Desambiguação Lexical de Sentido com uso de Informação Multidocumento por meio de Redes de Co-ocorrência

Fernando Antônio A. Nóbrega, Thiago A. Salgueiro Pardo

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

{fasevedo,taspardo}@icmc.usp.br

Abstract. *Word Sense Disambiguation (WSD) aims at determining the appropriate sense of a word in a particular context. This is a very important area, especially to provide resources to other areas of Natural Language Processing. Many applications are engaged in the multidocument context, where the processing is performed on a set of texts, however, there are no reports of applications and experiments of WSD in this context. In this paper, we present two methods of WSD, a single and another multi document, motivated by evidence of corpus. The methods were designed to Portuguese language and compared with other works for the language, and stood higher in their respective fields.*

Resumo. *A Desambiguação Lexical de Sentidos (DLS) visa determinar o sentido adequado de uma palavra em determinado contexto. Trata-se de uma área bastante importante, sobretudo, para prover recursos para outras áreas do Processamento de Língua Natural. Diversas aplicações desta área engajam-se no cenário multidocumento, onde o processamento é efetuado em um conjunto de textos, porém, não há relato de aplicações/avaliações da DLS neste contexto. Neste trabalho, apresentam-se dois métodos de DLS, um mono e outro multidocumento, motivado por evidências de córpus. Os métodos foram destinados a língua portuguesa e comparados com outros trabalhos para o idioma, e apresentaram-se superiores em diferentes cenários.*

1. Introdução

A Desambiguação Lexical de Sentidos (DLS) objetiva desenvolver e avaliar métodos automáticos que determinam o sentido correto de uma palavra (Palavra Alvo - PA) em um contexto especificado, por meio de um Repositório de Sentidos (RS) [Agirre and Edmonds 2006]. A DLS é uma área muito relevante, sobretudo, por prover recursos para diminuir problemas de ambiguidade lexical, caracterizada como um fenômeno linguístico em que palavras assumem diferentes significados conforme o contexto em que são empregadas [Piruzelli and da Silva 2010], para contribuir na melhoria de desempenho de outras aplicações do Processamento de Língua Natural (PLN).

Wordnets são RS frequentemente empregados em pesquisas nessa área, pois são consideradas mais eficazes para a DLS [Miller 1995]. Esse tipo de RS que organiza os sentidos por meio de uma hierarquia de conjuntos de sinônimos, denominados *synsets*. Cada *synset* é constituído por quatro informações básicas: 1) conjunto de palavras sinônimas; 2) glosa, uma definição informal do *synset*; 3) conjunto de exemplos de sentenças que possuem alguma palavra do item 1; e 4) relações linguísticas com outros *synsets*.

Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pages 138–147,
Fortaleza, CE, Brazil, October 21–23, 2013.

Tendo em vista que a desambiguação de substantivos comuns proporciona melhores resultados em diversas aplicações do PLN [Plaza and Diaz 2011], que palavras desta classe morfossintática são mais frequentes e, consequentemente, podem proporcionar maiores fenômenos de ambiguidade, a WordNet de Princeton (Wn-Pr) [Fellbaum 1998], foi empregada como RS neste trabalho, visto que sua utilização mostrou-se factível para o PT-BR por meio do método multilíngue com recursos de tradução automática proposto por [Nóbrega and Pardo 2012].

Normalmente, métodos de DLS empregam informações de contexto local, palavras que estão em torno da PA, para determinar o sentido mais adequado. Contudo, diversas aplicações do PLN inserem-se no cenário multidocumento, no qual a computação ocorre sobre uma coleção de textos, tais como Sumarização Multidocumento, Agrupamento de Textos, Recuperação de Informação, etc.

No cenário supracitado, embora alguns trabalhos de DLS possibilitem empregar informação multidocumento, como [Agirre and Soroa 2009], que fazem uso de grafos para representar os contextos das palavras por meio de relações entre seus respectivos sentidos, não há relato de experimentos ou análises de DLS nesse cenário.

