UM MODELO DE SINTAGMA NOMINAL LEXICAL NA RECUPERAÇÃO DE INFORMAÇÕES

Claudia OLIVEIRA (IME-RJ) [1]
Maria Cláudia de FREITAS (PUC-Rio) [2]

RESUMO: O objetivo desse trabalho é mostrar o importante papel do Processamento de Linguagem Natural no desenvolvimento e no aprimoramento das tecnologias de Recuperação de Informação. Particular atenção é dispensada à modelagem e à extração de sintagmas nominais para indexação e busca de documentos. Conceitos básicos de Recuperação de Informação são apresentados para motivar os aspectos lingüísticos da extração automática de termos. Um modelo de sintagma nominal – chamado sintagma nominal lexical – foi elaborado tendo em vista aplicações que lidem essencialmente com o conteúdo da informação, como recuperação de informação e indexação. Por isso, uma de suas principais características é a presença de um núcleo lexical. Experimentos com o aprendizado automático deste tipo de sintagma nominal são descritos.

ABSTRACT: This paper shows the important role of Natural Language Processing in the development and the improvement of the technologies related to Information Retrieval. Special attention is given to the formalization and the automatic extraction of noun phrases to be used in document indexing and search. Some basic concepts of Information Retrieval are presented in order to motivate the linguistic aspects involved in automatic term extraction. A model of noun phrase – the lexical noun phrase – was defined to conform to the needs of computer applications that deal essentially with the information content of nominal expressions, such as information retrieval applications. As such, one of the main restrictions of the defined noun phrase is that it must have a lexical head. Some experiments with machine learning of this type of noun phrase are presented.

1. Introdução

Recuperação de Informações lida com a representação, o armazenamento, a organização e o acesso a itens de informação. Na recuperação 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. O objetivo desse trabalho é mostrar o importante papel do Processamento de Linguagem Natural no desenvolvimento e no aprimoramento das tecnologias de Recuperação de Informação. Particular atenção é dispensada à modelagem e à extração de sintagmas nominais para indexação e busca de documentos.

Desde a última década o Aprendizado de Máquina (AM) tem sido uma ferramenta valiosa para a realização de tarefas centrais do Processamento Automático de Linguagem Natural (PLN), como etiquetagem morfossintática (Brill,1995; Ratnaparkhi,1998), identificação de sintagmas nominais básicos (*base noun phrase*) (Ramshaw e Marcus, 1995; Cardie e Pierce, 1998; Tjong, 2000), análise sintática superficial (Ramshaw e Marcus, 1995; Osborne, 2000; Tjong, 2002; Megyesi, 2002), entre outros, economizando tempo e mão de obra.

Na área de processamento e compreensão de textos, uma das aplicações do AM é a extração automática de termos para a indexação, com ênfase na escolha de seqüências de palavras altamente discriminatórias para um determinado tópico – os sintagmas nominais. Sendo assim, o *chunking* de SNs é uma solução rápida e robusta para a identificação de SNs em textos, sem o alto custo de uma análise sintática completa.

A aplicação de algoritmos de AM a tarefas de identificação de SN tem produzido ótimos resultados experimentais. Embora, em sua maioria, os métodos de AM sejam gerais o suficiente para serem aplicados a uma ampla variedade de línguas, existem especificidades lingüísticas que podem influenciar seu desempenho. Essas características – no nosso caso, a estrutura do SN – não têm recebido a atenção merecida porque, acreditamos, a maioria dos trabalhos desenvolvidos têm como alvo o inglês.

O modelo de SN que descrevemos foi desenvolvido para superar dificuldades decorrentes da transposição da língua-alvo do inglês para português. Em português, por exemplo, os SNs são, em média, maiores e contêm mais preposições. Nossa principal motivação é ampliar o conceito de *chunk* a fim de

[1] Departamento de Engenharia de Sistemas, Instituto Militar de Engenharia. Email: cmaria AT de9.ime.eb.br
[2] Departamento de Letras, Pontifícia Universidade católica do Rio de Janeiro. Email: claudiaf AT let.puc-rio.br

considerar um conjunto mais significativo de SNs mantendo a eficácia do método para textos em língua portuguesa.

