1
Introdução

O objetivo deste trabalho é apresentar subsídios para a elaboração automática de ontologias específicas quanto ao domínio. Especificamente, busco investigar até que ponto é possível a elaboração automática de ontologias diretamente a partir de corpus, sem a determinação a priori das categorias que a compõem.

Vivemos na sociedade da informação, uma sociedade na qual o volume de informação nunca foi tão grande. Nossa capacidade de compreender, selecionar e organizar o conhecimento não consegue acompanhar a velocidade com que a informação, que aparece principalmente sob a forma de textos, é disponibilizada. Nesse contexto, é fundamental o desenvolvimento de ferramentas capazes de processar esse vasto material disponível; ferramentas capazes de extrair conhecimento de textos e transformá-lo em uma codificação da informação que seja armazenável, reutilizável e recuperável. E, mais ainda, de ferramentas voltadas para a língua portuguesa.

Objetivando auxiliar as tarefas de "gerenciamento" e "manipulação" da informação contida em textos, sistemas de recuperação e extração de informação têm se tornado populares. Porém, como afirma Vossen (2003), "para processar informação é preciso informação" (2003:464). Essa informação, por sua vez, pode ser mínima, vinda de um léxico que contenha apenas indicação sobre classes de palavras, ou pode ser de grande complexidade, quando originada de alguma base que contenha formalizações sobre conhecimento de mundo.

De fato, léxicos computacionais vêm assumindo crescente importância em sistemas de processamento automático de linguagem natural (PLN) (Boguraev e Pustejovsky, 1996). Um léxico computacional pode assumir tanto a estrutura linear de um dicionário quanto uma estrutura hierárquica, e, nesse caso, se aproximaria de uma taxonomia. Pode, também, fornecer outros tipos de relação, além da hiperonímia/hiponímia presentes em taxonomias. Quando o tipo de informação codificada é de natureza "mais lingüística", como a indicação sobre

classes de palavras, é comum o uso do termo léxico; para fazer referência a alguma base que contenha formalizações sobre conhecimento de (ou de algum) mundo, o termo ontologia costuma ser mais utilizado. Porém, como lembra Vossen (2003), a diferença entre léxicos e ontologias está longe de ser clara e há, sem dúvida, uma grande sobreposição sobre a informação que ambos veiculam.

Ao lidar com ontologias – descrições do mundo ou de porções do mundo – esta tese lida, indiretamente, com significado. Com isso, dialoga com a semântica, "domínio de investigação de limites movediços" e para o qual não há jargões bem estabelecidos (Ilari e Geraldi, 1985:6). Além disso, o trabalho se insere na área essencialmente interdisciplinar que é o Processamento automático de Linguagem Natural (PLN). Termos como ontologias, tesauros, léxicos semânticos e taxonomias são amplamente utilizados quando se quer fazer referência a bases que contêm informação sobre a língua necessárias ao processamento de textos, mas sua definição difere conforme o interesse e formação dos grupos de pesquisa, havendo pouca concordância sobre o que sejam.

Um léxico semântico, por exemplo, pode ser tanto uma lista de palavras com rótulos relativos à categoria semântica – a palavra *carro* pode ser rotulada como *veículo* (Phillips e Riloff, 2002; Riloff e Shepherd, 1997), quanto uma ferramenta responsável pela normalização entre termos e conceitos (Buitelaar, 2001). Isto é, assumindo que a língua é redundante, e que diferentes termos podem fazer referência ao mesmo objeto no mundo, a função de um léxico semântico seria realizar o mapeamento entre termos similares e conceito. Esta normalização deve considerar tanto informação relativa à classe semântica, definindo o tipo de objeto que um determinado termo ou conjunto de termos similares representa (por exemplo, *sinagoga*, *igreja* e *catedral* podem ser relacionadas à classe semântica *prédio religioso*); quanto informação relativa à estrutura semântica, definindo com quais outros objetos, atributos e ações tal objeto pode co-ocorrer (Buitelaar, 2001).

Dias-da-Silva (2004) apresenta seis definições para o termo tesauro:

- um inventário de itens do vocabulário de uma língua particular;
- um inventário de palavras tematicamente organizadas, isto é, um dicionário onomasiológico;
- um inventário de sinônimos e antônimos;
- um inventário que constitui um índice para a informação armazenada em um computador; uma lista de assuntos

relacionados à informação que deve ser recuperada por meio de palavras-chave;
— um inventário eletrônico, isto é, um arquivo de computador que armazena sinônimos que aparecem para o usuário durante o processo de correção automática;
— um inventário eletrônico de sinônimos e antônimos.

