Explorando Métodos de Uso Geral para Desambiguação Lexical de Sentidos para a Língua Portuguesa

Fernando Antônio A. Nóbrega[1], **Thiago A. Salgueiro Pardo**[1]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Caixa Postal 668 – 13560-970 – São Carlos – SP – Brasil

Abstract. *We present in this paper the investigation of some simple and general-use word sense disambiguation methods for common nouns in news texts in Brazilian Portuguese language, using Princeton Wordnet as a sense repository. We report the efforts for building a sense-tagged corpus and its use for evaluating the methods. Our results show that, in general, the baseline method that selects the most frequent senses is better than the traditional method proposed by Lesk, but that the latter is better for more ambiguous words.*

Resumo. *Apresenta-se, neste artigo, a exploração de alguns métodos simples e de uso geral para a desambiguação lexical de sentidos de substantivos comuns em textos jornalísticos escritos em Português do Brasil, usando-se a Wordnet de Princeton como repositório de sentidos. Relatam-se os esforços para a construção de um córpus anotado com sentidos e seu uso para avaliar os métodos explorados. Os resultados mostram que, em geral, o método baseline de selecionar os sentidos mais frequentes é melhor do que o método tradicional proposto por Lesk, mas que este último tem melhor desempenho para palavras mais ambíguas.*

1. Introdução

Diversas áreas de pesquisa desenvolvem aplicações que necessitam processar dados representados em língua natural. Como exemplo, podem-se citar sistemas de Auxílio à Escrita (AE), Recuperação de Informação (RI), Simplificação Textual (ST), Sumarização Automática (SA) e Tradução Automática (TA), dentre outros. Algumas destas aplicações, como a RI e algumas das linhas de SA, inserem-se no contexto multidocumento, no qual o processamento dos dados ocorre a partir de um conjunto de textos ou documentos, agregando obstáculos maiores para bons resultados destas aplicações.

Aplicações deste cenário, além de terem que abordar dificuldades específicas de cada área, devem lidar com características inerentes à língua natural, das quais destaca-se a ambiguidade lexical, caracterizada pela possibilidade de uma palavra assumir diferentes significados conforme o contexto em que é empregada. Este fenômeno da ambiguidade é considerado um dos principais obstáculos para a melhoria de aplicações que exigem Processamento de Linguagem Natural (PLN) [Piruzelli e da Silva 2010]. Por exemplo, em TA, a palavra "banco" presente na sentença "o banco financiou a compra", poderia ser traduzida como "*chair*", no sentido de assento, ou como "*bank*" no sentido de instituição financeira. Faz-se necessário conhecer o sentido correto da palavra para determinar sua tradução mais apropriada (no caso, o sentido de instituição financeira).

A Desambiguação Lexical de Sentidos (DLS) corresponde à tarefa de determinar o sentido mais adequado de uma palavra, dados o contexto em que esta foi empregada (sentença, texto, documento, etc.) e um conjunto finito de possíveis sentidos/significados (em dicionários, ontologias, etc.) [Agirre e Edmonds 2006]. Geralmente, os métodos de DLS são aplicados em etapas de pré-processamento de outras aplicações, para que os problemas com ambiguidade sejam amenizados e, consequentemente, resultados melhores sejam alcançados [Mitkov 2004].

Para o Português do Brasil (PT-BR), há poucos trabalhos no âmbito da DLS, o que, consequentemente, pode influenciar negativamente no resultado de aplicações que exigem PLN. [Specia 2007] apresenta um método de DLS voltado para a tarefa de tradução automática Inglês-Português, que se constitui na desambiguação de uma seleção de 10 verbos, considerados mais ambíguos, do inglês. [Machado et al. 2011] apresentam métodos de desambiguação que podem ser aplicados ao PT-BR, entretanto, específicos para o contexto geográfico, por exemplo: determinar se "São Paulo" se refere ao estado, cidade ou time de futebol. Estes trabalhos são bastante específicos, o que diminui sua efetividade para uso com a língua geral. Assim, o cenário atual da DLS para o PT-BR apresenta-se deficiente, dada a falta de métodos de uso geral que poderiam ser utilizados em diversas aplicações.

Assim, neste artigo, descrevem-se: (1) a construção de um córpus com anotação de sentidos para o PT-BR; e (2) a exploração de métodos simples de DLS, de uso geral e não específicos de domínio e aplicação, avaliados sobre o córpus compilado. Os métodos apresentados visam, em princípio, somente desambiguar palavras classificadas como substantivos comuns, visto que esta é a classe gramatical mais frequente em textos (vide Seção 3).