Um dos requisitos fundamentais da modalidade supervisionada do AM é um corpus anotado (cf. seção 3.2). O modelo de SN subjacente à anotação – chamado sintagma nominal lexical – foi elaborado tendo em vista aplicações que lidem essencialmente com o conteúdo da informação, como recuperação de informação e indexação. Por isso, uma de suas principais características é a presença de um núcleo lexical.

O restante do artigo está organizado da seguinte maneira: na seção 2, apresentamos os principais conceitos da Recuperação de Informação, tendo em vista a contribuição da Lingüística nas tarefas básicas dessa aplicação computacional; na seção 3, tratamos da identificação automática dos sintagmas nominais do português segundo um modelo que privilegia as características informacionais dessas estruturas; na seção 4 tecemos algumas considerações sobre as diferenças qualitativas entre o SN e o termo, unidade de informação em um contexto de recuperação de documentos; na seção 5, apresentamos nossas considerações finais.

2. Recuperação de Informações: principais conceitos

Os sistemas computacionais de busca de documentos começaram a ser desenvolvidos já na década da 60. Naquela época, ainda dependia-se totalmente do indexador humano para a escolha das palavras chaves retiradas de cada texto. A busca tinha de ser feita dentro deste pequeno conjunto de ponteiros.

Devido ao número reduzido de "clientes" – bibliotecas e arquivos – a área não atraía grande interesse até o surgimento da *Internet*, mais especificamente da *World Wide Web*. O novo conceito de repositório universal de conhecimento, de acesso livre ou de baixíssimo custo, sem um corpo editorial central, gerou uma verdadeira explosão na produção e consumo de textos.

Os maiores problemas dessa vasta biblioteca digital estão relacionados à organização e recuperação de itens de seu acervo, portanto são questões cujas soluções encontram-se dentro do escopo da área Recuperação de Informações (RI) (Baeza-Yates e Ribeiro Neto, 1999). Com o contínuo crescimento da Internet, a área de Recuperação de Informações (RI) vem desenvolvendo métodos cada vez mais poderosos e completos de indexação e busca. Muitas dessas novas tecnologias são baseadas no Processamento de Linguagem Natural (PLN), principalmente no processamento automático de textos. De maneira geral, o objetivo da RI é possibilitar, através de uma organização apropriada dos itens de informação, a satisfação da necessidade de informação do usuário. O problema de compreensão e especificação dessa necessidade não é simples. Por exemplo:

> O usuário deseja obter documentos contendo informações sobre pesquisadores em lingüística computacional que atuam na área de recuperação de informação com produção bibliográfica recente, envolvidos com projeto de software.

A questão de como transformar essa formulação do problema em um conjunto preciso de descrições para os textos alvos não é resolvida trivialmente (Sparck Jones e Willet, 1997). As ferramentas mais conhecidas de busca na Internet exigem que a necessidade de informação seja codificada por um conjunto de palavras. Em alguns casos, estão disponíveis operações *booleanas* sobre as palavras de modo a possibilitar maior número de combinações, mas ainda assim é muito difícil para qualquer usuário exprimir em uma seqüência de palavras, todas as possibilidades que resultarão em uma resposta adequada à sua consulta.

Portanto, a ênfase dos processos de organização e recuperação deve ser na informação e não nos dados. Para a recuperação de dados, o usuário deve perguntar "que documentos contêm esse conjunto de palavras?", quando "que documentos contêm informação a respeito desse tópico?" parece ser a questão mais adequada. Nessa abordagem a noção de *relevância* é extremamente importante, pois, a partir de um tópico, não se podem classificar documentos dicotomicamente como relacionado ou não relacionado. É muito mais pertinente que os documentos sejam organizados em uma ordenação de relevância, dentro de um espectro contínuo.

Nos últimos 20 anos, RI ultrapassou os objetivos primários de indexação e busca. As pesquisas mais atuais têm se voltado para a modelagem, classificação e categorização de documentos, interface com usuários, visualização de dados, filtragem e perfil de usuário, enfim, buscam-se maneiras de aprimorar o acesso à informação que se encontra camuflada como uma agulha em um palheiro.