Neste trabalho, a fim de investigar a DLS no cenário multidocumento e direcionada ao PT-BR, dada a carência de trabalhos de DLS nessas áreas, apresentam-se dois métodos de DLS para substantivos comuns que empregam o método de recuperação de *synsets* da Wn-Pr para palavras de idiomas diferentes do inglês proposto por [Nóbrega and Pardo 2012]. O primeiro método é uma adaptação do algoritmo de [Mihalcea and Moldovan 1999], que, após experimentos, mostra-se eficaz na desambiguação de palavras consideradas mais ambíguas. O segundo é uma adaptação para aumentar a eficiência e desempenho do melhor método de [Nóbrega and Pardo 2012] no cenário multidocumento.

Este artigo está dividido em mais 5 seções. Na Seção 2, disserta-se sobre trabalhos de DLS relacionados. A análise de córpus, que direcionou e proporcionou a avaliação dos métodos propostos, será apresentada na Seção 3. Os métodos desenvolvidos, bem como os artefatos técnicos e teóricos, serão descritos na Seção 4. Na Seção 5 são discutidas as avaliações dos métodos e a comparação de seus resultados com o trabalho de [Nóbrega and Pardo 2012]. Por fim, a conclusão deste trabalho é apresentada na Seção 6.

2. Trabalhos Relacionados

O algoritmo de [Lesk 1986] atribui rótulos às palavras no contexto e adota que o sentido mais adequado para uma PA é aquele mais similar com esses rótulos. Um rótulo, dada uma palavra e suas possíveis definições em um dicionário, é um conjunto lexical extraído dessas definições. A métrica de similaridade utilizada é quantificada pelo número de palavras sobrepostas, ou seja, número de palavras presentes na definição dos sentidos e nos rótulos das palavras no contexto da PA. [Kilgarriff et al. 2000] descrevem uma simplificação desse algoritmo, a fim de diminuir seu tempo de execução computacional, e [Banerjee 2002] apresentam a adaptações, na proposta original de Lesk, para utilização da Wn-Pr.

[Mihalcea and Moldovan 1999] fazem uso da Wn-Pr como RS e apresentam um algoritmo de desambiguação pautado em pares de palavras, sendo uma destas a PA e a

outra, uma palavra do contexto (palavra-contexto). Nesse algoritmo, a PA é desambiguada com o *synset* que ocorre mais vezes com a palavra-contexto. Para tanto, por meio de padrões de *queries* (chaves de consulta) entre as palavras do conjunto de sinônimos dos *synsets* e palavra-contexto na ferramenta de busca Altavista®[1], realizam pesquisas na Web para contabilizar a quantidade de páginas retornadas.

[Nóbrega and Pardo 2012] apresentam uma abordagem multilíngue para a DLS, cujo princípio é um método para recuperação de *synsets* da Wn-Pr para palavras de idiomas que não sejam o inglês por meio de uma etapa de tradução automática. Com isso, é possível desambiguar palavras de diferentes línguas por meio da Wn-Pr. É importante ressaltar que os autores relatam deficiências, principalmente, pelo fato de lacunas lexicais, que são ausências ou especificações de conceitos entre línguas diferentes. Além disso, para o PT-BR, os autores apresentam a utilização desse método nos algoritmos de [Kilgarriff et al. 2000], [Banerjee 2002] e um procedimento heurístico (que atribui o sentido mais frequente).

Dentre os métodos apresentados pelos autores supracitados, destacam-se duas adaptações do algoritmo de [Kilgarriff et al. 2000], que obtiveram resultados semelhantes ao do método heurístico. Nessas adaptações, os rótulos das palavras no contexto são formados por suas respectivas traduções e a distinção entre elas é a informação do *synset* usada no cálculo de similaridade. Na primeira, utiliza-se a glosa de cada *synset* (algoritmo GT, de glosa + tradução) e na segunda, são usados os exemplos (método ST, de *sample* + tradução).

3. Análise de Ambiguidade Multidocumento no Córpus CSTNews

O CSTNews [Aleixo and Pardo 2008, Cardoso et al. 2011] é um córpus constituído por 140 textos, organizados em 50 coleções, com dois ou três textos cada uma. Esse córpus, além de outros níveis de anotação, disponibiliza 10% dos substantivos mais frequentes de cada coleção desambiguados com *synsets* da Wn-Pr [Nóbrega and Pardo 2012]. Essa anotação foi realizada por um processo manual (com auxilio de uma ferramenta), cujo valor de concordância Kappa [Carletta 1996] em relação à escolha dos *synsets* entre grupos distintos de anotadores foi de 0,77.

