8
Conclusões

Apresentei aqui subsídios para a elaboração automática de ontologias específicas quanto ao domínio. Embora a metodologia, em si, não seja nova, pois a correlação entre relações de hiponímia e a ocorrência de determinados padrões léxico-sintáticos em textos foi sugerida por Hearst (1992), acredito que as principais contribuições deste trabalho estão

(i) na proposta de novos padrões para a identificação da hiperonímia;

(ii) na adaptação e refinamento dos padrões existentes para o português;

(iii) na indicação de que o cruzamento das informações extraídas com os padrões, gerando inferências (produzindo conhecimento), é um processo válido e produtivo, desde que seja realizado em um corpus de domínio;

(iv) na adoção de uma perspectiva relativista com relação ao significado, que tem como conseqüência principalmente a análise de relações semânticas pouco convencionais, que poderiam ser consideradas "erro". Uma perspectiva relativista se mostra produtiva na medida em que legitima os dados vindos do corpus e as relações de significado que nele aparecem.

Com relação aos itens (i) e (ii), os padrões "tipos de" e "chamado" apresentaram um alto índice de precisão, embora tenham identificado poucas relações. A análise cuidadosa da estrutura "tais como" levou à identificação da estrutura variante "como", de alta freqüência no corpus, e a ajustes nas regras relacionados à presença de vírgula nas expressões. Uma análise minuciosa dos resultados iniciais dos padrões levou à criação da regra HHiper/ HHipo, que considera como sintagma hiperônimo / hipônimo apenas o último substantivo em SNs que contém sintagmas preposicionados – estruturas sintáticas altamente ambíguas na língua. Com isso, os resultados obtidos na extração foram muito positivos, principalmente se comparados aos obtidos em outros estudos (Hearst,

1998; Cederberg e Widdows, 2003). Porém, como já dissemos antes, a comparação deve ser vista com cautela, pois tanto a forma de avaliação – julgamento humano – é subjetiva, quanto as condições em que os trabalhos foram realizados foram diferentes (número de relações avaliadas, técnica de identificação das relações). Além disso, é preciso considerar que boa parte do sucesso na identificação é dependente de um fator "externo" – a etiquetagem de classes de palavras e de sintagmas nominais. Neste trabalho, o corpus etiquetado passou por uma revisão manual, na tentativa de minimizar a interferência de outras variáveis na identificação das relações, principalmente porque a estrutura do SN em português é mais complexa (tendo em vista a identificação automática) do que a do inglês (Oliveira e Santos, 2005). Ainda assim, os resultados da comparação servem como ilustração do potencial das regras.

Por fim, em favor da regras apresentadas aqui, lembro que, no padrão "como/tais como", 29% dos erros foi decorrente da presença de uma oração no sintagma hiperônimo / hipônimo[38], e que o extrator automático de sintagmas nominais subjacente à identificação das estruturas não reconhece SNs com orações. Conseqüentemente, é razoável supor que os resultados poderiam ser ainda melhores utilizando um modelo de SN que admita a identificação automática de orações.

Já no caso da regra "e outros", como 20% dos erros é decorrência de uma estratégia discursiva em que o hiperônimo retoma apenas o último elemento sa coordenação, uma forma de melhorar a precisão seria ajustar a regra para considerar apenas o último substantivo. Nesse caso, embora haja alguma perda na abrangência, a maior precisão pode ser útil, por exemplo, para uma etiquetagem de corpus de treino para sistemas de aprendizagem automática.