O córpus compilado trata-se de um importante recurso para pesquisas em DLS para o PT-BR, pois há poucos trabalhos nesta área para o idioma. Com base nesse córpus, dois métodos de DLS são explorados neste trabalho. O primeiro consiste em uma abordagem heurística, de adotar o sentido mais frequente para uma palavra. O segundo método desenvolvido é uma adaptação do método de Lesk [Lesk 1986], que corresponde a um método simples com grande abrangência e bastante difundido na literatura.

Este artigo é organizado em 6 seções. Na Seção 2, é abordada a área de DLS e suas principais abordagens e tarefas. Na Seção 3, são discutidos possíveis repositórios de sentidos e os recursos usados neste trabalho. Na Seção 4, apresentam-se as adaptações de métodos de DLS para o PT-BR. Por fim, nas Seções 5 e 6, apresentam-se os resultados avaliados e as considerações finais deste artigo, respectivamente.

2. Desambiguação Lexical de Sentido

O problema abordado pela DLS, de encontrar o sentido mais adequado de uma palavra, é classificado como um problema Completo da Inteligência Artificial (IA), que somente será solucionado quando aplicações da IA "compreenderem" completamente a língua natural e forem "capazes de assimilar" conhecimento enciclopédico, que evolui com o tempo (no contexto da DLS, tem-se que novas palavras e significados podem surgir ou cair em desuso), para serem resolvidos [Ide e Véronis 1998].

Geralmente, a DLS é tratada como um problema de classificação, onde uma palavra alvo (palavra que se deseja desambiguar) possui n classes, que representam seus

possíveis significados. Assim, dada uma janela de palavras, que pode ser representada por uma lista com m palavras à esquerda e à direita da palavra alvo, busca-se determinar qual classe (significado) é ativada (mais adequado) para a palavra alvo [Mitkov 2004].

Os métodos de DLS podem ser categorizados pela tarefa em que se aplicam e pela abordagem que adotam. As tarefas indicam quais palavras devem ser desambiguadas, sendo as principais, segundo [Jurafsky e Martin 2009]: (1) *lexical sample*, quando apenas um conjunto pré-determinado de palavras deve ser desambiguado; (2) *all words*, quando todas as palavras de um texto ou sentença devem ser desambiguadas; e (3) transferência, quando a desambiguação é realizada traduzindo a palavra de um idioma para outro. Ressalta-se que variações destas tarefas podem ser realizadas, como a desambiguação de palavras por etiqueta morfossintática, que restringe o conjunto de palavras a serem desambiguadas por meio de suas etiquetas, por exemplo, desambiguar apenas palavras que sejam classificadas como substantivos. Já as abordagens de métodos de DLS indicam como os métodos são implementados, podendo-se citar: (1) abordagens baseadas em conhecimento; (2) abordagens baseadas em córpus; e (3) abordagens híbridas, que agregam características das duas abordagens anteriores.

Abordagens baseadas em conhecimento usam diversas fontes de informação (repositórios de significados, documentos web, regras manuais, etc.) para desambiguar as palavras em um texto ou sentença. Em geral, são abordagens que demandam menos tempo para serem desenvolvidas, pois não necessitam de uma etapa de treinamento de método de Aprendizado de Máquina (AM), e abrangem uma maior quantidade de palavras que são capazes de desambiguar. Geralmente, os métodos desta abordagem são aplicados à tarefa de *all words*.

Abordagens baseadas em córpus (geralmente, anotado com os significados das palavras) aplicam métodos de AM para desenvolver um classificador. Esses métodos demandam mais tempo para serem desenvolvidos, pois exigem etapa de treinamento e, algumas vezes, compilação de recursos para treinamento e teste (criação e anotação de córpus). Normalmente, são aplicados à tarefa de *lexical sample*, portanto, apesar de normalmente alcançarem resultados melhores do que métodos baseados em conhecimento, são métodos menos abrangentes.

Dadas as características supracitadas, neste trabalho são descritos métodos da abordagem baseada em conhecimento. Esta abordagem é adotada porque é mais abrangente, o que visa contribuir para diversas aplicações do PLN para o PT-BR, independente de gênero e domínio textual. Quanto à tarefa, assim como apresentado na Seção 1, desambíguam-se apenas palavras da classe de substantivos comuns, por esta classe morfossintática ser a mais frequente em textos.

