Um Modelo de Identificação e Desambigüização de Palavras e Contextos

Christian Nunes Aranha[1], Maria Cláudia de Freitas[2], Maria Carmelita Pádua Dias[2], Emmanuel Lopes Passos[1]

[1]Departamento de Engenharia – Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) Rio de Janeiro – RJ – Brasil;

[2]Departamento de Letras – Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) Rio de Janeiro – RJ – Brasil;

{chris_ia, emmanuel}@ele.puc-rio.br, {claudiaf, mcdias}@let.puc-rio.br

Abstract. *This paper focus on the interpretation of polisemic words and contexts. It suggests that this interpretation is context-sensitive, regardless of the inner representation of the words themselves. Context is viewed as a cluster formed by a target-word and the words co-ocurring with it. This cluster is achieved by means of statistical treatment of texts as well as graphs representing data obtained in the processing. Preliminary results show that this kind of approach is promising in terms of polisemic words disambiguation.*

Resumo*. Este trabalho aborda a questão da interpretação de palavras e contextos polissêmicos, e sugere que apenas o contexto é necessário para tal interpretação, independente de uma representação interna e autônoma das palavras. O contexto é visto como um aglomerado de palavras coocorrentes com uma palavra-alvo. Esse aglomerado é decorrente de um tratamento estatístico de textos, bem como de grafos que representam os dados obtidos. Os resultados preliminares mostram que este tipo de abordagem é promissor em termos de desambigüização de palavras polissêmicas.*

1. Introdução

Atribuir um significado a uma palavra ou distinguir os diferentes significados de uma palavra polissêmica em um dado contexto são tarefas corriqueiras para qualquer falante de uma língua. Porém, do ponto de vista do Processamento em Linguagem Natural (PLN), a situação não é tão simples. Paradoxalmente, é cada vez mais evidente a necessidade de programas capazes de lidar com a desambigüização de palavras – na recuperação de informações, por exemplo, uma busca por palavra-chave que eliminasse os documentos que trazem esta palavra com um significado não apropriado seria altamente desejável.

O presente trabalho apresenta resultados preliminares provenientes de um projeto que consiste em pesquisar e elaborar modelos relacionados ao Processamento de Linguagem Natural em um único módulo, chamado Cortex. Apresentamos aqui alguns resultados referentes ao processamento de itens lexicais, especificamente à representação automática do significado de palavras inseridas em um contexto. O

princípio subjacente é o de que aspectos das propriedades de uma palavra podem ser capturados por meio de dados estatísticos relativamente às outras palavras que tendem a ocorrer próximas à palavra alvo. Em termos teóricos, buscamos respaldo em teorias vinculadas às correntes pragmáticas do significado – como as de Firth (1957), de inspiração wittgensteiniana -, segundo as quais o significado de uma palavra só pode ser determinado quando inserido em um contexto de uso. O Cortex utiliza uma abordagem puramente estatística da informação lexical. Com um processamento estatístico das palavras de um corpus, o programa é capaz de distinguir os diferentes contextos em que uma determinada palavra pode acontecer, explicitando assim seus possíveis significados e desfazendo ambigüidades.

2. Delimitação do problema: a representação de palavras polissêmicas

O léxico, cada vez mais, vem sendo reconhecido como um dos pontos-chave de programas que visam lidar com PLN. Nesse âmbito, assumem importância fundamental as questões relativas à representação e à aquisição lexicais – de um lado, como representar uma palavra e seu(s) significado(s); e, de outro, como construir um léxico capaz de adquirir novas palavras.

Com relação à representação do significado, a polissemia aparece como um problema teórico, mas não prático. Ou seja, embora de difícil definição em termos teóricos, na prática a polissemia não apresenta qualquer dificuldade – o que Taylor (2003) chama de "o paradoxo da polissemia".

