DOI: 10.5433/1981-8920.2009v14nesp.p125

DETECÇÃO E EXTRAÇÃO DE CANDIDATOS A ACEPÇÕES BASEADAS EM UM TESAURUS DE COLOCADOS

DETECCIÓN Y EXTRACCIÓN DE CANDIDATOS A LAS ACEPCIONES BASADAS EN UN TESAURUS DE COLOCADOS

J. L. De Lucca - joluc@doctor.upv.es
Pesquisador acadêmico da Universidade Católica de Portugal e da Universidad Politécnica de Valencia.
Doutor em Linguística Geral pela Universidade de São Paulo.
Autor da Elsevier Science.

Resumo

A desambiguação envolve a determinação de todos os distintos sentidos de cada palavra sob considerações que dependem da anotação manual do sentido de cada palavra atribuindo o sentido apropriado a cada ocorrência de uma palavra. Este artigo descreve um programa para a desambiguação automática dos múltiplos sentidos de uma mesma palavra (ambiguidade semântica) identificando em um *corpus* textual – CHADES (Corpus Hispano-Americano de Español), de modo que as frases, sem restrição, estão semanticamente relacionadas ao mesmo grupo de sentidos encontrado neste *corpus.* O modelo de ferramenta apresentado buscou solucionar os problemas de lexicografia, recuperando exemplos relevantes ao usuário, por meio da análise de contexto. Reforça a necessidade do desenvolvimento de recursos linguísticos e linguístico-computacionais volumosos que permitam a representação do conhecimento expresso em texto de línguas naturais.

Palavras-chave

Tesauro de colocados. Indexação. Acepção.

1 INTRODUÇÃO

A palavra não existe sozinha, já dizia Vygotsky (1987). Ela adquire significado apenas quando se encontra em um contexto, na companhia de outras palavras, pois há sempre uma diferença muito grande entre o significado de uma palavra no dicionário, onde está sozinha, e o que ela adquire quando se encontra na companhia de outras

palavras. O significado dicionarizado de uma palavra nada mais é do que uma peça no xadrez do sentido, não passa de uma potencialidade que se realiza de diversas formas, na fala e na escrita.

É o contexto que determina o significado da palavra. Por exemplo, nas frases: "*Se sentó en uno de los últimos bancos de la iglesia*" e "*en el Atlántico Norte hay bancos de atunes*", a palavra banco adquire significados diferentes. No primeiro exemplo, banco significa lugar para sentar e no segundo, banco significa grande quantidade.

Para levar a cabo nosso projeto, utilizamos como '*training corpus'* (um corpus menor, próprio para testes) um *corpus* em língua espanhola, com cerca de cinco milhões de palavras, o *Corpus Hispanoamericano de Lengua Española* (CHADES), desenvolvido entre 1996 e 2006 pelo autor deste artigo. O projeto visa a extração automática de palavras polissêmicas (aquelas que se caracterizam por mais de uma acepção) em língua espanhola.

Nosso objetivo é o desenvolvimento de um modelo de ferramenta para solucionar um dos problemas da lexicografia ou tratamento de *corpora* ou agrupamento de contextos, isto é, qualquer informação que possa ser usada para caracterizar a situação de uma entidade em torno de *nódulos*, identificados como fornecedores de informação sobre uma entidade particular ou uma coleção de entidades. Jackendoff (1990), recuperando os exemplos mais relevantes do *corpus* para cada nódulo, através da análise de contexto. Para tanto, partimos do desenvolvimento de ferramentas próprias, pois nenhum *software* no mercado, tais como, Wordsmith®, Concordancer®, Hyperbase® e outros, atenderia a todas as exigências requeridas pelo projeto em questão. Assim, desenvolvemos um algoritmo que nos possibilitou extrair do *corpus*, candidatos a sentidos lexicais, baseados em um *HiperThesaurus* (aqui entendido como estrutura capaz de armazenar palavras ou repositório de conhecimento e elos ou ligações entre palavras que constituem determinado contexto). Uma estrutura de *HiperThesaurus* é capaz de armazenar palavras e as relações entre elas, permitindo a recuperação de informações contextuais, não apenas por palavras-chave mas, também, por um conjunto de palavras inter-relacionadas que caracterizam determinado contexto.

Decidimos tomar como dicionário padrão da língua, o *Diccionario Salamanca de la Lengua Española* (DSLE), ressalvando que, para aquelas frases cujo sentido não

estivesse entre as acepções registradas pelo DSLE, nós agregaríamos como acepções adicionais.

Inúmeros testes foram realizados levando em consideração nossa experiência em lexicografia e, também, os modelos para desambiguação de outros autores. Os resultados aqui apresentados referem-se a um modelo desenvolvido por nós – mas não inédito – baseado em análise co-ocorrencial.

2 METODOLOGIA

Antes de prosseguirmos, uma primeira pergunta que não pode ficar sem resposta: Que modelo de desambiguação lexical seguir? Os modelos são divididos em: baseados unicamente em *corpus* ou baseados em *corpus* e dicionários. Ambos podem ser: anotado ou sem anotação. O modelo pode, ainda, ser voltado para a desambiguação de todas as palavras: um modelo geral, independente da palavra a desambiguar, ou um modelo empregado apenas para subconjuntos de palavras, geralmente modelos específicos para cada palavra, com característica e estrutura totalmente distintas dos modelos empregados para outras palavras.