2.1 Manipulação do documento

O ponto de partida de um sistema de RI é a definição do conjunto de documentos disponíveis a serem representados e armazenados. Um documento é, em princípio, qualquer item contendo informação. Apesar do

crescente interesse em recuperação de informações multimídia, os documentos que contêm informações expressas em linguagem natural são considerados como foco do nosso trabalho.

Cada documento da coleção é representado, em geral, por um conjunto de palavras ou expressões, indexadas, os chamados **termos de indexação**. Um **índice** é a organização de palavras ou expressões, pela associação de cada uma a ponteiros que permitam a recuperação rápida de informação. A indexação de uma expressão parte do pressuposto de que há uma associação direta entre esta e o significado do conteúdo do documento, ou seja, espera-se que a expressão represente o documento em um certo nível.

Tradicionalmente, a escolha dos termos representantes dos textos é feita por especialistas em Ciência da Informação. Os índices eram, e na maioria das bibliotecas tradicionais ainda são, criados manualmente como hierarquias de categorização.

Com o aumento da capacidade de armazenagem e processamento dos computadores modernos, tornou-se possível incluir todas as palavras, e até mesmo seqüências de um determinado tamanho máximo, no índice. Com isso espera-se superar, com a força bruta, as dificuldades de seleção de termos em um sistema de RI.

Para determinar qual a unidade de texto que deverá ser indexada, são utilizados conceitos lingüísticos de delimitação de unidades lexicais. Ao invés de considerar como palavra as seqüências separadas por espaços em branco e pontuações, um sistema de recuperação de documentos pode optar por unidades menores – lemas – ou unidades maiores – palavras compostas e sintagmas.

A representação de um documento por meio de seu conjunto de termos é chamada em (Baeza-Yates e Ribeiro Neto, 1999) de visão lógica do documento. O documento primário, ou sua estrutura interna (segmentos como título, parágrafos, seções, etc) é processada de modo a retirarem-se espaços, pontuações, ou outros elementos gráficos que sejam irrelevantes. A partir daí, as palavras sem poder discriminatório são eliminadas – as chamadas *stopwords* – que, em geral, são palavras funcionais ou expressões muito freqüentes.

O tratamento lingüístico mais acurado pode ser empregado a partir dessa fase, com a identificação de sintagmas nominais e eliminação do restante das expressões do texto. Os SNs são extraídos com a utilização de análise sintática, muitas vezes utilizando gramáticas de SNs, em vez de gramáticas completas da língua alvo.

3. Extração automática de sintagmas nominais

As expressões de um texto podem ser classificadas de acordo com seu poder discriminatório. Na ponta inferior do espectro encontram-se as palavras gramaticais, como preposições advérbios e conjunções. As de maior poder discriminatório são, em geral, aquelas de sentido substantivo que podem realizar funções temáticas, como sujeito e objeto, e certas funções semânticas, como agente e instrumento. As expressões desse tipo são em grande parte sintagmas nominais (SNs), daí a sua importância nos estudos lingüísticos relacionados a RI. Os SNs são os indexadores mais promissores de um texto (Kuramoto, 1995).

O modelo de SN proposto neste trabalho, para ser utilizado em sistemas de RI, é denominado SN_L (*Sintagma Nominal lexical*), por privilegiar expressões substantivas autônomas e com potencial de exercer a função de termos de indexação.

3.1. O Modelo SN_L

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. 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 et al. (1994), a estrutura interna do SN 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 do SN_L aproxima-se da definição de Mateus quanto à estrutura interna do sintagma: é composto por núcleo (+complementos) (+especificadores). Por outro lado, diferentemente do modelo de Mateus et al., o núcleo do SN_L é restrito aos substantivos. Essa escolha se justifica, principalmente, tendo em vista as

necessidades da RI, cujo maior interesse está na identificação de unidades de informação, isto é, termos de alto poder discriminatório. Nosso modelo 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.

A seguir discutimos as principais características do SN_L . Essa descrição pode ser encontrada com mais detalhes em (Freitas et al., 2005).