As *ontologias*, tema central desta tese, são de definição ainda mais variada. Como são objeto de estudo de diferentes áreas (filosofia, ciências cognitivas, inteligência artificial, semântica lexical, lexicografia e ciência da informação), é natural que haja uma multiplicidade de acepções que, não coincidentemente, corresponderão a diferentes tipos de ontologia. Brewster et al. (2005) chegam a afirmar que ontologias têm sido vendidas para a comunidade acadêmica como uma "panacéia" (2005:1). O termo pode fazer referência a taxonomias, como as do Yahoo, a bases de dados lexicais, como a WordNet (Fellbaum, 1998) e a construtos logicamente coerentes sobre os quais sistemas de raciocínio podem operar. Brewster e Wilks (2004) sugerem que tanto ontologias como taxonomias e tesauros estão dispostas em um continuum: em um extremo estariam as ontologias completamente explícitas, elaboradas de modo a facilitar o cálculo de inferências lógicas. Em outro extremo, estruturas que se organizam como mapas conceituais, que envolvem algum esforço de interpretação humana para que possam ser consideradas uma representação de conhecimento. Em algum ponto entre esses extremos estão taxonomias e hierarquias navegáveis na Internet, como os diretórios do Yahoo, claramente menos rigorosas do que uma ontologia completamente especificada. Os autores acreditam ainda que essas taxonomias "meio-termo" são, exatamente por não pretenderem total rigor teórico, mais fáceis de serem construídas de forma automática ou semi-automática.

Neste trabalho, utilizo a definição de Hovy (2002), segundo a qual uma ontologia é um conjunto de termos, associados a definições em linguagem natural, que utilizam, se possível, relações formais e restrições, sobre algum domínio de interesse, usado por humanos, bases de dados e programas de computador[1].

No âmbito do PLN, ontologias são úteis em uma série de tarefas. Na recuperação de informação e de documentos, ontologias permitem expansão do

[1] "For generality we define an ontology rather loosely as a set of terms, associated with definitions in natural language (say, English) and, if possible, using formal relations and

termo da busca, tanto por sinônimos quanto por hipônimos. Porém, é preciso considerar que este tipo de expansão, se, por um lado, leva a um aumento no número de documentos recuperados, por outro, leva a um declínio na precisão, isto é, mais documentos irrelevantes são recuperados.

Na sumarização automática, assume-se que frases que possuem palavras diferentes, mas relacionadas por meio de relações de hiperonímia ou sinonímia podem estar relacionadas, contribuindo para o cálculo de relevância de palavras em um texto. Ainda na área de geração de textos, a utilização de termos hiperônimos contribui para a coesão textual e maior fluidez do texto, evitando a repetição de palavras.

A resolução de anáforas é uma tarefa que também se beneficia de uma ontologia. Em um par de sentenças como "*Maria comprou pêssegos lindos. As frutas estavam doces e suculentas*", a relação de hiperonímia entre *pêssego* e *fruta* possibilita uma "compreensão" da sentença.

Atualmente, boa parte das aplicações de PLN que necessita de informação semântica utiliza a WordNet (Fellbaum, 1998). Porém, a WordNet é feita para a língua inglesa e, para o português brasileiro, embora o projeto Wordnet.Br (Dias-da-Silva, 2004) esteja em andamento, os resultados ainda não estão disponíveis para uso (a seção 4.1 trata detalhadamente das wordnets). Além da limitação imediata relativa ao idioma, outras restrições fazem com que o uso da WordNet como ontologia seja visto com ressalvas.

A primeira delas refere-se à presença freqüente de sentidos raros. A WordNet inclui, por exemplo, o sentido de *computador* como "*aquele que computa, que realiza cálculos*" e isso é um problema quando o objetivo é a expansão dos termos de uma busca, por exemplo, já que a expansão de *computador* incluirá sinônimos como *calculista* (Pantel e Ravichandran, 2004).

Outra limitação é a ausência de jargões, de termos específicos de determinadas áreas, bem como a presença esparsa de nomes próprios.