Nosso modelo baseia-se em *corpus* sem anotação, cujo modelo é empregado para todas as palavras a desambiguar. Este método é semelhante aos empregados por Geffroy (1975), Dolan (1994), Hongyan e Tzoukermann (1999). Este *corpus* é conhecido como *corpus* de treinamento, que fornece as informações de que necessitamos para criar uma descrição do vocabulário da língua. Para uma dada palavra, podemos descobrir uma gama de informações, incluindo:

Semântica - Evidencia os diferentes sentidos de uma palavra. Neste projeto, as informações semânticas foram utilizadas para a anotação semântica.

Sintática - Como uma palavra se combina gramaticalmente com outras. Neste projeto, as informações sintáticas serviram para determinar as *stopwords*[1] a serem extraídas das frases.

Contextual - O contexto, no qual uma palavra aparece e as frases em que faz parte. Neste projeto o contexto é cada frase onde o 'Nódulo' aparece.

[1] Stop words são artigos, proposições, pronomes, palavras curtas e comuns.

Estilística - Os tipos de texto onde uma palavra ocorre com maior frequência, se textos literários, poéticos, jornalísticos, técnicos, *chats* ou *email*. No projeto, o *corpus* é composto por uma variedade de textos para evitar que haja uma tendência para algum domínio do conhecimento.

Nosso corpus é constituído por cinco milhões de palavras extraídas de textos modernos e clássicos da literatura hispano-americana e espanhola, assim como de artigos publicados em revistas científicas e em jornais da América Latina e Espanha. Os textos foram digitados, digitalizados e também extraídos da Internet.

Escolhemos o *nódulo* dentro da frase ou oração em que ocorre, respeitando exclusivamente a pontuação forte. Entendemos por pontuação forte os pontos finais, bem como o ponto de exclamação, interrogação e reticências, desde que após estes, haja uma nova frase ou período. Frase é todo enunciado linguístico capaz de transmitir uma ideia. Oração é todo enunciado linguístico que se estrutura ao redor de um verbo.

Separamos os itens lexicais das *stopwords* e definimos que seriam no máximo cinco (5) as unidades lexicais dos colocados da esquerda, e também cinco (5), no máximo, os colocados da direita do *Nódulo*. Embora em muitas frases haja mais do que cinco unidades lexicais à esquerda ou à direita foram consideradas apenas as cinco primeiras.

Consideramos a frase como unidade de significação segmentável, pelo que privilegiamos, por um lado, a pontuação forte e por outro, apenas de momento, as relações internas, em nível de estrutura frásica entre os vocábulos de um mesmo grupo.

No projeto, a evidência estatística relativa de diferentes palavras ou sentidos ou construções gramaticais, serviu para selecionar as co-ocorrentes mais próximas do *Nódulo*, assim como aquelas palavras que se identificam mais com um determinado *Nódulo*

O registro do corpus foi desenvolvido em um banco de dados em MSDOS e teve o seguinte formato:

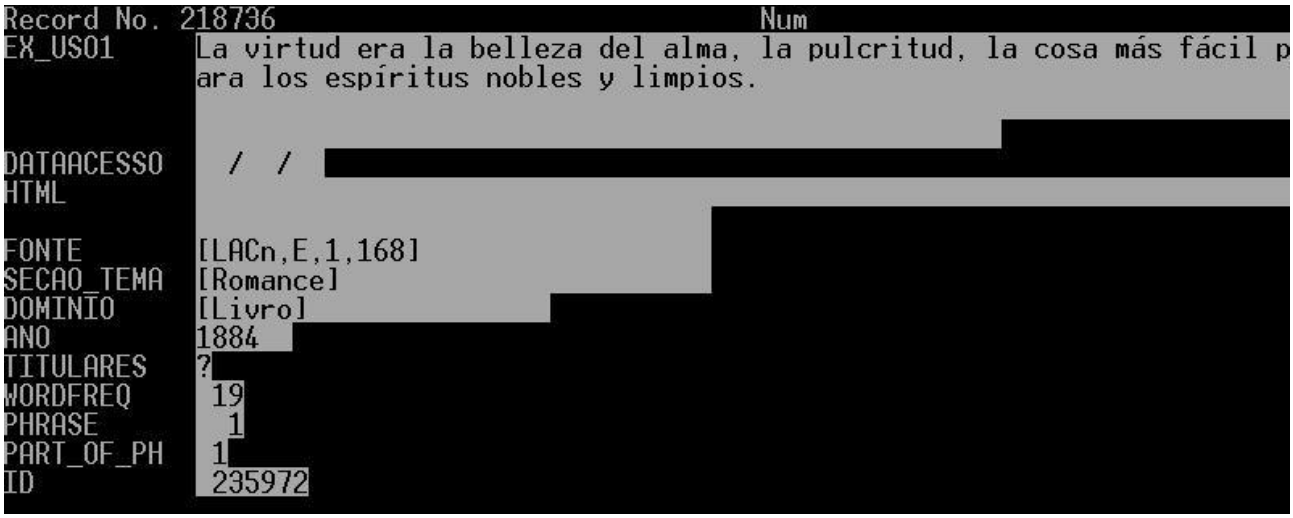


Figura 1 – Registro do *corpus*

O campo EX_USO1 refere-se à frase extraída do *corpus*. O campo DATAACESSO destina-se apenas àquelas frases extraídas de textos provenientes da Internet. HTML contém a URL completa da fonte, quando esta tratar-se de texto proveniente da Internet. FONTE é a referência bibliográfica em código, que se encontra decodificado em outro banco de dados. SECAO_TEMA é o assunto/tema do texto de onde foi extraída a frase. DOMINIO refere-se ao tipo de produção: Livro, Revista ou Jornal. ANO é o ano em que foi publicado pela primeira vez. TITULARES, quando se tratar de manchetes. WORDFREQ, refere-se à quantidade de palavras da frase. Em PHRASE todos os registros são identificados com o número 1. PART_OF_PH refere-se a qual parte da frase, pois há muitas frases que tiveram de ser fragmentadas devido à sua extensão, algumas com mais de 1000 caracteres. ID é o número de controle.