3.1.1. 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 SN_L é a exigência de um núcleo único substantivo, não elíptico. Com relação à primeira característica, SNs cujo núcleo são nomes coordenados serão tratados aqui como dois SN_Ls distintos, diferentemente do que propõe (Mateus et al., 1994). Uma motivação para esta escolha está na dificuldade de percepção da coordenação quando se trata de SNs com estruturas complexas. Deste modo, se em (1a) não há 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 SN_L, encontram-se, tanto em (1a) quanto em (1b), dois SN_Ls, como indicado pelos colchetes.

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

Além disso, a escolha do núcleo unitário se justifica quando consideramos as tarefas de RI, pois parece mais produtivo lidar com a coordenação de modo a explicitar a presença de duas ou mais dessas unidades de informação. Porém, estamos conscientes de que essa escolha não é livre de problemas. No exemplo (1c), ao extrair 2 SN_Ls ("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 complicada 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**.

Outra peculiaridade do núcleo do SN_L é a presença obrigatória de um núcleo lexical. Pronomes pessoais, pronomes substantivos e numerais, quando exercendo a função de núcleo do sintagma, são descartados por terem referência anafórica a um outro elemento lexical ou oracional no discurso. Assim, em (1f,g), "ela" (nominativo) e "isso" são exemplos de pronomes substantivos excluídos do SN_L.

(1f) ela atuará junto a **o Conselho Monetário Nacional.**
(1g) Felizmente isso não foi necessário.

Quanto aos 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)". Assim sendo, ao encontrarmos um numeral na posição de núcleo de um SN, este deve ser descartado por não representar SN lexical, como em (1h).

(1h) apenas três estavam presentes a **a sessão**

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

(1i) fica em vigor até 30 de **dezembro de este ano.**

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 (1j) e (1k). No exemplo (1j) deve ser descartado o numeral, já em (1k), o SN_L tem "juros" como núcleo e o numeral é um pós-modificador.

(1j)**resposta** foi afirmativa em 28% de os casos
(1k)a **URV** mais **juros de 3% a o ano**

3.1.2. Complementos, Especificadores e outras Considerações

Os complementos do SN_L são sintagmas adjetivais e preposicionais; orações relativas e apostos são descartados. Os especificadores são artigos, pronomes demonstrativos, possessivos e indefinidos; ou quantificadores, como numerais.

Outra exigência é a necessidade de continuidade do SN_L, o que exclui casos como (1l), onde o advérbio intercorrente desmembrou o SN "a elevação da margem extra..." em dois SN_Ls.

(1l) Isso explica **a elevação**, principalmente, de **a margem extra para ocorrências excepcionais.**

Por fim, uma última consideração relevante diz respeito ao particípio. Tradicionalmente classificado como uma das formas nominais do verbo, o particípio será sempre um adjetivo no SN_L, exceto quando ocorrer com um verbo auxiliar, como fica claro pelo contraste dos exemplos (1m) e (1n).

(1m) **dois soldados israelenses** foram atingidos por tiros.
(1n) **as pequenas agremiações formadas a partir de a divisão de o PLD**

