ONTOLOGIAS LINGÜÍSTICAS APLICADAS AO PROCESSAMENTO AUTOMÁTICO DAS LÍNGUAS NATURAIS: O CASO DAS REDES *WORDNETS*[1]

Ariani Di FELIPPO[2] (UNESP/Ar.)

RESUMO: O objetivo geral deste trabalho é o de ilustrar o uso de ontologias lingüísticas no Processamento Automático das Línguas Naturais (PLN). Para tanto, definir-se-á o termo "ontologia lingüística" e se ressaltará a importância desse tipo de recurso para o PLN. Tomando-se as redes em formato *wordnet* como um tipo de ontologia lingüística, descrever-se-ão sua arquitetura e seus pressupostos psicolingüísticos e se ilustrará o uso da WordNet de Princeton em algumas aplicações de PLN. Por fim, algumas propostas de refinamento desse tipo de base serão apresentadas, bem como o estado atual da rede wordnet para o português do Brasil, a WordNet.Br.

ABSTRACT: This paper aims at illustrating the use of linguistic ontologies in Natural Language Processing (NLP). Accordingly, we define the term "linguistic ontology" and emphasize the relevance of this type of lexical resource for NLP. Focusing on lexical databases in *wordnet* format as a special type of linguistic ontology, we define a general wordnet architecture and its psycholinguistic underlying assumptions, and illustrate the use of Princeton WordNet in some NLP applications. In addiction, some wordnet lexical database refinement and expansion proposals are showed as well as the actual status of the ongoing work on Brazilian Portuguese wordnet, which is called WordNet.Br.

1. Introdução

No âmbito do Processamento Automático das Línguas Naturais (PLN), área de pesquisa multidisciplinar em que se busca simular computacionalmente competências lingüísticas como *sumarização* e *tradução*, os sistemas de PLN geralmente possuem componentes em que estão armazenadas as chamadas "bases[3] de conhecimento estático", que são: base gramatical, base lexical e base conceitual. A base gramatical contém representações das regras sintáticas. A base lexical armazena uma coleção de unidades lexicais associadas a feixes de traços morfológicos, sintáticos, semânticos e até mesmo pragmático-discursivos.

As bases conceituais, em especial, contêm um "modelo do mundo" ou uma abstração da realidade, em que são descritos tipos de objetos, eventos, propriedades e relacionamentos entre esses tipos (ALLEN, 1995; REITER, DALE, 2000). Esse tipo de base desempenha um papel fundamental nos sistemas de PLN porque limita a "visão de mundo" simulada pelo sistema (DIAS-DA-SILVA, 1996). Em outras palavras, uma base conceitual comumente armazena o que se denomina "ontologia". O objeto "ontologia" é questão controversa em várias áreas, havendo, portanto, uma grande flutuação definicional e terminológica, como demonstram Guarino e Giareta (1995). Uma das definições mais usuais é a de "uma especificação de uma conceitualização (ou seja, "uma visão simplificada do mundo") caracterizada por propriedades formais (explícitas) e propósitos específicos" (GRUBER, 1993). No PLN, as ontologias têm sido amplamente empregadas em várias aplicações (p.ex.: sumarização, resolução anafórica, recuperação de informação, desambiguação (lexical) de sentido, etc.) com o objetivo de melhorar o desempenho dos sistemas computacionais nessas tarefas. Neste trabalho, especificamente, ilustra-se o uso no PLN de um tipo especial de "ontologia lingüística", as redes wordnets. Assim, na Seção 2, defini-se "ontologia lingüística" e ressalta-se sua relevância para o PLN. Na Seção 3, apresenta-se a arquitetura geral de uma rede wordnet por meio da descrição da "mãe de todas as wornets", a WordNet (FELLBAUM, 1998). Na Seção 4, ilustra-se o uso da WNP, em algumas aplicações de PLN. Na Seção 5, salientam-se algumas propostas de refinamento e expansão das redes em formato *wordnet*. Na Seção 6, apresenta-se o estado atual de desenvolvimento da

[1] Este trabalho está relacionado a um projeto de doutorado financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq.

[2] E-mail para contato: arianidf@uol.com.br

[3] O termo "base" é empregado no sentido computacional de "base de dados" ("*database*"), ou seja, "coleção de dados armazenada de modo sistemático na memória do computador".

wordnet para o português do Brasil, a WordNet.Br (DIAS-DA-SILVA et al, 2006). E, por fim, na Seção 7, apresentam-se algumas considerações finais sobre este trabalho.

2. Definindo "ontologia lingüística"

Como bem salientam inúmeros autores, a definição do objeto denominado "ontologia" é questão de controvérsia em muitas áreas do conhecimento (GUARINO, GIARETA, 1995). Mesmo sem um consenso sobre sua definição, esses objetos apresentam características comuns que permitem certas classificações dos mesmos. No âmbito do Processamento Automático das Línguas Naturais e da Inteligência Artificial, dois tipos de ontologia podem ser claramente identificados: as chamadas *ontologias lingüísticas* e as *ontologias conceituais* (VOSSEN, 1998a, PALMER, 2001, FARRAR, BATEMAN, 2005).

As *ontologias lingüísticas* caracterizam-se por armazenar apenas conceitos lexicalizados (em uma determinada língua), isto é, conceitos expressos por uma ou mais palavras de uma língua. Sob esse ponto de vista, uma ontologia é um inventário dos sentidos de uma dada língua, ou seja, é um inventário somente daqueles conceitos compartilhados por uma comunidade lingüística. Nesse sentido, uma ontologia lingüística do holandês, por exemplo, não armazenaria o conceito *"container"*, já que este não é lexicalizado nessa língua (VOSSEN, 1998a). As ontologias lingüísticas mais difundidas no PLN são Mikrokosmos (VIEGAS et al., 1996), SENSUS (HOVY, 1998) e WordNet (FELLBAUM, 1998). A hipótese para o emprego desse tipo de ontologia no tratamento computacional das línguas natural reside exatamente no fato de que uma ontologia lingüística "emerge" da semântica e do léxico de uma língua específica.

