4
Trabalhos relacionados à extração automática de hiperonímia

Neste capítulo, relato os principais trabalhos que tratam da extração automática de relações de hiperonímia a partir de textos. Começo com uma apresentação detalhada das wordnets que, embora não sejam elaboradas automaticamente, são freqüentemente vistas como um modelo a ser atingido.

4.1.
WordNet, EuroWordNet e Wordnet.Br

A WordNet (Fellbaum, 1998) é um léxico semântico relacional desenvolvido para a língua inglesa, disponível para uso online[15], cujos principais objetivos são (i) oferecer uma combinação de dicionário e tesauro que seja mais utilizável de um ponto de vista intuitivo; e (ii) dar suporte a tarefas que envolvem a análise automática de textos. Na WordNet, as palavras estão agrupadas em conjuntos de sinônimos chamados *synsets*. Sua forma de organização se baseia nos resultados de experimentos psicolingüísticos e busca reproduzir a estrutura do nosso léxico mental.

A WordNet distingue entre substantivos, verbos, adjetivos e advérbios. Cada *synset* contém um grupo de palavras ou expressões sinônimas. A maioria dos *synsets* se conecta a outros *synsets* por meio de relações semânticas como hiperonímia, hiponímia, coordenação, holonímia e meronímia (substantivos); hiperonímia, troponímia e acarretamento (verbos); substantivos relacionados e formas de particípio de verbos (adjetivos); adjetivos-raiz (advérbios).

As relações de hiperonímia/hiponímia entre os *synsets* de substantivos podem ser compreendidas como relações entre categorias conceituais, o que permite que a WordNet seja interpretada (e utilizada) como uma ontologia lexical

[15] A WordNet está disponível em http://wordnet.princeton.edu/.

na IA. No nível mais alto, as hierarquias estão organizadas em 25 primitivos nominais.

Atualmente (2006), a WordNet, "a mãe de todas as wordnets" (Fellbaum, 1998), conta com mais de 150.000 palavras organizadas em mais de 115.000 synsets, constituindo-se em um modelo para outras línguas e tornando-se um dos recursos de maior impacto no PLN. A WordNet vem se desenvolvendo desde 1985 na universidade de Princeton, em um projeto que já recebeu mais de 3 milhões de dólares para seu desenvolvimento.

Tomando como modelo a WordNet de Princeton, foi desenvolvida a EuroWordNet (Vossen, 1998), uma base de dados multilingüe que integra wordnets de diversas línguas européias. A diferença fundamental entre a WordNet de Princeton e a EuroWordNet é o fato de a segunda ser multilingüe – as wordnets das diversas línguas são relacionadas pelo "Inter-Lingual-Index" (ILI), uma lista de *synsets* que corresponde aos *synsets* da WordNet de Princeton.

A Wordnet.Br é a versão brasileira da WordNet, que conta atualmente com cerca de 11.000 verbos, 17.000 substantivos, 15.000 adjetivos e 1.000 advérbios, num total de 44.000 palavras e 18.500 *synsets* (Dias-da-Silva et al., 2006). Para sua elaboração, a Wordnet.Br reaproveita material disponível em outras fontes, como as versões eletrônicas dos dicionários *Aurélio* e *Michaelis*, dicionários de sinônimos e antônimos, um dicionário analógico e um dicionário de verbos do português. Porém, a maior parte do trabalho de elaboração é feita manualmente (Dias-da-Silva et al., 2006). Na fase atual de desenvolvimento, os lingüistas que participam do projeto têm realizado (i) a análise da consistência semântica dos *synsets*; (ii) a coleta e seleção das frases-exemplo, extraídas de corpus.