Por fim, e não menos importante: a WordNet é feita manualmente, o que implica um trabalho lento e dependente de vasta mão de obra. E, como o sucesso de um sistema é em grande parte dependente do tamanho da base, conseqüentemente é necessária uma grande equipe de pesquisadores e

---

constraints, about some domain of interest, used in their work by human, data bases, and computer

lexicógrafos para que ela seja efetivamente utilizada. Em conseqüência, sua atualização, um aspecto fundamental se admitimos que o conhecimento que se quer capturar está em constante fluxo, é mais custosa. Além disso, o caráter manual também esbarra nas limitações sofridas por dicionários: as definições estão sujeitas à subjetividade de lexicógrafos; ontologias e taxonomias refletem uma visão particular de mundo – a visão de quem as constrói, mesmo que corroborada por especialistas (Kilgarriff, 2003; Wilks, 2002).

Tendo em vista as restrições apresentadas, tem-se investido recentemente em formas de automatizar o processo de aquisição de informação lexical, desenvolvendo-se metodologias para a construção automática de bases de conhecimento, taxonomias ou ontologias (Hearst, 1992, 1998; Widdows, 2003; Snow et al., 2005; Phillips e Riloff, 2002; Caraballo, 1999; Maedche e Staab, 2000, entre outros), ancorando na informação contida em textos o conhecimento a ser representado (Buitelaar et al., 2005).

Uma ontologia como uma forma de representação do conhecimento é um modelo abstrato do que um indivíduo ou uma comunidade acreditam ser verdadeiro sobre o mundo. Nessa visão, textos seriam a única fonte concreta de informação com relação a esse conhecimento, na medida em que é possível sua análise, manipulação e extração de determinados tipos de informação (Brewster et al., 2005).

Dentre as propostas de construção automática de ontologias a partir de textos, não há investigações voltadas para a língua portuguesa. Este trabalho visa a suprir esta lacuna, apresentando subsídios para a construção automática de uma ontologia específica de domínio que auxilie o desempenho de tarefas de processamento automático de linguagem natural. Para tanto, proponho, seguindo os trabalhos para a língua inglesa desenvolvidos por Marti Hearst (1992, 1998), a extração de relações de hiperonímia em um corpus da área de saúde, por meio da identificação de determinados padrões léxico-sintáticos. Proponho, também, que os resultados obtidos nessa extração sejam cruzados de modo a possibilitar a realização de inferências – aumentando as informações disponíveis na ontologia.

Do ponto de vista teórico, assumo uma postura compatível com uma visão pragmática "radical" do significado, expressa sobretudo nas *Investigações*

---

programs" (Hovy 2002:92).

Filosóficas de Wittgenstein (1953), segundo a qual os significados não existem enquanto entidades autônomas.

Esta perspectiva também é compatível com a investigação do uso da língua em grandes corpora. A utilização de corpus para pesquisas lingüísticas pode ser compreendida tanto como uma metodologia quanto como uma teoria. Esta divisão encontra respaldo na distinção entre lingüística baseada em corpus e lingüística dirigida por corpus, notada em Sinclair (1996, apud Oliveira, 2006), como mostra o quadro 1, retirado de Oliveira (2006).

Lingüística baseada em corpus	Lingüística dirigida por corpus
o corpus é utilizado para validar, verificar e melhorar observações lingüísticas que já tenham sido realizadas	um corpus é de importância essencial no surgimento de novas idéias de como examinar os dados
o lingüista não questiona posições teóricas pré-estabelecidas ou categorias descritivas aceitas; sua posição com respeito à estrutura da língua já se estabilizou	o lingüista acredita que é possível conciliar o tipo de evidência que emerge do corpus com as posições estabelecidas; ele deixa abertas as possibilidades de mudanças radicais na teoria para lidar com as evidências
o corpus é utilizado para ajudar a estender e melhorar a descrição lingüística	a evidência do corpus é soberana, portanto o lingüista minimiza os pressupostos sobre a natureza das categorias teóricas e descritivas
um exemplo de questão relevante: WHOM ainda é utilizado em inglês? Como?	um exemplo de questão relevante: a distinção entre gramática e léxico é necessária?