As *ontologias conceituais*, por sua vez, caracterizam-se pelo armazenamento de conceitos para os quais não há lexicalizações, ou seja, não há unidades lexicais que os representem, por exemplo: os conceitos *"coisa parcialmente temporal"* e *"partes do corpo humano"* (VOSSEN, 1998a, PALMER, 2001). Os níveis artificiais são inseridos para que se alcance uma estruturação mais controlada dos conceitos dada a aplicação para a qual a ontologia foi feita. Além de apresentar níveis particulares para conceitos que não são lexicalizados, as ontologias conceituais podem negligenciar conceitos lexicalizados que não são relevantes para seus propósitos. Segundo Palmer (2001), a ontologia conceitual mais difundida no PLN é CYC (LENAT, GUHA, 1990).

Na seqüência, apresenta-se com detalhes a organização da WordNet, ressaltando as características que fazem dela uma "ontologia lingüística".

3. WordNet de Princeton e o formato *wordnet*

Em meados da década de 1980, os pesquisadores do Laboratório de Ciência Cognitiva da Universidade de Princeton (EUA), impulsionados por pressupostos (psico)lingüísticos sobre a organização do léxico mental (ou seja, parte do conhecimento lexical do falante delimitada por sua língua), decidiram construir uma base de dados lexicais (do inglês, *"lexical database"*) que tivesse certa organização conceitual e não alfabética, ou seja, que fosse organizada em função do significado a não da forma ou expressão lingüística (MILLER, FELLBAUM, 1991). Essa iniciativa deu origem, no início da década de 1990, à base de dados lexicais para o inglês americano denominada WordNet. Essa base, também denominada WordNet de Princeton (doravante, WNP), é, na verdade, uma rede de palavras em que unidades lexicais (palavras e expressões), pertencentes às categorias dos substantivos, verbos, adjetivos e advérbios, organizam-se sob a forma de *synsets* (abreviação do termo em inglês *"synonym set"*, isto é, conjunto de unidades sinônimas). Os synsets, inclusive, relacionam-se entre si por meio de cinco relações lógico-conceituais: antonímia, hiponímia, meronímia, acarretamento e causa (CRUSE, 1986; FELLBAUM, 1998). A Figura 1 ilustra a noção de synset e algumas relações lógico-conceituais.

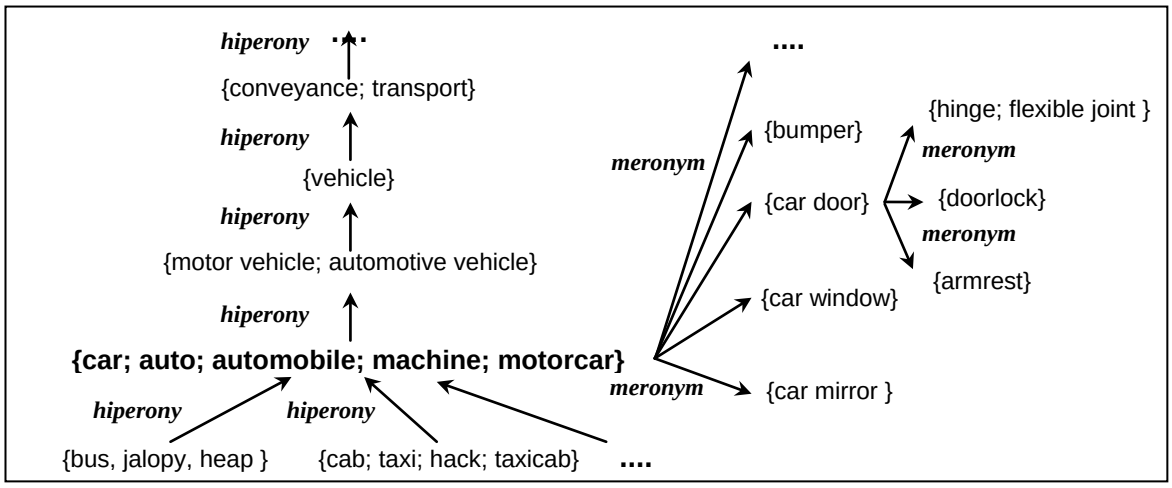

Figura 1: Alguns relacionamentos do synset {car; auto; automobile; machine; motorcar} na WordNet 2.1.

Nessa figura, cujo exemplo foi extraído da WNP (version 2.1), observa-se que o synset {car; auto; automobile; machine; motorcar} está relacionado a:

•conceito mais geral ou synset hiperônimo: {motor vehicle; automotive vehicle};
•conceito mais específico ou synset hipônimo, p.ex.: {cruiser; squad car; patrol car; police car; prowl car} e {cab; taxi; hack; taxicab};
•partes que o compõem ou synsets merônimos, p.ex.: {bumper}, {car door}, {car mirror} e {car window}.

Observa-se ainda que cada synset relaciona-se novamente a outros synsets, por exemplo: {motor vehicle; automotive vehicle} está relacionado à {vehicle}, e {car door}, por sua vez, está relacionado aos synsets {hinge; flexible joint}, {doorlock} e {armrest}.

A organização de uma wordnet assemelha-se à de um dicionário onomasiológisgo e/ou à de um *thesaurus*, como o idealizado por Roget (1972). No entanto, ao contrário de um *thesaurus*, as relações (i) entre synsets (conceitos), (ii) entre unidades lexicais e (iii) entre synsets e unidades lexicais são explicitamente codificadas em uma base de dados no formato *wordnet* (cf. Figura 1).

