintáticas superficiais binárias entre as palavras. As relações sintáticas consideradas importantes para as línguas utilizadas são: relações entre o verbo e seu sujeito, seus complementos e adjuntos; relações entre o substantivo e seus complementos e adjuntos e relações entre os adjetivos e advérbios e seus modificadores.

A partir do córpus de L2, são coletados exemplos de sentenças que contêm as possíveis traduções da palavra ambígua. Todos esses exemplos são então submetidos ao procedimento de análise sintática superficial, de modo que sejam identificadas suas relações sintáticas superficiais. Idealmente, deve ser empregado um analisador sintático similar, para que sejam extraídas relações sintáticas que possam ser comparadas às de L1. Caso isso não seja possível, um mecanismo para mapear relações diferentes, mas equivalentes, produzidas por analisadores diferentes, precisa ser criado.

As relações são utilizadas como filtro no processo de seleção da palavra-alvo, pois são descartados todos os exemplos coletados de L2 que ocorrem em relações sintáticas diferentes daquele em que a palavra ambígua é usada na sentença em L1. Como resultado dessa fase, obtém-se um subconjunto de ocorrências das possíveis traduções da palavra ambígua na L2. Se essa palavra só ocorre com uma tradução em L2, nas várias relações sintáticas da L1 em que ela aparece, então essa tradução é automaticamente escolhida. Caso contrário, é preciso recorrer à fase estatística.

Na fase estatística, um procedimento inicialmente conta o número de ocorrências de cada uma das possíveis traduções da palavra ambígua na L2 coletadas na fase lingüística. Esse número é utilizado como parâmetro inicial para um modelo probabilístico simples que se baseia, principalmente, na freqüência de ocorrências de cada uma das possíveis traduções, restritas pela relação sintática, conforme mencionado. A seleção de uma palavra-alvo para cada palavra ambígua leva em consideração, também, a ocorrência dessa palavra ambígua em mais de uma relação sintática (por exemplo, um verbo irá ocorrer na relação com o seu sujeito e com os seus objetos). A escolha da tradução deve ser consistente com todas as relações em que a palavra ocorre. Além disso, diferentemente da maioria dos outros trabalhos, a seleção para uma palavra ambígua considera também as demais palavras na sentença já desambiguadas, como um processo de propagação de restrições.

Para avaliar seu trabalho, os autores realizam um experimento considerando 103 palavras ambíguas do hebreu e 54 palavras ambíguas do alemão como L1 e o inglês como L2. Nessa simulação, várias etapas do mecanismo de escolha são realizadas manualmente. Por exemplo, são eliminadas as possíveis traduções que ocorrem em relações sintáticas distintas. Essa simplificação, segundo os autores, foi necessária porque eles não dispunham de processos e recursos lingüísticos eletrônicos apropriados, como um *parser* para o hebreu.

O córpus do inglês utilizado para a análise de freqüências possui 25 milhões de palavras. As escolhas do mecanismo foram comparadas às escolhas indicadas por um tradutor humano, medindo-se a cobertura e a precisão do modelo. Para o hebreu, os resultados foram: cobertura = 68% e precisão = 91%. Para o alemão, cobertura = 50% e precisão = 78%. Segundo os autores, a cobertura do mecanismo proposto é baixa principalmente pela insuficiência de exemplos no córpus monolingüe. Já com relação à precisão, os autores reconhecem que um dos problemas é a falta do emprego de conhecimento mais profundo no processo de escolha.

O trabalho dos autores representa o primeiro esforço na utilização de recursos multilingües para a desambiguação e apresenta como grande vantagem o fato de não ser necessário um córpus etiquetado com sentidos, tampouco córpus paralelos corretamente alinhados para a o treinamento do modelo. Contudo, ele apresenta vários problemas. Primeiramente, o filtro das relações sintáticas exige que as duas línguas sejam sintaticamente similares, de modo que possam ser analisadas pelo mesmo *parser* ou, pelo menos, que as relações de uma língua possam ser mapeadas nas relações da outra. Os autores reconhecem, também, que pode haver ambigüidade entre as relações: uma relação de uma língua ter mais de uma equivalente em outra língua. Contudo, eles não apontam para estratégias que possam contornar esse problema. Além disso, o fato de terem realizado parte do processo manualmente pode ter deixado de revelar outros problemas.

Por fim, apesar de os autores afirmarem que seu objetivo é a desambiguação monolingüe, em L1, seus trabalhos indicam apenas parte desse processo, que corresponde à desambiguação multilingüe, na tradução de L1 para L2. O mecanismo proposto, segundo os próprios autores, tem, de fato, aplicação direta para a desambiguação na TA. Outros autores, como Ng & Zelle (1997), por exemplo, classificam o trabalho como multilingüe.

Com relação ao restante do processo, para a desambiguação monolingüe, Dagan et al. apenas indicam que seria necessário um dicionário bilíngüe do tipo L1-L1 → L2, ou seja, que fornecesse, para cada entrada da L1, sua definição (sentido) em L1 e sua tradução para L2. Dessa maneira, poderia ser identificado qual sentido de L1 corresponde à tradução indicada pelo mecanismo em L2. Contudo, esse processo pode não ser direto, uma vez também pode haver ambigüidade na verificação dessa correspondência. Outro problema não analisado nos dois trabalhos diz respeito às ambigüidades que se mantêm entre as línguas. Para esses casos, os autores apenas mencionam que seria necessário os uso de uma terceira língua para a desambiguação monolingüe.

Yarowsky (1992) apresenta um trabalho não-supervisionado para a desambiguação de sentidos em textos irrestritos utilizando um modelo estatístico sobre as categorias mais genéricas do *thesaurus* Roget, localizando o contexto da palavra ambígua na descrição dessas categorias no *thesaurus*.

A noção de sentido no trabalho de Yarowsky corresponde às categorias genéricas do *thesaurus*. Assim, um sentido é atribuído a uma palavra ambígua a partir da identificação de qual a categoria na qual ela se enquadra. Essa identificação se dá pela análise do contexto da palavra ambígua, seguida da sua comparação com palavras indicativas de cada uma das categorias. Para levantar um contexto de palavras indicativas de cada categoria do *thesaurus*, são extraídos de um córpus todos os possíveis exemplos em que pelo menos uma palavra (seu lema) da categoria aparece, juntamente com uma janela 100 palavras (50 à direita e 50 à esquerda), denominada “assinatura” dessa palavra. Nessa busca, somente as palavras da categoria gramatical adequada são consideradas (um etiquetador morfossintático é utilizado no pré-processamento). Os exemplos coletados podem incluir palavras ambíguas, que pertencem a outras categorias. Para minimizar a

possível influência dessa ambigüidade, o peso de cada ocorrência da palavra buscada (inicialmente, 1) é dividido pelo número de vezes que ela aparece em todos os exemplos.

A partir dos exemplos, é utilizado um algoritmo que estima a informação mútua para identificar as palavras mais comuns em cada categoria e, portanto, mais indicativas dessa categoria. Esse algoritmo simplesmente divide a probabilidade da palavra aparecer na categoria para a qual foi buscada pela probabilidade de ocorrer em todas as categorias. Com isso, obtém-se uma medida da "importância" da palavra para identificar uma categoria do *thesaurus*. Como resultado, a lista de palavras importantes ou indicativas de cada categoria inclui, em média, 3.000 palavras que têm maior probabilidade de co-ocorrer com os membros da categoria. Segundo o autor, essa quantidade é muito mais significativa que as pequenas definições de palavras disponibilizadas nos dicionários.

Para desambiguar uma nova palavra ambígua, o sistema verifica o seu contexto de 100 palavras. De acordo com um modelo bayesiano, é identificada a categoria cuja classe de palavras indicativas contém o maior número de palavras coincidentes e essa é escolhida como a categoria e, portanto, o sentido da palavra.

Yarowsky realiza testes com 12 substantivos ambíguos e reporta uma acurácia média de 92%, sendo que as distinções envolvem, em média, três sentidos. O autor afirma que seu trabalho é mais adequado à tarefa de extração de informações sobre um contexto global e, portanto, à desambiguação de substantivos. Além disso, as categorias genéricas do *thesaurus* utilizadas correspondem a sentidos muito distintos.