Quadro 1: Lingüística baseada em corpus vs. Lingüística dirigida por corpus (Oliveira, 2006)

Segundo Oliveira (2006),

> "a distinção entre abordagens baseadas em corpus e dirigidas por corpus se assemelha ao contraste entre as abordagens top-down e bottom-up de resolução de problemas. No primeiro caso, o processo é analítico e os conceitos mais gerais da teoria do problema, suas abstrações de mais alto nível, são utilizadas para iniciar a análise. Os dados são os utilizados em última instância, na confirmação, extensão ou rejeição da teoria. Por outro lado, a abordagem bottom-up inicia-se com os dados e, em processos de síntese, formulam a teoria que abstrai e generaliza a informação inerente aos dados. Na prática da pesquisa lingüística, embora não na teoria, uma mistura das duas metodologias é invariavelmente necessária. No caso de uma pesquisa interdisciplinar, que busca meios lingüísticos de atingir objetivos computacionais, assim como prover meios computacionais para adicionar aos instrumentos de analise lingüística, a convergência das metodologias pode se acentuar"
>
> (Oliveira 2006:16)

Essa "mistura" das duas metodologias a que Oliveira se refere é o que Biber et al. (1998) chamam de abordagem baseada em corpus, uma abordagem que assume a complementaridade dos dois tipos de conhecimento.

Nesta tese, a abordagem baseada em corpus privilegia o trabalho de observação sobre o corpus na busca por determinados padrões léxico-sintáticos – isto é, privilegia o processo de síntese. Por outro lado, é inegável que o próprio *insight* sobre que tipo de padrão buscar, bem como sua formulação lingüística, só foram possíveis, ou melhor, foram bastante facilitados pela intuição da lingüista.

Em suma, a ontologia a ser desenvolvida apresenta as seguintes características:

- é baseada em língua;
- é totalmente baseada em corpus e não em dicionários ou outras bases preexistentes;
- é potencialmente infinita, pois novos termos podem ser constantemente acrescentados;
- é construída automaticamente.

A possibilidade de construção automática evidencia uma grande aproximação entre a metodologia proposta neste trabalho e técnicas da área de Extração de Informação.

A extração de informação (EI) pode ser considerada um tipo de recuperação de informação cujo objetivo é a retirada automática e seletiva de informações de documentos (textos). Trata-se de um processo que tem como entrada uma coleção de textos e que produz como saída dados em formato estruturado, que podem ser utilizados para povoar algum tipo de base de dados.

Vista desse modo, a tarefa de construção automática de ontologias pode ser considerada decorrência de técnicas de EI, pois se busca extrair, do texto, informação estruturada a respeito de determinadas relações entre as palavras. No caso específico dos padrões de hiperonímia apresentados aqui, uma grande vantagem é sua generalidade, que permite sua aplicação em diferentes domínios e gêneros textuais.

1.1. Organização da tese

No capítulo 2, trato dos fundamentos teóricos desta tese. Apresento o ponto de vista adotado para lidar com a questão dos significados e das relações semânticas entre as palavras. Ainda no capítulo 2, discuto as implicações das diferentes perspectivas sobre o significado para o entendimento do que são ontologias e analiso a visão tradicional a respeito de taxonomias e da relação de hiperonímia.

Os capítulos 3 e 4 constituem uma resenha da literatura sobre ontologias. No capítulo 3 apresento critérios formulados na tentativa de padronização do que sejam ontologias, dedicando atenção especial à proposta de Brewster e Wilks (2004), por tratar de ontologias construídas a partir de corpus. Além disso, examino as formas de avaliação que vêm sendo utilizadas na tentativa de aferição do sucesso e de comparação entre ontologias construídas automaticamente. No capítulo 4, descrevo os principais trabalhos que tratam da extração automática de relações de hiperonímia a partir de textos, apresentando de maneira mais detalhada as wordnets (ainda que estas não sejam feitas automaticamente) e o trabalho de Marti Hearst (1992, 1998).

Os capítulos 5 e 6 são o cerne deste trabalho. No capítulo 5, descrevo a metodologia: o corpus e os padrões utilizados na identificação da hiperonímia; e no capítulo 6 apresento os resultados obtidos.

Por fim, no capítulo 7, reflito sobre a proposta inicial – a possibilidade de elaboração automática de ontologia específica de domínio a partir de corpus – à luz dos resultados obtidos e apresento sugestões de trabalhos futuros.