Do modo como foi elaborada, cada synset da WNP é, por definição, construído de modo a codificar ou representar um único conceito lexicalizado por suas unidades constituintes. Dessa forma, a WNP armazena apenas os conceitos lexicalizados na variante americana da língua inglesa, o que a caracteriza como uma "ontologia lingüística". A WNP também registra outras informações, a saber: (i) para cada unidade lexical, há uma ou mais frases-exemplo para ilustrar o seu contexto de uso, p.ex.: para car, no referido synset, há a frase-exemplo "he needs a car to get to work" (em português, "ele necessita de um carro para conseguir trabalhar"); (ii) para cada synset, há uma glosa que especifica, de modo informal, o conceito por ele lexicalizado, p.ex.: para o synset {car; auto; automobile; machine; motorcar}, há a glosa "a motor vehicle with four wheels; usually propelled by an internal combustion engine" (em português, "um veículo com quatro todas; usualmente impulsionado por um motor de combustão interno").

Em virtude da potencial geração de dicionários de sinônimos e antônimos e da potencial aplicação no domínio do PLN, a WNP desencadeou a construção de redes no formato *wordnet* para várias línguas. Atualmente, há redes wordnets para a maioria das línguas[4], inclusive para o árabe, o sânscrito e o tâmil[5].

Na próxima seção, são exemplificadas algumas aplicações da WNP no domínio do PLN.

[4] As "wordntes pelo mundo" estão listadas no endereço http://www.globalwordnet.org/gwa/wordnet_table.htm.
[5] Língua dravídica, que é língua oficial no estado de Tâmil Nadu (S.E. da Índia), também falada no Sri Lanka e na Malásia, em Cingapura e na Indonésia (FERREIRA, 2004).

4. A WordNet de Princeton e algumas de suas aplicações no PLN

Como mencionado, as redes wordntes passaram a ser aplicada em várias tarefas de PLN. Tal aplicação objetiva melhorar o desempenho dos sistemas computacionais que processam língua natural por meio da inserção de informações de natureza léxico-semântica e semântico-conceitual. Atualmente, as wordnets têm sido especificamente utilizadas nas seguintes tarefas/aplicações: (a) *recuperação de informação* (do inglês, *"information retrieval"*), (b) *sumarização automática* (do inglês, *"automatic summarization"*), (c) *desambiguação de sentido* (do inglês, *"word sense desambiguation"*), (d) *categorização de textos,* (e) *resolução anafórica*, entre outras (FELLBAUM, 1998; VOSSEN, 2003; MORATO et al, 2004).

Para exemplificar tal aplicação, detalha-se, a seguir, o emprego da WNP nos processos de *recuperação de informação, desambiguação de sentido* e *sumarização automática*, ressaltando seus principais problemas e contribuições[6]. Vale salientar, aliás, que o amplo emprego da WNP no PLN está principalmente vinculado aos fatores: *acessibilidade*, *reutilização* e *pertinência lingüística* e *abrangência.*

(a) **Acessibilidade:** arquivo-fonte disponibilizado integralmente via web[7];
(b) **Reutilização:** implementação em formato de *base de dados* garante a reutilização dos dados (nota 3);
(c) **Pertinência lingüística:** embasamento teórico oriundo da Psicolingüística e da Lingüística e estratégias metodológicas baseadas na reutilização de dicionários corroboram a pertinência/qualidade das informações;
(d) **Abrangência:** armazenamento de 117.049 substantivos, 11.488 verbos, 22.141 adjetivos e 4.601 advérbios do inglês americano (versão 2.1)[8, 9]; robustez necessária para o PLN;

4.1. Recuperação de informação

Um sistema de recuperação de informação (ou simplesmente de RI) pode ser superficialmente entendido com um mecanismo para que informações possam ser encontradas em uma coleção de documentos (p.ex.: documentos de áudio, imagem, texto). Comumente, essa tarefa envolve a recuperação de um subconjunto de documentos da coleção considerados relevantes pelo sistema diante de uma consulta (do inglês, *"query"*) formulada pelo usuário (SANFILIPPO et al., 1999). Uma consulta pode ser formada por apenas uma palavra ou mesmo um grupo desordenado de palavras. Geralmente, a estratégia empregada nos sistemas de RI é a de associar um conjunto de termos (do inglês, *"index terms"*) a cada documento da coleção e cruzar tais termos com as palavras da consulta. Por fim, os documentos que satisfizeram esse cruzamento são selecionados e retornados ao usuário.

No entanto, o vocabulário usado pelo usuário na descrição da consulta difere, na maioria das vezes, do conjunto de índices associados aos documentos (XU, CROFT, 1996), o que prejudica a seleção dos documentos a serem retornados. Diante desse problema, os pesquisadores têm adotado a estratégia de "expansão da consulta". Normalmente, essa expansão pode ser feita por meio da associação das palavras da consulta a (i) seus sinônimos e/ou hipônimos ou a (ii) palavras co-ocorrentes, aumentando, assim, a possibilidade de cruzamento entre a consulta e os termos. Nesse contexto, vários autores (p.ex: Richardson e Sweaton (1995), Voorhees (1998), entre outros) têm discutido o uso da WNP na expansão da consulta por meio da inclusão de sinônimos e/ou hipônimos das palavras que compõem a consulta. Por exemplo, dada a consulta *"personal computer"* (em português, "computador pessoal"), o sistema de RI, munido da WNP, realiza a expansão pela inclusão dos sinônimos *"PC"* e *"microcomputer*" (em português, "microcomputador") e dos hipônimos *"desktop computer"* e *"portable computer"* (em português, "computador portátil").