4.2. Extração automática de hiperonímia

Condamines e Rebeyrolle (2000) classificam em métodos top-down ou bottom-up as diversas técnicas desenvolvidas para a extração automática de relações semânticas a partir de textos. Métodos top-down utilizam padrões lingüísticos pré-definidos; as técnicas para a aquisição de relações semânticas se baseiam em regras criadas manualmente para a extração dos dados. O trabalho de Hearst (1992, 1998) se enquadra nesta abordagem – e a desvantagem da técnica

consiste justamente na tarefa manual de codificação das regras, que pode requerer um grande trabalho. Nos métodos bottom-up não é fornecida nenhuma informação sobre os dados que serão extraídos. As palavras são agrupadas (ou classificadas) por meio de técnicas de agrupamento (clusterização) que se baseiam na similaridade entre contextos de palavras. De maneira geral, o problema desta abordagem é que freqüentemente os grupos de palavras (clusters) não são rotulados – trata-se de aglomerados semânticos, o que pode ser um problema para determinadas aplicações. Freqüentemente, essa técnica é utilizada na extração de associações entre palavras (Lin e Pantel, 2002; Widdows 2003) e, de maneira mais rara, na elaboração de tesauros (Kilgarriff, 2003). Alguns trabalhos apresentam uma combinação de abordagens top-down e bottom-up, conjugando técnicas de clusterização e codificação de regras (Caraballo, 1999; Cerderberg e Widdows, 2003; Snow et al., 2005; Morin e Jacquemin, 2004).

4.2.1. Os padrões de Marti Hearst

Marti Hearst (1992, 1998) foi a primeira a utilizar a idéia de que determinados padrões léxico-sintáticos poderiam, sistematicamente, expressar determinadas relações semânticas.

Nesse contexto, relações de hiponímia seriam especialmente úteis às tarefas de PLN porque permitiriam a expansão de léxicos existentes, como a WordNet. Com isso, um dos objetivos da metodologia é auxiliar, de maneira automática ou semi-automática, o trabalho de lexicógrafos e construtores de bases de conhecimento dependentes de domínio.

Especificamente, Hearst (1992, 1998)[16] propõe métodos de extração automática de relações léxico-sintáticas e compara os resultados obtidos automaticamente com os obtidos manualmente pela equipe de lexicógrafos da WordNet.

Hearst propõe a identificação, no corpus, de padrões léxico-sintáticos que codifiquem a relação de hiperonímia na língua inglesa e que obedeçam aos seguintes critérios:

- Ocorrência freqüente e em diferentes tipos de texto;
- Indicação (quase) sempre constante da relação de interesse;
- Pouca ou nenhuma necessidade de conhecimento pré-codificado.

Seguindo esses critérios, os padrões encontrados para o inglês foram:

(i) NP_0 such as NP_1 {, NP_2 ... , (and | or) NP_i}
(ii) such NP_0 as {NP ,}* {(and | or)} NP
(iii) NP {, NP}* {,} or other NP_0
(iv) NP {, NP}* {,} and other NP_0
(v) NP_0 {,} including { NP ,}* {or | and} NP
(vi) NP_0 {,} especially { NP ,}* {or | and} NP

onde NP_0 corresponde a um sintagma nominal (SN) hiperônimo e os demais NPs (NP_1, NP_2...NP_i) a SNs hipônimos:

$$SN_0 > SN_1, SN_2, SN_3 ... SN_i$$

Os padrões (i), (iii) e (iii) foram descobertos manualmente, por meio de observação no corpus. Porém, para que a abordagem seja mais abrangente, Hearst sugere um procedimento-padrão de descoberta, por meio do qual os demais padrões foram identificados, e que consiste basicamente de 4 etapas:

- decidir qual a relação lexical de interesse;
- derivar, por meio da WordNet, uma lista de pares de palavra na qual a relação esteja expressa: por exemplo, para a relação de meronímia, o par *carro-volante;*
- extrair sentenças do corpus em que ambas as palavras (*carro* e *volante*) apareçam, registrando o contexto lexical e sintático em que foram encontradas;
- encontrar semelhanças entre esses contextos e tentar generalizar: contextos comuns levam a padrões que indicam a relação de interesse.