A própria definição tradicional de polissemia – a existência de significados distintos, porém relacionados, em uma mesma palavra – traz consigo questões nada triviais, como qual é a natureza do significado de uma forma lingüística, e o que se entende por relações entre os significados.

A polissemia é um fenômeno inerente a todas as línguas naturais e ela raramente se constitui em um problema de comunicação entre as pessoas. Nenhum falante experimenta qualquer dificuldade para interpretar palavras polissêmicas em seu cotidiano. No entanto, quando se trata relacionar os diversos significados de uma palavra, os falantes apresentam uma grande variação, em relação ao número de acepções bem como a como representá-las. Especialmente no caso de palavras polissêmicas, falantes tendem a discordar quanto às distinções entre significados, e às vezes o mesmo falante pode divergir em suas opiniões sobre o(s) significado(s) de uma palavra.

Um outro exemplo da dificuldade de se lidar com a polissemia é a variação existente, entre dicionários de uma mesma língua, para enumerar e definir significados de uma mesma palavra. Conseqüentemente, propostas de tratamento do significado baseadas em *machine readable dictionaries* (MRD) também costumam apresentar problemas, como demonstram Fillmore & Atkins (2000).

Do ponto de vista computacional, como representar o(s) significado(s) igualmente se mostra um desafio ainda maior, tendo em vista as necessidades de formalização e delimitação necessárias ao meio eletrônico. Taylor resume esse problema, afirmando que

> "a sentence containing *n* words each of which is *m*-times polysemous will in principle have *n* x *m* potential readings. (…)It is not surprising, therefore, that

disambiguation is a major issue en natural language processing." (Taylor 1995:647-648)[1]

Com relação à aquisição lexical, o problema principal consiste em como representar o léxico - um conjunto com um número potencialmente infinito de elementos – de forma a permitir o acréscimo de itens sem comprometer ou modificar o sistema. Outro problema é o fato de novos significados poderem ser incorporados a palavras já existentes (fato comum no caso de terminologias técnicas), o que, de certa forma, nos faz retornar à questão da polissemia.

O interesse na desambigüização de palavras não é recente, e remonta a 1950. (cf *Computational Linguistics* 1998 volume especial sobre desambigüização). Os primeiros trabalhos consistiam em elaborar classificadores "especialistas" que fossem capazes de enumerar os diferentes contextos em que uma palavra pudesse aparecer. Tais classificadores, porém, eram construídos manualmente, o que apresenta um problema para o processamento automático. Como já foi dito, o léxico é um "sistema" aberto: palavras novas são criadas, bem como novos significados para palavras já existentes são cunhados a todo momento. Alimentar manualmente uma base lexical seria um trabalho infindável que, além de tempo, também dependeria de uma vasta mão de obra, o que parece pouco vantajoso. Posteriormente, à medida que *machine readable dictionaries* (MRD) e bases lexicais do tipo WordNet (Fellbaum 1998) se popularizaram, passaram a ser utilizados no fornecimento de informações para a desambigüização automática. Do mesmo modo, porém, a utilização de bases lexicais "prontas" também não parece uma boa solução, pois só há um deslocamento do problema, uma vez, na maioria das vezes, estas são alimentadas manualmente.

Um tratamento realmente automático de dados lexicais, com vistas à interpretação semântica de palavras polissêmicas, pode ser vislumbrado com abordagens estatísticas, como veremos a seguir.

3. Abordagens estatísticas no tratamento lexical

Tentando eliminar o trabalho humano das tarefas de aquisição/ representação lexical, tem-se investido em abordagens estatísticas do léxico, que vêm trazendo resultados promissores e a possibilidade de tratamento de fenômenos como a polissemia (Schütze 1998, Widdows 2002, 2003, Farkas & Li 2002).

As abordagens estatísticas podem ser baseadas em aprendizagem supervisionada e aprendizagem não-supervisionada. No primeiro caso, o processo de desambigüização faz uso de um corpus de treinamento já rotulado. Cada ocorrência de uma palavra ambígua é anotada com um rótulo semântico. Na aprendizagem não-supervisionada, o que está disponível para o treinamento é um corpus não rotulado.