Segundo Vossen (2003), os resultados dos trabalhos que empregam a WNP indicam que a expansão da consulta por meio da inclusão de palavras semanticamente relacionadas (i) aumenta a cobertura do sistema, ou seja, aumenta o número de documentos selecionados e retornados ao usuário e (ii) diminui a precisão, ou seja, aumenta o número de documentos irrelevantes dentre aqueles selecionados e retornados. Ainda segundo Vossen (2003), um dos motivos da queda de precisão reside no fato de que as palavras que compõem a

[6] Todos os exemplos vinculados à WNP serão dados em inglês, seguidos de suas devidas traduações.
[7] A base de dados da WNP está integralmente disponível no endereço http://wordnet.princeton.edu/man/wnstats.7WN.
[8] Dados estatísticos descritos em http://wordnet.princeton.edu/man/wnstats.7WN.
[9] O tratamento computacional das línguas naturais requer, na maioria dos casos, recursos lexicais robustos do ponto de vista quantitativo.

consulta são, na maioria das vezes, ambíguas e, por isso, aparecem em mais de um synset na WNP. Nesse caso, o sistema de RI, munido da WNP, expande a consulta com a inclusão dos sinônimos e hipônimos de todos os synsets nos quais a(s) palavra(s) da consulta aparece(m), diminuindo o número de documentos relevantes dentre os retornados. Em alguns casos, esse problema é amenizado pela especificação, *à priori*, do sentido da(s) palavras(s) da consulta pelo usuário.

4.2. Desambiguação

Um dos problemas mais discutidos no âmbito do PLN é a questão da *desambiguação de sentido*, que ocorre quando as duas ou mais opções de sentido (ou tradução) de uma dada palavra têm a mesma categoria gramatical. Por exemplo: a palavra *"know"* pode ser traduzida como *"saber"* ou *"conhecer"* (polissemia) e a palavra *"light"*, como *"leve"* ou *"luz"* (homonímia). Assim, a tarefa de desambiguação é problema comum a muitas aplicações de PLN, como a recuperação de informação, tradução automática, sumarização automática, entre outras, e consiste na escolha por um dos possíveis sentidos de uma palavra quando da sua interpretação. Para a resolução desse problema, os pesquisadores do PLN têm lançado mão de vários métodos e/ou estratégias de desambiguação. Dentre os métodos baseados em conhecimento, está o que se utiliza de léxicos computacionais, dicionários eletrônicos, ontologias e/ou *thesauri* (SPECIA, NUNES, 2004).

A aplicação da WNP na tarefa de desambiguação de sentido, em especial, pauta-se na hipótese de que palavras semanticamente relacionadas ou de um mesmo campo semântico tendem a co-ocorrer em um documento (VOSSEN, 2003). Dessa forma, a estratégia adotada é, de um modo geral, (i) identificar os sentidos/synsets que contêm as palavras em foco na WNP, (ii) identificar as relações entre sentidos/syntes e, por fim, (iii) identificar qual o sentido mais provável das palavras em foco. Por exemplo: dado um texto em que apareçam os nomes ambíguos *"organ"* e *"bass"*, o módulo de desambiguação de um sistema de tradução automática, por exemplo, munido da WNP, identifica que *"organ"* aparece em 6 synsets e *"bass"* em 8 syntes. Dentre esses synsets, o sistema identifica que o synset 3 {electric organ, electronic organ, Hammond organ, organ} de *"organ"* tem como segundo hiperônimo imediato {musical instrument, instrument} e que o synset 7 {bass} de *"bass"* tem como hiperônimo imediato {musical instrument, instrument} (Figura 2). Dessa forma, o módulo ou sistema de desambiguação é capaz de identificar que o sentido "instrumento musical" é mais provável que os demais.

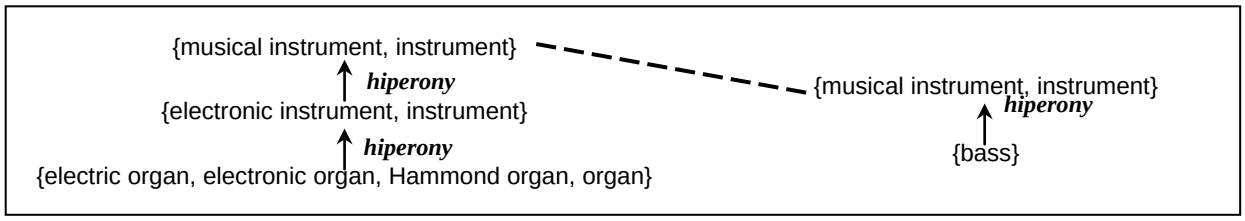

Figura 2: Relacionamento entre hiperônimos de synsets distintos que auxilia na desambiguação de sentido.

Naturalmente, o exemplo descrito é bastante simplificado. O emprego da WNP na desambiguação de sentido leva em consideração uma série de critérios para a identificação do sentido mais provável das palavras em foco (LI, ABE, 1995, LEACOCK, CHODOROW, 1998).

Os resultados do emprego da WNP na tarefa de desambiguação de sentido, no entanto, são bastante frustrantes. Kilgarrif (1997) e Véronis (1998) afirmam que o problema reside no refinamento dos sentidos, uma vez que foram delimitados manualmente. Em outras palavras, os autores afirmam que a existência de sentidos semanticamente muito relacionados, como ilustrado em (1), dificultam a tarefa automática de desambiguação.

(1)1. {indulgent} – (showing or characterized by or given to indulgence; "indulgent grandparents")
2. {indulgent, lax, lenient, soft} – (tolerant or lenient; "indulgent parents risk spoiling their children")

Palmer (2000) argumenta que os melhores resultados na desambiguação ocorrem quando são utilizadas informações vinculadas às restrições selecionais, posto que, nesses casos, a distinção de sentido baseia-se em critérios mais explícitos.

4.3. Sumarização

A sumarização no domínio do PLN consiste na produção automática de um sumário (ou resumo) a partir de um ou mais textos-fonte, preservando seu conteúdo informacional (MANI, 2001). Uma das abordagens utilizadas consiste na extração de sentenças completas do texto-fonte para a construção do sumário. Mas quais sentenças comporão o sumário? Para responder a tal questão, os pesquisadores do PLN comumente lançavam mão do tratamento numérico dos textos, explorando, por exemplo, a *freqüência* das palavras no texto, *frases indicativas* (p.ex.: "em conclusão", "este artigo descreve"), *posição* dos parágrafos (BARSILAY, ELHADAD, 1997), etc.