Um problema do trabalho de Yarowsky é que o algoritmo falha quando um sentido se distribui por várias categorias, ou seja, quando uma distinção de sentido deve ser realizada independentemente da categoria. Este é o caso, por exemplo, do sentido "vantagem" da palavra *interest* (como em *self-interest*). Esse sentido pode ocorrer independentemente da área (finanças, música, etc.).

Resnik (1997) propõe um algoritmo não-supervisionado para desambiguar os sentidos de substantivos, com base em preferências de seleção e em uma hierarquia conceitual (como a da WordNet). Segundo o autor, o conceito tradicional de preferências de seleção, em que as preferências são definidas por traços booleanos (+humano ou -humano, por exemplo), limita a abrangência dos trabalhos, já que tais traços precisam ser manualmente codificados. No trabalho proposto, esse problema é eliminado, pois as preferências de seleção são automaticamente obtidas a partir de um córpus, de medidas estatísticas e de informações da hierarquia conceitual.

Um modelo probabilístico captura co-ocorrências de relações sintáticas no córpus e sentidos da WordNet. Para tanto, o córpus usado para o treinamento do sistema é previamente analisado por um *parser*, que extrai as relações sintáticas superficiais envolvendo substantivos, como sujeito-verbo, verbo-objeto, núcleo-modificador, etc. Uma distribuição estimada a priori captura a probabilidade de um sentido ocorrer como argumento de uma relação predicado-argumento, independentemente do predicado. Para estimar as probabilidades a priori, é utilizada uma metodologia que identifica os sentidos das palavras do córpus, definida pelo autor em outros trabalhos (Resnik, 1995a). Quando o predicado passa a ser conhecido, essa probabilidade pode mudar. A diferença entre essa probabilidade condicional e a probabilidade a priori determina a preferência de seleção.

Um modelo da Teoria da Informação, baseado em entropia, quantifica essa diferença definindo o peso de uma preferência de seleção de um predicado. Intuitivamente, esse peso representa a medida da quantidade de informação que o predicado fornece sobre o sentido do seu argumento. Quanto maior a probabilidade do sentido ocorrer independentemente do predicado, menor é a força do predicado como preferência de

seleção. Assim, os elementos que co-ocorrem com o substantivo nas relações sintáticas são usados como preferências de seleção.

Para caracterizar a adequação de um sentido como argumento para um predicado, é analisada a sua contribuição relativa ao peso global da preferência de seleção. Sentidos adequados têm uma probabilidade condicional maior que a probabilidade a priori.

O autor realiza experimentos de avaliação, considerando como córpus de treinamento um subconjunto do córpus Brown, com as estruturas sintáticas geradas manualmente. Uma parte desse córpus é manualmente etiquetada com os sentidos da WordNet para a comparação com a etiquetação automática. Várias relações sintáticas são testadas. O autor relata uma acurácia média entre todos os tipos de relação de 40%. Esse resultado é inferior ao de outros trabalhos não-supervisionados, mas vale notar que esse trabalho realiza a desambiguação considerando distinções refinadas.

Lee (1997) apresenta um trabalho de DLS para um sistema de TA do inglês para o coreano, que segue o método direto de tradução por palavras, empregando técnicas estatísticas para a seleção lexical, incluindo a DLS, e a re-ordenação das palavras na língua-alvo.

O trabalho de TA utiliza córpus paralelos entre as duas línguas e dicionários bilíngües. Para a DLS, com base nos documentos paralelos, é criado um dicionário de tradução para cada palavra da língua-fonte, que consiste de todas as suas possíveis traduções na língua-alvo, extraídas do córpus. A partir desses dicionários, o problema de DLS é estruturado como um problema de classificação. Para tanto, são usadas como características co-ocorrências chamadas pelo autor de "*collocations* esparsas". Essas co-ocorrências correspondem a todas as combinações de palavras (tomadas de duas a duas) na sentença a ser traduzida. O algoritmo de aprendizado supervisionado empregado é o SNoW, que aprende, como modelo, uma rede de funções lineares com regras de atualização.

Como córpus paralelo, é utilizado um conjunto de 689 documentos (17.846 sentenças) manualmente traduzidos do inglês para o coreano e manualmente alinhados por palavras. Assim, os dicionários de tradução são diretamente extraídos desses documentos. Para avaliar o trabalho, são selecionados exemplos de 121 substantivos ambíguos que possuem mais de 50 exemplos no córpus. Os resultados da classificação foram comparados à *baseline* da escolha pelo sentido mais freqüente e à classificação utilizando um algoritmo Naive Bayes. O classificador gerado pelo algoritmo SNoW apresenta uma precisão média de 57.46%, superior à *baseline* (53.87%) e à precisão do classificador Naive Bayes (47.49%).

O autor também avalia seu trabalho considerando o conjunto de teste da tarefa de desambiguação de um conjunto de palavras do SENSEVAL-2, para o inglês e para o coreano. Para tanto, a tarefa de DLS do seu sistema, que é multilingüe, é mapeada (manualmente) para os resultados fornecidos pelo SENSEVAL, que são monolingües. Além disso, o teste é realizado somente com as palavras do conjunto do SENSEVAL coincidentes com as palavras usadas para o treinamento do sistema. Os resultados obtidos são muito próximos aos da *baseline* do sentido mais freqüente, mas estão bem abaixo dos obtidos pelos sistemas mais bem colocados no exercício. Esse desempenho baixo, segundo o autor, se deve ao uso apenas das palavras vizinhas como características. O uso apenas de *collocations* é justificado pelo autor em função do seu objetivo obter um modelo portável, que possa ser diretamente utilizado em outras línguas.

Uma questão interessante citada pelo autor, ao tentar relacionar as possíveis traduções das palavras coincidentes no seu córpus com aquelas do conjunto de testes do SENSEVAL, é que não há correlação direta entre o número de sentidos monolingües e o

número de possíveis traduções para outra língua, pelo menos entre as duas línguas em questão. Por exemplo, a palavra do inglês *day* apresenta 7 possíveis sentidos, de acordo com os dados do SENSEVAL, mas possui 12 possíveis traduções para o coreano.

Karov & Edelman (1998) propõem um trabalho híbrido baseado em similaridades de palavras e contextos (sentenças), que também pode ser considerado um trabalho de aprendizado semi-supervisionado. Essencialmente, duas palavras são similares se aparecem em contextos similares; contextos são similares se possuem palavras similares. Contudo, como o trabalho se baseia em definições de dicionário, isso pode levar a uma medida transitiva de similaridade, na qual dois contextos são considerados similares mesmo que não compartilhem as mesmas palavras, e duas palavras são consideradas similares mesmo que não compartilhem palavras vizinhas similares.

O trabalho utiliza conhecimento codificado em um dicionário eletrônico (alternativamente, poderia ser em um *thesaurus*) e um algoritmo de aprendizado a partir de córpus. Inicialmente, é empregado um córpus não-supervisionado, no qual as palavras ambíguas vão sendo automaticamente etiquetadas, em um processo iterativo, a partir da análise dos exemplos que não contêm essas palavras, mas contêm palavras relacionadas extraídas das suas definições no dicionário. O sistema aprende, a partir do córpus, um conjunto de usos típicos de cada sentido da palavra ambígua, listada no dicionário eletrônico. A um novo caso de uma palavra ambígua é atribuído o sentido associado ao uso típico mais similar ao seu contexto.

A partir dos exemplos de treinamento, o cálculo da similaridade é feito com base nas semelhanças do uso as palavras ambíguas nas sentenças, e não com base no seu significado. Assim, são analisados, basicamente, o contexto de ocorrência das palavras e as relações de distância nesse contexto. Nesse trabalho, palavras como *doctor* e *health* podem ser consideradas similares, pois tendem a aparecer nos mesmos contextos, apesar de elas apresentarem categorias gramaticais distintas e não serem consideradas semanticamente próximas em hierarquias como a da WordNet.