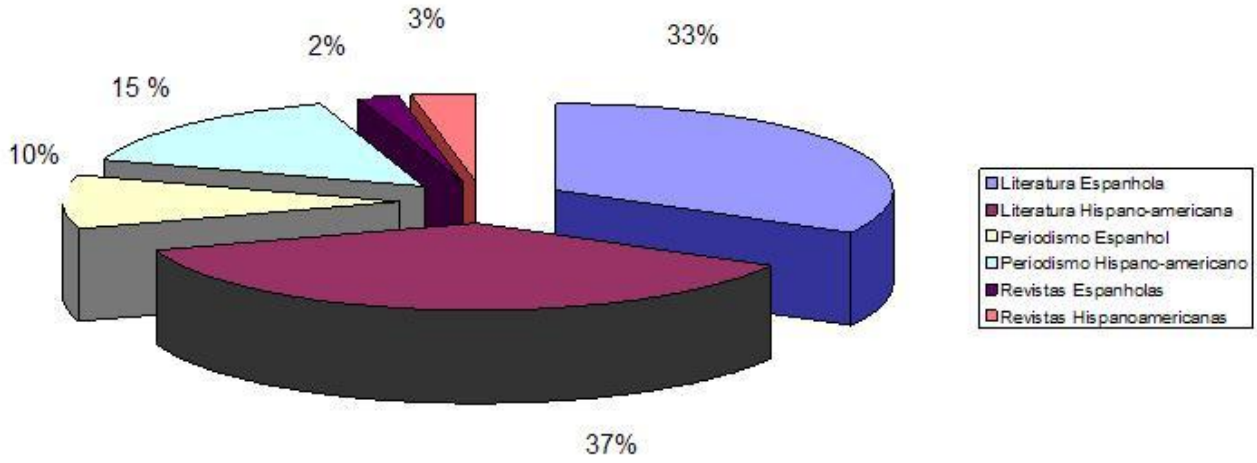


Figura 2 - Distribuição do *Corpus*

Conforme a figura 2, a maior parte do *corpus* é constituída pelas literaturas hispanoamericana e espanhola, respectivamente 37% e 33%; os jornais da América Latina e Espanha ocupam o terceiro e quarto lugares com 15% e 10% respectivamente. As revistas da América Latina e da Espanha ocupam modestos quinto e sexto lugares, ficando estas com 3 e 2%, respectivamente.

Etapas

1ª etapa: anotação semântica

A primeira etapa foi separar do *corpus* – mais de 300.000 frases - aquelas que possuem os Nódulos-alvo selecionados: *banco*, *acoger*, *querella* e *hambriento*.

Selecionadas as ocorrências de cada palavra alvo, passou-se então à anotação semântica, ou seja, a sua desambiguação, que consistiu em atribuir a cada palavra (*Nódulo*) o sentido mais apropriado do dicionário. O dicionário, no caso, foi o *Diccionario Salamanca de la Lengua Española* (1996), como já dito. Este processo realizou-se manualmente por apenas um anotador, embora tenha sido feito por três vezes para cada

Nódulo. O recomendado, contudo, é que tal tarefa seja feito por uma equipe, buscando um acordo final de aproximadamente 90% entre os anotadores.

2ª etapa: determinação do significado das palavras

A divisão das palavras em diferentes sentidos e entradas em diferentes dicionários é arbitrário (ATKINS; LEVIN, 1988; ATKINS, 1991; KILGARRIFF, 1993). Estes autores defendem que o significado deve ser decidido pelo contexto da palavra. Outros autores, contudo, defendem que uma palavra tem um número fixo e determinado de significados como definido no dicionário. No entanto, para uma mesma palavra, diferentes dicionários podem registrar diferentes números de significados. Um significado no dicionário A pode corresponder a dois significados no dicionário B. Isto mostra que a determinação e o número de significados, por verbete, não são absolutos. A ordem também pode não ser a mesma. Uma rápida comparação entre dois dicionários deixará isto mais claro. Por exemplo, consideremos a palavra *banco*, no quadro 1, abaixo. Ela tem oito significados no *Diccionario Planeta*, mas apenas sete no *Diccionario Salamanca*.

Diccionario Salamanca	***Diccionario Planeta***
Banco	***Banco***
1. Asiento para varias personas	1. Asiento largo y estrecho para varias personas
2. Establecimiento financiero y de crédito	2. Soporte, caballete o mesa que sirve para diferentes usos en numerosos oficios
3. Establecimiento sanitario donde se conservan sangre órganos y líquidos orgánicos	3. Bajío más largo que ancho
4. Conjunto numeroso de peces que se desplazan juntos	4. Multitud de peces que van juntos
5. Elevación del fondo en mares, ríos o lagos	5. Estrato de gran espesor, veta de una cantera
6. Conjunto de nubes o de niebla	6. Organismo público de crédito
7. Mesa de trabajo de algunos artesanos	7. Local donde se ejercen las operaciones de banca
	8. Organismo encargado de la conservación de órganos o líquidos fisiológicos

Quadro 1 – As acepções de *banco*

O segundo significado no *Diccionario Salamanca* corresponde a dois significados no *Diccionario Planeta* (6 e 7). O significado 6 do *Diccionario Salamanca* não existe no *Diccionario Planeta*, enquanto o sentido 5 do *Diccionario Planeta* não existe no *Diccionario Salamanca*.

