A anotação de um corpus para o aprendizado supervisionado de um modelo de SN

Maria Claudia de Freitas[1], Milena Uzeda-Garrão[1], Claudia Oliveira[3], Cícero Nogueira dos Santos[2], Maria Cândida Silveira[1]

[1]Departamento de Letras e [2]Departamento de Informática
Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro

[2]Departamento de Engenharia de Sistemas
Instituto Militar de Engenharia, Rio de Janeiro

claudiaf@let.puc-rio.br, migarrao@uol.com.br, cmaria@centroin.com.br, nogueirati@yahoo.com.br, mariacan@uol.com.br

Abstract. *This paper describes the construction of the main linguistic resource associated with the task of training an automatic NP extractor: an annotated corpus. The annotation is specifically designed to be used in the training phase of a rule learning algorithm: Transformation Based error-driven Learning (TBL). The NP model used in the annotation has been developed to overcome performance related problems encountered when the target language changed from English to Portuguese.*

Resumo. *Esse artigo descreve a construção do recurso lingüístico mais importante para a tarefa de treinamento de um extrator automático de SNs: um corpus etiquetado. A etiquetagem é especialmente projetada para o uso na fase de treinamento de um algoritmo de aprendizado de regras: Transformation Based error-driven Learning (TBL). O modelo de SN utilizado foi desenvolvido para superar problemas de desempenho encontrados quando a língua alvo foi transposta do inglês para o português.*

1. Introdução

Durante a última década, o Aprendizado de Máquina (AM) tem demonstrado ser uma ferramenta bastante eficaz na viabilização de tarefas lingüísticas, que de outro modo seriam impossibilitadas devido à enorme quantidade de tempo e mão-de-obra necessários. AM tem sido aplicado a problemas centrais em PLN, como etiquetagem morfossintática, identificação de sintagmas nominais, análise sintática, total e parcial, desambiguação da ligação do sintagma preposicional, etiquetagem de atos de fala, correção ortográfica, análise sintática parcial, desambigüação do significado das palavras, identificação dos limites das sentenças e determinação de papéis semânticos, para citar os problemas mais ativos nos últimos anos.

Na extração de informações a partir de textos, a idéia é que o sistema saiba escolher seqüências de palavras com alto poder discriminatório e potencial informativo. Para isso os sintagmas nominais (SN) apresentam-se como candidatas naturais, pois, de um ponto de vista lingüístico, elas tipicamente carregam significado substantivo, desempenham papéis semânticos e geralmente trazem o tema do enunciado.

Esse artigo descreve a construção do recurso lingüístico mais importante para a tarefa de treinamento de um extrator automático de SNs: um corpus etiquetado. A

etiquetagem é especialmente projetada para o uso na fase de treinamento de um algoritmo de aprendizado de regras: Transformation Based error-driven Learning (TBL) [Brill, 1995]. O modelo de SN utilizado foi desenvolvido para superar problemas de desempenho encontrados quando a língua alvo foi transposta do inglês para o português.

As técnicas de AM supervisionado exigem como entrada um corpus de treino com exemplos corretamente identificados do problema que se deseja aprender a resolver. De acordo com os experimentos descritos em [Ngai e Yarowsky, 2000], na identificação de SNs em textos em inglês é mais vantajoso utilizar recursos humanos para fazer a anotação do corpus e utilizá-lo para treinar um identificador de SNs do que utilizar recursos humanos para criar manualmente regras de transformações para uma gramática de identificação. Dentre as vantagens listadas no trabalho por Ngai e Yarowsky, destacam-se:

i) a aquisição distribuída de conhecimento, pois com a utilização de AM fica mais fácil a combinação de esforços de um grupo de pessoas. Corpora de treino criados por pessoas diferentes podem ser combinados facilmente para formarem um corpus maior, como o caso da nossa experiência. Em contraste, é muito difícil, ou quase impraticável, a combinação de listas de regras criadas manualmente por pessoas diferentes;
ii) a robustez do conhecimento empíricas, pois o desempenho de sistemas que utilizam regras codificadas manualmente tende a apresentar uma maior variação, enquanto que os resultados de sistemas treinados com corpora anotados são mais uniformes;
iii) independência dos mecanismos de inferência, pois, uma vez construído um corpus de treino, o aprendizado pode ser realizado por diversas técnicas, e subseqüentes progressos nos algoritmos de treinamento podem trazer melhorias nos resultados sem a necessidade de alterações no corpus. Em contraste, o desempenho obtido por um conjunto de regras codificado manualmente é definitivo, a não ser que haja uma revisão humana das regras.

