Programação

Agenda Atividade

Área de Conhecimento
Léxico e Semântica

Tipo de Atividade
comunicação coordenada

Título
O desenvolvimento de recursos léxico-computacionais para o processamento automático do português

Resumo
Inserindo-se na temática "Léxico e Semântica" e ilustrando, com exemplos de pesquisas em desenvolvimento, o alcance dos estudos do léxico para "além dos espaços da linguística", a proposta é discutir a construção de um tipo específico de léxico: a construção de bases de conhecimento lexical para integrar sistemas de processamento automático de línguas naturais (PLN), campo de pesquisa que reúne linguistas e cientistas da computação no empreendimento de modelar computacionalmente o conhecimento linguístico (HANDKE 1995). No âmbito do PLN, a meta é desenvolver sistemas computacionais capazes de processar automaticamente informações codificadas em língua natural como, por exemplo, os sistemas de correção ortográfica, de tradução automática e de recuperação de textos/informação (JURAFSKY e MARTIN 2008). Os sistemas que são baseados em conhecimento linguístico, em oposição aos sistemas de base estatística (MANNING e SCHÜTZE 2003), não abordados nesta sessão, incluem necessariamente uma base de conhecimento lexical, módulo que lhes fornece o acervo dos itens lexicais da língua para a qual são propostos (ALLEN 1994). Nessa base, são especificadas, para cada item lexical, uma ou mais das seguintes propriedades: fonéticas, morfológicas, sintáticas, semânticas e pragmático-discursivas. Nesse universo de construção de bases de dados lexicais, destaca-se, devido a sua adequação linguístico-tecnológica e a sua ampla adoção pela comunidade científica, o desenvolvimento das redes wordnets, ponto de convergência das pesquisas em debate na sessão. As wordnets são um tipo específico de léxico, em que, à semelhança a um dicionário analógico (AZEVEDO 2010), os itens lexicais estruturam-se em função de categorias gramaticais, de tipos semânticos gerais e de relações de sentido que se estabelecem entre eles. Idealizada por pesquisadores da Universidade de Princeton para o inglês norte-americano (FELLBAUM 1998), a primeira wordnet, hoje conhecida como a WordNet de Princeton. A wordnet de uma língua caracteriza-se por estruturar o léxico de língua geral em função de quatro categorias sintáticas (substantivo, verbo, adjetivo e advérbio) e de um rol de relações de sentido (CRUSE 2004) que se verificam entre eles (sinonímia, antonímia, hiponímia, meronímia, acarretamento e causa). A estrutura assim se desenha: a sinonímia agrupa os itens lexicais de cada categoria em synsets (isto é, em "conjuntos de sinônimos"), cada um dos quais representando simultaneamente um conceito (compartilhado pelos itens lexicais que formam o synset) e um nó da rede; as demais relações de sentido, excetuando-se a antonímia, que, dependendo das características do item léxico, conecta tanto synsets quanto itens lexicais, conectam synsets, que representam as conexões entre os nós da rede. Além de estabelecer os cânones teórico-metodológicos para a construção desse tipo de base de dados lexicais, a WN.Pr fomentou a proposição da wordnet multilíngue EuroWordNet (VOSSEN, 1998), iniciativa de pesquisadores europeus que aglutinou esforços intelectuais e de fomento para a produção do conjunto de wordnets para cada uma das línguas da União Europeia (EU) e, sobretudo, para a especificação de critérios operacionais que especificam o complexo alinhamento semântico conecta, synset a synset, cada wordnet construída para cada língua da UE à rede WN.Pr. No contexto de construção de recursos léxico-computacionais dessa natureza, os três trabalhos

aqui propostos discutem, nesta ordem, (i) a construção de wordnets e, em especial, a construção da WordNet.Br, que é o wordnet para o português brasileiro, e o seu alinhamento semântico à WN.Pr (DIAS-DA-SILVA 2010); (ii) uma proposta de descrição dos adjetivos e da relação de antonímia que se verifica entre eles em uma wordnet com um estudo piloto, a partir de corpus, dos adjetivos descritivos do português brasileiro (BARROS 2010); e (iii) estratégias de identificação automática da sinonímia em corpus especializado no processo de construção de uma wordnet terminológica para o domínio da Educação a Distância (DiFELIPPO 2010).