Tais estratégias ou critérios, no entanto, são muito superficiais e dependentes de gênero. Assim, foram desenvolvidas estratégias mais sofisticadas, que incorporam ao tratamento numérico das informações textuais também o processamento de informações lingüísticas dos textos-fonte. A proposta de Barzilay e Elhadad (1997), por exemplo, utiliza como critério a *coesão lexical* (isto é, o encadeamento de itens lexicais no texto), identificando nos textos-fonte as possíveis *cadeias lexicais* (isto é, seqüência de palavras semanticamente relacionadas). Aquelas cadeias mais fortemente conectadas indicam as sentenças significativas para compor o sumário.

Essa proposta tornou-se factível do ponto de vista computacional principalmente pela disponibilidade, dentre outras ferramentas, de recursos como a WNP, capazes de indicar as relações semânticas entre as palavras (RINO, PARDO, 2003). Mais especificamente, o método de Barzilay e Elhadad (1997) consiste na (i) seleção das palavras candidatas (substantivos e compostos nominais) e na (ii) construção das cadeias lexicais, que envolve a identificação das relações léxico-semânticas e lógico-conceituais entre as palavras. Para ilustrar tais etapas, considera-se o texto do Quadro 1, adaptado de Barzilay e Elhadad (1997).

Mr. Kenny is the **person** that invented an anesthetic **machine** which uses **micro-computers** to control the rate at which an anesthetic is pumped into the blood. Such **machines** are nothing new. But his **device** uses two **micro-computers** to achieve much closer monitoring of the pump feeding the anesthetic into the patient.[10]

Quadro 1: Texto utilizado na exemplificação do processo de identificação de cadeias lexicais.

Por meio da associação das palavras em negrito (do texto no quadro anterior) aos synsets da WNP e da identificação das relações entre esses synsets, o método é capaz de identificar, ao final do processo, as cadeias lexicais descritas na Figura 3 (extraída de Barzilay e Elhadad (1997)).

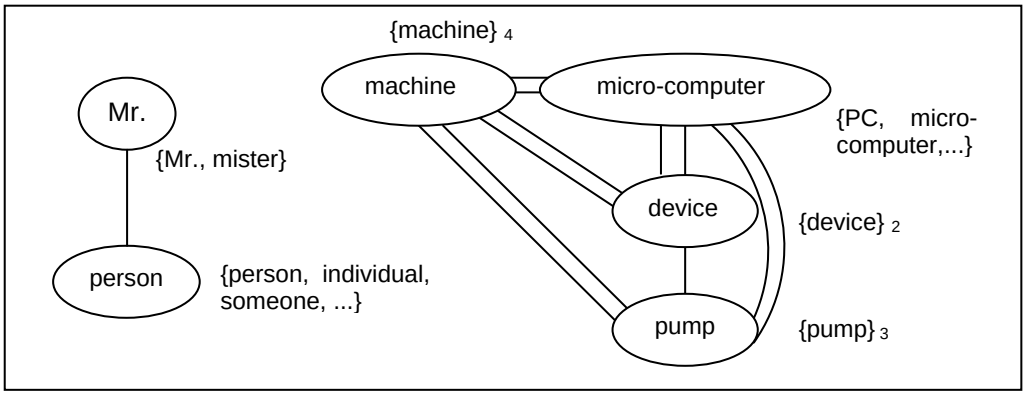

Figura 3: Cadeias lexicais identificadas no texto-exemplo por meio das relações da WNP.

O cômputo do relacionamento semântico das cadeias de palavras é feito de diversas formas: pela identificação de palavras idênticas ou palavras com mesmo significado, por sinonímia e por relações ontológicas de herança ou "parentesco". No caso do exemplo da Figura 2, a cadeia mais forte é a da direita.

Por fim, para identificar as sentenças que comporão o sumário, os autores propõem três regras distintas que podem ser aplicadas: a que focaliza a seleção da primeira sentença no texto-fonte em que aparece pela primeira vez um dos conceitos de cada cadeia forte; a que seleciona as sentenças que possuem

[10] "Mr. Kenny é a pessoa que inventou uma máquina anestésica que usa microcomputadores para controlar o taxa em que um anestésico é bombeado para o sangue. Tais máquinas não são nada novas. Mas seu dispositivo usa dois microcomputadores para alcançar um monitoramente melhor da bomba alimentando o anestésico ao paciente" (tradução nossa).

os membros mais representativos das cadeias fortes; e a que seleciona sentenças das regiões do texto em que há maior concentração de cadeias fortes. Na avaliação realizada por Barzilay e Elhadad, a utilização da WNP melhorou a tarefa de sumarização automática, já que houve uma taxa alta de concordância entre os juízes humanos que avaliaram a qualidade dos sumários automáticos em relação aos sumários produzidos por humanos.

A seguir, são apresentadas algumas proposta de expansão para as bases do tipo *wordnet.*

5. Algumas propostas de expansão/refinamento

No âmbito do PLN, há vários os trabalhos que apresentam propostas de refinamento, enriquecimento e/ou expansão das redes *wordnet*. Com base no tipo de informação a ser agregada às wodnets, é possível classificar algumas dessas propostas. Em outras palavras, nota-se que alguns trabalhos objetivam integrar às wordnets ora ontologias "top-level" ora informação de natureza valencial.

5.1. Integração de ontologias "top-level"

De um modo geral, as ontologias ditas *top-level* são hierarquias independentes de línguas, cujos elementos descritos são conceitos abstratos ou gerais como *objeto, lugar* e *ação* (GANGEMI et al, 2001). Dentre os trabalhos que visam à integração desse tipo de ontologia, estão os de Vossen (1998b), Hovy (1998), O'Hara et al (1998), Niles e Pease (2003), entre outros. A seguir, ilustra-se esse tipo de ontologia com a descrição da proposta de Vossen (1998).