A partir desses padrões, quando uma relação de hiponímia é descoberta, o SN encontrado é considerado uma unidade atômica, indivisível. São retirados apenas o que Hearst chama de "modificadores indesejados", como alguns adjetivos comparativos ("*smaller*", "*important*"). Um problema já observado por

[16] A principal diferença entre os dois trabalhos está no corpus utilizado: em 1992, os padrões foram extraídos de *Grolier's Encyclopaedia*; em 1998, de seis meses do jornal *New York Times*.

Hearst, e que produz erros também para a língua portuguesa (cf. seção 6.1) diz respeito à determinação do referente de um sintagma preposicional (SPrep). Para o inglês, Hearst nota que, na maioria das vezes, o substantivo final no SPrep que precede o "*such as*" (no padrão (i)) é o hiperônimo da relação, como no exemplo (1), embora existam inúmeras exceções, como ilustra a frase (2):

*(1) Agar is **[**a substance prepared from a mixture of red algae**]**, such as Geldium, for laborary or industrial use.*
*(2) A bearing is a structure that supports a **[**rotating part of a machine**]**, such as shaft, axle, spindle, or wheel.*

Isto é, para as frases (1) e (2) seriam extraídas, respectivamente, as relações (1') e (2'), em que a relação (1') está errada pois o sintagma hiperônimo é apenas *red algae*. Já as relações extraídas em (2') estão corretas.

(1´) Geldiu < substance prepared from a mixture of red algae

(2') shaft < rotating part of a machine
axle < rotating part of a machine
spindle < rotating part of a machine
wheel < rotating part of a machine

Com relação à ambigüidade do SPrep nos outros padrões, Hearst comenta apenas que, no padrão "*and other*", diferentemente do "*such as*", freqüentemente o SN completo corresponde ao hiperônimo (3), o que ilustraria a dificuldade de se trabalhar com textos, principalmente de jornais, por sua diversidade, em contraste com as estruturas textuais relativamente previsíveis de dicionários e enciclopédias. Como resposta a essas dificuldades, Hearst sugere que uma solução simples seria descartar as orações em que a ambigüidade é possível, buscando-se apenas SNs simples.

(3) *Temples, treasuries, and other important**[**civic buildings**]**.*

Como um dos objetivos de seu trabalho é, automaticamente, aumentar as relações da WordNet, a análise dos resultados é feita por meio de uma comparação entre as relações identificadas automaticamente e as relações de hiperonímia presentes na WordNet. Em geral, Hearst observa que as relações obtidas a partir do corpus de jornal tendem a ser menos taxonômicas, ou prototípicas, do que as encontradas em textos enciclopédicos; são mais

influenciadas pelo contexto em que aparecem, e refletem de forma mais sistemática julgamentos subjetivos e usos metafóricos do que afirmações estabelecidas que constam de enciclopédias. Como exemplo, uma afirmação como "*Casablanca* é um *clássico*" pode ser considerada decorrente de um julgamento de valor (embora Hearst reconheça que enciclopédias muitas vezes afirmam que determinados atores são estrelas, o que não é tão diferente). Do mesmo modo, a declaração "*AIDS* é um *desastre*" pode ser entendida mais como uma relação metafórica do que taxonômica.

Além disso, como a maioria dos termos da WordNet são nomes sem modificadores ou nomes com um único modificador, os algoritmos de Hearst extraem apenas relações que consistem de nomes sem modificadores, tanto no sintagma hiperônimo quanto no hipônimo. A utilidade dessa restrição estaria na dificuldade de se encontrar um procedimento transparente capaz de determinar quais modificadores são importantes. Acrescente-se a isso que, para fins de avaliação, na maioria dos casos é mais fácil julgar a correção de uma relação com substantivos sem modificadores.