3. Repositório de Sentidos e Recursos

Métodos de DLS necessitam de Repositórios de Sentidos (RS) para representar os possíveis significados das palavras, como dicionários, *thesaurus*, *wordnets*, ontologias, etc. O método de Lesk [Lesk 1986] e algumas de suas variações, por exemplo, fazem uso de dicionários tratáveis por computador. Entretanto, [Miller 1995] aponta que dicionários, mesmo sendo manipuláveis por computador, são estruturas idealizadas para manipulação humana e não para máquinas, o que, consequentemente, proporciona menos efetividade à DLS.

Repositórios mais elaborados e direcionados à manipulação computacional são as *wordnets*, que representam os significados por meio de conjuntos de palavras sinônimas, denominados *synsets*, e as relações linguísticas entres estes [Miller 1995]. Os *synsets* correspondem à união de quatro elementos: (1) conjunto de palavras sinônimas; (2) glosa, descrição informal do significado do synset; (3) conjunto de possíveis exemplos de uso dos sinônimos; e (4) conjunto de relações com outros *synsets*. Ressalta-se que o item (3) e algumas relações do item (4) podem não ser especificadas, ou seja, há *synsets* que não possuem exemplos de uso, bem como há *synsets* que não possuem algumas relações com outros *synsets*.

A primeira *wordnet*, e também a mais difundida, é a WordNet de Princeton[1](Wn-Pr), destinada à língua inglesa [Fellbaum 1998]. Para o PT-BR, tem-se a WordNet-BR[2] (Wn-BR)[Silva 2005]. A Wn-Br, em seu estado atual de desenvolvimento, possui somente *synsets* para a classe gramatical de verbos. Portanto, para se atingir os objetivos deste trabalho, adota-se a Wn-Pr como RS, o que se mostra factível (mesmo sendo um RS compilado para outro idioma) devido à utilização de ferramentas de tradução automática e córpus anotado.

Outras características da Wn-Pr que justificam sua utilização são as variedades de aplicações que ela possibilita, tais como: (1) migrar para outras *wordnets*, visto que estas, inclusive a Wn-BR, possuem *synsets* indexados aos da Wn-Pr, e (2) utilizar a estrutura e relações da Wn-Pr para fazer inferências e relacionar palavras em aplicações variadas de PLN.

Ao aplicar a Wn-Pr, compilada para o idioma Inglês, torna-se necessária a utilização de ferramentas de TA, fazendo com que a busca por *synsets* ocorra em dois passos: (1) buscar todas as possíveis traduções da palavra alvo; e, (2) com estas traduções, buscar os possíveis *synsets* na Wn-Pr. Este processo possibilita a ocorrência de *gaps* lexicais, que decorrem de ausências, generalizações ou especificações de sentidos entre línguas distintas [Di Felippo 2008]. Por exemplo, não há tradução direta da palavra "caipirinha" para o Inglês, e a palavra "dedo" no PT-BR pode ser traduzida como "*finger*", no sentido de dedo da mão, ou "*toe*", significando dedo do pé.

No caso de ausência de tradução, o conceito da palavra pode ser generalizado e, posteriormente, traduzido. No exemplo anterior, a palavra "caipirinha" pode ser generalizada para "bebida", o que permite o uso da tradução "*drink*". No caso de especialização de conceitos, deve-se observar o contexto de utilização da palavra para sua correta desambiguação, o que remete ao processo inerente à DLS.

Estes processos de generalização e especificação, durante a anotação do córpus descrita posteriormente, são realizados manualmente pelos anotadores. Em seguida, estes dados serão aplicados para geração de um módulo que agregue informação para ferramentas de tradução automática. De forma automatizada, a generalização é necessária para palavras que o tradutor automático não consegue traduzir. Já a especificação, torna-se um processo inerente à desambiguação, porém, contribui para dados mais esparsos, visto que mais *synsets* são retornados.

Ressalta-se que outros RS podem ser empregados, assim como um *thesaurus* (por

[1]http://wordnet.princeton.edu/.

[2]http://www.nilc.icmc.usp.br/wordnet/.

exemplo, o TEP 2.0 [Maziero et al. 2008]), que é um dicionário no qual as palavras são organizadas conforme seus sentidos [Sardinha 2005]. A estrutura de um *thesaurus* permite a busca por palavras sinônimas, assim como nas *wordnets*, porém, não se apresentam relações linguísticas mais elaboradas entre os significados armazenados.

Para avaliação de métodos de DLS, em geral, usa-se córpus ou conjunto de exemplos anotados. Neste trabalho, foi anotado e empregado o córpus CSTNews [Aleixo e Pardo 2008, Cardoso et al. 2011], composto por 140 textos jornalísticos agrupados por assunto em 50 grupos, com 3 textos em média por grupo.