A anotação distribuída do corpus exigiu a definição de um modelo de SN consistente, que pudesse ser compartilhado entre os anotadores. As duas principais motivações para o novo modelo são: i) estender o conceito de chunk de SN [Abney, 1991] para abranger um conjunto mais significativo de SNs no PB e ii) restringir esse mesmo conjunto para incluir apenas SNs lexicais – aqueles cujo núcleo é uma palavra lexical. O trabalho situa-se na área de Terminologia e Lexicografia Computacional, por isso a importância dos SNs lexicais, em detrimento dos "mais gerais" – cujo núcleo pode ser pronominal ou elíptico. Nosso objetivo não é a extração de SNs para a elaboração de árvores sintáticas mas sim para a identificação de termos enquanto unidades de informação.

O restante do artigo está organizado da seguinte maneira: na seção 2, explicitamos a definição de SN utilizada na etiquetagem; na seção 3, apresentamos, em linhas gerais, o aprendizado baseado em transformações; a seção 4 descreve o processo empregado na derivação do corpus de aprendizado; a seção 5 apresenta alguns resultados obtidos no aprendizado de regras e na seção 6 discutimos esses resultados e futuros desdobramentos da pesquisa.

2. O Modelo SNr

De acordo com [Koch, 1985], um sintagma consiste em um conjunto de elementos que constituem uma unidade significativa dentro da sentença e que mantêm entre si relações

de dependência e de ordem. Organizam-se em torno de um elemento fundamental, denominado núcleo. Quando o núcleo for um verbo, tem-se um sintagma verbal (SV) e, quando o núcleo for um nome, um SN. Funcionando como modificador de um SN ou de um SV, temos o sintagma preposicional (SP), que combina preposições e substantivos. Para o português, duas propostas interessantes de delimitação formal de SN são as de [Mateus, 1994] e [Perini, 1995].

Segundo Perini, "o SN pode ser definido de maneira muito simples: é o sintagma que pode ser sujeito de alguma oração" [Perini, 1995:92]. Com respeito à sua estrutura interna, a definição de SN máximo de Perini possui um forte traço posicional. Visto que as possibilidades de variação da ordem interna dos componentes de um SN são reduzidas, o esquema formal do SN máximo pode ser dado delimitando-se uma área à esquerda e uma área à direita do núcleo. À esquerda situam-se determinantes, possessivos, reforços, quantificadores, numeradores e um conjunto limitado de modificadores pré-núcleo. À direita encontram-se os modificadores, com uma certa estruturação funcional, mas nesse ponto Perini admite que a descrição mais profunda dessa estrutura ainda se encontra em pesquisa. Embora atraente, a proposta de Perini apresenta maior riqueza no detalhamento dos elementos à esquerda do núcleo, justamente aqueles de menor conteúdo informativo.

Para [Mateus, 1994], a categoria sintática SN é a projeção de um nome. Sua estrutura interna abrange um núcleo obrigatório e dois tipos de constituintes opcionais: complementos e especificadores. O núcleo pode ser ocupado por um substantivo ou pronome; o complemento pode ser composto por sintagmas adjetivais, preposicionais, oracionais e epítetos (ou apostos); os especificadores podem ser determinantes, quantificadores e expressões qualitativas.

O modelo proposto neste trabalho, doravante SNr (*Sintagma Nominal reduzido*), aproxima-se da definição de Mateus no que concerne à estrutura interna do sintagma – núcleo (+complementos) (+especificadores) – e se distingue deste principalmente pelas características do núcleo. Tendo em vista a perspectiva lexicográfica-computacional adotada, interessada principalmente na identificação de unidades de informação, consideramos SNr apenas aqueles cujo núcleo é um substantivo, o que é compatível com a definição de *SN lexical* de [Radford, 1981 apud Crystal, 1988]: os SNs lexicais, diferentemente das anáforas e dos SNs pronominais, são livres em todas as posições da sentença, isto é, sua referência é tipicamente independente dos outros SNs.