Além de eliminar a necessidade de etiquetação manual prévia dos exemplos de treinamento, o trabalho minimiza o problema de dados esparsos, ou seja, da inexistência de exemplos (ou existência de poucos exemplos) de cada sentido de uma palavra. Isso é feito considerando-se como exemplos para uma palavra, além dos existentes no córpus, aqueles provenientes dos exemplos nas definições para essa palavra no dicionário. Esses exemplos passam, então, a fazer parte do córpus de treinamento. Além disso, diferentemente dos trabalhos que se baseiam apenas na quantidade de co-ocorrências por sentido, todas as sentenças são consideradas, de modo que não deixam de ser analisados exemplos de sentidos pouco freqüentes. Segundo Karov & Edelman, na sua avaliação, esse trabalho leva a resultados melhores que os dos trabalhos baseados apenas em co-ocorrência, principalmente quando os dados são esparsos. Contudo, os autores não apresentam resultados numéricos dessa avaliação.

Os trabalhos de Wilks & Stevenson (1998) e Stevenson & Wilks (1999; 2000; 2001) representam aperfeiçoamentos de seus trabalhos anteriores já citados (Wilks & Stevenson, 1997a; 1997b) (Seção 3.1.2). Esses trabalhos, assim como os de McRoy (1992) e Ng & Lee (1996), entre outros, apesar de não serem específicos para a TA, são de grande importância para o modelo que se pretende propor, uma vez que utilizam diversas fontes de conhecimento. Além disso, eles utilizam diferentes técnicas de AM para combinar essas fontes de conhecimento e visam realizar a desambiguação em larga escala.

Além das informações anteriormente consideradas, os autores empregam outros tipos de conhecimento, provenientes, principalmente, do dicionário eletrônico LDOCE.

Como principal diferencial dos outros trabalhos, utilizam uma técnica baseada em córpus para combinar os diferentes tipos de conhecimento. Como nos seus trabalhos anteriores, a desambiguação é realizada com base nos sentidos de homógrafos do LDOCE e engloba todas as palavras de conteúdo de um texto. Contudo, nos trabalhos mais recentes, os autores focalizam a DLS, especificamente, e não mais a tarefa etiquetação de sentidos. Os resultados também são analisados para a desambiguação tanto de homógrafos quanto de sentidos mais refinados.

Além de uma fase de pré-processamento, todos os trabalhos dos autores empregam conhecimento pré-codificado em um filtro, que elimina alguns dos possíveis sentidos de cada palavra ambígua, e em diversos etiquetadores parciais, que sugerem alguns sentidos para tal palavra. Esses processos são realizados isoladamente, de maneira independente. O resultado de todos os processos, agrupado, é um conjunto reduzido de possíveis sentidos para cada palavra ambígua, incluindo todos os sugeridos pelos etiquetadores parciais, exceto os eliminados pelos filtros, em um determinado contexto. Esse contexto é indicado por 10 palavras na vizinhança da palavra ambígua na sentença, em determinadas posições relativas e com determinadas etiquetas gramaticais, ou seja, 10 *collocations*. Em seguida, é utilizado um algoritmo de AM para classificar cada sentido desse conjunto como "apropriado" ou "não-apropriado".

Dentre os trabalhos de Stevenson & Wilks citados, todos empregam os mesmos tipos de conhecimento e fontes de informação, usando praticamente os mesmos filtros e etiquetadores. As variações desses trabalhos dizem respeito principalmente ao algoritmo de aprendizado empregado e à configuração dos testes realizados. A descrição a seguir é baseada principalmente em Stevenson & Wilks (2000), sendo que os demais trabalhos são bastante similares.

Com relação às fontes de informação e aos tipos de conhecimento, os autores empregam: (a) a identificação de nomes de entidades para excluir nomes próprios da desambiguação; (b) as categorias gramaticais, identificadas por um etiquetador morfossintático, para eliminar os sentidos de palavras de classes gramaticais diferentes da classe da palavra ambígua; (c) a verificação da sobreposição entre as definições de dicionário dos sentidos da palavra ambígua e as definições dos sentidos das palavras vizinhas na sentença; (d) a verificação da sobreposição entre os códigos de área (do dicionário) dos diversos sentidos da palavra e dos sentidos das palavras vizinhas à palavra a ser desambiguada; (e) restrições de seleção, com base nas relações identificadas por um *parser* superficial e nos traços semânticos e restrições estabelecidas no dicionário; e (f) 10 *collocations* para indicar o contexto de ocorrência da palavra ambígua (primeiro substantivo, primeiro verbo e primeira preposição da direita e da esquerda, e primeira e segunda palavras da direita e da esquerda).

No pré-processamento dos textos a serem desambiguados, além das funções básicas como a separação do texto em palavras e sentenças e a lematização, é realizada a etiquetação morfossintática das palavras, por meio do etiquetador Brill (Brill, 1992), e a identificação de nomes de entidades, por meio do sistema LaSIE (Gaizauskas et al., 1996). Esse último recurso é empregado para identificar, neste caso, nomes próprios (pessoas, lugares, etc.), que não precisarão, portanto, ser desambiguados. Além disso, essa informação será utilizada em um dos etiquetadores parciais.

O primeiro processo aplicado sobre os dados pré-processados é um filtro baseado na etiqueta gramatical das palavras. As etiquetas identificadas pelo etiquetador Brill (48 etiquetas) são, primeiramente, manualmente mapeadas em etiquetas do LDOCE, que são mais genéricas (17 etiquetas). Todos os sentidos cuja categoria gramatical não corresponde à categoria retornada para a palavra na sentença são, então, eliminados. Para minimizar a influência de erros de etiquetação gramatical, nos casos mais graves, nos quais nenhum dos

possíveis sentidos de uma palavra tem a etiqueta gramatical associada a ela na sentença, todas as etiquetas possíveis são mantidas.

O segundo processo (*Tagger 1*) é um etiquetador parcial baseado na identificação de sobreposições nas definições textuais do dicionário LDOCE dos sentidos da palavra a ser desambiguada e dos sentidos das suas palavras vizinhas na sentença, conforme mencionado anteriormente. Para minimizar o esforço computacional exigido para testar as várias combinações entre as definições, os autores utilizam o algoritmo de otimização de Cowie et al. (1992), que elimina a necessidade de testar todas as combinações de sentidos.

O terceiro processo (*Tagger 2*) utiliza códigos pragmáticos do LDOCE, também chamados de "códigos de área". Conforme mencionado, esse dicionário contém uma hierarquia de códigos pragmáticos, os quais indicam a área dos sentidos das palavras. Diferentemente de Guthrie et al (1991), Stevenson & Wilks utilizam ambos os níveis dos códigos de área. O processo consiste da identificação da sobreposição dos códigos de área (nos dois níveis) dos possíveis sentidos da palavra a ser desambiguada com os códigos de área das suas palavras vizinhas. A vizinhança considerada inclui todo o parágrafo, e não apenas a sentença, como na sobreposição de definições (*Tagger 1*). Também é utilizado, aqui, o algoritmo de otimização para minimizar o número de testes de combinações.

O quarto processo (*Tagger 3*) utiliza restrições de seleção, com base em informações também contidas no LDOCE. Cada sentido de uma palavra de conteúdo apresenta, nesse dicionário, traços semânticos e/ou restrições de seleção simples, como H (humano), M (humano masculino), P (planta) e S (sólido). Ao todo, são usadas 35 classes semânticas no dicionário. Os sentidos de um substantivo apresentam um subconjunto desses traços. Os sentidos de advérbios, adjetivos e verbos, por sua vez, apresentam a lista dos traços que eles exigem nos substantivos que eles modificam ou que os complementam. Para utilizar as restrições é necessário, portanto, identificar relações sintáticas superficiais entre verbos, adjetivos e advérbios e os substantivos que são o núcleo dos seus argumentos. Isso é feito por meio das relações providas por um analisador sintático superficial.

Além das classes semânticas, o processo de resolução das restrições de seleção faz uso das informações provenientes do identificador de nomes de entidades, realizado no pré-processamento. Elas podem ajudar a desambiguar as outras palavras na sentença. Por exemplo, se um dos possíveis sentidos de verbo ambíguo exige um objeto com o traço "humano" e o objeto, na sentença, é um nome próprio de pessoa, o sistema seleciona esse sentido como mais apropriado.