A tarefa de anotação constituiu-se pela desambiguação de 10% dos substantivos mais frequentes em cada grupo. Este valor foi adotado ao se verificar que palavras abaixo deste limiar apresentam, normalmente, frequências baixas, e que, consequentemente, possuem menor impacto para outras aplicações do PLN. Cada grupo de textos do córpus foi anotado manualmente por uma dupla ou trio de anotadores, nunca por somente uma pessoa. O processo ocorreu em reuniões diárias (cinco dias por semana) com duração de um hora, totalizando 5 semanas, sendo: uma semana de treinamento; e 4 semanas para a anotação do córpus. Para auxiliar este processo, foi desenvolvido e empregado um *software* de anotação, o NASP (NILC - Anotador de Sentidos para o Português).

A opção pelo CSTNews se fez por suas duas principais características: textos jornalísticos e cenário multidocumento (coleção de documentos). Textos jornalísticos apresentam uma linguagem mais comum, pois são destinados a um público variado de leitores. Assim, desenvolvendo e avaliando métodos de DLS nestes textos, acredita-se que o grau de generalização dos métodos seja maior. Já a segunda característica, cenário multidocumento, possibilita avaliar o comportamento dos significados das palavras, bem como dos métodos de DLS, neste cenário, comum em diversas aplicações como RI e SA multidocumento.

O córpus foi anotado com os *synsets* da Wn-Pr e a concordância entre os anotadores, aferida pela medida kappa [Carletta 1996] para alguns textos, foi de: 0,853 para concordância nas traduções; 0,729 para os *synsets*; e 0,697 para tradução e synset juntos. Em geral, concordâncias acima de 0,6 (a medida kappa vai até 1,0) indicam que a tarefa de anotação é sistemática o suficiente para que haja confiança nos dados anotados e que, portanto, pode-se automatizar a tarefa em questão. De fato, para a tarefa de DLS, a concordância mais importante é a do *synset*, entretanto, avaliou-se também a tradução devido à necessidade de se traduzir as palavras em PT-BR para, posteriormente, buscar seus possíveis significados na Wn-Pr.

Tabela 1. Concordância entre anotadores

	Total	Parcial	Nula
Synset	62,22%	22,42%	14,36%
Tradução	82,87%	11,08%	6,05%
Tradução–*Synset*	61,21%	24,43%	14,36%

Foram também aferidas três outras medidas de concordância: (1) concordância total, quando todos os anotadores concordavam em uma palavra; (2) concordância parcial, quando a maioria dos anotadores concordava; e (3) concordância nula, quando menos do que a maioria dos anotadores concordavam. Os resultados obtidos com estas métricas

são apresentados na Tabela 1. Tem-se que a concordância total foi sempre superior a 60%, mostrando-se maior para concordância entre as traduções. A concordância parcial, que se trata de um valor complementar à concordância total, não apresentou valores inferiores aos valores da concordância nula.

Por meio do CSTNews, também foi realizado um experimento a fim de substanciar a desambiguação apenas de substantivos comuns. Este experimento constituiu-se em etiquetar morfossintaticamente o córpus por meio do MXPOST [Ratnaparkhi 1986], que possui 97% de acurácia para o PT-BR [Aires 2000], e calcular a frequência de ocorrência das etiquetas morfossintáticas no córpus. Após este processo, verificou-se que, em um total de 23.458 palavras (excluindo-se *stopwords*), a classe gramatical mais presente é a de substantivos comuns, como se apresenta na Figura 1.

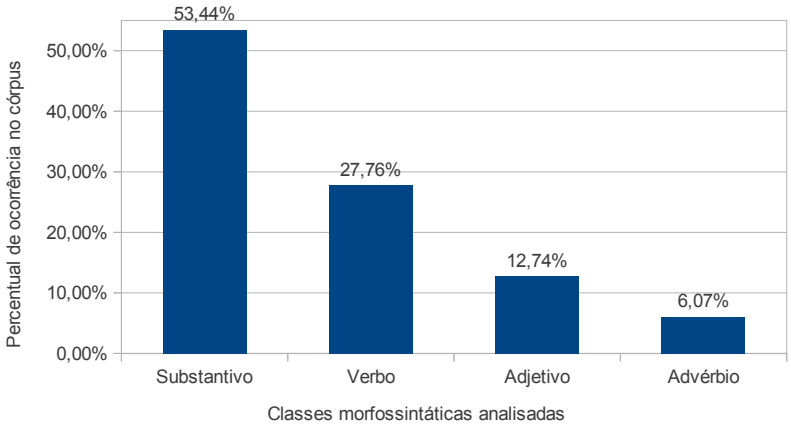

Figura 1. Percentual de ocorrência das etiquetas morfossintáticas no córpus CSTNews (somente classes gramaticais abertas)