2.1. Núcleo

O núcleo é o elemento fundamental do SN, determinando a concordância interna da expressão. Tradicionalmente, tanto nomes como pronomes podem ser núcleos de SNs, no entanto, o conceito de SN lexical dá ao nome a exclusividade dessa posição.

A primeira peculiaridade do SNr é a exigência de um núcleo unitário. Ou seja, não basta que o núcleo seja um substantivo, mas é preciso que ele seja composto por um único substantivo. Assim, por exemplo, SNs cujo núcleo são nomes coordenados serão tratados aqui como dois SNrs distintos, diferentemente do que propõe [Mateus, 1994] . Uma motivação para esta escolha está na dificuldade de percepção da coordenação quando se trata de SNs com estruturas complexas.

Embora o exemplo (1a) de [Mateus, 1994:185] para SNs com núcleos coordenados não apresente problemas na segmentação ("papel e caneta"), a adição de complementos e

modificadores aos mesmos substantivos dificulta esta tarefa, como mostra a seguinte variante (1b) ("papel que seja pautado como antigamente e todas as suas canetas"). De acordo com o modelo SNr, encontram-se, em (1a) e em (1b), dois SNrs.

(1a) Dá-me $[papel]_{SNr}$ e $[caneta]_{SNr}$ para escrever uma carta.
(1b) Dá-me $[papel]_{SNr}$ que seja pautado como antigamente e $[todas\ as\ suas\ canetas]_{SNr}$ para escrever uma carta.

A segunda motivação para a o núcleo unitário é que, se a meta da identificação dos SNrs é auxiliar a tarefa de identificação de unidades de informação, parece mais produtivo lidar com a coordenação de modo a explicitar a presença de duas ou mais dessas unidades. No entanto, essa opção não é livre de problemas. No exemplo (1c), ao extrair 2 SNrs ("filmes" e "comerciais bucólicos") não é possível a leitura ("filmes bucólicos" e "comerciais bucólicos"), uma das duas alternativas disponíveis nessa construção ambígua. No entanto, a situação é mais problemática quando há discordância morfossintática nessa segmentação (de número em 1d e de gênero em 1e).

(1c) Ela é mais sórdida, mais contrastante com as radiosas imagens perpetuadas em **filmes e comerciais bucólicos** sobre refrigerantes, cigarros e «fast food» .
(1d) Paula Milhim Monteiro Alvarenga, 27, deverá ser acusada de participar de orgias na frente das crianças e de usá-las em sessões de **filmes e fotos pornográficos** .
(1e) Ele está sempre com o macacão completamente vestido, **barba e cabelo impecáveis**.

Pronomes no núcleo

Outra peculiaridade do núcleo do SNr é decorrência de sua identificação com o SN lexical. Pronomes pessoais, pronomes substantivos e numerais não são considerados núcleo de um SNr - são descartados por terem referência anafórica a um outro elemento lexical ou oracional no discurso. Assim, em (1f,g,h), "ela" (nominativo), "outras" e "isso" são exemplos de pronomes substantivos excluídos do SNr.

(1f) ela_PROPESS atuará_V junto=a_PREP **o_AR Conselho=Monetário=Nacional_NPROP**
(1g) puderam_VAUX visitar_V los_PROPESS **outras_PROADJ vezes_N**
(1h) Felizmente_ADV isso_PROSUB não_ADV foi_V necessário_ADJ

Numerais no núcleo

Para classificar os numerais, tomamos como base a seguinte descrição de [Azeredo, 2000:120]: "O numeral é sempre constituinte de um sintagma nominal, ora ocupando a posição de núcleo – numerais fracionários e multiplicativos –, ora ocupando a posição de termo adjacente – numerais cardinais e ordinais. Colocado após o substantivo, sempre na mesma forma, o numeral cardinal produz sentido ordinal: página seis (= sexta página); item dez (= décimo item)".