O processo seguinte também é um etiquetador parcial (*Tagger 4*), o último do sistema. Esse etiquetador considera um contexto mais amplo para desambiguar cada palavra. Para tanto, são selecionadas 50 palavras de cada lado da palavra ambígua e é utilizado um modelo estatístico dos códigos de área primários dessas palavras no LDOCE. Esse modelo estatístico é similar ao modelo bayesiano de probabilidades e visa estimar a probabilidade de um determinado código de área predominar, num determinado contexto, dados os códigos de área de todas as palavras nesse contexto (que consiste das 100 palavras vizinhas da palavra ambígua). Definido o código de área predominante para cada palavra ambígua, são selecionados como candidatos todos os sentidos que apresentam, no LDOCE, aquele código.

Como resultado, são obtidos os diferentes conjuntos de sentidos "sugeridos" pelo filtro e pelos etiquetadores parciais para cada palavra, juntamente com exemplos do seu uso em um contexto formado pelas 10 *collocations*. Com relação aos algoritmos de aprendizado empregados nos diferentes trabalhos, todos são supervisionados, mas variam no que diz respeito ao paradigma de aprendizado: simbólico (listas de decisão) e baseado em instâncias.

Em Wilks & Stevenson (1998), os autores empregam listas de decisão. Em testes realizados com o córpus SEMCOR, os autores relatam uma acurácia de 83.4% na identificação dos sentidos para todas as palavras ambíguas. O nível de refinamento dos sentidos é apresentado pelo LDOCE, que não é tão refinado quanto o nível da WordNet, mas não chega ao nível da homografia. Os autores apresentam também os resultados da mesma avaliação combinando todos os filtros e etiquetadores parciais por meio de um sistema de votação simples, sem a utilização do algoritmo de aprendizado de máquina. O teste realizado resultou em 59% de acurácia. Com isso, eles concluem que há um benefício considerável na utilização do aprendizado de máquina na tentativa de otimizar a combinação entre as diversas evidências de desambiguação.

Em Stevenson & Wilks (1999; 2000; 2001), os autores empregam um algoritmo baseado em instâncias, denominado TiMBL (Daelemans et al., 1998). Na avaliação descrita em Stevenson & Wilks (2000), por exemplo, considerando o córpus SEMCOR e um mapeamento entre os seus sentidos e os do LDOCE, cada processo (filtros e etiquetadores parciais) apresenta, isoladamente, uma acurácia na desambiguação de sentido que varia entre 44% e 79%. Com a integração dos resultados dos processos, utilizando o algoritmo de AM, essa acurácia chega a 90% para a desambiguação de todas as palavras ambíguas com um nível refinado de distinção de sentidos, e 94% no nível de homografia. Tal resultado é pouco menor que a acurácia de outros sistemas de DLS para vocabulários restritos, mas é significativamente alto, considerando-se que esse sistema procura desambiguar todas as palavras ambíguas de um texto e pode, portanto, ser aplicado em larga escala.

Paliouras et al. (1999) empregam um algoritmo de aprendizado simbólico de árvores de decisão (C4.5) para a DLS. O objetivo dos autores é produzir um modelo genérico, capaz de desambiguar todas as palavras de conteúdo de um texto, para a aplicação de Extração de Informações.

Um subconjunto do SEMCOR (Seção 2.7.1), constituído apenas por artigos de notícias financeiras, é utilizado para o treinamento. Contudo, o sistema de extração de informações ao qual o módulo de DLS deve ser acoplado é baseado nos sentidos do LDOCE. Assim, as etiquetas de sentido da WordNet do SEMCOR são convertidas em etiquetas do LDOCE.

Para minimizar o problema da representação esparsa, são utilizadas apenas características locais. Segundo os autores, esse problema se agrava à medida que características relativas a contextos maiores são contempladas. Essas características incluem: lema da palavra ambígua, o escore de freqüência do sentido no LDOCE, a etiqueta gramatical da palavra ambígua e 10 *collocations* (primeiro substantivo/verbo/preposição à esquerda e à direta da palavra ambígua, primeira e segunda palavras à esquerda da palavra ambígua). A palavra ambígua é representada como característica porque os autores pretendem analisar se é possível usar o mesmo modelo para todas as palavras, ou seja, se é possível gerar um modelo de desambiguação mais genérico.

A avaliação do trabalho foi realizada com base em textos do SEMCOR do mesmo domínio que os usados no treinamento, com os sentidos igualmente mapeados para as etiquetas do LDOCE. Contudo, foram considerados apenas verbos e substantivos. Os resultados são apresentados separadamente para verbos (precisão de 71.6%, cobertura de 66.2%) e substantivos (precisão de 58.5%, cobertura de 39.7%). De modo geral, segundo os autores, os resultados mostram que abordagens baseadas em árvores de decisão são adequadas para o problema de DLS. Contudo, vale lembrar que os valores precisão e cobertura podem ser devidos, em parte, à delimitação do domínio.

Com relação à utilização do mesmo modelo de representação para todas as palavras, os autores concluem que é uma alternativa viável, uma vez que o algoritmo automaticamente cria subárvores para cada uma das palavras ambíguas.

Em um trabalho posterior (Paliouras et al., 2000), os autores procuram confrontar o desempenho da abordagem simbólica empregada com o desempenho de outros trabalhos supervisionados. Para tanto, realizam um experimento com a mesma configuração que a do trabalho anterior, ou seja, com o mesmo córpus de treinamento e teste e as mesmas características, considerando agora vários algoritmos de classificação disponíveis no ambiente de aprendizado de máquina Weka, de três diferentes paradigmas de aprendizado: simbólico (C4.5 e ao C4.5rules, para a geração de árvores e regras de decisão, respectivamente), estatístico (Naive Bayes) e baseado em instâncias (KNN e tabela de decisões).

Os experimentos de teste realizados levam em consideração ambas as classes gramaticais (verbos e substantivos) e mostram que os algoritmos do paradigma simbólico apresentam resultados superiores, seguidos do algoritmo Naive Bayes e, por fim, dos dois algoritmos baseados em instâncias. Em primeiro lugar está o algoritmo C4.5, com 77.4% de cobertura e 82.6% de precisão. Em último, o algoritmo KNN, com 49% de cobertura e 66.3% de precisão. Segundo os autores, os algoritmos simbólicos apresentam um desempenho melhor que os demais em função da grande quantidade de valores distintos que cada característica pode ter na representação usada. Essa característica favorece os algoritmos que são capazes de realizar uma seleção e ordenação de características de maneira mais flexível, como é o caso das arvores de decisão.

Vale notar a inversão na ordem dos resultados dessa comparação, no que diz respeito aos algoritmos Naive Bayes e baseados em instâncias, com relação aos resultados apresentados por Escudero et al. (2000b). Entretanto, os autores não citam os parâmetros utilizados em cada algoritmo. Além disso, as configurações gerais de teste são bastante distintas. Novamente, um fator de influência pode ter sido a delimitação do domínio do córpus a textos de notícias financeiras.

Os trabalhos baseados em córpus descritos são listados na Tabela 5, de acordo com o modo e o paradigma de aprendizado, o(s) recurso(s) lingüístico(s) usado(s), o conjunto de palavras para as quais forma criados e/ou testados, o nível de refinamento das distinções entre os sentidos e a acurácia apresentada.