Na anotação semântica as acepções ou sentidos que não foram registrados no *Diccionario Salamanca*, foram adicionadas com base em outros dicionários de língua espanhola ou em nosso conhecimento da língua. Desse modo, a anotação semântica para *acoger*, *banco*, *hambriento* e *querella* ficou, em resumo, distribuída conforme figuras 3, 4, 5 e 6:

A primeira coluna identifica o sentido, *tagged sense*; a segunda coluna, o nódulo e a terceira coluna, a frase extraída do *corpus*.

1	acoger	Juli Clavijo: "El castigo consistía en acoger a una familia de refugiados"
2	acoger	Vegadeo acogerá unas jornadas sobre la madera y el sector forestal
3	acoger	Ahora bien, la unión es difícil de sostener en la actualidad por la imprescindible especialización a la que el
4	acoger	El Pleno acogerá varias peticiones para el problema de la perrera
5	acoger	Se puso por esto a su servicio, se decidió a imitarle y a acoger sus advertencias en lo tocante a las práctic
6	acoger	Y si acaso tanta acción arrebatada consiguiera suplir las necesidades de techo, la permanencia de un ent
7	acoger	Washington señala a Somalia como «país potencial» para acoger otra acción militar
8	acoger	La hospitalidad es un modo de acoger la diferencia, en el intercambio, en el juego de coexistir en la diversi
9	acoger	Por eso, el acoger propio de la tolerancia es un acto que trasciende incluso la hospitalidad domiciliaria.

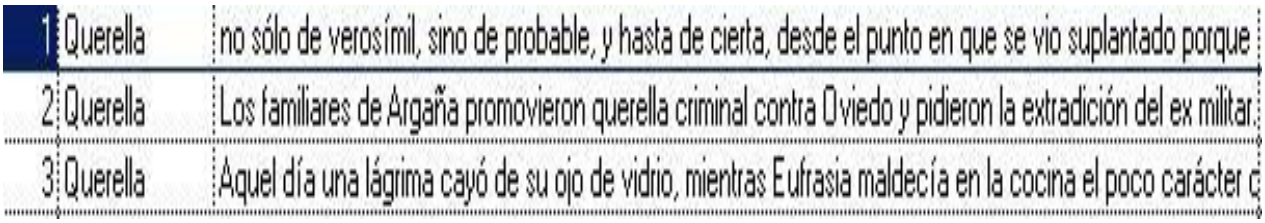
Figura 3 – *acoger*

1	banco	Se los alcanzaba por debajo del banco.
2	banco	empobrecimiento generalizado que ello representa; al mismo tiempo, rebajar las activas que cobran los bancos, pero siempre manteniéndolas
3	banco	Los bancos de niebla persisten en Zamora
4	banco	China abre su primer banco de cerebros dedicado a investigación
5	banco	El campeón sin Riquelme y con Jorge Bermúdez en el banco
6	banco	Novedoso banco de pruebas en Galicia
7	banco	Sin embrago, estas estadísticas se obtienen de los llamados bancos de árboles (tree-bank, en inglés), es decir, corpus analizados y marcado
8	banco	concentrados en los países industrializados; lo mismo con las investigaciones, las patentes, los bancos y bases de datos, los sistemas de red
9	banco	Un gran tropel hacía vibrar la pampa, y otros vaqueros atravesaron el "banco" antes que la yeguada apareciera a mi vista, de cuyo grupo de
10	banco	Su hijo y Andrés le reemplazaban en el banco de la paciencia (así llamaba él al escritorio a la antigua).

Figura 4 – *banco*

1	hambriento	No era ella seguramente la hambrienta sino los cachorros; ni se explicaba él de otro modo tan corteses mo
2	hambriento	Es la tiniebla blanca, nebulosa y hambrienta como el deseo.
3	hambriento	La deuda emergente ha tenido un buen desempeño desde principios del año 2002, ya que los inversores l

Figura 5 – *hambriento*



1	Querella	no sólo de verosímil, sino de probable, y hasta de cierta, desde el punto en que se vio suplantado porque
2	Querella	Los familiares de Argaña promovieron querella criminal contra Oviedo y pidieron la extradición del ex militar.
3	Querella	Aquel día una lágrima cayó de su ojo de vidrio, mientras Eufrasia maldecía en la cocina el poco carácter c

Figura 6 - *querella*

3ª etapa: Análise dos co-ocorrentes e a estatística sintagmática

Nossa perspectiva de co-ocorrência pode ser definida por (DUBOIS et al., 1973, p. 153) “Diz-se que os elementos B, C, D são co-ocorrentes de outro elemento A quando figuram com A para produzir um enunciado, vindo cada um deles numa posição determinada”.

Na estatística sintagmática – estudo da co-presença das formas no interior do mesmo texto –, os elementos caracterizam-se pela continuidade e contiguidade, obrigando-nos a não esquecer o contexto em que as formas estão inseridas.

O *Nódulo* ou palavra de busca, que é a palavra acompanhada do texto ao seu redor ou co-texto, pode constituir-se de uma ou mais palavras. Selecionamos quatro *nódulos* para esta análise: *banco*, *acoger*, *hambriento* e *querella*. A escolha foi aleatória. Em nosso *corpus* de cinco milhões de palavras encontramos: 461 frases contendo a unidade lexical *banco*; 105 para *acoger*; 73 para *querella* e, finalmente, 56 para *hambriento*.