Por meio do CSTNews, investigou-se a presença da ambiguidade no cenário multidocumento. Para tanto, contabilizou-se a quantidade de sentidos atribuídos para uma mesma palavra em uma coleção. Nessa análise, 677, 42 e 8 palavras foram anotadas, respectivamente, com um, dois e três *synsets* diferentes, ou seja, ocorrências de uma mesma palavra tendem a assumir o mesmo significado em uma coleção de textos relacionados.

Ao contabilizar o nível de ambiguidade, ou seja, a quantidade de possíveis *synsets* atribuíveis para cada palavra anotada no córpus CSTNews, observou-se que 361 (77%) palavras distintas apresentaram, no mínimo, duas possibilidades de desambiguação e 196 ($\approx$ 42%) ficaram acima da média (com $\approx$ 6 *synsets* atribuíveis).

4. Métodos de DLS Desenvolvidos

Neste trabalho, por meio da proposta multilíngue [Nóbrega and Pardo 2012], apresentam-se dois algoritmos de desambiguação. O primeiro (Seção 4.1) trata-se de uma adaptação

[1] http://www.altavista.com/

do algoritmo proposto por [Mihalcea and Moldovan 1999], que emprega a Web como córpus. Já o segundo (Seção 4.2), é constituído por alterações no melhor método de DLS avaliado em [Nóbrega and Pardo 2012], direcionando-o ao cenário multidocumento por meio de uma Rede de Co-ocorrência Lexical (RCL).

É importante ressaltar que os métodos descritos neste trabalho adotam a mesma etapa de pré-processamento proposta em [Nóbrega and Pardo 2012], que é composta por: 1) tokenização; 2) etiquetagem morfossintática por meio do MXPOST; 3) extração de *stopwords* e sinais de pontuação; e 4) lematização das palavras restantes.

4.1. Método Irrestrito Adaptado de [Mihalcea and Moldovan 1999]

[Mihalcea and Moldovan 1999], após diversos experimentos, sugerem a utilização de verbos como contexto para desambiguação de substantivos comuns. Assim, tendo em vista os objetivos deste trabalho, a configuração anterior foi selecionada, adotando como contexto o verbo mais próximo da PA.

Após a etapa de pré-processamento (vide Seção 4), devido à característica multilíngue deste trabalho, cada palavra-contexto (verbo mais próximo da PA), durante o processo de desambiguação, foi traduzida automaticamente. Assim, pela possibilidade de duas ou mais traduções, o processo de criação de *queries* empregado em [Mihalcea 2006] foi repetido para cada possível tradução.

Além da alteração supracitada, neste trabalho, foi empregado o buscador Microsoft Bing®em vez do Altavista®. O Bing foi escolhido por ser gratuito e possuir API (Interface de Programação de Aplicativo, do inglês *Application Programming Interface*), que possibilita a criação de programas de computador que utilizem suas funcionalidades. Outras opções foram cogitadas, como Google Search®e Yahoo Search®, porém, são recursos com API comercial e/ou limitações de utilização gratuita.

4.2. DLS Multidocumento

Tendo em vista que ocorrências de uma mesma palavra tendem a assumir o mesmo sentido em uma coleção de textos relacionados (vide Seção 3), optou-se por empregar um algoritmo DLS que atribui somente um sentido para todas as ocorrências de uma palavra. Decisão essa que é adotada em alguns métodos heurísticos [Mihalcea 2006]. Com esse objetivo, em vez de empregar informação local (palavras em torno da PA), empregou-se uma Rede de Co-ocorrência Lexical (RCL) para capturar as relações lexicais em uma coleção de textos, que possibilita uma melhor representação do cenário multidocumento, e empregar o algoritmo de DLS. Essa abordagem baseia-se na hipótese de que ao encontrar relações léxicas mais relevantes em uma coleção, o resultado da DLS pode ser melhorado.

Uma RCL representa relações (ponderadas ou não) entre palavras. Em geral, uma RCL é modelada por meio de um grafo $G(V, A)$ composto por um conjunto de vértices V e um conjunto de arestas A, de forma que todo $v \in V$ representa uma palavra na rede e toda $a(v', v'')^w \in A$ indica uma aresta (ou relação) valorada com peso w entre duas palavras.