O primeiro problema encontrado na classificação dos numerais diz respeito à sua etiquetagem morfossintática. No corpus derivado do Mac-Morpho (descrito na seção 4) a anotação morfossintática encontrada é N se o numeral encontra-se em posição de núcleo de SN, e NUM se estiver como modificador de um substantivo. Isso fica claro (1i), onde "três" é anotado como N e "uma" é anotado como NUM.

(1i) de_PREP três_N para_PREP **uma_NUM vez_N por_PREP semana_N**

Nossa decisão foi por alterar essa anotação e rotular consistentemente todos os numerais com NUM. Assim sendo, ao encontrarmos um numeral na posição de núcleo de um SN, este deve ser descartado por não representar SN lexical, como em (1j).

(1j) apenas_PDEN três_NUM estavam_V presentes_ADJ a_PREP **a_ART sessão_N**

Nas datas por extenso o numeral acaba por funcionar como núcleo, o que faz com que seja descartado do SNr. No exemplo (1k), o numeral "30" estaria modificando um núcleo elíptico "dia", na ausência deste.

(1k) fica_V em=vigor_ADV até_PREP 30_NUM de_PREP **dezembro_N de_PREP este_PROADJ ano_N**

[Mateus 1994] considera que em (1l), "metade dessa quantia" contém um quantificador com estrutura [QUANT+de][1], assim como, em (1m), "vários dos empresários". Se tomássemos esse ponto de vista, os quantificadores em construções como essas seriam incluídos no SNr. Entretanto, julgamos mais conveniente concordar com a análise proposta pelo PALAVRAS [Bick, 2000], em que "metade" e "vários" seriam núcleos dos SNs destacados. Sendo assim, fica valendo o critério do núcleo lexical, portanto "metade de" e "vários de" são descartados.

(1l) com_PREP metade_NUM de_PREP **essa_PROADJ quantia_N**
(1m) estiveram_V detidos_PCP vários_PROSUB de_PREP **os_ART empresários_N**

No caso em que o numeral é seguido do símbolo de porcentagem ("%"), optamos por manter o conjunto NUM+% como um único numeral. Para manter a consistência com a nossa concepção de núcleo, consideramos distintos os casos (1n) e (1o). O exemplo (1n) é da mesma natureza que o (1l), por isso deve ser descartado o numeral. Já em (1o), o SNr tem "juros" como núcleo e o numeral é um pós-modificador.

(1n) **resposta_N** foi_V afirmativa_ADJ em_PREP 28=%_NUM de_PREP os_ART casos_N
(1o) a_**ART URV_N** mais_PROADJ **juros_N de_PREP 3=%_NUM a_PREP o_ART ano_N**

Voltaremos a discutir a anotação dos numerais como complemento e especificador.

2.2 Complementos

Segundo [Mateus, 1994], os complementos de um SN podem ser sintagmas adjetivais, preposicionais e oracionais, e epítetos (apostos). Desses, são considerados complementos na estrutura interna do SNr apenas sintagmas adjetivais e preposicionais.

Um aposto explicativo de um SNr é considerado externo a ele, isto é, é considerado um outro SNr, desde que seu núcleo seja um substantivo (2a).

(2a) **Vicente=Paulo=da=Silva_NPROP**, **o_ART Vicentinho_NPROP**, **presidente_N de_PREP o_ART sindicato_N**, havia_VAUX se_PROPESS responsabilizado_PCP por_PREP **o_ART uso_N de_PREP o_ART caminhão_N**

No caso dos complementos que são sintagmas oracionais, extraem-se apenas os SNrs internos à oração, como (2b). Em uma possível variante, em que o complemento fosse um sintagma preposicional, seriam identificados os SNrs como em (2c).

(2b) não_ADV há_V **qualquer_PROADJ evidência_N** de_PREP que_PRO-KS a_ART inflação_N esteja_VAUX voltando_V
(2c) não_ADV há_V **qualquer_PROADJ evidência_N de_PREP chuva_N**

[1] A estrutura [QUANT + de] corresponde à presença de um elemento quantificador seguido imediatamente pela preposição "de". No exemplo "metade dessa quantia", teríamos o quantificador "metade" seguido pela preposição "de" ("metade de essa quantia").