Existem algumas palavras presentes em um documento que são utilizadas com o intuito de conectar as frases: *stopwords*. Essas e outras palavras pertencentes a classes de palavras cuja finalidade é auxiliar a estruturação da linguagem (tais como conjunções e preposições), não foram incluídas na estrutura de índice. Além dessas, também existem palavras cuja frequência na coleção de documentos é muito alta - palavras que aparecem em praticamente todos os documentos de uma coleção não são capazes de discriminar documentos e também não devem constar na estrutura de índice (FOX, 1992). Todas essas palavras consideradas sem valor para a busca, devido à sua natureza, frequência ou semântica, são denominadas palavras negativas (ou *stopwords*) (KOWALSKI, 1997; KORFHAGE, 1997; SALTON, 1983). Essas palavras dificilmente são utilizadas em uma

consulta, pois sua indexação somente tornaria o índice maior do que o necessário. Conforme quadro 2 abaixo.

DOCUMENTO NORMALIZADO	**DOCUMENTO SEM *STOPWORDS***
La amistad de Cuba se daba por descontada;	amistad Cuba daba descontada;
la del Perú, favorecida por la Revolución militar,	Perú, favorecida Revolución Militar,
había sin embargo que cultivarla y protegerla de las	había embargo cultivarla protegerla
viejas querellas históricas, latentes todavía en	viejas querellas históricas, latentes todavía
algunos sectores del ejército y de la opinión pública.	algunos sectores ejército opinión pública.

Quadro 2 – *Stopwords*

No tratamento estatístico dos dados obtidos, qualquer ocorrência de uma determinada forma lexical num texto homogêneo, mantém relações com as formas que a precedem ou que a seguem. Essas relações podem traduzir-se numericamente e são de vários tipos:

a) Lateralidade

Os dois tipos de colocados (esquerda e direita) foram tratados independentemente. Qualquer item recolhido foi anexado à expansão correspondente. Exemplos:

*Pero a mí clarito ma latía **querella** pero como al señor no le gustaba contestar.*

*El Supremo admite a trámite la **querella** por prevaricación contra los jueces del narco.*

*PSOE e IU presentan una **querella** contra Matas en el Supremo por delito electoral.*

Embora em todos os três exemplos existam mais do que cinco (5) colocados à direita, somente foram consideradas os cinco (5) primeiros, pois testes revelaram que se trabalharmos com um número maior, os resultados finais serão alterados e perder-se-á o nível de acertos atualmente existente.

$$\sum\left[\frac{(x+\mu)}{\delta}\right]^2$$

Figura 7 - Cofrequência observada (cfo)

b) cofrequência observada: número de vezes que uma forma lexical ocorre à volta do *Nódulo*

Este índice de cofrequência poderia permitir uma fácil classificação numericamente hierarquizável. Assim, seria considerada como a forma mais próxima do *Nódulo* a que aparecesse um maior número de vezes nos seus colocados à esquerda ou à direita. Diríamos então que essa forma era uma co-ocorrente do *Nódulo*. Não esqueçamos, contudo, que esta cofrequência depende diretamente do tamanho dos colocados considerados. Assim, essa forma poderá ser vizinha – encontrar-se nos colocados à esquerda ou à direita do *Nódulo* –, mas não co-ocorrente – para tal é necessário que seja atraída por esse *Nódulo*.

c) Proximidade (1/d)

Para medirmos a distância entre o Nódulo e os seus colocados, basta contarmos o número de itens lexicais que se interpõem. O índice assim obtido (d), seria o inverso da proximidade. Quanto mais próximo estivesse um item do *Nódulo*, tanto mais baixo seria esse item. Ora, quanto mais afastado estiver do *Nódulo*, tanto menor será a importância do item em relação ao *Nódulo*. Se esse item encontrar-se perto do *Nódulo* há que se valorizar essa proximidade. Para tal, o item imediatamente contíguo ao *Nódulo* será valorizado em 1/1, o segundo em 1/2 e assim sucessivamente.

d) Coeficiente de vizinhança (V)

O Coeficiente de Vizinhança (V) leva em conta os dois índices. Desta forma, a fórmula final agrega o cálculo da Proximidade e o cálculo da cofrequência observada, pois não se pretende privilegiar nem a cofrequência em relação à proximidade, nem aquela em relação a esta.

Procurando corrigir os dados da cofrequência em função da proximidade recorremos à seguinte fórmula para determinar o coeficiente de vizinhança.

$$V = \frac{\Sigma\left[\frac{(x+\mu)}{\delta}\right]^2 \sqrt{\log\left[\Sigma\frac{1}{d}\right]}}{\log_\lambda}$$

Figura 8 - Coeficiente de vizinhança

4ª etapa: HyperThesaurus

A palavra *thesaurus* é mais familiar ao público em geral no contexto da ferramenta desenvolvida originalmente por Roget. Este tipo de *HyperThesaurus* é semelhante a um thesaurus de recuperação de informação, pois contém alguma organização hierárquica, seu propósito básico é dar equivalentes. Porém, o propósito básico de um *thesaurus* de recuperação da informação (*information retrieval*) é bastante diferente. Um *Thesaurus*, como o de Roget, foi projetado para ajudar os escritores a acharem uma palavra ou frase que expressasse precisamente o significado que eles buscavam. Em contraste, um *thesaurus* de recuperação de informação concentra equivalentes próximos. Sua finalidade é autorizar um subconjunto de língua natural que reunirá informação relacionada o mais próxima possível para facilitar a recuperação de informação.