Tabela 5. Lista dos trabalhos de DLS híbridos

Trabalho	**Modo**	**Paradigma**	**Recurso**	**Conj. palavras**	**Nível de refinamento dos sentidos**	**Acurácia**
(Dagan et al., 1991) e (Dagan & Itai, 1994)	não-supervisio-nado	-	-	103 palavras do hebreu, 54 palavras do alemão	?	hebreu: 91% (precisão), 68% (cobertura), alemão: 78% (precisão), 50% (cobertura)
(Yarowsky, 1992)	não-supervisio-nado	-	*thesaurus* Roget	12 substantivos	baixo	92%
Resnik (1997)	não-supervisio-nado	-	WordNet	alguns substantivos	alto	40%
(Lee, 1997)	supervisio-	funções	MRD	121	?	57.46%

	nado	lineares	bilíngüe	substantivos		(precisão)
(Karov & Edelman, 1998)	semi-supervisio-nado	?	MRD	?	?	?
(Wilks & Stevenson, 1998) e (Stevenson & Wilks, 1999; 2000; 2001)	supervisio-nado	simbólico e baseado em instâncias	MRD LDOCE	palavras do SEMCOR	médio	83.4%: simbólico, 90%: baseado em instâncias
(Paliouras et al., 1999)	supervisio-nado	simbólico	MRD LDOCE	subconjunto de verbos e substantivos de um domínio do SEMCOR	alto	65% (precisão), 53% (cobertura)
(Paliouras et al., 2000)	supervisio-nado	comparação entre simbólico, estatístico e baseado em instâncias	MRD LDOCE	subconjunto de verbos e substantivos de um domínio do SEMCOR	alto	algoritmo simbólico - melhor desempenho: 82.6% (precisão), 77.4% (cobertura)

4 Considerações Finais

Conforme apresentado neste relatório, vários trabalhos vêm sendo desenvolvidos para a DLS, os quais possuem variações em muitos aspectos, incluindo os discutidos na Seção 2, como o método de codificação do conhecimento, os tipos de conhecimento utilizados, a abrangência e precisão, a possibilidade de generalização, etc. Os aspectos que se mostram mais relevantes em alguns trabalhos são retomados por diversos outros, as quais procuram aprimorá-los ou estendê-los para outros cenários. Como se pode perceber, a maioria dos trabalhos foram propostos e utilizados na DLS monolingüe.

Apesar da ilustração de resultados de avaliações individuais ou constrastivas de alguns trabalhos, uma análise comparativa entre vários trabalhos não foi realizada, pois, conforme discutido na Seção 2, uma comparação dessa natureza não poderia ser realizada diretamente, já que tais sistemas são muito diferentes entre si.

Com relação aos sistemas de TA, a maioria não possui módulos específicos de DLS. Nesses trabalhos, a DLS é realizada implicitamente, durante a tradução. Os sistemas não comerciais, desenvolvidos em pesquisas acadêmicas, procuram empregar conhecimentos profundos, manualmente especificados, para resolver, efetivamente, o problema da ambigüidade lexical de sentido. Contudo, esses sistemas são restritos a domínios e gêneros específicos e não são facilmente generalizáveis. Os sistemas comerciais, por outro lado, para cobrir qualquer gênero e domínio, empregam conhecimentos muito superficiais ou em quantidades insuficientes para a resolução da ambigüidade lexical de sentido, em geral.

Em se tratando da TA envolvendo o português, além dos sistemas comerciais, o único trabalho voltado especificamente para a DLS é o de Leffa (1998), que procura mostrar a importância do uso de *collocations* para a desambiguação, mas não apresenta uma proposta significativa de DLS. Com relação às ferramentas comerciais para essa língua, não se tem conhecimento de sistemas que empreguem mecanismos profundos e bem-sucedidos de DLS. Por essas razões, conforme mencionado, pretende-se propor, em um trabalho posterior, uma abordagem de DLS voltada especificamente para a TA do inglês para o português. Essa abordagem deverá considerar vários aspectos relevantes levantados pelas abordagens monolingües, incluindo o uso de vários tipos de conhecimentos, profundos e superficiais, mas deverá adquirir esses conhecimentos automaticamente, eliminado a necessidade de codificação manual. Essa abordagem parece representar uma alternativa bastante viável para a DLS efetiva na TA em larga escala envolvendo essa língua.

Referências Bibliográficas

Agirre, E., Martínez, D. (2000). Exploring Automatic Word Sense Disambiguation with Decision Lists and the Web. In Proceedings *of the COLING Workshop on Semantic Annotation and Intelligent Content*. Saarbrücken.

Agirre, E.; Martínez, D. (2001). Knowledge Sources for Word Sense Disambiguation. *Proceedings of the 4th International Conference TSD*. Lecture Notes in Computer Science Series, Springer Verlag, Plzen.

Agirre E., Martínez, D. (2004). Unsupervised WSD Based on Automatically Retrieved Examples: The Importance of Bias. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Barcelona.

Agirre, E.; Rigau, G. (1996). Word Sense Disambiguation Using Conceptual Density. In *Proceedings of the 16th International Conference on Computational Linguistics*. Copenhagen.

Arnold, D.J.; Balkan, L.; Humphreys, R.L.; Meijer, S.; Sadler, L. (1993). *Machine Translation: An Introductory Guide*. Blackwells-NCC, London.

Beale, S. (1997). *HUNTER-GATHERER: Applying Constraint Satisfaction, Branch-and-Bound and Solution Synthesis to Computational Semantics*. PhD Thesis, Language Technologies Institute, Carnegie Mellon University.

Black, E. (1988). An Experiment in Computational Discrimination of English Word Senses. *IBM Journal of Research and Development*, 32(2), pp. 185-194.

Boguraev, B. (1979). *Automatic Resolution of Linguistic Ambiguities*. Report 11, Computer Laboratory, University of Cambridge, Cambridge.

Borba, F.S. (1996). Uma Teoria de Valências para o Português. Ática, São Paulo.

Bräscher, M. (2002). A Ambigüidade na Recuperação da Informação. *DataGramaZero - Revista de Ciência da Informação*, 3(1).

Brill, E. (1992). A Simple Rule-based Part-of-speech Tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, pp. 152-155. Morgan Kaufmann, San Mateo.

Brill, E. (1995). Transformation Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, 21(4), pp. 543-565.

Briscoe, T. (1991). Lexical Issues in Natural Language Processing. In E. Klein and F. Veltman (eds), *Proceedings of the Symposium on Natural Language and Speech*, pp. 39-68. Springer-Verlag, Berlin.

Brown, P.F.; Cocke, P.F.; Della Pietra, S.A.; Della Pietra, V.J.; Jelinek, F.; Lafferty, J.D.; Mercer, R.L.; Roossin, P.S. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, 16, 79-85.

Brown, P.F.; Della Pietra, S.A.; Della Pietra, V.J.; Mercer, R.L. (1991). Word Sense Disambiguation Using Statistical Methods. In *Proceedings of the 29th Annual Meeting of Association for Computational Linguistics*, pp. 264-270. Berkley, CA.

Bruce, R.; Wiebe, J. (1994). Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association. for Computational Linguistics*, pp.139-145. Las Cruces.

Brun, C. (2000). A Client/Server Architecture for Word Sense Disambiguation. In *Proceedings of the 17th conference on Computational Linguistics*, 1, pp. 132-138. Saarbrücken.

Burnard, L. (2000). *Reference Guide for the British National Corpus (World Edition)*. Oxford University Press.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: the Kappa Statistics. *Computational Linguistics*, 22(2), pp. 249- 254.

Carlson, A.J.; Cumby, C.M.; Rizzolo, N.D.; Rosen, J.L.; Roth, D. (1999). *SNoW User Manual*. Computer Science Department, University of Illinois, Urbana-Champaign (http://l2r.cs.uiuc.edu/~cogcomp/software/snow-userguide/ [01/04/2004]).

Copeland, C.; Durand, J.; Krauwer, S.; Maegaard, B. (1991). The Eurotra Formal Specifications. *Studies in Machine Translation and Natural Language Processing*, 2. Commision of European Communities.

Correia, M. (1996). Terminologia e Lexicografia Computacional. In *Jornada Panllatina de Terminologia*, pp. 83-91. IULA / Universitat Pompeu Fabra, Barcelona.

Cost, S.; Salzberg, S. (1993). A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. *Machine Learning*, 10(1), pp. 57-78.

Cottrell, G. W. (1989). *A Connectionist Approach to Word Sense Disambiguation*. Research Notes in Artificial Intelligence. Morgan Kaufmann, San Mateo.

Cottrell, G. W.; Small, S.L. (1983). A Connectionist Scheme for Modelling Word Sense Disambiguation. *Cognition and Brain Theory*, 6, pp. 89-120.

Cowie, J.; Guthrie, J.A.; Guthrie, L. (1992). Lexical Disambiguation Using Simulated Annealing. In *Proceedings of COLING'92*, 1, pp. 359-365. Nantes.