Numerais como complementos

Em posição pós-núcleo, o numeral não apresenta problemas, sendo classificado como integrante do SNr:

(2d) Mesquita_NPROP defendeu_V **a_ART resolução_N 1.401_NUM**
(2e) **a_ART ampliação_N** ... deve_VAUX ser_VAUX terminada_PCP em_PREP 1995_NUM

2.3 Especificadores

Ainda de acordo com [Mateus, 1994] os especificadores podem ser determinantes e quantificadores. Os determinantes compõem uma classe limitada de elementos que precedem o nome: os artigos e pronomes adjetivos possessivos, demonstrativos e indefinidos. No modelo proposto, os determinantes são considerados parte integrante do SNr, como nos exemplos (3a).

(3a) segundo_PREP **a_ART mesma_PROADJ pesquisa_N**

Com respeito aos quantificadores, assim como os especificadores, a maioria se comporta, simplesmente, como pré-modificadores. O exemplo (3b) ilustra os casos mais comuns. A estrutura [QUANT+de], que apresenta problemas na anotação, foi resolvida conforme a seção 2.1 acima.

(3b) Benito=Mussolini_NPROP fez_V "**algumas_PROADJ coisas_N boas**_ADJ em_PREP a_ART Itália_NPROP"

Numerais como especificadores

Em posição pré-núcleo, não há dúvidas na maioria das situações em que o numeral está dentro do SN, como em:

(3c) existem_V **350_NUM índios_N terenas_ADJ e_KC guaranis_ADJ**

Nas situações em que o numeral ocorre acompanhando uma unidade de medida, de tempo ou monetária, ele também é considerado como parte integrante do SNr, pois seria igualmente um modificador pré-nominal :

(3d) cerca=de_PREP **70_NUM km_N de_PREP Petrolina_NPROP**
(3e) **o_ART macaco_N** , com_PREP **12.000_NUM anos_N** e_KC **25_NUM kg_N**
(3f) fazer_V **hora_N extra_ADJ de_PREP quatro_NUM horas_N diárias_ADJ**
(3g) **126_NUM reais_N** e **KC 50_NUM centavos_N** é_V **o_ART valor_N de_PREP a_ART multa_N**

Quando o numeral vem acompanhado de uma unidade de medida, sem que haja separação entre os *tokens*, consideramos um caso de pré-modificador. Então, em (3h), "19" é um modificador de "horas", que é o núcleo. Observa-se que o grupo "20h" já é um só token, rotulado com N. O mesmo pode ser dito para unidades como km, kg, etc.

(3h) fica_V aceso_PCP de_PREP **as_ART 19h_N** a_PREP **as_ART 20h_N**

2.4 Outras considerações

Continuidade do SNr

O SNr é necessariamente contínuo, o que exclui casos como (4a), onde o advérbio intercorrente desmembrou o SN "a elevação da margem extra..." em dois SNrs.

(4a) Isso_PROSUB explica_V **a_ART elevação_N**, principalmente_ADV, de_PREP **a_ART margem_N extra_ADJ para_PREP ocorrências_N excepcionais_ADJ**

Expressões Multi-vocabulares

O corpus MacMorpho vem com alguns tipos de expressões multivocabulares (EMV) - ou polilexicais - reconhecidas, como nomes próprios, advérbios (em=geral), preposições (cerca=de), pronomes (os=quais), numerais (730=mil) e palavras denotativas (quer=dizer). As EMVs verbais (dar=certo) não são etiquetadas como um só item, o que faz com que alguns SNrs sejam extraídos impropriamente. No exemplo (4b) seria mais consistente extrair o NP "as Gardenberg sisters", considerando a expressão "entrou em contato" uma EMV verbal. No entanto, sem a anotação apropriada só é possível extrair "contato com as Gardenberg sisters"

(4b) entrou_V em_PREP **contato_N com_PREP as_ART Gardenberg_NPROP sisters_N**

Particípio

O particípio é tradicionalmente considerado uma das formas nominais do verbo. [Pimenta-Bueno, 1986] utiliza a denominação "forma deverbal [V+do]" para caracterizar a não uniformidade do comportamento dessas formas quanto ao seu caráter verbal ou nominal. Por um lado, há uma série de propriedades do particípio que os aproxima dos adjetivos e os difere de verbos. Por outro lado, há também propriedades de formas V+do que as aproximam de verbos e as diferem de adjetivos.