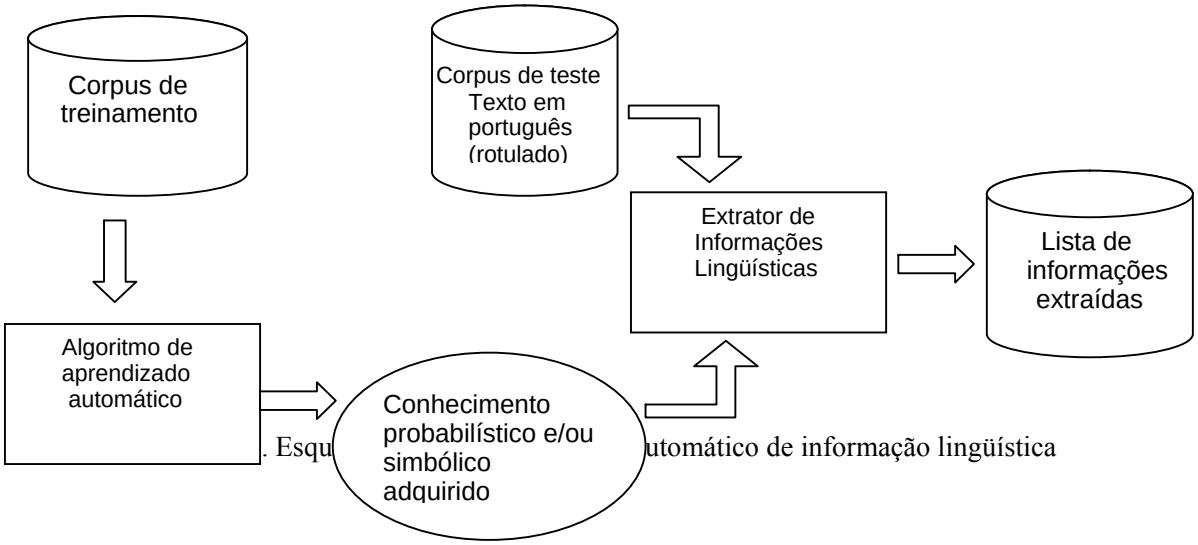
3.2. O Aprendizado Automático do SN_L

Um modelo de aprendizado automático de informação lingüística encontra-se esquematizado na figura 1. A partir de corpora de treinamento, um algoritmo de aprendizado automático sintetiza um determinado conhecimento lingüístico, seja de forma probabilística ou simbólica. Esse conhecimento passa a fazer parte de um programa extrator do tipo de informação lingüística selecionada para o aprendizado. Assim, o extrator será capaz de reconhecer, em novos textos, as características desejadas.

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 baseado em observações 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.



. Esqu utomático de informação lingüística

Realizamos experimentos no aprendizado automático do SN_L utilizando um algoritmo de aprendizado automático de regras (simbólico): Transformation Based error-driven Learning (TBL) (Brill, 1995). Os corpora de treino e teste utilizados nesse estudo foram derivados do Mac-Morpho, um corpus de 1,1 milhão de palavras retiradas do jornal brasileiro Folha de São Paulo[3], 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 SN_L. O percentual de tokens corrigidos, entre etiquetas SN (a grande maioria das correções), etiquetas morfossintáticas e adições de tokens para corrigir erros no enunciado, foi de aproximadamente 7%.

3.2.1. Experimentos e resultados

Para produzir um conjunto de regras identificadoras de SN_Ls 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 SN_Ls identificados corretamente / total de SN_Ls identificados); abrangência (total de SN_Ls identificados corretamente / total de SN_Ls no corpus) e $F_{\beta=1} = ((\beta^2+1)* \text{precisão}*\text{abrangência})/(\beta^2 * \text{precisão} + \text{abrangência})$.

Nossos resultados são apresentados na tabela 1, em comparação com as experiências de Santos, feitas sobre um corpus semelhante ao nosso (MacMorpho+PALAVRAS), com maior número de tokens, mas sem a correção manual que caracterizou os SN_Ls. Essa diferença de tamanho nos levou a optar pela técnica de validação cruzada para a avaliação dos resultados com o corpus SN_L. 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.

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

Índice	Resultados: corpus SN_L – 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. 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 SN_L tão grande exigiria um enorme esforço de re-anotação manual.

4. Do SN ao Termo

Dentre os SNs que ocorrem em um texto existe um grupo de maior importância por ser constituído do vocabulário específico de um determinado domínio de conhecimento, ou seja, sua terminologia. Termos são unidades arbitrárias de indexação, que buscam representar o conteúdo de um texto – ou de uma determinada área, correspondendo, de um ponto de vista lingüístico, aos sintagmas nominais. Portanto, o interesse da extração automática de SNs é motivado, não só para o aprimoramento da recuperação de informações, como também para a aquisição automática de terminologia e construção automática de tesauros, dicionários e ontologias.