Trabalhos

Título
A construção de wordnets: uma modelagem linguístico-computacional de léxicos

Autor(es)
BENTO CARLOS DIAS-DA-SILVA

Resumo
Neste trabalho, discutindo-se os fundamentos teóricos e metodológicos de construção da WordNet.Br (WN.Br) (DIAS-DA-SILVA 1998, 2003, 2007, 2010), abordam-se os principais aspectos de uma das áreas mais vibrantes nos domínios da semântica lexical e lexicografia computacional (ZAMPOLLI, CIGNONI e PETERS 1990; SAINT-DIZIER e VIEGAS 1995; EYNDE e GIBBON 2000) e do processamento automático de língua natural (PLN) (ALLEN 1994, MITIKOV 2002; DIAS-DA-SILVA 1996, 2006; JURAFSKY e MARTIN 2008): a modelagem linguístico-computacional de léxicos. Parte-se da contextualização da discussão, caracterizando esse tipo de modelagem do conhecimento lexical dos falantes como uma tarefa que envolve tanto a seleção, a descrição e a formalização dos diferentes tipos de informações associadas aos itens léxicos (FILLMORE 1971; HANDKE 1995), quanto o gerenciamento constante e a manipulação rápida e seletiva dessas informações. Ressalta-se que os produtos dessa tarefa, as bases de dados lexicais (BOGURAEV e BRISCOE 1989; NUGUES 2006), que constituem o módulo lexical de sistemas de PLN, podem também subsidiar os estudos do léxico e a compilação de obras lexicográficas, pois são recursos léxico-computacionais de grande porte em que são sistematicamente armazenadas parcelas significativas do acervo lexical de uma língua. Após a contextualização, descrevem-se os componentes e pontuam-se as repercussões de um tipo específico de base de dados lexicais, cujos cânones teórico-metodológicos subsidiam a construção de outras bases semelhantes, incluindo a WN.Br. Trata-se da rede lexical wordnet, uma base de dados lexicais específica, idealizada por pesquisadores da Universidade de Princeton (MILLER 1986; MILLER e FELLBAUM 1991) e instanciada na WordNet de Princeton (WN.Pr), que modela parcelas do léxico do inglês norte-americano (FELLBAUM 1998). Em linhas gerais, uma rede wordnet, de modo semelhante aos dicionários analógicos (AZEVEDO 2010), organiza os itens lexicais da língua para a qual é construída em função de categorias gramaticais (substantivo, verbo, adjetivo e advérbio), tipos semânticos gerais (Evento, Artefato, Localização, Processo, Emoções, entre outros) (GLIOZZO e STRAPPARAVA 2009) e relações de sentido (sinonímia, antonímia, hiponímia, meronímia, acarretamento e causa) (CRUSE 1986, 2004; MARRAFA 2000; GEERAERTS 2010). Sua estrutura de rede semântica hierarquicamente organizada, assim se desenha: a sinonímia agrupa os itens lexicais de cada categoria gramatical em synsets (isto é, em conjuntos de itens lexicais que são considerados sinônimos pelo falante em pelo menos um contexto de uso); cada synset, por sua vez, representa simultaneamente um conceito (expresso pelos itens lexicais que o formam) e um nó da rede; a antonímia, dependendo do tipo de adjetivo a que se aplica, estabelece as conexões entre synsets ou entre itens lexicais; as relações de hiponímia, meronímia, acarretamento e causa estabelecem conexões exclusivamente entre synsets. São essas as conexões que tecem uma rede wordnet. Dentre as repercussões, merece destaque a que ocorreu entre os

pesquisadores europeus, que aglutinaram esforços intelectuais e de fomento para a produção de uma wordnet multilíngue: a EuroWordNet (VOSSEN, 1998), que resulta da conexão, synset a synset e por meio de relações de (quase)equivalência semântico-conceitual, denominadas EQ_RELATIONS (VOSSEN et al. 1998; PETERS et al. 1998), das diferentes wordnets em construção para cada uma das línguas da União Europeia aos synsets da WN.Pr, acrescidos dos synsets construídos para outras línguas e não constantes da WN.Pr. A partir dessa iniciativa europeia complementar sistematizaram-se os métodos e os critérios operacionais de especificação do complexo alinhamento semântico-conceitual que permite conectar à rede WN.Pr as wordnets das demais línguas. Por fim, e complementando as discussões anteriores com exemplos ilustrativos, apresentam-se o método de específico de construção da WN.Br (DIAS-DA-SILVA 2010), a rede wordnet em fase de desenvolvimento para o português brasileiro, amostras de alinhamentos semânticos aos synsets da WN.Pr, as estatísticas atuais da base lexical da WN.Br e os desdobramentos futuros desse empreendimento brasileiro.