Pimenta-Bueno propõe a categorização dos particípios de acordo com as seguintes hipóteses: nos contextos posteriores aos verbos "ter" e "haver" eles são verbos; em todos os outros contextos são adjetivos, exceto nos contextos V__N ("Zé foi nomeado síndico") e V__Adj ("Zé foi declarado incompetente"). Nesses casos, a forma é classificada como *particípio passivo*. Tendo em vista essa análise, consideramos o particípio como forma verbal ou particípio passivo sempre que estiver em contextos posteriores a verbos auxiliares ou de ligação, como em (4c). Nos outros casos, o particípio é um adjetivo e pode ser complemento de um SNr, como em (4d).

(4c) **dois_NUM soldados_N israelenses_ADJ** foram_VAUX atingidos_PCP por_PREP tiros_N
(4d) **as_ART pequenas_N agremiações_ADJ formadas_PCP a=partir=de_PREP a_ART divisão_N de_PREP o_ART PLD_NPROP**
(4e) o_ART diálogo_N contrasta_V com_PREP **o_ART apoio_N dado_PCP por_PREP o_ART governo_N**

3. O Aprendizado Baseado em Transformações

A idéia central do algoritmo TBL [Brill, 1995] está na geração de uma lista ordenada de regras que corrigirão progressivamente erros de anotação em um corpus de treino, produzidos por uma imprecisa classificação inicial [Santos, 2005]. O TBL é um algoritmo guloso, visto que, a cada iteração, a regra escolhida para ingressar na lista de regras aprendidas é aquela que provocar maior redução de erros na classificação atual dos itens do corpus de treino.

A anotação de um corpus é vista como um problema de categorização de cada um de seus itens lexicais. As categorias-alvo serão representadas por um conjunto de etiquetas. No caso da anotação de SNs, essas etiquetas são: I – dentro de SN e O – fora de SN. As entradas do algoritmo são: i) duas instâncias de um corpus, uma corretamente anotada com as etiquetas da categoria alvo e outra que permanece não anotada; ii) um classificador básico (*baseline system*), utilizado para atribuir uma classificação inicial aos itens do corpus, geralmente baseada em freqüências verificadas no corpus de treino; iii) um conjunto de moldes de regras (*templates*), que indicam as combinações de traços,

na vizinhança de um item, que possam determinar a classificação desse item. Os moldes são os elementos que provocam maior impacto no comportamento do aprendizado com TBL, visto que eles devem exprimir exatamente as informações contextuais importantes para o problema em questão.

O aprendizado começa com a atribuição de uma classificação inicial aos itens do corpus de treino. Em seguida, a classificação corrente é comparada com a classificação correta e, em cada ponto em que houver erro, todas as regras que o corrigem serão geradas a partir da instanciação dos moldes de regras com o contexto do item atualmente analisado. Normalmente, uma regra irá corrigir alguns erros, mas também poderá provocar outros pela alteração de itens que estavam classificados corretamente. Dessa forma, depois de computados os valores para a pontuação de todas as regras candidatas, a regra que tiver maior pontuação será selecionada e colocada na lista de regras aprendidas. A regra selecionada é então aplicada ao corpus, e o processo de geração de regras será reiniciado enquanto for possível gerar regras com pontuação acima de um limite especificado. Na classificação de novos textos com uso dessa técnica necessitamos apenas submeter o texto ao classificador inicial, e logo em seguida aplicar a lista de regras na seqüência em que foram aprendidas. A característica mais atraente do TBL é que as regras aprendidas são interpretáveis pelos humanos, em contraste com a saída de etiquetadores estocásticos.