Em termos gerais, modelos estatísticos baseados em coocorrência funcionam da seguinte maneira: a partir de um vasto corpus textual, conta-se, para uma dada palavra-alvo, o número de palavras que aparecem ao seu lado em uma janela de tamanho pré-determinado – por exemplo, 15 palavras. Na etapa seguinte, cada palavra é representada por meio das freqüências cumulativas das ocorrências no escopo da janela. Palavras

[1] "Uma sentença que contenha *n* palavras, cada uma delas *m* vezes polissêmica terá em princípio n x m leituras potenciais. (...) Não é de surpreender, então, que a desambigüização seja uma importante questão no processamento de linguagem natural". (Tradução dos autores)

com significados similares tenderão a ocorrer em contextos similares e palavras polissêmicas tenderão a ocorrer em contextos diferentes.

Subjacente a esses modelos, está a idéia de que o significado de uma palavra corresponde ao seu padrão de uso, e não ao significado considerado autonomamente. Porém, muitas vezes a decisão de não considerar o significado propriamente dito – ou intrínseco – das palavras é tomada por praticidade, pois, como diz Schütze, "(...) *providing information for sense definitions can be a considerable burden*."(1998: 97).[2] Segundo o autor (Schütze 1998), para se definir o significado "verdadeiro" das palavras – o que ele chama de etiquetagem de significados (*sense labeling*) –, é necessário se levar em conta uma fonte externa de conhecimento, que pode ser tanto um dicionário, um corpus bilíngüe, *thesauri* ou conjuntos de treinamento de etiquetados manualmente.

Schütze, assim, aponta para a dificuldade de se chegar ao significado "verdadeiro" de uma palavra; entretanto, ele não nega a sua existência. Do mesmo modo, Widdows (2003), que também apresenta um modelo de aquisição e desambigüização lexical baseado em informação contextual, afirma que o significado pode ser descrito de forma "clara, flexível e acurada", através de um pensamento científico cuidadoso e de investigação empírica. Ainda segundo Widdows (2003), métodos estatísticos, embora tenham trazido enormes contribuições, apenas adivinham o significado das palavras.

Embora o Cortex também desconsidere o significado propriamente dito, ou intrínseco, das palavras, o faz motivado teoricamente. No âmbito de uma teoria de inclinação pragmática como a de Firth, o significado de uma palavra é compreendido justamente como decorrência das suas relações com o contexto. Seguindo a linha wittgensteiniana, Firth afirma que "you shall know the meaning of a word by the company it keeps" (1957: 194-6); e, de acordo com Cruse (1986), "o significado de uma palavra é constituído por suas relações contextuais". Ou seja, parte-se do pressuposto de que não há significado fora de um contexto. No caso específico do processamento realizado pelo Cortex, "contexto" corresponde estritamente ao ambiente lingüístico em que uma palavra pode ocorrer, e nada mais além disso. De forma mais específica, contexto corresponde a uma janela cujo limite é o ponto final.

No Cortex, o significado é compreendido como uma rede de relações entre as palavras; o significado de uma palavra *p* é determinado pelas relações entre *p* e as outras palavras que coocorrem com *p*. Especificamente, a cada significado de *p* corresponde uma rede de relações diferente. Assim, por exemplo, a palavra *ataque* pode vir numa relação com *jogador* e *futebol*, em que é possível depreender o seu significado inserido em uma situação de *esportes*. Em outro contexto, pode vir acompanhada de *bombas* e *terrorismo*, o que reflete o significado de *agressão*, e ainda pode aparecer coocorrendo com *sintoma* e *medicamento*, incorporando o significado de *acesso de doença*.