4. Adaptação de Métodos de DLS para O PT-BR

Os métodos descritos são aplicados à desambiguação de substantivos comuns, visto que: (1) alguns autores, como [Plaza e Diaz 2011], no âmbito da SA, apontam que a desambiguação de substantivos é suficiente para galgar melhores resultados em aplicações de PLN; (2) a DLS é uma tarefa computacionalmente complexa e que, geralmente, utiliza outras ferramentas de PLN (etiquetadores morfossintáticos, lematizadores, buscadores, etc.), demandando muito tempo para desambiguar todas as palavras de um texto [Mitkov 2004]; e (3) esta é a classe morfossintática mais presentes em textos, como descrito na Seção 3.

Para a tradução automática foi aplicado o WordReference®[3], que se trata de um dicionário bilíngue para diversos idiomas. Esta ferramenta foi adotada por apresentar resultados consistentes ao listar as possíveis traduções de uma palavra. Outras ferramentas bastante difundidas na Web, como o Google Translator®[4], não foram aplicadas por serem serviços pagos[5] ou por não terem apresentado bons resultados em alguns testes preliminares.

[3]http://www.wordreference.com/

[4]http://translate.google.com.br/

[5]O serviço pago trata-se da API de desenvolvimento, não a interface online para uso humano.

Nas próximas subseções, são descritos os dois métodos avaliados neste artigo. Na Subseção 4.1, descreve-se um método heurístico baseado em adotar o synset mais usual para a palavra alvo. Na Subseção 4.2, é apresentada uma adaptação do método de Lesk e algumas variações de utilização.

4.1. Método Heurístico

Métodos de DLS, normalmente, apresentam grande complexidade computacional [Mitkov 2004]. Assim, diversas abordagens heurísticas são aplicadas, principalmente para aferir valores *baseline*, tais como: adotar o primeiro synset da Wn-Pr, visto que seus *synsets* são ordenados pela frequência de uso; e adotar o sentido mais frequente em um córpus.

Neste artigo, é apresentado um método heurístico baseado na escolha do primeiro synset da Wn-Pr, ou seja, o sentido mais usual. Assim como na adaptação do método de Lesk, descrito na próxima subseção, este método adota o melhor synset em duas etapas. Primeiramente, para uma palavra alvo, busca-se a tradução mais frequente (ou seja, a primeira que aparece) segundo o dicionário WordReference®. Em uma segunda etapa, retorna-se o synset mais usual para a tradução adotada, e assim, desambígua-se a palavra alvo com sua tradução e respectivo sentido da Wn-Pr mais frequentes.

4.2. Adaptação do método de Lesk

No método de Lesk, cada palavra na janela recebe um rótulo, constituído por uma lista de palavras extraídas de seus possíveis significados (suas definições) em um dicionário. Após a criação destes rótulos, para cada palavra a ser desambiguada, computa-se a quantidade de palavras em comum entre cada um de seus possíveis significados e os rótulos das palavras em sua janela. Posteriormente, o significado que obtiver maior número de sobreposições de palavras é selecionado [Lesk 1986].

Neste trabalho, o método de Lesk foi adaptado por meio do uso de ferramentas de tradução e pela variação na criação dos rótulos. De forma mais precisa, essa adaptação ocorreu sobre o trabalho de [Banerjee e Pedersen 2002], que apresentam uma variação do método de Lesk para o uso da Wn-Pr como repositório.

Primeiramente, realiza-se uma etapa de pré-processamento, que inclui: (1) etiquetagem morfossintática por meio do MXPOST; (2) eliminação de todas as *stopwords*; (3) lematização das palavras restantes; (4) tradução destas palavras usando as formas lematizadas; e, (5) com suas respectivas traduções, busca dos possíveis *synsets* na Wn-Pr. Após esta etapa, constroem-se os rótulos das palavras e, posteriormente, calcula-se o número de sobreposições de palavras entre os significados das palavras alvo e os rótulos das palavras em suas respectivas janelas.

Os rótulos para cada palavra são conjuntos de palavras lematizadas (sem repetição e *stopwords*) extraídas de 3 fontes (sendo as duas primeiras também aplicadas por [Banerjee 2002]): (F1) glosas dos *synsets*; (F2) exemplos de uso dos *synsets*; e (F3) possíveis traduções para cada palavra. Ressalta-se que, durante a criação dos rótulos, estas fontes podem ser aplicadas de forma individual ou em conjunto.