No trabalho, 200 instâncias do padrão "*e outros*" foram avaliadas manualmente. Os avaliadores deveriam classificar os resultados de acordo com oito categorias, como mostra a tabela 2, retirada de Hearst (1998):

Freqüência	**Explicação**
38	Alguma versão dos SNs e sua relação correspondente foi encontrada na WordNet
31	A relação não apareceu na WordNet e foi considerada uma relação ótima (em alguns casos ambos os SNs estavam presentes, em outros casos não)
35	A relação não apareceu na WordNet e foi considerada uma relação pelo menos boa (em alguns casos ambos os SNs estavam presentes, em outros casos não)
19	Relação muito geral
8	Relação muito subjetiva, ou que continha referentes inapropriados (e.g., "these")
34	Os SNs envolvidos eram muito longos, muito específicos e/ou muito dependentes de contexto
12	As relações eram repetições dos casos acima
22	As frases não continham a forma sintática apropriada (e.g., "all of the above, none of the above, or other")

Tabela 2: Resultado da avaliação de 200 frases com o padrão "e outros" (Hearst, 1998)

Consciente do alto grau de subjetividade deste tipo de avaliação, e assumindo uma abordagem "cautelosa" na avaliação, 63% das relações extraídas foram consideradas corretas, isto é, passíveis de serem inseridas na WordNet.

4.2.2. Outros trabalhos

Morin e Jacquemin (2004) apresentam um sistema – *Prométhée* – que extrai e utiliza padrões léxico-sintáticos no estilo Hearst a partir de corpus. O processo de extração automática de padrões é realizado em sete etapas:

(a) seleção manual da relação semântica que se deseja identificar;
(b) coleta de uma lista de pares de termos que participam na relação. Esses pares podem ser extraídos de um tesauro, de uma base de conhecimento ou ainda especificados manualmente;
(c) descoberta de frases em que os pares de termos ocorram – as frases são representadas como expressões léxico-sintáticas;
(d) descoberta de contextos comuns que generalizem as expressões léxico-sintáticas – estes contextos são calculados utilizando funções de similaridades e processos de generalização;
(e) Validação dos padrões por um especialista;
(f) Uso dos padrões validados para a extração de outros pares de termos;
(g) Validação dos pares candidatos por um especialista.

De um conjunto inicial, criado manualmente, de 40 pares de termos relacionados por hiperonímia, o sistema Prométhée identificou 11 padrões léxico-sintáticos[17] que consistem de pequenas variações dos padrões identificados em Hearst (1992, 1998). Os padrões estão descritos abaixo, e SN_1 corresponde ao SN hiperônimo e:

(1) {deux | trois...} SN_1 (Lista de SNs)
(2) {certain | quelque | de autre...} SN_1 (Lista de SNs)
(3) { deux | trois...} SN_1 : (Lista de SNs)
(4) { certain | quelque | de autre...} SN_1 : (Lista de SNs)
(5) SN_1 tel que Lista de SN

---

(1) [17] {dois | três...} SN_1 (Lista de SNs)
(2) {certos | alguns | outros...} SN_1 (Lista de SNs)
(3) { dois | três...} SN_1 : (Lista de SNs)
(4) {certos | alguns | outros...} SN_1 : (Lista de SNs)
(5) SN_1 tais como Lista de SN
(6) SN_1, particularmente SN_2
(7) SN_1 como Lista de SNs
(8) SN_1 tais como Lista de SNs
(9) SN_2 {e|ou} outros SN_1
(10) SN_1, e em particular SN_2
(11) Dentre SN2, SN_1, (esse padrão parece não se aplicar ao português)

(6) SN_1, particulièrement SN_2
(7) SN_1 comme Lista de SNs
(8) SN_1 tel Lista de SNs
(9) SN_2 {et lou} de autre SN_1
(10) SN_1 et notamment SN_2
(11) Chez le SN2, SN_1,

Esses padrões foram aplicados em um corpus constituído de resumos e títulos de artigos científicos produzidos por pesquisadores, engenheiros e técnicos das áreas de agricultura e indústria alimentícia, o corpus "[AGRO-ALIM]". O corpus possui 427.482 palavras, com uma média de 316 palavras por resumo (Jacquemin et al., 2002).