Daelemans, W.; Zavrel, J.; Van Der Sloot, K.; Van Den Bosch, A. (1998). *TiMBL: Tilburg Memory Based Learner*. Technical Report 98-03. Tilburg.

Dagan, I.; Itai, A. (1994). Word Sense Disambiguation Using a Second Language Monolingual Corpus. *Computational Linguistics*, 20, pp. 563-596.

Dagan, I.; Itai, A.; Schwall, U. (1991). Two Languages are More Informative than One. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 130-137. Berkeley.

Dagan, I.; Marcus, S.; Markovitch, S. (1993). Contextual Word Similarity and Estimation from Sparse Data. In *Proceedings of the 31st Meeting of the Association for Computational Linguistics*. Columbus.

Diab, M.; Resnik, P. (2002). An Unsupervised Method for Word Sense Tagging using Parallel Corpora. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, Philadelphia.

Dihn, D.; Kiem, H.; Hovy, E. (2003). BTL: a Hybrid Model for English-Vietnamese Machine Translation. In *Proceedings of the MT Summit IX*, pp. 23-27. New Orleans.

Dini, L.; Di Tomaso, V.; Segond, F. (1998). Word sense disambiguation with functional relations. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*. Granada

Dorr, J. B. (1993). *Machine Translation: A View from the Lexicon*. The MIT Press, Cambridge.

Dorr, B.J.; Jordan, P.W.; Benoit, J.W. (2000). A Survey of Current Paradigms in Machine Translation. In M. Zelkowitz (ed), *Advances in Computers*, 49, pp. 1-68. Academic Press, London.

Dorr, B. J.; Katsova, M. (1998). Lexical Selection for Cross-Language Applications: Combining LCS with WordNet. In *Proceedings of AMTA'1998*, pp. 438-447. Langhorne.

Duda, O.R.; Hart, P.E. (1973). *Pattern Calssification and Scene Analysis*. Wiley, New York.

Dyvik, H. (2002). Translations as semantic mirrors: From Parallel Corpus to Wordnet. In Proceedings of the *23rd International Conference on English Language Research on Computerized Corpora of Modern and Medieval English*, Gothenburg..

EAGLES Lexicon Interest Group (1998). *Preliminary Recommendations on Semantic Encoding*. Interim Report (http://www.ilc.cnr.it/EAGLES96/rep2/rep2.html [01/04/2004]).

Edmonds, P.; Cotton, S. (2001). SENSEVAL-2: Overview. In *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems*, pp. 1-5.

Egedi, D.; Palmer, M.; Park, H.S.; Joshi, A.K. (1994). Korean to English Translatin Using Synchronous TAGs. In *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*, pp. 48-55. Maryland.

Escudero, G.; Màrquez, L.; Rigau, G. (2000a). Naive Bayes and Exemplar-Based Approaches to Word Sense Disambiguation Revisited. In *Proceedings of the 14th European Conference on Artificial Intelligence.* Berlin.

Escudero, G.; Màrquez, L.; Rigau, G. (2000b). Boosting Applied to Word Sense Disambiguation. In *Proceedings of the 12th European Conference on Machine Learning*. Barcelona.

Escudero, G.; Màrquez, L.; Rigau, G. (2000c). A Comparison between Supervised Learning Algorithms for Word Sense Disambiguation. In *Proceedings of CoNLL-2000 and LLL-2000*, pp. 31-36. Lisboan.

Escudero, G.; Màrquez, L.; Rigau, G. (2001). Using LazyBoosting for Word Sense Disambiguation. In *Proceedings of the 2nd International Workshop on evaluating Word Sense Disambiguation Systems*. Toulouse.

Fernández, J.; Castilho, M.; Rigau, G.; Atserias, J.; Turmo, J. (2004). Automatic Acquisition of Sense Examples using ExRetriever. In *Proceedings of the International Conference on Language Resources and Evaluation*, pp. 25-28. Lisbon.

Flanagan, M.; McClure, S. (2002). *SYSTRAN and the Reinvention of MT*. IDCBulletin #26459 - Jan 2002 (http://www.systransoft.com/IDC/26459.html [10/03/2004]).

Florian, R.; Cucerzan, S.; Schafer, C.; Yarowsky, D. (2002). Combining Classifiers for Word Sense Disambiguation. *Natural Language Engineering*, 1(1), pp. 1-14. Cambridge University Press, Cambridge.

Francis, W. M.; Kucera, H. (1979). *Brown Corpus – Manual of Information*. Department of Linguistics, Brown University (http://helmer.aksis.uib.no/icame/brown/bcm.html [03/04/2004]).

Frederking, R.E.; Brown, R.D. (1996). The Pangloss-Lite Machine Translation System. In *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas*, pp. 268-272. Montreal.

Freund, Y.; Schapire, R.E. (1996). Experiments with a New Boosting Algorithm. In *Proceedings of the 13th International Conference on Machine Learning*, pp. 148-156.

Firth, J.R. (1957). *Papers in Linguistics 1934-1951*. Oxford University Press, London.

Fujii, A.; Inui, K.; Tokunaga, T.; Tanaka, H. (1998). Selective Sampling for Example-based Word Sense Disambiguation. *Computational Linguistics*, 24 (4), pp.573-597.

Gaizauskas, R.; Wakao, T.; Humphreys, K.; Cunningham, H.; Wilks, Y. (1996). Description of the LaSIE System as Used for MUC-6. In *Proceeedings of the 6th Message Understanding Conference*, pp. 207-220. Morgan Kaufmann, San Mateo.

Gajek, O. (1991). The METAL system. *Communications of the ACM*, 34 (9), pp. 46-47.

Gale, W.A.; Church, K.W. (1991). A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of ACL-91*, Berkeley CA.

Gale, W.A.; Church, K.W. Yarowsky, D. (1992a). Using Bilingual Materials to Develop Word Sense Disambiguation Methods. In *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 101-112.

Gale, W.A.; Church, K.W. Yarowsky, D. (1992b). Estimating Upper and Lower Bounds on the Performance of Word Sense Disambiguation Programs. In *Proceedings of the 30th Annual Meeting of Association for Computational Linguistics*, pp. 249-256. Newark.

Gale, W., K. Church, and D. Yarowsky (1992c). One Sense Per Discourse. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pp. 233-237.

Gale, W.A.; Church, K.W. Yarowsky, D. (1992d). A Method for Disambiguating Word Senses in a Large Corpus. *Computers and the Humanities*, 26, pp. 415-439.

Goodman, K; Nirenburg, S. (1991). *The KBMT Project: A case study in Knowledge-Based Machine Translation*. Morgan Kaufmann Publishers, California.

Guthrie, J., Guthrie, L., Wilks, Y., and Aidinejad, H. (1991). Subject-Dependent Co-Occurrence and Word Sense Disambiguation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 146-152. Berkeley.

Hayes, P.J. (1976). *A Process to Implement Some Word Sense Disambiguation*. Working paper 23, Institut pour les Etudes Sémantiques et Cognitives, Universitè de Genève. Gèneve.

Hearst, M. (1991). Noun Homograph Disambiguation using Local Context in Large Text Corpora, *Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora*. Oxford.

Hirst, G. (1987). Semantic Intepretation and the Resolution of Ambiguity. *Studies in Natural Language Processing*. Cambridge Universisty Press, Cambridge.

Ide, N. (1999). Parallel Translations as Sense Discriminators. In *Proceedings of the SIGLEX99 Workshop: Standardizing Lexical Resources*, pp. 52-61. Maryland,.

Ide, N.; Erjavec, T.; Tufi, D. (2002). Sense Discrimination with Parallel Corpora. In *Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pp. 54-60. Philadelphia.

Ide, N.; Véronis, J. (1998). Word Sense Disambiguation: The State of the Art. *Computational Linguistics*, 24 (1).

Jackendoff, R. (1990). *Semantic Structures*. The MIT Press, Cambridge.

Karov, Y.; Edelman, S. (1998). Similarity-based Word Sense Disambiguation. *Computational Linguistics*, 24(1), pp. 41-59.

Katz, J.J.; Fodor, J.A. (1963). The Structure of a Semantic Theory. *Language*, 39, pp. 170-210.