Para satisfazer as necessidades comunicativas da melhor maneira possível, termos devem obedecer a três critérios (Jacquemin, 2001): economia, precisão, adequação. O critério de economia influencia a criação de novos termos, na direção da reutilização de material lexical em composições (justaposições e supercomposições). Por exemplo, a partir da combinação dos termos *futebol* e *vôlei*, obtém-se a especialização *futevôlei*. Por outro lado, o fator economia também se faz presente na formação de siglas e na supressão de material lexical recuperável pragmaticamente, como em *restaurante (de comida) vegetariano*.

A influência do critério de precisão é na direção oposta, ou seja, a necessidade de eliminação das ambigüidades faz com que aos termos sejam adicionados modificadores, como na adjetivação de *banco de dados* gerando *banco de dados textual*.

O critério de adequação consiste no equilíbrio entre a economia e a precisão, de acordo com o contexto, o conhecimento do domínio e o nível técnico do documento.

Assumindo que o tratamento computacional do SN em português está equacionado do ponto de vista sintático, ainda que a precisão obtida nos experimentos possa ser melhorada, passamos à descrição de cunho semântico do SN para que a identificação automática do termo possa ser empreendida.

4.1. Variações de um termo

Um termo admite variações formais. Os dois níveis lingüísticos onde essas variações mais ocorrem são a morfologia e a sintaxe.

Na morfologia, as variantes de um termo podem ser flexionais ou derivacionais. Na sintaxe, as possibilidades de variações de um termo maior que uma palavra são mais numerosas e interessantes. Em (Jacquemin, 2001) encontram-se resumidas algumas propostas para a distinção entre SNs comuns e compostos. Barkema (1994) propõe três critérios.

- *Composicionalidade*: indica o quanto do significado pode ser depreendido do conteúdo lexical das partes e da estrutura sintática do todo.
- *Colocabilidade*: indica o quanto é aceitável a substituição de um item lexical do sintagma por um sinônimo ou antônimo.
- *Flexibilidade:* indica a possibilidade de variações sintáticas ou morfológicas dentro do sintagma.

Os compostos apresentam baixa composicionalidade, baixa colocabilidade e baixa flexibilidade. Já Nunberg, Sag e Wasow (1994) apresentam seis critérios de distinção de compostos.

- *Convencionalidade* e *inflexibilidade*: os opostos à composicionalidade e à flexibilidade, de Barkema.

- *Figuração*: o grau de influência de figuras de linguagem (metáfora, hipérbole) envolvido na construção do sintagma.
- *Proverbialidade, informatividade* e *afetação*: características dos registros lingüísticos em que expressões idiomáticas são normalmente utilizadas.

O critério de flexibilidade é o mais diretamente relacionado à sintaxe e à morfologia, apesar de haver uma correlação entre propriedades semânticas e versatilidade sintática, e entre o critério de composicionalidade e o de flexibilidade.

Existem duas abordagens hegemônicas para a caracterização do grau de flexibilidade de construções sintáticas. A primeira, defendida por Barkema (1994), propõe um estudo estatístico baseado em corpus, para estabelecer a freqüência de uma construção e suas variantes, gerando o perfil de flexibilidade de uma expressão. As variações consideradas por Barkema são classificadas em *modificações externas* (por exemplo, na adjetivação de **guerra fria *incipiente***) e *modificações internas* (por exemplo, coordenação de **guerra fria** e **guerra civil** gerando ***guerra fria e civil***).

Já Maurice Gross (1988, apud Jacquemin, 2001) propõe uma caracterização introspectiva da flexibilidade sintática por meio de testes de aceitabilidade realizados pelo lingüista. Dentre as variações propostas por Gross, que devem ser possíveis para sintagmas livres, mas não para compostos, as seguintes podem ser destacadas.

- *Predicatividade*: aceitabilidade da construção atributiva correspondente. Por exemplo, **esta saia é justa** não é uma variante aceitável para a expressão **saia justa**.
- *Nominalização*: aceitabilidade da nominalização do adjetivo. Por exemplo, **a justeza da saia** não pode se referir à qualidade de uma **saia justa**.
- *Restrição de seleção no adjetivo*: aceitabilidade de outros adjetivos na mesma posição. Por exemplo, sendo **saia justa** uma situação difícil, **saia larga** não se refere a uma situação fácil.
- *Variação de número* (**saias justas**), *adjunção de advérbio* (**saia muito justa**), remoção do adjetivo (**saia**), *seleção de restrição para o substantivo núcleo* (**camisa justa**).