Um *thesaurus* pode ser definido como "um vocabulário controlado organizado em uma ordem conhecida" (EBECKEN; LOPES; COSTA, 2003, p. 337-370) com tipos especificados de relações, identificados por indicadores de relação unificados. Um vocabulário controlado é um subconjunto de língua natural, consistindo de termos preferidos e não preferidos. Os propósitos primários de um *HyperThesaurus* são identificados como promoção de consistência na indexação de documentos e facilitação de busca. A definição no padrão internacional é semelhante. Enquanto o contexto de padrões de *thesaurus* é *tags* para etiquetação do conteúdo de documentos, os princípios de desenvolvimento de *thesaurus* podem ser aplicados a uma grande variedade de tarefas de organização de informação.

Inicialmente construímos um *HyperThesaurus* com apenas seis palavras: dois substantivos (*banco* e *querella*); um verbo (*acoger*) e um adjetivo (*hambriento*).

As palavras que compõem o *HyperThesaurus* foram denominadas de Colocados, que são os traços que determinam os relacionamentos semânticos que caracterizam cada sentido ou acepção.

Como exemplo, construímos a seguinte Figura 9:

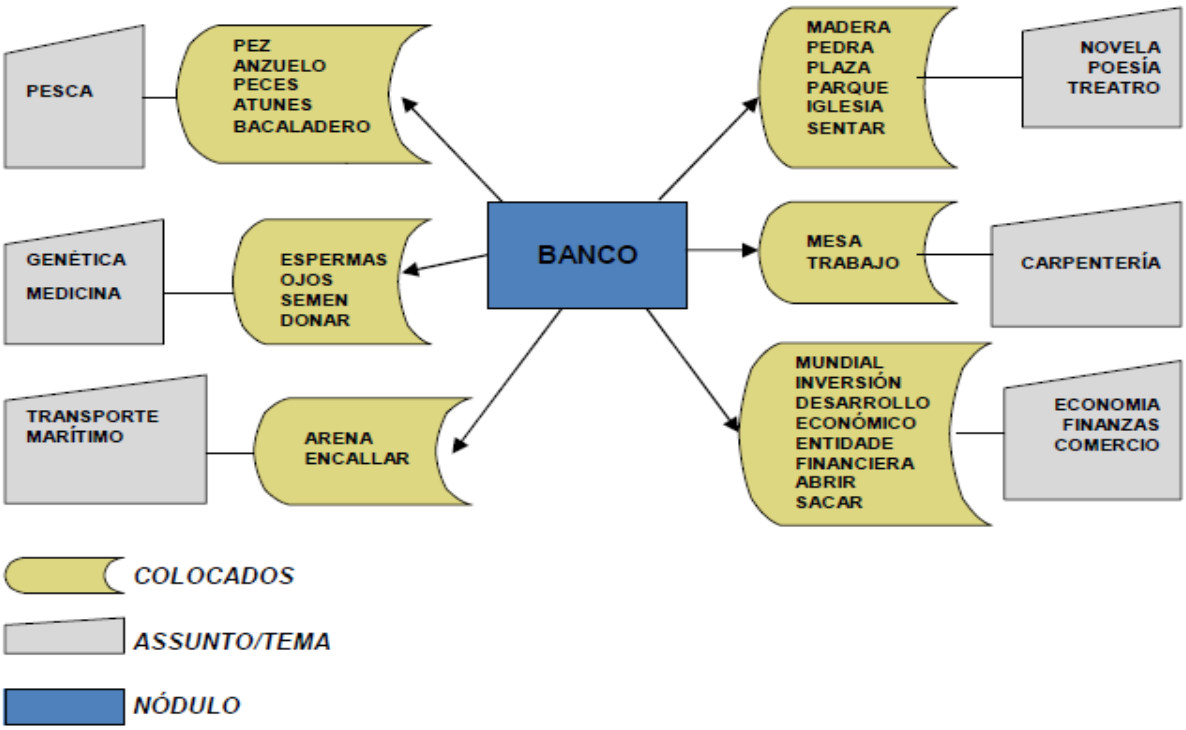


Figura 9- Exemplo de Colocados

Nesta figura vemos diversos colocados relativos ao *nódulo BANCO*. Em amarelo vemos os colocados e, em azul, o assunto ou tema. No exemplo acima, *MADERA*, *PEDRA*, *PLAZA*, *PARQUE*, *IGLESIA*, *SENTAR* são os colocados de um sentido de *BANCO, ou seja,* lugar para sentar. Novela, poesia e teatro são os tipos de textos onde ocorrem com maior frequência no *corpus*.

De posse de um *HyperThesaurus* é possível identificar, nas sentenças baseadas em *corpus*, quais são as palavras que pertencem a um determinado sentido/acepção, e isto é feito pelas ferramentas desenvolvidas neste projeto.

A ferramenta agora existe. Embora seu *HyperThesaurus* seja limitado, seu algoritmo e metodologia foram desenvolvidas e o que temos que fazer agora é ampliar o *HyperThesaurus* e adicionar mais algumas ferramentas, como um lematizador próprio para espanhol.

3 RESULTADOS

Antes de iniciarmos a apresentação dos resultados, é procedente definir *Precision* e *Recall*, de acordo com a desambiguação de sentido *(word sense disambiguation*):

Recall mede o grau de abrangência do processo de recuperação, ou seja, o total de frases relevantes que contêm a palavra polissêmica (de busca) que foi recuperada, dividido pelo total de frases em que a palavra polissêmica (de busca) aparece no *corpus*.