Kilgarriff, A. (1992). *Polysemy*. PhD Thesis, University of Sussex, UK.

Kilgarriff, A. (1995). Inheriting Polysemy. In Patrick Saint-Dizier and Evelyne Viegas (eds.), *Computational Lexical Semantics*. Cambridge University Press.

Kilgarriff, A. (1997a). I Don't Believe in Word Senses. *Computers and the Humanities*, 31 (2), pp. 91-113.

Kilgarriff, A. (1997b). What is Word Sense Disambiguation Good For? In *Proceedings of the NLP Pacific Rim Symposium'97*. Phuket.

Kilgarriff, A.; Palmer, M. (2000). Introduction to the Special Issue on SENSEVAL. *Computers and the Humanities*, 34(1-2), pp. 1-13. Kluwer Academic Publishers, The Netherlands.

Krovetz, R. (1998). *More than One Sense Per Discourse*. Research Memorandum, NEC Research Institute, Princeton.

Leacock, C.; Chodorow, M.; Miller, G.A. (1998). Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24 (1), pp. 147-165.

Leacock, C.; Towell, G.; Voorhees, E.M. (1993). Corpus-Based Statistical Sense Resolution. In *Proceedings of the ARPA Human Language Technology Workshop*, pp. 260-265. Morgan Kaufmann Publishers, San Francisco.

Leacock, C.; Towell, G.; Voorhees, E.M. (1996). Towards building contextual representations of word senses using statistical models. In *Corpus Processing for Lexical Acquisition*, pp. 97-113. The MIT Press.

Lee, Y.K.; Ng, H.T. (2002). An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 41-48. Philadelphia.

Leffa, V. J. (1998). Textual constraints in L2 lexical disambiguation. *System*, 26(2), pp. 183-194, Great Britain.

Lenci, A.; Busa, F.; Ruimy, N.; Gola, E.; Monachini, M.; Calzolari, N.; Zampolli, A. (1999). *SIMPLE - Semantic Information for Multifunctional Plurilingual Lexica: Linguistic Specifications*. Internal Report, University of Pisa and Institute of Computational Linguistics of CNR, Pisa.

Lesk, M. (1986). Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 1986 SIGDOC Conference*, pp. 24-26. Toronto.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Lyons, J. (1977). Semantics. Cambridge University Press, Cambridge.

Magnini, B.; Strapparava, C.; Pezzulo, G.; Gliozzo, A. (2002). The Role of Domain Information in Word Sense Disambiguation. *Natural Language Engineering*, 8(4), pp. 359-373. Cambridge University Press, Cambridge.

Manning, C.D.; Schütze, H. (2001). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge.

Martínez, D.; Agirre, E. (2000). One Sense per Collocation and Genre/Topic Variations. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 207-215, Hong Kong.

Martínez D.; Agirre E.; Màrquez L. (2002). Syntactic Features for High Precision Word Sense Disambiguation**.** *Proceedings of the 19th International Conference on Computational Linguistics*. Taipei.

Masterman, M. (1957). The Thesaurus in Syntax and Semantics. *Mechanical Translation*, 4, pp. 1-2.

Masterman, M. (1961). Semantic Message Detection for Machine Translation Using an Interlingua. In *1961 International Conference on Machine Translation of Languages and Applied Lnguage Analysis*, pp. 437-475. London.

McCarthy, D.; Koeling, R.; Weeds, J.; Carroll, J. (2004). Using Automatically Acquired Predominant Senses for Word Sense Disambiguation. Accepted for publication in *Proceedings of the ACL SENSEVAL-3 Workshop.* Barclona.

McRoy, S. (1992). Using Multiple Knowledge Sources for Word Sense Discrimination. *Computational Linguistics*, 18(1), pp. 1-30.

Mihalcea, R. (2004). Co-training and Self-training for Word Sense Disambiguation. In *Proceedings of the Conference on Natural Language Learning*. Boston.

Mihalcea, R.; Moldovan, D.I. (1999). A Method for Word Sense Disambiguation of Unrestricted Text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland.

Miller, G.A.; Beckwith, R.T.; Fellbaum, C.D.; Gross, D.; Miller, K. (1990). Wordnet: An On-line Lexical Database. In *International Journal of Lexicography*, 3(4), pp. 235-244.

Miller, G.A.; Chorodow, M.; Landes, S.; Leacock, C; Thomas, R.G. (1994). Using a Semantic Concordancer for Sense Identification. In *Proceedings of the ARPA Human Language Technology Workshop - ACL*, pp. 240-243. Washington.

Monard, M.C.; Baranauskas, J.A. (2003). Conceitos sobre Aprendizado de Máquina. In S.O. Rezende (org.), *Sistemas Inteligentes: Fundamentos e Aplicações*, pp. 89-114. Manole, Barueri.

Montoyo, A.; Romero, R.; Vazquez, S.; Calle, M.; Soler, S. (2002). The Role of WSD for Multilingual Natural Language Applications. In *Proceedings of TSD'2002*, p. 41-48. Czech Republic.

Mooney, R.J. (1996). Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning. In *Proceedings of the Conference on Empirical Methods in NLP*, pp. 82-91. Somerset, New Jersey.

Mowatt, D. (1999). Types of Semantic Information Necessary in a Machine Translation Lexicon. In *Proceedings of the TALN'99*, Cargèse.

Ng, H.T. (1997a). Exemplar-Based Word Sense Disambiguation: Some Recent Improvements. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*. Providence.

Ng, H. T. (1997b). Getting Serious about Word Sense Disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, pp. 1-7. Washington.

Ng, H.T.; Lee, H.B. (1996). Integrating Multiple Knowledge Sources to Disambiguate Word Senses: An Exemplar-Based Approach. In *Proceeedings of the 34th Annual Meeting of Association for Computational Linguistics*, pp. 40-47. Somerset.

Ng, H.T.; Zelle, J. (1997). Corpus-Based Approaches to Semantic Interpretation in Natural Language Processing. *AI Magazine*, 18(4), pp. 45-64.

Ng, H.T.; Wang, B.; Chan, Y.S. (2003). Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 455-462. Sapporo.

Paliouras, G.; Karkaletsis, V.; Spyropoulos, C.D. (1999). Learning Rules for Large-Vocabulary Word Sense Disambiguation. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pp. 674-679. Morgan Kaufmann Publishers, San Francisco.

Paliouras, G.; Karkaletsis, V.; Androutsopoulos, I.; Spyropoulos, C.D. (2000). Learning Rules for Large-Vocabulary Word Sense Disambiguation: a comparison of various classifiers. In *Proceedings of the 2nd International Conference on Natural Language Processing*, pp. 383-394. Lecture Notes in Artificial Inteligence, Springer, Patra.

Palmer, M. (1998). Are WordNet sense distinctions appropriate for computational lexicons? In *Proceedings of Senseval, Siglex98*. Brighton.

Palmer, M. (2000). Consistent Criteria for Sense Distinctions**.** *Computers and the Humanities*, 34 (1-2), Kluwer Academic Publishers.

Pantel, P.; Lin, D. (2002). Discovering Word Senses from Text. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 613-619. Edmonton.

Park, S.; Zhang, B.; Kim, Y.T. (2003). Word Sense Disambiguation by Learning Decision Trees from Unlabeled Data. *Applied Intelligence*, 19, pp. 27-38. Kluwer Academic Publishers, The Netherlands.

Patrick, A.B. (1985). *An Exploration of Abstract Thesaurus Instanttiation*. M.Sc.Thesis, University of Kansas, Kansas.

Pedersen, B.S. (1997). *Lexical Ambiguity in Machine Translation: Expressing Regularities in the Polysemy of Danish Motion Verbs*. PhD Thesis, Center for Sprogteknologi, Copenhagen.

Pedersen, T. (2000). A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambiguation. In *Proceedings of NAACL*, pp. 63-69, Seattle.

Pedersen, T. (2002a). A Baseline Methodology for Word Sense Disambiguation. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City.

Pedersen, T. (2002b). Evaluating the Effectiveness of Ensembles of Decision Trees in Disambiguating Senseval Lexical Samples. In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pp. 81-87. Philadelphia.