Na identificação de SNs básicos do inglês com TBL, [Ramshaw e Marcus, 1995] obtiveram 92,3% de abrangência e 91,8% de precisão utilizando um corpus de treino contendo 200 mil itens e um corpus de teste com 50 mil itens. Em [Megyesi, 2002] são reportados resultados de 99,39% de precisão e 99,12% de abrangência na identificação de SNs básicos em textos em lingua sueca com o uso de TBL, sendo que a identificação é feita como parte da tarefa de análise sintática parcial. Megyesi usou um corpus de treino contendo 200 mil itens e um corpus de teste contendo aproximadamente 100 mil itens. Na seção 5, apresentamos alguns dos nossos resultados.

4. A derivação do corpus

Os corpora de treino e teste utilizados nesse estudo foram derivados do Mac-Morpho [Marchi, 2003], um corpus de 1,1 milhão de palavras retiradas do jornal brasileiro Folha de São Paulo[2], no ano de 1994, e etiquetado morfossintaticamente com o conjunto de etiquetas do projeto Lacio-Web.

Para a geração das etiquetas SN, seria necessária a identificação de todos os SNs presentes no Mac-Morpho, o que seria impraticável manualmente. A melhor forma encontrada para automaticamente se realizar essa tarefa foi a utilização do analisador sintático PALAVRAS [Bick, 2000]. Os SNs foram então identificados, visto que a anotação do PALAVRAS é rica o suficiente para prover as informações sintáticas que delimitam os constituintes da sentença.

Cada SN reconhecido no corpus Mac-Morpho recebeu a etiqueta SN. Em seguida, um grupo de quatro lingüistas revisou um fragmento de aproximadamente 140 mil tokens do corpus para eliminar erros da etiquetagem automática e discrepâncias entre o SN concebido para o PALAVRAS e o SNr. O percentual de tokens corrigidos,

[2] O Mac-morpho está disponibilizado via web pelo projeto Lacio-Web (www.nilc.icmc.usp.br/lacioweb/), do Núcleo Interinstitucional de Lingüística Computacional (NILC).

entre etiquetas SN (a grande maioria das correções), etiquetas morfossintáticas e adições de tokens para corrigir erros no enunciado, foi de aproximadamente 7%.

5. Experimentos e resultados

Para produzir um conjunto de regras identificadoras de SNrs, no padrão TBL, utilizamos a ferramenta *catTBL*, apresentada em [Santos, 2005]. Os índices de avaliação dos experimentos realizados foram: precisão geral (total de itens classificados corretamente / total de itens); precisão (total de SNrs identificados corretamente / total de SNrs identificados); abrangência (total de SNrs identificados corretamente / total de SNrs no corpus) e $F_{\beta=1} = ((\beta^2+1)* \text{precisão}*\text{abrangência})/(\beta^2 * \text{precisão} + \text{abrangência})$.

Nossos resultados, apresentados na tabela 1, foram obtidos utilizando o conjunto de moldes de regras que conseguiram o melhor desempenho relatado no trabalho de Santos, incluindo moldes lexicalizados e TAs com restrições. As experiências de Santos foram feitas sobre um corpus semelhante ao nosso (MacMorpho+PALAVRAS),com maior número de tokens, mas sem a correção manual que caracterizou os SNrs. Essa diferença de tamanho nos levou a optar pela técnica de validação cruzada para a avaliação dos resultados com o corpus SNr. Para tal, foram geradas 10 amostragens diferentes do corpus, cujas sentenças foram escolhidas de forma aleatória. Em cada amostragem, o corpus foi particionado em 70% para treinamento e 30% para teste. As médias, mínimos, máximos e desvios padrões para a precisão, abrangência e $F_{\beta=1}$encontram-se na tabela.

Índice	Resultados: corpus SNr – 93k				Resultados: corpus Santos – 200k	Resultados: corpus Santos – 500k
	Méd.	Min.	Máx.	D.P.		
Precisão	83,8%	83,0%	84,7%	0,57	84,6%	85,9%
Abrangência	84,2%	83,4%	84,9%	0,50	85,2%	86,6%
$F_{\beta=1}$	84,0%	83,5%	84,7%	0,36	84,9%	86,2%

Tabela 1: Resultados da aplicação das regras aprendidas.