Um projeto de identificação e extração de termos para o português deve levar em consideração as especificidades lingüísticas, tanto gramaticais quanto lexicais, do nosso idioma. O SN_L encontra-se em uma etapa intermediária do processamento automático de grupos nominais para que sejam utilizados em sistemas de recuperação de informações como termos.

5. Considerações Finais

O processamento de linguagem natural se aplica decisivamente em sistemas automáticos de recuperação de informações. Os efeitos do uso de técnicas de PLN não são percebidos somente na melhoria da eficiência dos sistemas, mas também na viabilização de funções extremamente úteis, tais como a compilação e o enriquecimento de tesauros.

A definição de sintagma nominal deve ser cercada de restrições que dialogam com a Ciência 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 SN_L — 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.

6. Referências bibliográficas

BAEZA-YATES, R., RIBEIRO NETO, B.: *Modern Information Retrieval*. Addison Wesley, 1999.

BARKEMA, H. Determining the Syntactic Flexibility of English Idioms. In: G. Tottie and U. Fries: *English Corpus Linguistics. Papers from the 14th ICAME conference in Zürich*. Amsterdam: Rodopi 1994.

BICK, E. *The Parsing System Palavras: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, Aarhus University, Dinamarca, 2000.

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

CARDIE, C & PIERCE, D. *Error-driven pruning of treebank grammars for base noun-phrase identification*. In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, Ithaca-NY**,** p.218-224. 1998.

FREITAS, M., UZEDA-GARRÃO, M., OLIVEIRA, C., SANTOS, C., SILVEIRA, M. A anotação de um corpus para o aprendizado supervisionado de um modelo de SN. In: *Anais do XXV Congresso da Sociedade Brasileira de Compútação*, São Leopoldo, Brasil, 2005.

JACQUEMIN, C. *Spotting and Discovering Terms Through Natural Language Processing*. The MIT Press, 2001.

KURAMOTO, H.: Uma abordagem alternativa para o tratamento e a recuperação de informação textual : os sintagmas nominais. *Revista Ciência da Informação,* v.5*,* n. 2, 1996.

MATEUS, M.H., BRITO, A.M., DUARTE, I., FARIA, I.H. *Gramática da língua portuguesa.* 4ª edição, Ed. Caminho, Lisboa, 1994.

MEGYESI, B. Shallow Parsing with PoS Taggers and Linguistic Features. *Journal of Machine Learning Research 2,* 639-668, 2002.

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

NUNBERG, G., SAG, I. E WASOW, T. Idioms. *Language*. Vol. 70, No. 3. pp 491--538. 1994.

OSBORNE, M. Shallow Parsing as Part-of-Speech Tagging. In: *Proceedings of CoNLL-2000*, Lisboa, Portugual,. pgs. 145-147, 2000.

PERINI, M. *Gramática Descritiva do Português*. Editora Ática, 2000.

RATNAPARKHI, A., *Maximum Entropy Models for Natural Language Ambiguity Resolution***.** Ph.D. thesis, University of Pennsylvania, 1998.

RAMSHAW, L. A. e MARCUS, M. Text Chunking Using Transformation-Based Learning. Em: *Proceedings of the Third ACL Workshop on Very Large Corpora*, Association for Computational Linguistics, 1995.

SANTOS, C. N. *Aprendizado de máquina na identificação de sintagmas nominais: o caso do português brasileiro***.** Dissertação de Mestrado, Instituto Militar de Engenharia, Rio de Janeiro, 2005.

SPARCK JONES, K. E WILLET, P. (eds*.)***.** *Readings in Information Retrieval*, Morgan Kaufmann Publishers, 1997.

TJONG, E. F. Memory-Based Shallow Parsing. *Journal of Machine Learning Research 2* 559-594, 2002.