Tendo a possibilidade de permutar a forma de construção dos rótulos, foram adotadas 6 variações deste método: (G-T) compara a glosa do synset com rótulos em F3;

(S-T) compara exemplos do synset com rótulos de F3; (GS-T) compara a glosa e os exemplos do synset com rótulos em F3; (G-G) compara a glosa do synset com rótulos em F1; (S-S) compara exemplos do synset com rótulos em F2; e (GS2) compara glosa e exemplos do synset com rótulos em F1 e F2. Segundo [Banerjee 2002], essa variação na construção de rótulos se faz pertinente devido à diferença entre as especificações de cada synset. Por exemplo, alguns *synsets* possuem glosas mais extensas do que outros, assim, ao usar rótulos constituídos somente por palavras da glosa, tais *synsets* seriam privilegiados.

5. Avaliação

A avaliação dos métodos descritos, apresentados nas Subseções 4.1 e 4.2, ocorreu sobre 50 grupos de textos do córpus CSTNews, descrito na Seção 3, totalizando 140 textos e 4.366 palavras desambiguadas (sendo 466 palavras distintas).

As métricas utilizadas foram: (1) abrangência, ou seja, quantidade de palavras que o método consegue classificar, corretamente ou não; (2) cobertura, que corresponde ao número de palavras corretamente classificadas em relação a todas as palavras do conjunto de teste; (3) precisão, calculada pelo número de palavras corretamente classificadas em relação a todas as palavras que tiveram significados atribuídos pelo método; e (4) acurácia, segundo [Specia 2007], número de palavras corretamente desambiguadas em relação ao total de palavras, usando o sentido majoritário (mais frequente) quando o método não consegue desambiguar alguma palavra. Estas métricas são comuns na avaliação de trabalhos na área e em eventos específicos, como o Senseval[6].

Foram realizados dois experimentos. O primeiro constituiu-se na aplicação dos métodos de DLS sobre todo o córpus CSTNews, sendo necessário desambiguar todas as palavras que foram anotadas manualmente no córpus. O segundo experimento, também executado em sentenças do CSTNews, objetivou avaliar o desempenho dos métodos para uma lista de palavras composta por palavras que tinham pelo menos 3 significados (*synsets*) anotados no córpus, ou seja, palavras que se apresentaram mais ambíguas no córpus.

Para o primeiro experimento, todas as métricas supracitadas foram aplicadas, permitindo-se, assim, avaliar o comportamento dos métodos em todo o córpus. Estes resultados são apresentados na Tabela 2, na qual cada linha da tabela corresponde a uma métrica de avaliação, e as colunas correspondem aos métodos desenvolvidos. A coluna MF representa o uso do sentido mais frequente e as demais são variações da adaptação do método de Lesk.

Tabela 2. Avaliação do método de Lesk adaptado ao PT-BR no CSTNews

Métodos	MF	G-T	S-T	GS-T	G-G	S-S	GS2
Precisão	51.00	45.20	45.20	41.80	27.00	28.20	26.80
Cobertura	51.00	41.20	41.10	38.00	24.60	25.70	24.40
Abrangência	100.00	91.10	91.10	91.10	91.10	91.10	91.10
Acurácia	–	41.20	41.10	38.00	24.60	25.70	24.40

Já para o segundo experimento, que consiste em uma avaliação mais pontual para um conjunto pré-determinado de palavras, optou-se por apenas valores de precisão. Estes

[6]http://www.senseval.org/

Tabela 3. Avaliação do método de Lesk adaptado ao PT-BR em palavras específicas

Palavra	MF	G-T	S-T	GS-T	G-G	S-S	GS2
acordo	12.50	6.30	**12.50**	6.30	**18.80**	6.30	6.30
agência	41.70	**41.70**	**41.70**	**41.70**	0.00	0.00	0.00
ano	90.50	69.60	86.30	67.60	0.00	22.50	39.20
área	16.70	5.60	5.60	5.60	11.10	0.00	5.60
centro	61.50	0.00	**80.00**	0.00	20.00	**70.00**	30.00
competição	0.00	**6.70**	0.00	0.00	**46.70**	0.00	0.00
estado	33.30	**33.30**	16.70	16.70	**41.70**	0.00	8.30
filho	25.00	**25.00**	**25.00**	**25.00**	0.00	**25.00**	0.00
hora	50.00	**50.00**	**50.00**	**50.00**	**50.00**	0.00	25.00
investigação	74.10	61.50	73.10	61.50	38.50	73.10	69.20
local	30.00	0.00	**30.00**	0.00	0.00	0.00	0.00
obra	42.50	23.10	0.00	0.00	0.00	0.00	0.00
ouro	0.00	0.00	0.00	0.00	**30.00**	0.00	0.00
país	39.20	38.00	**40.00**	38.00	4.00	34.00	32.00
parte	10.00	0.00	**10.00**	**10.00**	0.00	**10.00**	0.00
partida	38.50	27.80	16.70	16.70	5.60	0.00	0.00
presidente	16.40	**19.00**	15.90	15.90	6.30	4.80	0.00
resultado	70.00	65.00	55.00	55.00	0.00	55.00	65.00
tempo	0.00	**14.30**	**28.60**	**35.70**	**7.10**	0.00	**7.10**
vez	0.00	0.00	**10.50**	**10.50**	0.00	0.00	0.00
vôo	3.60	0.00	0.00	0.00	0.00	**4.80**	**4.80**