A construção de uma RCL varia conforme a especificidade de cada tarefa e o tipo de relação desejada. Neste trabalho, dada uma coleção de textos D, cada $v \in V$ indica uma única palavra (extraindo *stopwords* e repetições) contido em D e cada aresta, pon-

derada com peso w, representa que duas palavras co-ocorreram em w Janelas de Palavra (JP), que corresponde a uma sequência de n palavras em um texto).

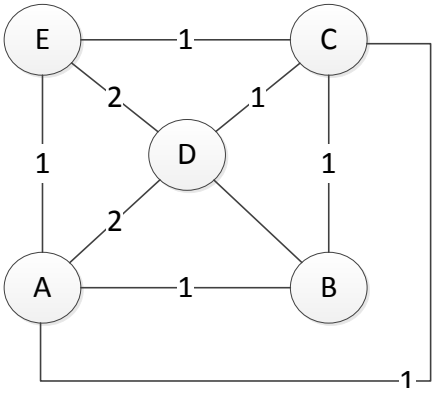

Figura 1. Exemplo de uma RCL para a sequência A,B,C,D,E,A e D

Por exemplo, para a sequência A,B,C,D,E,A e D têm-se as seguintes janelas de tamanho 3: ABC, BCD, CDE, D**EA**, **EA**D. É importante ressaltar que duas ou mais palavras podem co-ocorrer em janelas que se sobrepõem, por exemplo, as sequências marcadas em negrito (D**EA**, **EA**D), mesmo assim, estes casos devem ser contabilizados uma única vez. Portanto, as únicas letras que co-ocorrem duas vezes, nesse caso, são E-D e A-D e, consequentemente, as arestas que as relacionam são ponderadas com peso 2, como pode ser observado na Figura 1.

Após a criação de um RCL para uma coleção de documentos, conforme descrito anteriormente, para cada PA, utilizam-se as n respectivas palavras mais relacionadas (aquelas cuja arestas possuem maiores pesos) como contexto no algoritmo de desambiguação, que, nesse caso, foi a adaptação do algoritmo de [Kilgarriff et al. 2000] apresentada por [Nóbrega and Pardo 2012]. É importante ressaltar que o valor de n foi utilizado como parâmetro da RCL (tamanho da janela de palavras) e para a listagem das palavras mais relacionadas. Como experimentos, foi realizado configurações com redes com janelas de tamanho igual a 3 (R3) e 5 (R5).

5. Avaliação

Os métodos propostos neste trabalho foram comparados com os melhores resultados de [Nóbrega and Pardo 2012], que correspondem aos algoritmos (brevemente descritos na Seção 2: 1) heurístico (referenciado como H), que atribui o sentido mais frequente; e 2) método GT; e 3) algoritmo ST.

Foram adotadas quatro métricas de avaliação: P) Precisão, número de palavras corretamente desambiguadas sobre o total de palavras para as quais algum *synset* foi atribuído pelo método avaliado; C) Cobertura, número de palavras corretamente desambiguadas sobre o total de palavras anotadas no córpus; A) Abrangência, número de palavras desambiguadas correta ou incorretamente pelo método avaliado; e Ac) Acurácia, que corresponde à medida de precisão com auxílio do método heurístico para palavras que o método avaliado não foi capaz de atribuir algum *synset* [Specia 2007].

Com o propósito de avaliar estes métodos sob diferentes óticas, foram realizados três experimentos: 1) *all words*, na qual foi aferido o desempenho dos métodos na desam-

biguação de todas as palavras anotadas no CSTNews; 2) *lexical sample*, na qual aferiu-se a qualidade dos métodos na desambiguação de palavras mais ambíguas no córpus; e 3) avaliação multidocumento, onde foi verificado o desempenho da DLS ao empregar informações multidocumento.