No âmbito do projeto EuroWordNet[11], Vossen (1998b) propõe que cada synset das wordnets específicas sejam indexados a uma ontologia "top-level", no caso, a *"EuroWordNet's Top-Ontology"* (doravante, EWN TO). Antes de se descrever tal ontologia, ressalta-se que o projeto EuroWordNet reuniu, em uma BDL multilíngüe, as wordnets do inglês (britânico), holandês, espanhol, italiano, alemão, francês, tcheco e estônio[12] (VOSSEN, 1998b). Nessa base multilíngüe, os synsets específicos de cada wordnet, considerados semanticamente equivalentes, são identificados por um mesmo ILI, isto é, "*Inter-Lingual-Index*" (em inglês, "*Índice-Inter-Lingual*"). Cada ILI é identificado pelo número do registro de um synset (juntamente com todas as informações a ele associadas, ou seja, synset e glosa) da versão 1.5 da WordNet de Princeton (PETERS et al., 1998). O Quadro 2 ilustra essa identificação.

Bases lexicais	**Synsets**	**ILI-Nº= Synset/"Glosa" da WordNet 1.5**
Wordnet holandesa	{violoncel; cel; cello}	ILI-02411468 = {violoncello; cello}/"a large stringed instrument; seated player holds it upright while playing"
Wordnet espanhola	{chelo; violoncelista; violoncelo; violonchelo; cello}	
Wordnet italiana	{violoncello}	
WordNet americana	{violoncello; cello}	

Quadro 2: Equivalência conceitual entre três *synsets*, pertencentes a três *wordnets* distintas, por meio do ILI.

Com base no Quadro 1, pode-se dizer que o conceito identificado pelo ILI-02411468 é lexicalizado, no holandês, pelas formas {violoncel; cel; cello}; no espanhol, pelas formas {chelo; violoncelista; violoncelo; violonchelo; cello}; no italiano, pela forma {violoncello} e, no inglês, pelas formas {violoncello; cello}.

Como mencionado, a proposta de Vossen (1998) é a de indexar cada ILI à EWN TO. Essa ontologia é composta por 63 conceitos que, na verdade, funcionam como traços semânticos que podem ser aplicados disjuntivamente ou conjuntivamente. No Quadro 3, apresenta-se parte da WNN TO proposta para as "entidades de primeira ordem" (do inglês, *"first-order-entities"*) (qualquer entidade concreta perceptível e localizada no tempo e no espaço tridimensional) (LYONS, 1997) e ilustra-se quais os conceitos aos quais o ILI representado pelo synset {violoncello; cello} está vinculado, no caso, à conjunção {Artifact+Instrument+Object+} (*).

[11] http://www.dcs.shef.ac.uk/nlp/funded/eurowordnet.html

[12] A extensão da EuroWordNet para outras línguas consolida-se com a integração de línguas como o português europeu, sueco, basco, catalão, russo, grego e dinamarquês.

1st Order Entity	*Function*	*Origin*
Composition	Building	Artifact
Group	Comestible	*
Part	Container	Natural
Form	Covering	Living
Object *	Furniture	Animal
Substance	Garment	Creature
Gas	Instrument *	Human
LiquidLiquid Solid	….Vehicle	Plant
...		

Quadro 3: parte da WNN TO proposta para as "entidades de primeira ordem"

As propostas de expansão das redes wordnets visam, no geral, fazer dessas redes recursos lexicais cada vez mais úteis para o PLN. Com a inclusão de ontologias top-level, é possível, por exemplo, generalizar conceitos, minimizando, assim, o problema do refinamento demasiado dos synsets/conceitos nas wordnets (SANFILIPPO et al., 1999).

5.2. Integração de informação valencial

Como mencionado, os melhores resultados na tarefa de desambiguação de sentido ocorrem quando são utilizadas informações vinculadas às restrições selecionais (PALMER, 2000). As redes wordnets não apresentam explicitamente informações relacionadas à valência dos itens lexicais (grelha temática e restrições selecionais). Dessa forma, com o objetivo de melhorar o desempenho da aplicação das wordnets no PLN, algumas propostas de inserção de informações relacionadas à valência foram elaboradas. Dentre elas, estão as de Wagner (2000) e Di Felippo e Dias-da-Silva (2004).

Di Felippo e Dias-da-Silva (2004), em especial, propõem a inserção, na base da wordnet para o português do Brasil, a Wordnet.Br, da valência dos adjetivos qualificadores (QLs), pois estes, assim como os verbos, são os predicadores da língua por excelência (BORBA, 1996). O adjetivo *"esperto"* em (a) *"O garoto é esperto"* e em (b) *"A água estava esperta"*, por exemplo, projeta apenas um argumento, o A1, já que a semântica desse item implica "*x* é *esperto*". Na sentença (a), o adjetivo tem sentido de *"astuto"*, o qual é gerado pela combinatória de *"esperto"* com um argumento de traço semântico [+humano] (*"garoto"*). Já em (b), *"esperto"* tem sentido de *"quase quente"*, o qual é gerado pela combinatória desse adjetivo com um argumento de tipo semântico [*–animado*] *("a água"*). No caso de *"esperto"* em (a,b), o A1 é a entidade sobre a qual se verifica uma situação. A valência dos Ps tem sido amplamente representada por meio de um construto formal denominado *estrutura de argumentos*. Essa representação é composta por *papéis temáticos* (rótulos abstratos que representam as funções semânticas dos argumentos) e *restrições selecionais* (rótulos abstratos que restringem a semântica dos argumentos). No caso de *"esperto"* em (a), tem-se a estrutura <(A1)Tema[+*hum*]>.

Com base no pressuposto de que tanto o synset como a estrutura de argumentos representam um conceito (ou sentido), os autores propõem associar estruturas de argumentos a synsets. Em outras palavras, as estruturas de argumentos seriam uma espécie de "comentário", no sentido computacional desse termo, dos sentidos/synsets, refinando, assim, as informações relativas aos adjetivos qualificadores na base da WordNet.Br.