Os resultados do padrão "conhecido/a/os/as", que possibilitaria a inclusão de relações de co-referência na ontologia, foram desanimadores, pois apresentaram uma grande ambigüidade entre a expressão de co-referência e de hiperonímia. Em experimentos-piloto, não descritos neste trabalho, foram testadas também a identificação automática de apostos[39] e de orações explicativas[40] – construções

[38] Em "*fatores de risco como o hábito de fumar...*" é extraída a relação *fatores de risco* > *hábito*

[39] Exemplos de aposto:

(a) [*Metoprene, substância análoga a o hormônio juvenil de os insetos*,] que atua em as formas imaturas (larvas e pupas) , impedindo...

interessantes por também expressarem relações de co-referência. Porém, os resultados da identificação automática foram decepcionantes, o que levou à exclusão destas estruturas da metodologia. É importante salientar, contudo, que o problema não foi de ambigüidade das estruturas, como no caso do padrão "conhecido/a/os/as como", mas de natureza computacional: a identificação automática foi ineficaz. As estruturas são boas candidatas à expressão de co-referência, e merecem uma investigação detalhada quanto à possibilidade de identificação automática.

Com a exclusão dessas estruturas, que ofereceriam à ontologia relações de co-referência, a ontologia ficou apenas com as relações de hiperonímia, nisto se assemelhando a taxonomias.

Os resultados demonstraram, também, que freqüentemente nem todas as relações possíveis serão explicitadas na ontologia, indicando a necessidade de um trabalho humano complementar. Não há, por exemplo, nos resultados, uma relação entre a taxonomia de *animais* e a taxonomia de *mamíferos*. Isto nos faz ver com alguma cautela a afirmação de que "as categorias emergem do corpus" – sim, emergem, mas relações relevantes podem não emergir. Por outro lado, em uma visão otimista, é possível imaginar que em um corpus maior o problema seja minimizado.

A construção automática de ontologias a partir de grandes corpora é interessante tanto por reduzir a preocupação com o conhecimento a ser codificado, visto que esse conhecimento estaria no corpus, quanto por permitir a automação do processo, facilitando o trabalho de atualização. O que se tem, ao final, é um deslocamento do problema: em certa medida, passa-se para o corpus a "responsabilidade" de direcionar a construção da ontologia.

Investigações sobre a forma de avaliação de ontologias construídas automaticamente a partir de corpus são de fundamental importância, mas ainda não atingiram resultados satisfatórios. A versão simplificada da proposta de

---

(b) Estudos realizados em algumas áreas endêmicas de o estado de São Paulo utilizando a reação de imunofluorescência indireta, em comparação ao [*exame parasitológico de fezes, Kato-Katz,*] mostraram ..

[40]Exemplos de orações explicativas:

(a) Atualmente , a resistência à [*cloroquina, que é o antimalárico mais barato e mais amplamente usado*,] é comum em a África.

(b) ...foram devidas às [*doenças cardiovasculares, que são a primeira causa de morte em todas as grandes regiões de o país*,] com mortalidade proporcional...

avaliação de Brewster et al. (2004), que sugere uma comparação entre os termos relevantes presentes no corpus e os termos da ontologia, parece viável , justamente por prescindir de Wordnet e de um corpus semanticamente anotado. Porém, embora a metodologia verifique a adequação entre corpus e ontologia, não há como assegurar a correção das relações semânticas entre os termos. A proposta de Etzioni et al. (2005), de validação das relações por meio de busca por determinadas expressões na Web ("X é um Y"), pode ser um bom complemento nesse sentido. O principal problema desta abordagem é que, para a língua portuguesa, mecanismos de busca como o Google desconsideram acentos, o que leva a resultados indesejados.

A forma de avaliação utilizada aqui – validação manual – embora útil por permitir alguma comparação com outros trabalhos, é falha principalmente por não oferecer uma medida confiável nesta comparação. Julgamentos humanos são subjetivos, e um dos motivos para se sustentar a informação da ontologia em corpus é justamente a tentativa de minimizar esta subjetividade.

Retomando os critérios a que ontologias devem atender segundo Brewster e Wilks (2004) percebemos que todos foram atendidos, exceto o critério 5, que trata da origem dos dados para a construção da ontologia (documentos e uma taxonomia já existente), por razoes óbvias.