Em todos os experimentos, somente para o método heurístico, não foi realizada a etapa de pré-processamento. Para os demais métodos, a tarefa de DLS constituiu-se em: 1) encontrar corretamente a PA (etiquetando-a adequadamente como substantivo comum); e 2) desambiguar a PA em seu determinado contexto. Ressalta-se que algumas palavras não obtiveram tradução e/ou *synsets* automaticamente, o que influenciou na qualidade de todos os métodos. Portanto, para o cálculo da abrangência, estes casos foram desconsiderados.

Na apresentação dos resultados, os métodos desenvolvidos neste trabalho serão referenciados por siglas, da seguinte forma (os últimos três, são do trabalho de [Nóbrega and Pardo 2012]: (R3) Algoritmo de DLS proposto com uma RCL de janela com tamanho 3; (R5) análogo ao anterior, porém, com uma RCL de janela com tamanho 5; (M) adaptação do trabalho de [Mihalcea and Moldovan 1999]; (H) método heurístico; (G-T) variação com glosa e tradução; S-T) variação com exemplos e tradução.

5.1. All Words

Neste experimento, a tarefa de DLS caracteriza-se pela desambiguação de todas as palavras do CSTNews que foram anotadas com algum *synset*. Cada coleção do córpus foi processada separadamente pelos métodos e, posteriormente, verificaram-se os resultados obtidos.

Os resultados aferidos para cada método são apresentados na Tabela 1, onde, na primeira coluna, são dispostos os métodos avaliados e nas demais são listadas as métricas de avaliação P, C, A e Ac, cujo valores são exibidos em escala percentual. Por exemplo, tem-se que o método R3 obteve 49.56% de precisão.

Tabela 1. Avaliação Geral: All Words

Método	P(%)	C(%)	A(%)	Ac(%)
H	51.00	51.00	100	–
G-T	42.20	41.20	91.10	41.20
S-T	42.20	41.10	91.10	41.10
GS-T	41.80	38.00	91.10	38.00
M	39.71	39.47	**99.41**	39.59
R3	**49.56**	**43.90**	88.59	43.90
R5	46.87	41.80	87.65	41.80

Pode-se observar que o método heurístico (linha H) obteve a melhor precisão. Contudo, é importante ressaltar que o córpus CSTNews apresenta um cenário propício para este método, visto que a maioria das palavras foi anotada com um único *synset* e que, normalmente, foi o mais frequente. Outro fato importante é que a diferença estatística entre o método *baseline* e o algoritmo R3, segundo Teste T de Student com 95% de confiança, é irrelevante.

A melhor abrangência, excluindo H, foi atingida pela adaptação do trabalho de [Mihalcea and Moldovan 1999]. Tendo em vista que o valor de abrangência é influenciado diretamente pela qualidade da etapa de pré-processamento, igualmente aplicada em

todos os métodos, esse resultado era esperado, pois o algoritmo supracitado emprega menos informação advinda da etapa de pré-processamento, utilizando somente a PA e uma palavra-contexto.

Os valores de Acurácia, que representam o auxílio do algoritmo heurístico nos métodos avaliados, em geral, foram iguais aos valores de Cobertura. Esse fato indica que o método heurístico também não foi capaz de desambiguar, corretamente, palavras que os demais métodos não desambiguaram, o que, provavelmente, indica palavras difíceis para serem desambiguadas

5.2. Amostra de Palavras

Neste experimento (tarefa *lexical sample*), foram selecionadas somente palavras com dois ou mais sentidos anotados no córpus. Assim, por ser uma avaliação mais pontual, foi adotada apenas a métrica de Precisão.

Os resultados obtidos por cada algoritmo são apresentados na Tabela 2, não sendo diferenciado o algoritmo R3 do R5, visto que esses obtiveram resultados iguais. Nas linhas da tabela, são dispostas as 21 palavras consideradas e, nas colunas, os métodos avaliados. Por exemplo, tem-se que o método M obteve 94,83% de Precisão na desambiguação da palavra "ano". Nas últimas linhas da tabela, são listadas as quantidade de vezes em que os métodos foram superiores (Total > H), iguais (Total = H), maior ou igual (Total >= H) e inferiores (Total < H) ao método heurístico. Na última linha, é disposto o valor médio da Precisão aferida para cada algoritmo.