resultados são apresentados na Tabela 3, na qual, em cada linha, são dispostas as palavras selecionadas (21 no total), e, nas colunas, são apresentados os métodos. Alguns valores são apresentados em negrito, representando valores obtidos pelo método da adaptação de Lesk que são maiores ou iguais ao do método do sentido mais frequente, MF. Por exemplo, para a palavra "agência", as variações G-T, S-T e GS-T obtiveram valores de precisão (41.70) iguais aos do método heurístico.

De modo geral, a adaptação do método de Lesk obteve resultados inferiores ao do método heurístico do sentido mais frequente, como disposto na Tabela 2. Este fato ocorre porque o método heurístico mostra-se melhor na desambiguação de palavras que tiveram apenas um synset anotado, casos estes que ocorrem com maior frequência, como se pode observar na Figura 2. O método heurístico também se apresentou melhor na desambiguação de palavras que, apesar de ocorrerem com mais de dois sentidos anotados no CSTNews, tinham um sentido predominante, assim como a palavra "ano". Depois do método heurístico, o melhor método foi o de variação G-T.

Entretanto, de uma maneira mais pontual, analisando as palavras que apresentam mais de 2 *synsets* anotados no CSTNews, a adaptação do método de Lesk obteve bons resultados, sendo o método heurístico totalmente superior em apenas 6 casos dos 21 disponíveis, como disposto na Tabela 3. Dentre as variações apresentadas na Subseção 4.2, a S-T, que compara a glosa e os exemplos dos synset com traduções das palavras na janela da palavra alvo, obteve melhores resultados, sendo igual ou superior em 10 casos. É importante também ressaltar que a variação GS2, que usa a maior quantidade de informação, não obteve bons resultados, provavelmente por tratar dados mais esparsos devido às várias traduções.

Em ambos os experimentos, tem-se que a adaptação do método de Lesk necessita

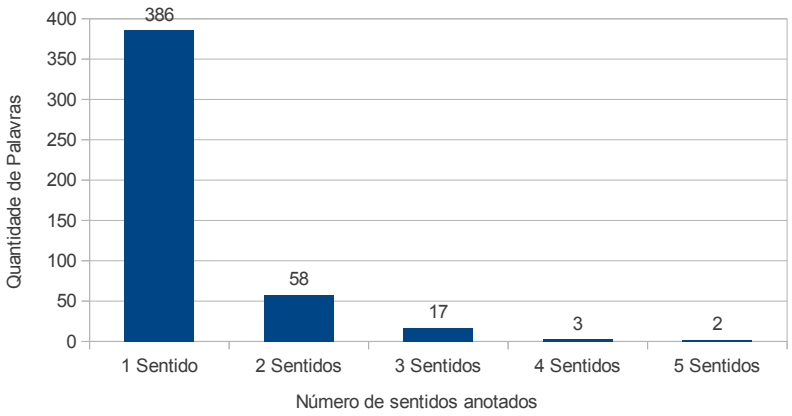

Figura 2. Número de palavras por quantidade de sentidos no CSTNews

mais do processo de tradução e posterior busca de *synsets*. Este processo é necessário para a palavra alvo e para as palavras em sua janela, o que não é realizado pelo o método heurístico. Assim, as traduções incorretas ou inexistentes das palavras na janela da palavra alvo interferem, de forma negativa, nos resultados obtidos.

6. Considerações Finais

Pelo que se tem conhecimento, o trabalho relatado neste artigo apresenta os primeiros resultados de aplicação de métodos de DLS de uso geral para a língua portuguesa. A anotação de sentidos no córpus CSTNews possibilita a experimentação e a avaliação de novos métodos de DLS para essa língua.