Palavras-Chaves
1 - Léxicos
2 - Relações de sentido
3 - Wordnets
4 - Redes semânticas
5 - Bases de dados lexicais

Título
Os adjetivos nas wordnets: um estudo tipológico e a especificação da antonímia

Autor(es)
CLÁUDIA DIAS DE BARROS

Resumo
Na rede WordNet de Princeton (WN.Pr) (FELLBAUM, 1998), os itens léxicos pertencentes às categorias de substantivos, verbos, adjetivos e advérbios são agrupados em conjuntos de sinônimos (os synsets), os nós da rede, e relacionados por meio de diferentes relações de sentido (sinonímia, antonímia, hiponímia, meronímia, causa e acarretamento), as conexões que unem os nós. Os adjetivos, em particular, são subcategorizados em duas grandes subclasses: a dos descritivos, porque preenchem os valores opostos de um atributo descritivo bipolar, organizando-se, assim, em termos de pares de antônimos diretos ("carro grande"/"carro pequeno") ou em termos de "aglomerados" de antônimos indiretos que se aglutinam em torno dos antônimos diretos "enorme=grande/pequeno=nanico"), e a dos relacionais, porque se relacionam semântica e morfologicamente aos substantivos de que são derivados e não possuem antônimos diretos ou indiretos ("crise mundial"). É, então, nesse contexto de construção de wordnets, em particular da construção da WordNet.Br (WN.Br), descrita em Dias-da-Silva et al. (2006 e 2010), e cujos 25 mil adjetivos podem ser consultados on line na base do base do TeP, um dicionário eletrônico de sinônimos e antônimos em construção para o português brasileiro (DIAS-DA-SILVA et al. 2000; DIAS-DA-SILVA e MORAES, 2003; MAZIERO, et al., 2008), que este trabalho tem como objetivos (a) propor uma terceira subclasse de adjetivos, a dos determinativos (BARROS, 2010), que inclui formas tradicionalmente classificadas como artigos, numerais ou pronomes ("várias palavras"), e (b) discutir a antonímia direta, que se caracteriza pela oposição entre itens léxicos, e a antonímia indireta, cuja oposição se dá entre conceitos, a partir da análise dos 100 adjetivos descritivos mais frequentes no corpus Mac-Morpho do projeto LacioWeb (ALUÍSIO et al., 2003), contendo 1.167.183 ocorrências do português brasileiro, formado por artigos jornalísticos retirados da Folha de São Paulo de 1994 e automaticamente etiquetado pelo analisador morfossintático automático Palavras (BICK, 2000) e revisto manualmente. Com o

auxílio do Unitex (PAUMIER, 2002), realizaram-se (i) uma análise preliminar de 100 adjetivos e (ii) uma análise exaustiva de 108 adjetivos. A análise preliminar revelou os seguintes números: 23 relacionais, 1 determinativo e 76 descritivos. A partir da análise preliminar, e concentrando o estudo na análise dos adjetivos descritivos (os que apresentam os antônimos diretos ou indiretos), identificaram-se mais 32 adjetivos, extraídos de dicionários de sinônimos e antônimos (BARBOSA, 1999), que eram os antônimos dos adjetivos extraídos do corpus, somando, assim, os 108 adjetivos que constituíram o universo da análise. Desse total, isolaram-se 17 pares de antônimos, como "alto/baixo". Como resultados desse estudo piloto, destacam-se (i) a proposição de uma subclassificação dos tipos de adjetivos mais detalhada que a prevista na WN.Pr, com o acréscimo da subclasse dos adjetivos (determinativos), (ii) o estabelecimento de tipos de tendências na formação dos pares antonímicos, através da observação e corpus de que os antônimos formados por itens léxicos são mais frequentes (47 pares do tipo "verdadeiro/falso") que aqueles formados por prefixos de negação (13 pares do tipo "disponível/indisponível"), (iii) a constatação da tendência de a antonímia direta ser mais frequente (61 adjetivos) que a indireta (15 adjetivos) e (iv) a seleção de contextos de uso dos pares antonímicos analisados (373 frases-exemplo). Para concluir, aponta-se que o estudo permitiu, por um lado, lançar mais luz sobre a descrição dos adjetivos do português; por outro, contribui para a especificação dos adjetivos da rede WN.Br, instrumentalizando, com critérios descritivos, seus desenvolvedores na tarefa de construção desse importante recurso léxico-computacional para o português brasileiro.