Tomamos também como ponto de partida que mesmo as palavras que não são tradicionalmente consideradas polissêmicas precisam ser "desambiguizadas", pois apenas o contexto de uso faz refletir a interpretação a ser tomada pela palavra em questão. Uma palavra como *jogador*, por exemplo, pode parecer tanto em um contexto que dirija a significação para *jogador de vôlei* quanto em um contexto que dirija a

[2] (...) fornecer informações para definições de sentido pode ser uma tarefa considerável". (Tradução dos autores)

significação para *jogador de futebol*. Essa característica fica especialmente evidente em PLN, uma vez que todas as palavras são potencialmente ambíguas (polissêmicas ou homônimas) e só o contexto desfaz a ambigüidade.

4. O Cortex

A abordagem do Cortex toma como inspiração o modelo de Schütze (1998), segundo o qual o significado de uma palavra ambígua pode ser distinguido a partir da análise dos seus padrões de contextualização. Neste modelo, tanto os significados quanto os contextos de uso de uma palavra ambígua são representados como direções em um espaço vetorial, e um contexto é atribuído a um significado quando ambos possuem a mesma direção. A idéia básica é que quanto mais vizinhos em comum duas palavras tiverem, mais similares elas serão; e quanto mais palavras similares aparecerem em dois contextos, mais similares os dois contextos serão. O modelo compreende duas etapas: treinamento e desambigüização. Na primeira etapa, em um corpus de treinamento não-rotulado, acontece a contagem da freqüência de coocorrência entre as palavras. Nesse momento são calculados os vetores das palavras, os vetores de contexto e os vetores de significado. Todos os contextos de uma palavra ambígua são coletados no corpus de treinamento. Numa etapa posterior, já no corpus de testagem, a partir das informações coletadas na etapa de treinamento, é possível desambiguizar uma determinada palavra-alvo.

No Cortex, assim como verificado em Schütze (1998), a hipótese subjacente à desambigüização é a de que o significado pode ser caracterizado em termos de padrões de contextualização. Por isso, no processamento realizado no Cortex também não há rótulo ou etiquetagem das palavras, ou seja, não há um valor intrínseco para cada palavra. A forma de se chegar ao "significado" é através das relações de coocorrência entre as palavras.

O Cortex contém um algoritmo estatístico que extrai conhecimento sobre o contexto das palavras em um corpus não rotulado. O algoritmo é aplicado diretamente ao corpus de testagem (no caso aqui apresentado, composto por dois meses de notícias de jornal).

A discriminação do contexto ocorre da seguinte maneira: no escopo de uma janela cujo limite é o ponto final[3], conta-se, ao longo do corpus, a quantidade de vezes que uma palavra coocorreu com todas as outras palavras. É realizado, então, um teste de hipótese para determinar a significância da relação entre duas palavras, isto é, se elas ocorreram ao acaso ou não. Um grafo é formado tomando-se como nós as palavras e arestas de todas as relações significativas entre os nós. Palavras funcionais não são computadas neste processo: um banco lexical com uma lista destes itens trata de eliminá-los a fim de reduzir o esforço computacional. Esta escolha deve-se ao fato de que palavras funcionais apresentam alta freqüência de ocorrência e também de coocorrência, o que as torna candidatas a estarem presentes no contexto por mero acaso.

A partir de uma palavra-alvo *p*, é utilizado um algoritmo que busca as palavras mais relacionadas a p e que têm, simultaneamente, uma grande quantidade de ligações

[3] Segundo Manning & Schütze (1999), cerca de 90% dos sinais de ponto de um texto correspondem realmente a pontos finais; logo, é razoável adotar o ponto como limite para tratamento estatístico de textos.

entre si. Essas palavras irão constituir um aglomerado, que, de certa forma, pode ser considerado um campo semântico. Palavras que contêm muitas ligações são consideradas fracas, e são dispensadas logo no início, já que suas ligações acabam sendo pouco representativas. Uma mesma palavra pode pertencer a diferentes aglomerados, o que seria indicativo de sua polissemia. Como já mencionado, não apenas palavras tradicionalmente consideradas polissêmicas apresentam diferentes aglomerados (isto é, diferentes contextos). Uma palavra como *jogador*, por exemplo, pode aparecer tanto no contexto vôlei como no contexto futebol. Quanto mais aglomerados forem detectados, mais refinada será a distinção entre as palavras.