Tabela 2. Avaliação lexical-sample

Palavra	H	G-T	S-T	GS-T	M	G(3,5)
acordo	12.50	6.30	**12.50**	6.30	**13.33**	**12.50**
agência	41.70	**41.70**	**41.70**	**41.70**	20.00	25.00
ano	90.50	69.60	86.30	67.60	**94.83**	47.22
área	16.70	5.60	5.60	5.60	0.00	5.56
centro	61.50	0.00	**80.00**	0.00	57.14	36.36
competição	0.00	**6.70**	**0.00**	**0.00**	**35.71**	0.00
estado	33.30	**33.30**	16.70	16.70	28.57	30.00
filho	25.00	**25.00**	**25.00**	**25.00**	**25.00**	**25.00**
hora	50.00	**50.00**	**50.00**	**50.00**	**50.00**	0.00
investigação	74.10	61.50	73.10	61.50	60.00	18.75
local	30.00	0.00	**30.00**	0.00	**33.33**	17.65
obra	42.50	23.10	0.00	0.00	**64.71**	26.83
ouro	0.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
país	39.20	38.00	**40.00**	38.00	**76.47**	17.46
parte	10.00	0.00	**10.00**	**10.00**	**14.77**	**10.00**
partida	38.50	27.80	16.70	16.70	12.50	**52.63**
presidente	16.40	**19.00**	15.90	15.90	14.46	9.41
resultado	70.00	65.00	55.00	55.00	**78.95**	66.67
tempo	0.00	**14.30**	**28.60**	**35.70**	**0.00**	**0.00**
vez	0.00	**0.00**	**10.50**	**10.50**	**0.00**	**0.00**
vôo	3.60	0.00	0.00	0.00	0.00	0.00
Total > H	-	3	4	2	**8**	1
Total = H	-	6	8	6	5	7
Total >= H	-	9	12	8	**13**	8
Total < H	-	12	9	13	8	13
Média	27.88	23.19	28.46	21.72	**32.37**	19.10

Em geral, as palavras mais ambíguas apresentam mais de um sentido por coleção de

documentos. Consequentemente, o método de DLS multidocumento obteve resultados pouco satisfatórios nesse experimento. Somente em um caso bem específico (a palavra "partida") o método multidocumento superou os demais. Essa palavra, apesar de ter sido anotada com dois *synsets* diferentes no córpus, foi anotada com um mesmo *synset* em uma das coleções, e, nesse caso, foi desambiguada corretamente apenas pelo método R3 e R5.

O método M obteve um bom resultado nessa avaliação, mesmo sendo inferior na anterior. Isso pode ser observado pela quantidade de vezes em que M foi melhor ou igual (13 casos) ao método heurístico. Além disso, seu valor médio de precisão foi superior a todos os outros algoritmos.

É importante também observar alguns casos particulares, como as palavras: "filho", com valor de precisão igual para todos os métodos; "hora", com todos os algoritmos, exceto o R3 e R5, com precisão igual a 50%; "ouro", em que todos os métodos determinaram incorretamente todas as suas ocorrências; e "vôo", em que somente o método heurístico obteve algum acerto.

5.3. Ganho de Informação Multidocumento

A qualidade da DLS por meio de RCL multidocumento, no primeiro experimento, mostrou-se superior aos demais, sendo, inclusive, equivalente ao método heurístico. Porém, no segundo experimento, seu desempenho não foi satisfatório. Assim, para analisar de forma mais eficiente a utilização desse tipo de informação para a DLS, especificamente no cenário multidocumento, verificaram-se as relações entre palavras desambiguadas pelos métodos G-T, M e R3 por meio de um Diagrama de Venn, ilustrado na Figura 2.

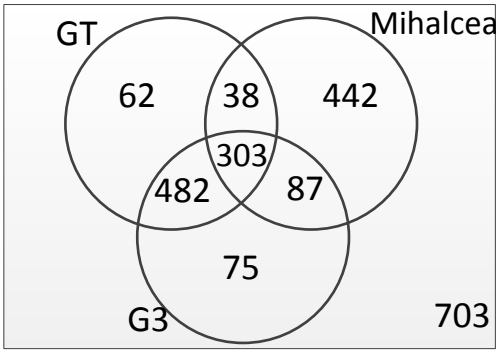

Figura 2. Relação de palavras desambiguadas pelos métodos

Nessa ilustração, cada círculo representa as palavras (aqui, referenciadas somente pela quantidade delas) corretamente desambiguadas pelos métodos. Assim, as áreas de sobreposições indicam casos em que dois ou os três métodos acertaram. Por exemplo, 482 palavras foram desambiguadas adequadamente pelos algoritmos GT e R3, e 303 palavras foram corretamente desambiguadas pelos três métodos. A área fora dos círculos compreende palavras que não foram corretamente desambiguadas por nenhum método, neste caso, 703.