A seguir, apresenta-se o estado atual do desenvolvimento da WordNet.Br.

6. A wordnet para o português brasileiro, a *WordNet.Br*

Iniciou-se em 2001 o empreendimento de construção da wordnet para o português do Brasil, a WordNet.Br (DIAS-DA-SILVA, et al. 2002). Atualmente, a base da WordNet.Br contém 44.678 unidades lexicais do português brasileiro (17.388 substantivos, 15.073 adjetivos, 11.078 verbos e 1.139 advérbios), distribuídas em aproximadamente 19.872 synsets, e registra apenas as relações de sinonímia e antonímia (DIAS-DA-SILVA et al, 2006). No seu desenvolvimento, no entanto, estão previstas as tarefas de (i) especificação das demais relações (hiponímia, meronímia, acarretamento e causa) e informações periféricas contidas na WNP (glosas e frases-exemplo) e (ii) indexação, nos moldes da EuroWordNet, da WNP à base da WordNet.Br.

Na etapa atual de desenvolvimento, estão sendo realizadas as seguintes tarefas lingüísticas e lingüístico-computacionais (DIAS-DA-SILVA et al, 2006):

(i) análise da consistência semântica dos synsets;
(ii) coleta/seleção e inserção das frases-exemplo para uma parcela da base dos verbos (baseada em córpus);
(iii) especificação e inserção de glosas para uma parcela dos verbos;
(iv) indexação semi-automática da base da WordNet.Br à WordNet de Princeton;
(v) herança automática das demais relações da WordNet de Princeton.

Vale ressaltar que a WordNet.Br está sendo construída com base em pressupostos oriundas tanto da Semântica Lexical, pura e computacional, como da Lexicografia Computacional e que, quando estável e mais completa, tornará factível certos tratamentos computacionais do português como os ilustrados neste trabalho.

7. Considerações finais

As redes wordnets são "bases de conhecimento estático" úteis para o processamento das línguas naturais, posto que fornecem uma descrição robusta da dimensão semântico-conceitual do léxico de uma língua natural. No geral, o emprego da WNP – entendida como uma ontologia lingüística – tem melhorado o desempenho de determinadas aplicações ou tarefas computacionais, por exemplo, a expansão de consulta em recuperação de informação e a sumarização de textos. Para determinadas aplicações, entretanto, a WNP se mostra um recurso deficiente ou problemático. Na tarefa de desambiguação de sentido, por exemplo, o emprego da WNP não se mostrou tão eficiente, pois o refinamento na distinção dos conceitos (synsets) ou a existência de conceitos semanticamente muito relacionados dificulta o processo automático de identificação do sentido de uma palavra. Com o intuito de aperfeiçoar esse tipo de base, várias propostas de refinamento e/ou expansão têm sido elaboras no âmbito PLN, tais como a inclusão de ontologias de "alto nível" e estrutura de argumentos.

8. Referências Bibliográficas

ALLEN, J. **Natural language understanding.** Addison Wesley, 1995.

BARZILAY, R.; ELHADAD, M. Using lexical chains for Text Summarization. In: INTELLIGENT SCALABLE TEXT SUMMARIZATION WORKSHOP - ISTS/ACL, 1997, Madrid. **Proceedings…** Madrid, 1997. Disponível em <http://www.cs.bgu.ac.il/~elhadad/lexical-chains.pdf>. Acesso em: 10 nov. 2006.

BORBA, F. S. **Uma gramática de valências para o português**. São Paulo: Editora Ática, 1996.

CRUSE, D. A. **Lexical Semantics**. Cambridge: Cambridge University Press, 1986.

DIAS-DA-SILVA, B. C. **A face tecnológica dos estudos da linguagem: O processamento automático das línguas naturais**. Araraquara, 1996. 272p. Tese (Doutorado em Letras) - Faculdade de Ciências e Letras, Universidade Estadual Paulista, Araraquara.

DIAS-DA-SILVA, B. C.; OLIVEIRA, M. F.; MORAES, H. R. Groundwork for the development of the Brazilian Portuguese Wordnet. In: INTERNATIONAL CONFERENCE PORTUGAL FOR NATURAL LANGUAGE PROCESSING – PorTAL, 3, 2002, Faro. **Proceedings…** Faro, 2002. p. 189-196.

DIAS-DA-SILVA, B.C.; DI FELIPPO, A.; HASEGAWA, R. Methods and tools for encoding the WordNet.Br sentences, concept glosses, and conceptual-semantic relations. In: INTERNATIONAL WORKSHOP ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE – PROPOR, 7, 2006, Itatiaia. **Proceedings…** Itatiaia, 2006. p. 120-130. ISBN 3-540-34045-9

DI FELIPPO, A.; DIAS-DA-SILVA, B. C. Edição de informações sintático-semânticas dos adjetivos na base da rede Wordnet.Br. In CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO (SBC), 24/ WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (TIL), 2, 2004, Salvador. **Anais...** Bahia, 2004. 1 CD-ROM. ISBN 85-88442-93-0.

FARRAR, S.; BATEMAN, J. Linguistic ontology baseline. **OntoSpace Internal Report I1-[OntoSpace]: D3. SFB/TR8**. Bremen: Collaborative Research Center for Spatial Cognition, 2005.

FELLBAUM, C. (Ed.). **Wordnet: an electronic lexical database**. Cambridge, The MIT Press, 1998.

FERREIRA, A.B.H. **Dicionário Aurélio eletrônico: novo dicionário Aurélio – século XXI** (versão 5.0). São Paulo: Positivo Informática Ltda., 2004. CD-ROM.

GANGEMI, A.; GUARINO, N.; MASOLO, C.; OLTRAMARI, A. Understanding top-level ontological distinctions. In: WORKSHOP ON ONTOLOGIES AND INFORMATION SHARING. 2001.