Além da possibilidade de desambigüização de qualquer palavra, e não apenas das classificadas como ambíguas, isto é, aquelas para as quais foram encontrados dois contextos distintos no corpus de treinamento, e da ausência de um número pré-determinado de contextos, o Cortex difere do modelo de Schütze por ser uma abordagem híbrida que utiliza, além de estatística, otimização por grafos. Nesse aspecto, o modelo se aproxima da abordagem de Widdows (2003) e Widdows & Dorow (2002), a qual busca, através de grafos, demonstrar relações semânticas entre as palavras. Após a montagem do grafo, basta uma condição inicial, ou uma palavra *p*, para que o sistema encontre automaticamente todos os diferentes contextos em que *p* aparece – ou seja, todos os seus "significados". Esses resultados podem variar em função de alguns parâmetros que devem ser configurados antes da busca. São eles:

N: quantidade máxima de sentidos a serem procurados = 100
A: quantidade de palavras armazenadas por contexto inicial = 10
L: limite permitido de ligações que uma palavra pode ter para participar de um contexto = 100%
S: fator de similaridade para unir dois contextos = começa com 90% e diminui até 70%

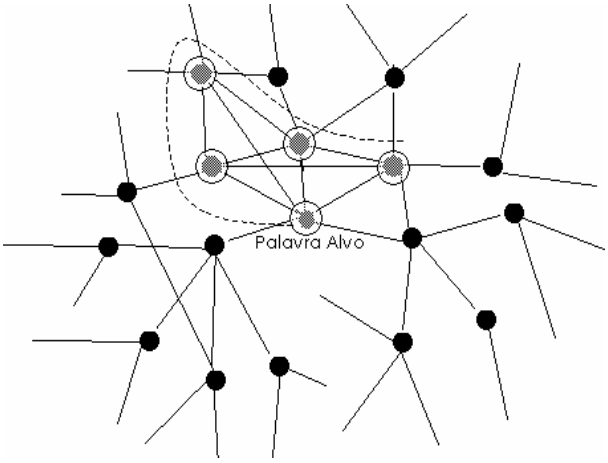

Figura 1. Criação dos contextos

O algoritmo de *clustering* é iterativo e funciona como uma aranha movimentando-se em uma teia. Inicialmente posicionada sobre a palavra-alvo *p*, a "aranha" passeia pelo grafo de acordo com as ligações mais fortes. É criado o "contexto 1", partindo da palavra mais forte, e, a cada passo, testamos se o novo nó (palavra) tem uma quantidade de ligações suficiente (L) para entrar no contexto. Em caso afirmativo,

a aranha adiciona a palavra ao contexto 1 e continua a passear na teia para a próxima palavra, e assim sucessivamente. Em caso negativo, o contexto 1 é fechado e é criado um novo contexto ("contexto 2"). O mesmo procedimento é realizado partindo da segunda palavra de relação mais forte. Os contextos vão sendo criados até acabarem as palavras.

Como o número de contextos encontrados inicialmente foi bastante grande, foi necessário limitar a procura em N contextos, e criou-se uma 2ª etapa para realizar um enxugamento dos contextos encontrados. Assim, uma vez detectados todos os contextos possíveis, o programa passa a re-agrupar da seguinte forma: primeiramente aglutina os conjuntos com similaridade de 90%, e, com esse resultado, aglutina novamente até não ser mais possível formar nenhum conjunto. Em seguida, o valor de S é diminuído para 80% e todo o procedimento se repete. Em uma última etapa, o valor de S cai para 70%.