Os resultados foram bastante uniformes. Como podemos ver, o desvio padrão foi abaixo de zero, o que indica que provavelmente o corpus é representativo, ou seja, tem bom tamanho. Comparando-se com os resultados de Santos, houve uma ligeira perda de desempenho, mais marcante comparando-se aos seus resultados com o corpus de 500k. A reprodução do treinamento com um corpus SNr tão grande exigiria um enorme esforço de re-anotação manual.

6. Considerações finais

Do ponto de vista terminológico e lexicográfico, a definição de sintagma nominal deve ser cercada de restrições que dialogam com a Teoria da Informação e excluem aquelas unidades lingüísticas com um suposto perfil nominal, mas que seriam mais precisamente identificadas como recursos discursivos (como, por exemplo, o pronome anafórico), desprovidos de teor informacional distintivo.

Portanto, a delimitação teórica e aplicação prática de SNr — ou seja, da mínima unidade lingüística com alto poder discriminatório — é um salto qualitativo para domínios computacionais que dependem da delimitação de unidades de

informação lingüísticas, como Recuperação de Informações, assim como algumas áreas importantes do Processamento de Linguagem Natural como a Sumarização Automática.

Do ponto de vista experimental, obtivemos resultados de treinamento com o corpus anotado de acordo com o modelo SNr, cuja qualidade parece bastante compatível com o estado da arte para SNs não básicos do português. Os próximos passos em nosso trabalho envolvem a análise, de um ponto de vista terminológico, dos SNrs identificados utilizando as regras TBL geradas.

7. Referências

Abney, S. (1991) Parsing by Chunks, em "Principle-Based Parsing: Computation and Psycholinguistics", editors Robert C. Berwick, Steven P. Abney e Carol Tenny, Kluwer Academic Publishers, Boston, p 257-278.

Azeredo, J. C. (2000) Fundamentos da Gramática do Português, Jorge Zahar, São Paulo.

Bick, E. (2000) "The Parsing System Palavras: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework", Aarhus University, Dinamarca.

Brill, E. (1995) "Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging". *Computational Linguistics*, 21(4):543-565.

Crystal, D. (1988) *Dicionário de Lingüística e Fonética.* Jorge Zahar Editor.

Koch, I. e Silva, M. C. (1985) Lingüística aplicada ao português: sintaxe. Cortez, São Paulo.

Lobato, L. (1986) Sintaxe Gerativa do português: da teoria padrão à teoria da regência e ligação. Editora Vigília, Belo Horizonte.

Mateus, M.H., Brito, A.M., Duarte, I., Faria, I.H. (1994) Gramática da língua portuguesa. 4ª edição, Ed. Caminho, Lisboa.

Marchi, A. R. (2003) "Projeto Lacio-Web: Desafios na Construação de um Corpus de 1,1 Milhão de Palavras de Textos Jornalísticos em Português do Brasil", em 51º. Seminário do Grupo de Estudos Lingüísticos do Estado de São Paulo, São Paulo.

Megyesi, B. (2002) Shallow Parsing with PoS Taggers and Linguistic Features, *Journal of Machine Learning Research*, v. 20, p. 639-668.

Ngai, G. e Yarowsky, D. (2000) Rule Writing or Annotation: Cost-efficient Resource Usage for Base Noun Phrase Chunking, em *Proceedings of the 38th Annual Meeting of the ACL*, Association for Computational Linguistics, Hong Kong.

Perini, M. (1995) Gramática Descritiva do Português. Editora Ática, São Paulo,

Pimenta-Bueno, M. (1986) As formas [V+do] em português: um estudo de classes de palavras. *D.E.L.T.A.*, 2(2):207-229.

Ramshaw, L. e Marcus, M. (1995) Text Chunking Using Transformation-Based Learning, em *Proceedings of the Third Workshop on Very Large Corpora*, editores D. Yarovsky e K. Church, Association for Computational Linguistics, EUA, p 82-94.

Santos, C. N. (2005) *Aprendizado de máquina na identificação de sintagmas nominais: o caso do português brasileiro*. Dissertação de Mestrado, IME, Rio de Janeiro.