GUARINO, N.; GIARETTA, P. Ontologies and knowledge bases: Towards a terminological clarification. In: MARS, NJ.I. (Ed.). **Towards very large knowledge bases**, IOS Press, 1995. p. 25-32. Disponível em: <http://www.loacnr.it/Papers/KBKS95.pdf>. Acesso em: 19 nov. 2006.

GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing. In: WORKSHOP ON FORMAL ONTOLOGY, Padua, 1993. Edited collection by Nicola Guarino.

HOVY. Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In: INTERNATIONAL CONFERENCE ON LANGUAGES RESOURCES AND EVALUATION (LREC), 1, 1998, Granada, Spain, 1998.

KILGARRIFF, A. I don't believe in word senses. **Computers and the Humanities,** v. 31, n. 2, p. 91-113, 1997.

LEACOCK, C.; CHODOROW, M. Combining local context and WordNet similarity for word sense identification. In: FELLBAUM, C. (Ed.). **WordNet: An electronic lexical database.** Cambridge: The MIT Press, 1998, p. 265-84.

LENAT, D., Guha, R. **Building large knowledge based systems: Representation and inference in the Cyc Project**. Addison-Wesley Publishing, 1990.
LI, H.; ABE, N. Generalizing case frames using a thesaurus and the MDL principle. In: CONFERENCE ON RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING (RANLP), 1995, Tzigov Chark, Bulgaria. **Proceedings …** Tzigov Chark, Bulgaria, 1995, p. 239-48.

LYONS, J. **Semantics**. Cambridge: Cambridge University Press, 1977.

MANI, I. **Automatic Summarization.** Amsterdam: John Benjamins Publishing Co., 2001.

MILLER, G. A.; FELLBAUM, C. Semantic networks of English. **Cognition,** v. 41, p. 197-229, 1991.

MORATO, J., MARZAL, M.A., LLORENS, J., MOREIRO, J. WordNet applications. In: GLOBAL WORDNET CONFERENCE, 2, 2004, Brno. **Proceedings...** Brno, 2004, p. 270-278.

NILES, I.; PEASE, A. Linking Lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged ontology. In: INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE ENGINEERING (IKE), 2003, Las Vegas. **Proceedings …** Las Vegas, 2003.

O'HARA, T.; MAHES, K.; NIRENBURG, S. Lexical acquisition with WordNet and the Mikrokosmos ontology. In: COLING-ACL - WORKSHOP ON "USAGE OF WORDNET IN NATURAL LANGUAGE PROCESSING SYSTEMS", 36, 1998, Montreal/Quebec. **Proceedings…** Montreal/Quebecl, 1998. p 94-101.

PALMER, M. Consistent criteria for sense distinctions**. Computers and the Humanities,** v. 34, n.1-2, 2000.

PALMER, M. Miltilingual resources. **Linguistica Computazionale,** v.14-15, 2001.

PETERS, W., VOSSEN, P., DÍEZ-ORZAS, P., ADRIAENS, G. Cross-linguistic alignment of wordnets with an inter-lingual-index. **Computers and the Humanities,** v. 32, p. 221-251, 1998.

REITER, E.;DALE, R. **Building natural language generation systems.** Cambridge: University Press, 2000.

RICHARDSON, R.; SWEATON, A. Using WordNet in a knowledge-based approach to Information Retrieval. In: BCS-IRSG COLLOQUIUM, 1995. Disponível em: <http://citeseer.ist.psu.edu/richardson95using.html>. Acesso em: 18 de nov. 2006.

RINO, L.H.M.; PARDO, T.A.S. A Sumarização Automática de textos: principais características e metodologias. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO (vol. VIII), 23/ JORNADA DE MINICURSOS DE INTELIGÊNCIA ARTIFICIAL, 3, 2003, Campinas. **Anais...** Campinas, 2003, p. 203-245.

ROGET, P. M. **Roget´s Thesaurus**. London: Galley Press, 1972.

SANFILIPPO, A.; CALZOLARI, N., ANANIADOUS, S., GAIZAUSKAS, R., SAINT-DIZIER, P., VOSSEN, P. (Eds.). Preliminary recommendations on Lexical Semantics encoding. **EAGLES Final Report, LE3-4244 ,** 1999.

SPECIA, L.; NUNES, M.G.V. Desambiguação lexical automática de sentido: um panorama. **Série de Relatórios Técnicos do NILC** TR-04-08, 2003, 117p.

VÉRONIS, J. A Study of polysemy judgements and inter-annotator agreement. In: SENSEVAL WORKSHOP, 1998, Sussex, **Advanced Papers…** Sussex, p. 2-4, 1998.

VIEGAS, E.; K. MAHESH; NIRENBURG. S. Semantics in action. In: PROCEEDINGS OF THE WORKSHOP ON PREDICATIVE FORMS IN NATURAL LANGUAGE AND IN KNOWLEDGE BASES, 1996, Toulouse. **Proceedings …**Toulouse, 1996, p. 108-115.

VOORHEES, E. M. Using WordNet for text retrieval. In: FELLBAUM, C. (Ed.). **WordNet: An electronic lexical database**. Cambridge: The MIT Press, 1998, p. 285-304.
VOSSEN, P. EuroWordNet: Linguistic ontologies in a multilingual database. **Communication and Cognition for Artificial Intelligence** (Special Issue), v. 15, n. (1-2), p. 37-80, 1998a.

VOSSEN, P. Introduction to EuroWordNet. **Computers and the Humanities,** v. 32, p. 73-89, 1998b.

VOSSEN, P. Ontologies. In: MITKOV, R. (Ed.). **The Oxford handbook of Computational Linguistics**. Oxford: Oxford University Press, 2003, p. 464-82.

XU, J.; CROFT, B. W. Query expansion using local and global document analysis". In: ACM CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (SIGIR), 19, 1996, Zurich. **Proceedings…** Zurich, 1996, p. 4-11.

WAGNER. Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis. In: WORKSHOP ON ONTOLOGY LEARNING AT THE 14TH ECAI, 1, 2000, Berlin. **Proceedings…** Berlin, 2000, p. 37-42.