Pedersen, T.; Bruce, R. (1997). Distinguishing Word Senses in Untagged Text. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*, pp. 197-207, Providence.

Pedersen, T.; Bruce, R. (1998). Knowledge Lean Word-Sense Disambiguation. In *Proceedings of the 15th National Conference on Artificial Intelligence*, pp. 800-805.

Pereira, F.; Tishby, N.; Lee, L. (1993). Distributional Clustering of English Words. In *Proceedings of the 31st Annual Meeting of Association for Computational Linguistics*, pp. 183-190. Ohio.

Pustejovsky, J. (1995). *The Generative Lexicon*. The MIT Press, Cambridge.

Quillian, M.R. (1961). A Design for an Understanding Machine. Presented in *Colloquium of Semantic problems in natural language*. Cambridge University, Cambridge.

Quinlan, J. R. (1988). *C4.5 – Programs for Machine Learning*. Morgan Kaufmann. 302p.

Ravin, Y; Leacock, C. (2000) Polysemy: An Overview. In *Polysemy: Theoretical and Computational Approaches*, pp. 1-29. Oxford University Press, Oxford.

Resnik, P. (1995a). Disambiguating Noun Groupings with Respect to WordNet Senses. In *Proceedings of the 3rd Workshop on Very Large Corpora*, pp. 54-68. Cambridge.

Resnik, P. (1995b). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 448-453, Montreal.

Resnik, P. (1997). Selectional Preferences and Sense Disambiguation. In *Proceedings of ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How*?, pp. 52-57. Washington.

Resnik, P.; Yarowsky, D. (1997a). Evaluating Automatic Semantic Taggers. In *ACL-SIGLEX Workshop Tagging Texts with Lexical Semantics: Why, What and How*?. Washington.

Resnik, P.; Yarowsky, D. (1997b). A Perspective on Word Sense Disambiguation Methods and their Evaluating. In *ACL-SIGLEX Workshop Tagging Texts with Lexical Semantics: Why, What and How*?. Washington.

Rivest, R.L. (1987). Learning Decision Lists. *Machine Learning*, 2(3), pp. 229-246.

Sanderson, M. (2000). Retrieving with Good Sense. *Information Retrieval*, 2(1), pp. 47-67. Kluwer Academic Publishers, The Netherlands.

Sanderson, M. (1994). Word Sense Disambiguation and Information Retrieval. In *Proceedings of the17th International ACM Special Interest Group on Information Retrieval*, pp. 49-57. Dublin.

Schütze, H. (1992). Dimensions of Meaning. In *Proceedings of Supercomputing'92*, pp. 787-796. IEEE Computer Society Press, Washington.

Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1), pp. 97-124.

Schütze, H.; Pedersen, J. (1995). Information Retrieval Based on Word Senses. In *Proceedings 4th Annual Symposium on Document Analysis and Information Retrieval*. Las Vegas.

Segond, F.; Schiller, A.; Grefenstette, G.; Chanod, J. (1997). An Experiment in Semantic Tagging Using Hidden Markov Model Tagging. In *Proceedings of the ACL/EACL'97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources*.

Shieber, S.; Schabes, Y. (1990). Synchronous Tree Adjoining Grammar. In *Proceedings of the 13th International Conference on Computational Linguistics*. Helsinki.

Small, S.L. (1980). *Word-expert Parsing, a Theory of Distributed Word-based Natural Language Based Understandig*. Technical Report 954, Departament of Computer Science, University of Maryland, Maryland.

Sparck-Jones, K.; Galliers, J.R. (1996). *Evaluating Natural Language Processing Systems: an Analysis and Review*. Springer-Verlag.

Stevenson, M.; Wilks, Y. (1999). Combining Weak Knowledge Sources for Sense Disambiguation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, p. 884-888. Stockholm.

Stevenson, M.; Wilks, Y. (2000). Large Vocabulary Word Sense Disambiguation. In Y. Ravin and C. Leacock, *Polysemy: Theoretical and Computational* Approaches, pp. 161-177. Oxford University Press, Oxford.

Stevenson, M.; Wilks, Y. (2001). The Interaction of Knowledge Sources for Word Sense Disambiguation. *Computational Linguistics*, 27(3), pp. 321-349.

Stokoe, C.; Oakes, M.P.; Tait, J. (2003). Word Sense Disambiguation in Information Retrieval Revisited. In *Proceedings of the ACM Special Interest Group on Information Retrieval*, pp. 159-166. Toronto.

Sussna, M. (1993). Word Sense Disambiguation for Free-text Indexing Using Massive Semantic Network. In *Proceedings of the 2nd International Conference on Information and Knowledge Base Managment*, pp. 67-74. Virginia.

Towell, G.; Voorhees, E.M. (1998) Disambiguating Highly Ambiguous Words. *Computational Linguistics*, 24(1), pp. 125-145.

Ullmann, (1964). *Semântica: uma Introdução à Ciência do Significado*. Fundação Calouste Gulbenkian, Lisboa.

Véronis, J. (1998). A Study of Polysemy Judgements and Inter-annotator Agreement. *Programme and Advanced Papers of the Senseval Workshop*, pp. 2-4. Herstmonceux Castle.

Véronis, J.; Ide, N.M. (1990). Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries. In *Proceedings of COLING'90*, 2, pp. 398-394. Helsinki.

Vossen, P. (1998). EuroWordNet: Building a Multilingual Database with WordNets for European Languages. *The ELRA Newsletter*, 3(1).

Vogel, S.; Och, F.J.; Tillmann, C.; Nießen, S.; Sawaf, H.; Ney, H. (2000). Statistical Methods for Machine Translation. In *Verbmobil: Foundations of Speech-to-Speech Translation*, pp. 377-393. Springer Verlag, Berlin.

Voorhees, E.M. (1993). Using WordNet to disambiguate word senses for text retrieval. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 171-180, Pittsburgh, PA.

Waltz, D.L.; Pollack, J.B. (1985). Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Processing. *Cognitive Science*, 9, pp. 51-74.

Wilks, Y. (1975). A Preferential, Pattern-seeking, Semantics for Natural Language Inference. *Artificial Intelligence*, 6, pp. 53-74.

Wilks, Y. (1997). Senses and Texts. *Computers and the Humanities*, 31(2).

Wilks, Y.; Fass, D.; Guo, C-M.; McDonald, J.E.; Plate, T.; Slator, BM. (1990). Providing Machine Tractable Dictionary Tools. *Journal of Machine Translation*, 5 (2), pp. 99-151.

Wilks, Y.; Stevenson, M. (1996). *The Grammar of Sense: Using Part-of-speech Tags as a First Step in Semantic Disambiguation*. Techical Report CS-96-05, University of Sheffield. Also published in Journal of Natural Language Engineering, 4(1), pp. 1-9, 1998.

Wilks, Y.; Stevenson, M. (1997a). Sense Tagging: Semantic Tagging with a Lexicon. In *Proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics: What, why and how"*. Washington.

Wilks, Y.; Stevenson, M. (1997b). Combining Independent Knowledge Sources for Word Sense Disambiguation. In *Proceedings of the 3rd Conference on Recent Advances in Natural Language Processing*, pp. 1-7. Tzigov Chark.

Wilks, Y.; Stevenson, M. (1998). Word Sense Disambiguation Using Optimised Combinations of Knowledge Sources. In *Proceedings of COLING'98*, pp. 1398-1402. Montreal, Canada.

Yarowsky, D. (1992). Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In *Proceedings of COLING'92*, pp. 454-460. Nantes.

Yarowsky, D. (1993). One Sense Per Collocation. In *Proceedings of the ARPA Human Language Technology Workshop*, pp. 266-271. Princeton.

Yarowsky, D. (1994). Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 88-95. Las Cruces.

Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189-196. Cambridge.

Yarowsky, D. (2000). Hierarchical Decision Lists for Word Sense Disambiguation. *Computers and the Humanities*, 34(1-2), pp. 179-186. Kluwer Academic Publishers, The Netherlands.

Zinovjeva, N. (2000). *Learning Sense Disambiguation Rules for Machine Translation*. Master's Thesis in Language Engineering. Department of Linguistics, Uppsala University.