O método M, isoladamente, acertou mais palavras que os demais. Essas palavras, normalmente, são aquelas mais ambíguas, que foram avaliadas no experimento de *lexical sample* (Seção 5.2 e, como dito anteriormente, o desempenho desse método foi superior

nessa avaliação. Contudo, em geral, percebe-se que o método M é inferior aos demais, pois a soma de todos os valores de seu círculo é inferior a soma dos demais conjuntos, o que evidencia seu desempenho pouco satisfatório no primeiro experimento (Seção 5.1.

Desconsiderando-se as palavras para as quais tanto R3 quanto GT obtiveram êxito mutuamente, R3 obteve 162 acertos contra 100 do algoritmo GT. Assim, considerando que R3 aplica o algoritmo GT, embora por meio de um RCL para representação deste cenário, o uso da informação multidocumento contribuiu efetivamente para a DLS neste cenário.

6. Conclusões

Neste trabalho, apresentou-se a análise semântica (de sentidos) multidocumento do córpus CSTNews, na qual se pôde concluir que as palavras tendem a manter o mesmo sentido em uma coleção. É importante ressaltar que o córpus analisado é constituído por notícias jornalísticas com apenas 140 textos. Portanto, não se pode concluir que tal afirmativa é genérica para qualquer cenário multidocumento. Tal generalização necessita de uma análise maior, por meio de córpus mais expressivos.

Além da análise do córpus, foram apresentados dois métodos de DLS e posterior comparação com os melhores algoritmos descritos em [Nóbrega and Pardo 2012]. Na primeira análise, tarefa *all words*, pôde-se verificar que os métodos multidocumento (R3 e R5) foram superiores ao trabalho supracitado. Além disso, os métodos R3 e R5 apresentam tempo de execução computacional menor, pois atribuem somente um sentido para todas as ocorrências de uma palavra na coleção, o que proporciona sua aplicação em cenários onde é necessário processamento de diversos documentos em um tempo curto.

Na segunda análise, verificou-se que M obteve melhores resultados, ou seja, apresentou-se melhor na desambiguação de palavras mais ambíguas. Isso ocorre pelo viés deste algoritmo, que, por empregar buscas na Web, funciona melhor para desambiguação de palavras com muitos *synsets* ou com vários sinônimos em cada sentido.

Como trabalhos futuros, pretende-se estender esta análise e métodos de DLS multidocumento para outros córpus e outros idiomas, a fim de generalizar os resultados obtidos. Pretende-se também, investigar formas de diferenciar ocorrências de palavras com sentidos diferentes por meio de uma RCL, no intuito de melhor o desempenho do algoritmo R3 e R5 para palavras muito ambíguas.

7. Agradecimentos

Ao CNPq e à FAPESP pelo auxilio financeiro.

Referências

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

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

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*. Master's thesis, Department of Computer Science University of Minnesota.

Cardoso, P. C. F., Maziero, E. G., Jorge, M. L. R. C., Seno, E. M. R., Felippo, A. D., Rino, L. H. M., Nunes, M. d. G. V., 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.

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

Kilgarriff, A., England, B., and Rosenzweig, J. (2000). English senseval: Report and results. In *Proceedings of 2nd International Conference on Language Resources and Evaluation*, pages 1239–1244.

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.

Mihalcea, R. (2006). *Word Sense Disambiguation: Algorithms and Applications*, chapter Knowledge-Based Methods for WSD, pages 107–131. Springer.

Mihalcea, R. and 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.

Nóbrega, F. A. A. e Pardo, T. A. S. (2012). Explorando métodos de desambiguação lexical de sentidos de uso geral para o português. In *Encontro Nacional de Inteligência Artificial*, Curitiba – Paraná – Brasil.

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

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