Palavras-Chaves
1 - Léxicos
2 - Adjetivos
3 - Antonímia
4 - Wordnets

Título
Os termos nas wordnets: estratégias de identificação automática da sinonímia em corpus especializado

Autor(es)
ARIANI DI-FELIPPO

Resumo
No estudo do processamento automático das línguas naturais (PLN), desenvolvem-se sistemas computacionais específicos que são capazes de processar informações codificadas em língua natural (p.ex.: os tradutores automáticos). Quando baseados em conhecimento linguístico, tais sistemas incluem uma base de conhecimento lexical que lhes fornece o acervo dos itens lexicais em que são especificadas, para cada item lexical, suas propriedades morfológicas, sintáticas, semânticas e pragmático-discursivas. Nesse universo de construção de bases de conhecimento lexical para sistemas de PLN, destaca-se, devido a sua adequação linguístico-tecnológica e a sua ampla adoção pela comunidade científica, o desenvolvimento das redes wordnets. Idealizadas por pesquisadores da Universidade de Princeton (FELLBAUM, 1998), que desenvolveram a WordNet de Princeton (para o inglês norte-americano), as wordnets caracterizam-se por estruturar o léxico de língua geral em função de quatro categorias sintáticas (substantivo, verbo, adjetivo e advérbio) e de um rol de relações de sentido que se verificam entre eles (sinonímia, antonímia, hiponímia, meronímia, acarretamento e causa). A sinonímia agrupa os itens lexicais de cada categoria em synsets (isto é, "conjuntos de sinônimos"), cada um dos quais representando simultaneamente um conceito (compartilhado pelos itens lexicais que formam o synset) e um nó da rede. Já as demais relações, conectando synsets, representam as conexões entre os nós. Dada a necessidade crescente de processamento automático de textos especializados, pesquisadores

têm proposto um tipo particular de extensão para as wordnets: as wordnets terminológicas (p.ex.: ArchiWordNet, JurWordNet, etc.). Essa extensão é, em geral, construída com a extração manual de informações contidas em recursos estruturados (p.ex.: dicionários, glossários, etc.). Diante da escassez desse tipo de recursos estruturados e da lenta coleta manual dessas informações, sobretudo para o português, Di Felippo (2010) propôs uma metodologia para o desenvolvimento de wordnets terminológicas que se caracteriza pela extração automática do conhecimento a partir de recursos especializados não-estruturados (corpora). Essa metodologia está sendo validada com a construção de uma wordnet terminológica (ou terminet) do domínio da Educação a Distância (EaD) em português do Brasil (WordNet.EaD), no âmbito do projeto TermiNet (2009-2011) (FAPESP 2009/06262-1/ CNPq 471871/2009-5). Uma das etapas do desenvolvimento de uma terminet previstas no projeto é a extração automática de termos sinônimos "potenciais", uma vez que o estatuto de sinônimo será validado em etapa posterior por especialistas do domínio da EaD. A extração automática dos candidatos a sinônimos que poderão formar os synsets da WordNet.EaD pauta-se em três abordagens: duas linguísticas e uma estatística (CABRÉ et al., 2005). Nas abordagens linguísticas, utilizam-se dois métodos que se pautam no reconhecimento de padrões léxico-sintáticos e no reconhecimento de variações terminológicas, respectivamente. O primeiro considera a evidência de que determinadas relações semânticas podem ser expressas na superfície do discurso por padrões léxico-sintáticos específicos, por exemplo, "be another word for", "also known as", etc., para a extração da sinonímia no inglês. O segundo considera o fato de que os conceitos são frequentemente expressos na superfície do texto por realizações distintas de termos preferenciais, as chamadas variantes terminológicas, por exemplo, as variações ortográficas (p.ex.: fotogerador/ foto-gerador/ foto gerador), morfológicas (p.ex.: cinematógrafo/ cinema), lexicais (p.ex.: método botton-up/ abordagem botton-up) e morfossintáticas (p.ex.: cromatografia em coluna/ cromatografia de coluna). Já na abordagem estatística, utiliza-se o método distribucional (FIRTH, 1957; HARRIS, 1968), que assume que similaridade distribucional maior entre os itens lexicais implica probabilidade maior de essas unidades serem sinônimas. Neste trabalho, a discussão focalizará apenas a utilização dos dois métodos linguísticos a partir de corpus no projeto TermiNet, enfatizando os subsídios linguísticos (conjunto de padrões e tipologia de variações), as etapas de processamento do corpus para a extração, os resultados obtidos e as vantagens e desvantagens de cada método linguístico na construção da WordNet.EaD.

Palavras-Chaves
1 - Léxicos
2 - Terminologia
3 - Sinonímia
4 - Wordnets

Palavras-Chaves
1 - Léxicos
2 - Terminologia
3 - Relações de sentido
4 - Wordnets
5 - Redes semânticas
6 - Bases de dados lexicais

fechar imprimir

I Congresso Internacional de Estudos do Léxico