A avaliação dos pares extraídos mostrou uma alta qualidade das relações produzidas, com uma precisão de 82%, mas uma abrangência de 56%. A avaliação foi feita por padrão extraído, e a tabela 3 reproduz os resultados de alguns padrões semelhantes aos descritos em Hearst.

Padrão	Qtde de relações	Precisão
(5) SN_1 tais como Lista de SNs	210	86%
(7) SN_1 como Lista de SNs	90	69%
(8) SN_1 tais como Lista de SNs	36	90%
(9) SN_1 elou outros SNs	17	59%

Tabela 3: Resultados de alguns padrões de Morin e Jacquemin (2004)

Após a descoberta dos padrões, e conseqüente extração de pares semanticamente relacionados, Morin e Jacquemin (2004) apresentam uma técnica para a aquisição incremental das relações extraídas por meio da exploração de relações sintáticas, morfossintáticas e semânticas entre os termos extraídos. Embora interessante, o método se apóia em uma ferramenta altamente sofisticada chamada FASTR (Jacquemin, 1999), um parser transformacional para o qual não há equivalente na língua portuguesa.

O trabalho de Cederberg e Widdows (2003) consiste na utilização de modelos matemáticos (Latent Semantic Analysis – LSA) para medir a similaridade semântica entre as palavras.

Os autores realizam três experimentos: no primeiro deles, constroem um sistema extrator de hiperonímia que utiliza as 6 regras de Hearst (1998). A partir de uma amostra de 430.000 palavras do *British National Corpus,* são extraídas 513 relações, das quais 100 foram selecionadas para avaliação manual. Na avaliação, cada relação deveria ser pontuada de acordo com os seguintes critérios:

> 4. As relações estão corretas da maneira como foram extraídas.
> 3. As relações estão corretas após uma ligeira modificação, como mudança plural-singular ou a remoção de artigo.
> 2. As relações estão "potencialmente corretas" mas requerem um processamento difícil para a obtenção da relação correta. Por exemplo, o substantivo está correto mas há problemas no sintagma preposicional.
> 1. A relação está correta de alguma forma, mas é muito geral ou muito específica para ser útil.
> 0. A relação está incorreta.

Após a avaliação, 40% das relações foram pontuadas como 3 ou 4 (relações corretas). A fim de melhorar os resultados, Cederberg e Widdows aplicaram um filtro utilizando uma variante do método LSA[18]. Os novos resultados mostraram um aumento das relações classificadas como 3 ou 4 de 40% para 58%, o que sugere a efetividade do filtro.

Em uma tentativa de aumentar o número de relações identificadas, já que os padrões de Hearst são considerados pouco freqüentes nos textos, Cederberg e Widdows utilizaram um método já descrito em Widdows e Dorow (2002), que consiste em considerar a pista fornecida pela estrutura de coordenação da língua (elementos que aparecem em listas tendem a ser semanticamente similares) aliada a um método de agrupamento (*clusterização*). Para tanto, assume-se que, em uma frase como

(4) Este não é o caso de açúcar, mel, cravos ***e outras*** *especiarias que...*

que leva à identificação da relação

cravos < especiarias,

e em uma frase como

(5) *Navios carregados com noz-moscada ou canela,* ***cravos ou*** *coentro enfrentaram...*

[18] O método LSA (Latent Semantic Analysis) avalia em que medida as palavras *x* e *y* aparecem em contextos similares por meio da representação de palavras como pontos em um espaço vetorial. Palavras com significados relacionados devem ser representadas como pontos próximos.

como a relação entre *cravo* e *especiarias* já foi identificada, a hipótese de coordenação levaria a identificação das relações

noz-moscada<especiarias;
canela<especiarias;
coentro< especiarias

Na etapa final do trabalho, Cederberg e Widdows (2003) aplicaram novamente o filtro LSA nos resultados da extração, que incluem aqueles obtidos com a utilização da pista de coordenação. De 260 relações avaliadas, 166 (64%) foram consideradas corretas (pontuação 3 ou 4), o que mostra o sucesso na combinação das técnicas.