Já *Precision* mede o grau de precisão do processo de WSD, ou seja, avalia o número de vezes que uma palavra polissêmica foi corretamente desambiguada, dividido pelo número total de resultados. Também medimos aqui, um segundo tipo de *Precision*, relativo a sentidos/acepções que foram integralmente ou não desambiguadas.

Os testes não foram realizados com todas as frases disponíveis que contivessem o *Nódulo*. Adotamos o critério de corte pela relevância, que no caso, era o número de unidades lexicais do *HyperThesaurus* encontradas em cada frase extraída do corpus ou seja 5%.

A avaliação da eficiência em *word sense disambiguation* e *information retrieval* é quantificada por *precision* (precisão) e *recall* (abrangência).

O algoritmo descrito na Metodologia foi aplicado a quatro palavras polissêmicas. Embora nossa intenção não seja comparar resultados, acreditamos ser relevante citar algumas pesquisas semelhantes desenvolvidas anteriormente.

Zemik (1990) relatou "*recall* e *precision*" de aproximadamente 70% para uma palavra apenas (*interest*) e observou que para outras duas palavras (*including* e *issue*), o resultado foi menos positivo. Clear (1989) registrou seus resultados para duas palavras (65% e 67%), aparentemente com *recall* de 85%. Lesk (1986) registrou 50-70% de *precision*, mas sem deixar claro sob que parâmetros e restrições. Black (1988) promoveu

um teste em cinco palavras, reportanto uma precisão média de 75% usando seu "optimal method on high entropy" distinguindo quatro sentidos. Na década de 1990, Hearst (1991) atingiu 84% enquanto Gale, Church e Yarowsky (1992) registraram 92% de *precision* em dois caminhos distintos.

Neste projeto foram determinados dois tipos de *precision*:

*Precision*1 =

$$\frac{\text{total de frases extraídas e identificadas corretamente com o sentido atribuído a ela}}{\text{total de frases extraídas}}$$

*Precision*2 =

$$\frac{\text{total de sentidos corretamente identificados (sem nenhum caso de erro)}}{\text{total de sentidos registrados}}$$

A diferença entre os dois critérios de *Precision* é que no primeiro são consideradas todas as frases. No segundo caso, apenas as frases cujos sentidos foram identificados 100% corretamente.

Recall:

$$\frac{\text{total de frases extraídas}}{\text{total de frases existentes com um determinado Nódulo}}$$

Temos assim:

BANCO

*Precision*1: 257/270 = 95,2%
*Precision*2: 4/10 = 60%
Recall: 270/460 = 59,5%

ACOGER

*Precision*1: 61/72 = 84,7%
*Precision*2: 4/8 = 50%
Recall: 72/104= 69,2%

QUERELLA

*Precision*1: 39/40 = 97,5%
*Precision*2: 2/3 = 66,7%
Recall: 40/63 = 64,5%

HAMBRIENTO
*Precision*1: 20/20 = 100%
*Precision*2: 3/3 = 100%
Recall: 20/55 = 37,04%

AVALIAÇÃO GLOBAL

*Precision*1: 377/402 = 93,8%
*Precision*2: 13/24 = 54,2%
Recall: 402/682 = 58,94%

O cálculo de *Precision*1 demonstra o alto nível de acertos do algoritmo, quase 100%. O fato de *Precision*2 ser bastante inferior ao *Precision*1, não invalida aquele teste; ainda pode ser considerado muito bom o fato de ter um percentual de 54% de acertos, pois aqui nos referimos ao número de casos em que todas as frases etiquetadas com uma acepção foram corretamente identificadas.

O *Recall* poderia ser bastante maior se não tivesse feita uma seleção. Foram escolhidas apenas as frases com um nível de relevância acima de 5%. O nível de significância é medido pela similaridade existente entre os colocados do *HyperThesaurus* e os colocados das frases.

4 CONSIDERAÇÕES FINAIS

Os resultados mostraram-se altamente significativos: *Precision*1 = 94% , *Precision*2 = 54% e *Recall* = 59%.

Embora a amostra tenha sido pequena com apenas quatro candidatos (*nódulos*), obtivemos um elevadíssimo nível de *Precision*. Dentre as 402 frases recuperadas – todas de palavras polissêmicas – acertamos o sentido em 377 delas ou 94%. Se considerarmos os casos em que todas as frases pertencentes a um mesmo sentido foram corretamente identificadas ou não pelo nosso algoritmo, temos ainda assim, um índice de 54%,

portanto, em mais da metade das frases acertamos todos os casos pertencentes a um mesmo sentido.

A taxa de 59% do *Recall* deve-se, em grande parte, à seleção das frases por relevância, o que reduziu significativamente o número de frases selecionadas, o que não significa que os restantes 41% fossem erroneamente identificadas pelo algoritmo. Significa apenas que houve um índice muito baixo de relevância – menos de 5% – o que só contribuiria para aumentar a dificuldade do usuário (lexicógrafo, terminógrafo) para selecionar as melhores opções para cada sentido ou acepção.

Esta pesquisa apresenta um modelo de ferramenta que busca solucionar um dos problemas da lexicografia baseada em *corpora*, recuperando exemplos realmente relevantes ao usuário, por meio da análise de contexto. Para que possa ter continuidade é necessário conciliar o desenvolvimento de recursos linguísticos e linguístico-computacionais volumosos com a criação e implementação de modelos computacionais que permitam a representação do conhecimento expresso em textos de línguas naturais. Este desafio, evidentemente, requer ainda a atuação de uma equipe de etiquetadores e colaboradores das áreas de Linguística, Informática e Matemática, além, é claro, de uma instituição que dê abrigo e condições de trabalho para o desenvolvimento de pesquisas em linguística computacional.