Há diversas possibilidades de trabalhos futuros. Sendo que uma das motivações deste trabalho é prover mecanismos de DLS para outras aplicações do PLN, e que várias destas aplicações se inserem no contexto multidocumento, pretende-se, primeiramente, avaliar o comportamento dos significados das palavras neste contexto. Com base na anotação do córpus, acredita-se, por exemplo, que os sentidos variam muito pouco dentro de grupos de textos de um mesmo assunto. Posteriormente, pretende-se investigar métodos mais sofisticados de DLS, como os apresentados por [Mihalcea e Moldovan 1999] e [Agirre e Soroa 2009].

7. Agradecimentos

Este trabalho contou com o apoio da FAPESP e do CNPq.

Referências

Agirre, E. e Edmonds, P. (2006). *Word Sense Disambiguation: Algorithms and Applications*, chapter Introduction, pages 1–28. Springer.

Agirre, E. e Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. In *Proceedings of 12th Conference of the European Chapter of the ACL*, pages 33–41.

Aires, R. V. X. (2000). Implementação, adaptação, combinação e avaliação de etiquetadores para o português do brasil. Dissertação de mestrado, Instituto de Ciências Matemáticas e de Computação – ICMC – USP.

Aleixo, P. e Pardo, T. A. S. (2008). CSTNews: Um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (cross-document structure theory). Technical Report 326, Instituto de Ciências Matemáticas e de Computação.

Banerjee, S. (2002). Adapting the lesk algorithm for word sense disambiguation to wordnet. Dissertação de mestrado, Department of Computer Science University of Minnesota.

Banerjee, S. e Pedersen, T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proceedings of 3th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 136–145.

Cardoso, P. C. F., Maziero, E. G., Jorge, M. L. R. C., Seno, E. M. R., Felippo, A. D., Rino, L. H. M., das Graças V. Nunes, M., e Pardo, T. A. S. (2011). CSTNews – a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Anais do III Workshop "A RST e os Estudos do Texto"*, pages 88–105, Cuiabá, MT, Brasil. Sociedade Brasileira de Computação.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22:249–254.

Di Felippo, A. (2008). *Delimitação e Alinhamento de Conceitos Lexicalizados no Inglês Norte-americano e no Português Brasileiro*. Tese de doutorado, Faculdade de Ciências e Letras, Universidade Estadual Paulista.

Fellbaum, C. (1998). *WordNet An Eletronic Lexical Database*. MIT Press.

Ide, N. e Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24:2–40.

Jurafsky, D. e Martin, J. H. (2009). *Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall. Pearson, 2nd edition.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of 5th Annual International Conference on Systems Documentation*, pages 24–26, New York, NY, USA. Association for Computing Machinery.

Machado, I. M., de Alencar, R. O., de Oliveira Campos Junior, R., e Davis, C. A. (2011). An ontological gazetteer and its application for place name disambiguation in text. *Journal of the Brazilian Computer Society*, 17:267–279.

Maziero, E. G., Pardo, T. A. S., Felippo, A. D., e da Silva, B. C. D. (2008). A base de dados lexical e a interface web do tep 2.0 — thesaurus eletrônico para o português do brasil. In *Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pages 390–392.

Mihalcea, R. e Moldovan, D. I. (1999). A method for word sense disambiguation of unrestricted text. In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*, pages 152–158, College Park, Maryland, USA. Association for Computational Linguistics.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38:39–41.

Mitkov, R. (2004). *The Oxford Handbook of Computational Linguistics*. Oxford University Press, Oxford, UK.

Piruzelli, M. P. F. e da Silva, B. C. D. (2010). Estudo exploratório de informações lexicais relevantes para a resolução de ambiguidades lexical e estrutural. In *Anais do Encontro do Círculo de Estudos Linguísticos do Sul*, Universidade do Sul de Santa Catarina, Palhoça, SC.

Plaza, L. e Diaz, A. (2011). Using semantic graphs and word sense disambiguation techniques to improve text summarization. In *Proceedings of Procesamiento del Lenguaje Natural*, number 47, pages 97–105.

Ratnaparkhi, A. (1986). A maximum entropy model for part-of-speech tagging.

Sardinha, T. B. (2005). *A Língua Portuguesa no Computador*. Mercado de Letras.

Silva, B. C. D. D. (2005). A construção da base da wordnet.br: Conquistas e desafios. In *Anais do XXV Congresso da Sociedade Brasileiro de Computação*.

Specia, L. (2007). *Uma Abordagem Híbrida Relacional para a Desambiguação Lexical de Sentido na Tradução Automática*. Tese de doutorado, Instituto de Ciências Matemáticas e de Computação – ICMC – USP.