O critério 1, coerência interna, é atendido uma vez que as relações são extraídas de um corpus específico do domínio e é razoável supor que, em um mesmo domínio, haja coerência entre os usos dos termos. O critério 2, herança múltipla, também foi atendido, já que um mesmo termo pode ter mais de um pai na ontologia. Como os algoritmos de extração são simples, imagino que não haja complexidade na computação, o que está de acordo com o critério 3. Por fim, como os rótulos das categorias são os próprios termos extraídos, o critério 4, que aponta para a necessidade de nós com rótulos únicos, e não com rótulos que são grupos de palavras, também está atendido.

Um último comentário com relação aos resultados diz respeito aos nomes próprios. Embora o objetivo inicial do trabalho não tenha sido a classificação semântica de nomes próprios, tarefa que pertence à área de Reconhecimento de Entidades Mencionadas (REM) (ou NER – Named Entity Recognition, subárea da Extração de Informação), quando a metodologia foi aplicada a um corpus geral, composto por notícias de jornal, o grande número de relações envolvendo essas

estruturas mostrou que as regras podem ser uma ótima ferramenta para a extração de entidades mencionadas. Lembro novamente, contudo, que o corpus passou por uma revisão manual, o que minimizou consideravelmente a quantidade de erros decorrentes de dificuldades no processo de segmentação (anterior ao processo de classificação semântica). Assumindo, novamente em uma visão otimista, que a tarefa de segmentação de nomes próprios já esteja resolvida, persistem outros problemas relativos à natureza gramatical da categoria, e que irão interferir em sua classificação semântica: *AIDS* é um nome próprio? Em caso afirmativo, subentende-se, portanto, que um critério para uma palavra ser considerada nome próprio é constituir uma sigla (pois o que mais difere *AIDS* de *sarampo*, *gripe* etc?) Mas até que ponto *AIDS* é ainda reconhecida como sigla, e não como palavra simples da língua (vide *aidético*)? E *doença de Chagas*, *Mal de Alzheimer* ? Também são nomes próprios?

Por fim, lembro que a metodologia se beneficiaria com a identificação de expressões multi-vocabulares (EMVs) nominais no corpus. Embora os critérios de identificação de EMVs sejam controversos (Oliveira et al. 2004), a percepção de que determinadas combinações nominais, principalmente as de estrutura *Substantivo + Preposição "de" + Substantivo*[41] devem ser consideradas um único item lexical tem implicações importantes sobretudo na aplicação das regras HHiper / HHipo. O fato de EMVs nominais poderem ser identificadas com sucesso por meio de testes estatísticos, já que suas estruturas são, muitas vezes, sintaticamente transparentes, torna a incorporação dessas estruturas viável a curto-prazo. A transparência sintática de EMVs nominais, porém, tem conseqüências na aplicação da regra HiperN. Em *dor de cabeça*, por exemplo, é interessante que a regra seja empregada, originando o hiperônimo *dor*. Já em *pé de atleta*, a criação do hiperônimo *pé* seria um problema. A aplicação, nas EMVs nominais, de uma medida de similaridade capaz de avaliar a transparência sintática dessas construções seria útil para a identificação de EMVs que não estariam sujeitas à aplicação da regra HiperN.

8.1. Desdobramentos

Embora o objetivo inicial da ontologia tenha sido auxiliar tarefas que envolvem o processamento automático de textos, os resultados mostraram que a metodologia também pode ser de grande valia para investigações lexicográficas e lingüísticas. Nesse sentido, o insucesso dos resultados das inferências no corpus genérico pode ser visto como conseqüência de um "efeito colateral" positivo, pois a aplicação das regras no corpus possibilitou dois importantes achados: um tratamento para a classificação semântica de nomes próprios e um auxílio para lexicógrafos na tarefa de elaboração de dicionários.