Snow, Jurafsky e Ng (2005), partindo da crítica de que os padrões léxico-sintáticos de Hearst (1998), embora amplamente utilizados em outros trabalhos, têm limitações quanto à abrangência (isto é, são poucos padrões) e quanto à forma de identificação (em geral, os padrões são identificados manualmente), propõem a utilização de aprendizagem de máquina para substituir este conhecimento construído manualmente. Em termos gerais, a abordagem de Snow et al.(2005) baseia-se em (a) coletar pares de substantivos, no corpus, que identifiquem relações de hiperonímia, utilizando a WordNet; (b) coletar, para cada par, frases em que ambos os substantivos apareçam; (c) realizar um *parsing* dessas frases para a extração automática de padrões; (d) treinar um classificador de hiperônimos utilizando esses resultados. Embora o trabalho pareça muito interessante, não temos como comparar os resultados de Snow et al. com os demais apresentados aqui, visto a forma de avaliação ser bastante diferente.

Em suma, embora diversos trabalhos venham propondo a identificação automática em textos de relações de hiperonímia, os padrões descritos originalmente em Hearst (1992, 1998) têm se mostrado os mais produtivos, sendo amplamente repetidos em combinação com outras técnicas.

A principal crítica à abordagem de Hearst é sua pouca abrangência, isto é, provavelmente nem todas as relações semânticas relevantes para uma ontologia são expressas por meio de pistas textuais – e talvez nem todas as relações de hiperonímia de um domínio. Por outro lado, a metodologia apresenta a grande

vantagem de oferecer grupos de palavras já rotulados com um hiperônimo, e não simplesmente aglomerados de palavras.

Trabalhos como os de Cederberg e Widdows (2003) e Snow et al. (2005) tentam conciliar os padrões com outras técnicas, a fim de aumentar a precisão e abrangência dos resultados, mas os dados, até o momento, sugerem que tais melhorias são pouco significativas.

Já o trabalho de Morin e Jacquemin (2004) apresenta um sistema capaz de extrair automaticamente do corpus padrões léxico-sintáticos para a expressão de relações semânticas. Para tanto, o sistema utiliza algoritmos e o cálculo estatístico de medida de similaridade. Porém, os padrões encontrados são muito semelhantes aos de Hearst, e resta saber se o processo extração automática de regras teria um desempenho tão eficaz na identificação outras relações semânticas, como a meronímia, que não têm sido tão exploradas.

As relações de hiperonímia identificadas por Morin e Jacquemin (2004) apresentam, em termos gerais, resultados bastante superiores aos de Hearst e de Cederberg e Widdows. Porém, uma comparação exata entre os trabalhos não é possível por diversas razões.

A primeira delas diz respeito ao tipo de avaliação realizada em cada trabalho. Hearst (1998) e Cederberg e Widdows (2003) avaliam a precisão das relações por meio de uma escala (parecida, mas não idêntica) de aceitação das relações identificadas, que vai do acerto total ao erro total; Morin e Jacquemin (2004) utilizam medidas de precisão e abrangência. Por outro lado, Hearst apresenta seus resultados por padrão léxico-sintático – especificamente, apresenta os resultados obtidos com apenas um padrão. Morin e Jacquemin também apresentam os resultados obtidos por padrão identificado, mas Cederberg e Widdows (2003) apresentam os resultados gerais, isto é, não sabemos o desempenho de cada regra.

O segundo obstáculo para uma comparação adequada diz respeito ao corpus: Hearst utiliza textos jornalísticos; Cederberg e Widdows (2003), uma amostra do *British National Corpus*, um corpus diversificado; e Morin e Jacquemin (2004) um corpus relativamente "controlado", composto por resumos de artigos técnicos, de um domínio específico.

Por fim, as diferenças quanto ao idioma também devem ser levadas em consideração: o trabalho de Morin e Jacquemin (2004) tem o francês como língua-alvo, e os trabalhos de Hearst e de Cederberg e Widdows voltam-se para o inglês.