REFERÊNCIAS

ATKINS, B. Building a lexicon: the contribution of lexicography*. International Journal of Lexicography*, n. 3, p.167-204, 1991.

ATKINS, B.; LEVIN, B. Admitting impediments. In: ANNUAL CONFERENCE OF THE UW CENTRE FOR THE NEW OED, 4., 1988, Oxford. *Proceedings*… Oxford, 1988

BLACK, E. An experiment in computational discrimination of english word Ssnses*. IBM Journal of Research and Development*, v 32, p. 185-194, 1988.

CLEAR, J. *An experiment in automatic word sense identification*. Oxford: Oxford University Press, 1989.

DICCIONARIO Planeta de la Lengua Española Usual. Barcelona: Planeta, 1989.

DOLAN, W. B. Word sense disambiguation: clustering related senses. In: CONFERENCE ON COMPUTATIONAL LIGUISTICS, 15., 1994, Kyoto. *Proceedings*… Morristown: Association for Computational Linguistics, 1994.

DUBOIS, J. et al. *Dictionnaire de linguistique*. Paris : Larousse. 1973.

EBECKEN, N. F., LOPES, M. C.S ., COSTA, M. C. A. Mineração de textos. In: REZENDE, S. de O. (Org.). *Sistemas inteligentes*. Barueri, SP: Manole, 2003. p. 337-370.

FOX, C. Lexical analysis and stoplists. In: FRAKES, W. B.; BAEZA-YATES, R. A. (Ed.). *Information retrieval*: data structures & algorithms. New Jersey: Prentice Hall PTR, p. 102-130. 1992.

GALE, W., Church, K.; YAROWSKY, D. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, n. 26, p.415-439, 1992.

HEARST, M. *Noun homograph disambiguation using local context in large text corpora*. Waterloo: University of Waterloo,1991.

HONGYAN, J.; TZOUKERMANN, E. Information retrieval based on context distance and morphology. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 22., 1999. *Proceedings*… 1999, p. 90-96.

JACKENDOFF, R. *Semantic structures*. Cambridge: MIT Press, 1990.

LESK, M. Automatic sense disambiguation: how to tell a pine cone from an ice cream cone. In: SIGDOC CONFERENCE, 1986, New York. *Proceeding*… New York: Association for Computing Machinery, 1986.

KILGARIFF, A. Dictionary word sentence distinctions: an enquiry into their nature. *Computers and the Humanities*, n. 26, p. 365-387, 1993.

KOWALSKI, G. *Information retrieval systems: theory and Implementation*. Massachusets: Kluwer Academic Publishers. 1997.

KORFHAGE, R. R. *Information retrieval and storage*. New York: John Wiley & Sons, 1997.

SALTON , G.; Mc Gill, M. J. *Introduction to modern information retrieval*. New York: Mc Graw-Hill Computer Series, 1983.

SANTILLANA EDUCACIÓN. *Diccionario salamanca de la lengua espanola*. Salamanca: Universidade de Salamanca, 1996.

VYGOTSKY, L. S. Thinking and speech. In: RIEBER, R.W.(Ed.). *Problems of general psychology*. New York: Plenum, 1987. (The collected works of Vygotsky, v. 1).

ZEMIK, U. Tagging word senses in a corpus: the Nee.die in the Haystack revisited. In: JACOBS, P. S. (Ed.). *Text-bated intelligenl systems*: currem research in text analysis, information extraction, and retrisval. New York, GE Research & Developmemt, 1990.

Title

Detection and extraction of word senses candidates based on a collocated thesaurus

Abstract

The Word Sense Disambiguation (WSD) involves the determination of all different senses for every word under considerations, which depends on manual sense annotation for every word and assigns for each occurrence of a word an appropriate sense. This paper describes a program that disambiguates Spanish word senses and automatically identifies which sentences in unrestricted text corpora are semantically related and which correspond to fundamentally some set of related senses applied to the CHADES (Corpus Hispano-Americano de Español). The tool's model presented sought to address the problems of lexicography, retrieving relevant examples to users, through the context's analysis. It reinforces the importance of developing language and computer-language's resources in order to allow the knowledge representation of its expression in natural language texts.

Keywords

Thesaurus of collocations. Indexation. Word sense.

Título

Detección y extracción de candidatos a las acepciones basadas en un tesaurus de colocados

Resumen

La desambiguación envuelve la determinación de todos los distintos sentidos de cada palabra bajo consideraciones que dependen de la anotación manual del sentido de cada palabra atribuyendo el sentido apropiado a cada ocurrencia de una palabra. Este artículo describe un programa para la desambiguación automática de los múltiples sentidos de una misma palabra (ambiguedad semántica) identificando en un corpus textual - CHADES (Corpus HispanoAmericano de Español), de modo que las frases, sin restricción, están semánticamente relacionadas al mismo grupo de sentidos encontrado en este corpus. El modelo de herramienta presentado recogió solucionar los problemas de lexicografía, recuperando ejemplos relevantes al usuarios, por medio del análisis de contexto. Refuerza la necesidad del desarrollo de recursos linguisticos y linguistico-computacionales voluminosos que permitan la representación del conocimiento expreso en texto de lenguas naturales.

Palabras claves

Tesaurus de Colocados. Indexación. Acepción.

Recebido em: 13.02.2009

Aceito em: 04.08.2009