8.1.1. Desdobramentos "mais" lingüísticos

De um ponto de vista lexicográfico, as relações entre os termos podem ser uma fonte valiosa para a observação dos contextos de ocorrência das palavras, contribuindo para a elaboração de dicionários e de léxicos específicos. A análise do comportamento das palavras ajuda na identificação dos seus múltiplos usos, fornecendo material para um processo preciso, empiricamente motivado e objetivo de atribuição de sentido.

Outro trabalho interessante relacionado à descrição do português é a caracterização formal, para posterior identificação automática, dos substantivos *relacionais*, aqueles que expressam relações entre indivíduos. como *pai*, *amigo*, *vizinho*, *adversário*, *concorrente*, *fundador*, *membro*, etc. A tarefa de classificação semântica de nomes próprios também se beneficiaria bastante deste tipo de informação.

A elaboração de critérios formais para a identificação automática de "adjetivos gerais", nos moldes da proposta de Oliveira (2006) de caracterização do substantivo-suporte, também seria de grande valia para tarefas de PLN.

[41] Alguns exemplos retirados do corpus: *prisão de ventre*, *atestado de óbito*, *taxa de natalidade*, *taxa de mortalidade*, *dor de cabeça*, *cinto de segurança*.

8.1.2.
Desdobramentos "mais" computacionais

Do ponto de vista do PLN, um trabalho interessante é aplicar técnicas de clusterização para distinguir grupos de palavras similares, utilizando como *seed words* palavras que já estão na ontologia, e verificar se o hiperônimo das *seed words* pode ser também hiperônimo das palavras do *cluster*. Com isso, haveria um aumento significativo da ontologia, com o acréscimo de co-hipônimos.

Outra possibilidade de trabalho é explorar de forma mais sistemática as técnicas de extração de informação na elaboração de ontologias. Por exemplo: excetuando-se os verbos auxiliares, os verbos mais freqüentes no corpus de saúde são *causar* e *evitar*. Supõe-se, portanto, que tais verbos expressem relações relevantes para o domínio de saúde. Em seguida, deve ser possível identificar, automaticamente, os sujeitos e objetos dos verbos, isto é, *X causa Y* e *X evita Y*. Desse modo, criam-se, semi-automaticamente, *templates* para a extração de mais informações.

Com relação aos padrões léxico-sintáticos utilizados neste trabalho, que podem ser considerados padrões de *templates* de EI, a principal vantagem está na generalidade: são padrões que podem ser aplicados a qualquer domínio, a qualquer tipo de texto – e o mesmo se aplica aos padrões referentes ao aposto e orações explicativas, não implementados.

8.2.
Considerações finais

Os resultados positivos da metodologia, tanto relativos ao corpus de domínio como ao corpus geral, indicam que sua aplicação pode ser uma importante aliada na elaboração de ontologias. Os resultados são decorrentes de uma análise lingüisticamente motivada e podem – devem – ser complementados com estratégias computacionais.

Uma estratégia utilizada, mas pouco vista em trabalhos de PLN, é a análise sistemática dos erros. Embora esta seja, sem dúvida, uma tarefa penosa, é de extrema valia para um entendimento de "*por que as coisas não estão acontecendo como o esperado*", principalmente quando estamos tratando de língua (em oposição a números). A elaboração das regras HHiper/HHipo, por exemplo, foi

decorrente de análise dos erros. A avaliação dos resultados – e dos erros – em termos das tradicionais medidas de precisão e abrangência não fornece pistas para aquilo que só a observação humana é capaz de descobrir, pois informam "apenas" o quanto os resultados obtidos ficaram distantes do ideal.

Em termos gerais, a metodologia apresenta como principais vantagens (i) a facilidade na automação do processo, minimizando a intervenção humana; (ii) facilidade na categorização de domínios especializados; (iii) maior dinamicidade, pois o fato de o corpus poder ser constantemente atualizado faz com que esteja menos sujeito a falhas. Suas principais desvantagens são a alta dependência de um corpus etiquetado e a dificuldade de avaliação sistemática (e de comparação) dos resultados.