O resultado da aplicação dos algoritmos é uma lista de palavras representantes de um contexto. Nos quadros 1 e 2, abaixo, estão os resultados de busca para as palavras *título* e *prova*.

Quadro 1: resultado de processamento do Cortex para a palavra *título*

1.		**2.**		**3**.
dívida	histórico	brasileira	Brasil	Milan
brasileira	novo	principal	vitória	sexto
Valor	US	brasileiro	time	europeu
principal	Brasil	final	estréia	
externa	subia	Milan	Roma	
face	consecutivo	Manchester	campeão	
Bond	histórica	domingo	espanhol	
alta	tarde	Lancepress	vaga	
cotado	lucros	competição	Federer	
negociado	imposto	conquistar	Santos	
final	inédito	Campeonato	inglês	
opera	internacionais	turno	gols	
valorização	positiva	conquistou	Fábio	
recorde	registrada	conquista	mantém	
cotação	segue	disputa	Emanuel	
dia	cai	seleção	conquistado	
risco	mantém	feira	Mantilla	
Real	disco	Liga	derrotar	
exterior	queda	jogos	casa	
tendência	investidores	mundial		
feira				

Quadro 2: resultado de processamento do Cortex para a palavra *prova*

1.	**2.**	**3**.	**4**.	**5**.	**6**.
Venceu domingo lugar brasileiro GP vencedor	piloto corrida brasileiro Indianápolis dia	balas coletes bala colete	especial domingo americano brasileiro Indianápolis Pan Americano agosto	pegada importando Americaninha pedofilia Memorando gente	exame legítima Jamilly

No Quadro 1, estão os resultados do processamento realizado tendo a palavra *título* como alvo. O sistema encontrou três contextos para ela. A partir das palavras presentes no contexto 1, é possível perceber que título está sendo utilizado em um contexto de economia. As palavras coocorrentes – *cotação*, *dívida, externa*, *investidores* e *lucros*, entre outras –sugerem tal interpretação. Já o contexto 2 atribui a *título* o significado de *prêmio em uma competição esportiva*, como sugerem as palavras *campeão*, *campeonato*, *competição*, *conquista*, *jogos*, *time*, entre outras. Por fim, o contexto 3 parece ser um refinamento do contexto anterior: trata-se de uma competição européia, como sugerem as palavras *europeu* e *Milan.*

Pelo Quadro 2, é possível observar que o sistema foi capaz de distinguir 6 contextos de uso de *prova*. A partir das palavras presentes no contexto 1, é possível atribuir a *prova* o significado de *competição*. Já o contexto 2 pode ser entendido como uma subdivisão (refinamento) do contexto 1: não se trata de uma competição qualquer, mas de uma *competição de corrida*, possivelmente de carros, como indicam as palavras coocorrentes *corrida* e *piloto*. No contexto 3*, prova* adquire o valor de *resistente a – à prova de balas*, no caso. A identificação do contexto 4 não é tão evidente, mas a presença da palavra *Pan* sugere que também se trate do contexto *competição*. Por fim, os contextos 5 e 6 parecem atribuir à prova o significado de *indício*: no contexto 5, as palavras coocorrentes *pedofilia* e *pegada* levam a esta interpretação; no contexto 6, *exame* e *legítima* também sugerem que *prova* está sendo interpretada como *indício*.

Um detalhe que chama a atenção em ambos os quadros é o grande número de nomes (substantivos e adjetivos), em oposição a verbos, presentes nos aglomerados. No grupo *prova*, por exemplo, das 32 palavras utilizadas na caracterização, apenas duas são verbo (e, assim mesmo, uma delas é uma forma nominal de verbo). No grupo *título*, das 83 palavras, apenas 8 são indubitavelmente verbos, o que pode ser indicativo da força da classe nominal na caracterização de contextos.

5. Considerações Finais e Direcionamentos Futuros

Apresentamos aqui os resultados preliminares do Cortex, um processador de linguagem natural. Tais resultados são relativos à atribuição de significado às palavras, e sua conseqüente desambigüização. O sistema é inspirado no modelo de Schütze (1998) e é capaz de identificar todos os contextos relativos à palavra-alvo que aparecerem no corpus.

Algumas questões permanecem em aberto. Uma delas é em que medida é vantajosa uma definição altamente refinada de uma série de contextos, como é possível observar na desambigüização de *prova*. Talvez não haja uma única resposta, e o grau de vantagem dependa diretamente dos objetivos do usuário.

De forma geral, os resultados, embora significativos, ainda precisam de ajustes.ë o caso, por exemplo, de contextos que parecem incluídos em outros, como aconteceu com a palavra *título*, em que o contexto 3 é um subconjunto do contexto 2. Algumas melhorias já estão sendo incorporadas para possibilitar um "enxugamento" nos resultados. Por exemplo, a duplicação de palavras decorrentes da flexão de plural nos nomes será eliminada. Assim, no contexto 3 de *prova*, por exemplo, *bala/balas* e *colete/coletes* darão lugar a somente uma instância de cada palavra: no resultado final aparecerá apenas *bala* e *colete*. Do mesmo modo a presença de formas flexionadas de um verbo será eliminada no resultado final. No contexto 2 de *titulo*, *conquistar* e *conquistou* darão lugar a uma instância apenas, possivelmente *conquistar*.

Outro ajuste diz respeito aos nomes próprios e compostos, que atualmente são considerados duas palavras distintas, como é o caso de *Estácio de Sá*, por exemplo, que apareceu em outro experimento como duas palavras (*Estácio* e *Sá*), quando na verdade é um nome próprio composto. Pretende-se refinar o processamento para incluir expressões compostas. Além disso, combinações de duas ou três palavras com um padrão muito alto de coocorrência passarão a ser consideradas apenas um item lexical. No contexto 4 do grupo prova, muito provavelmente as palavras *Pan* e *americano* se juntariam para formar o item *Pan americano*, e no grupo 1 de *título*, o mesmo aconteceria com *dívida* e *externa*, que passaria a ser considerada *dívida externa*.

Ajustes servirão também para definir melhor, ou eliminar, contextos pouco claros, como é o caso dos contextos 5 e 6 de *prova*.

Ainda assim, acreditamos que os primeiros resultados conseguidos pelo Cortex apontam para um possível caminho para a utilização de dados estatísticos para desfazer automaticamente a ambigüidade de palavras em contexto.

Referências Bibliográficas

Cruse, D. (1986) Lexical Semantics. Cambridge: CUP.

Farkas, I.& Li, P. (2002) "Modeling the development of lexicon with a growing self-organizing map", Proceedings of the 6th Joint Conference on Information Sciences, Research Triangle Park, NC, p. 553-556.

Fillmore, C. & Atkins, B. (2000) "Describing polisemy: the case of 'Crawl' ", In: Ravin & Leacock (eds.). Polysemy. Theoretical and computational approaches. Oxford: Oxford University Press.

Firth, J. R. (1957). Papers in Linguistics – 1934-1951. Oxford: Oxford University Press.

Manning, C & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts: The MIT Press.

Schütze, H. (1998). "Automatic Word Sense Discrimination", Computational Linguistics, 24(1), 97-123.

Schütze, H & J. Pedersen.(1995). “Information retrieval based on word senses”, Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, EUA, p. 161-175.

Taylor, J. (2003). “Polisemy’s paradoxes”, Language Sciences (25) 637-655.

Widdows, D. (2003). “A Mathematical Model for Context and Word-Meaning”, Fourth International and Interdisciplinary Conference on Modeling and Using Context, Stanford, California, June 2003, pp. 369-382.

Widdows, D & Dorow, B. (2002). “A Graph Model for Unsupervised Lexical Acquisition”, Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Taipei, Taiwan, p.1093-1099.