COMO UMA PERSPECTIVA RELATIVISTA PODE AUXILIAR A ELABORAÇÃO AUTOMÁTICA DE ONTOLOGIAS

Maria Cláudia de FREITAS (PUC-Rio)[1]
Violeta QUENTAL (PUC-Rio)[2]

RESUMO: Neste trabalho, propomos a elaboração automática de uma ontologia específica de domínio a partir da identificação de padrões léxico-sintáticos em um córpus. Nosso objetivo é duplo: tanto apresentar a metodologia e resultados quanto argumentar em favor da idéia de que uma abordagem não-representacionista do significado tem a vantagem de lidar com as imprevisibilidades das línguas naturais – compatibilizando a técnica de extração automática de informação a partir de córpus com uma perspectiva teoricamente motivada. Os resultados são encorajadores, e a ontologia resultante caracteriza-se principalmente por não conter categorias pré-definidas, já que categorias são abstrações que refletem uma perspectiva particular do mundo.

ABSTRACT: In this paper, we present the automatic building of a domain-specific ontology based on the identification of lexical-syntactic patterns in a corpus. Our purpose is twofold: to show the method and results and also to argue in favor of a non-representationalist treatment of word meaning – combining corpus-based information extraction techniques with a theoretically motivated approach. The results are encouraging, and the resulting ontology doesn't have pre-defined categories, since categories are abstractions that reflect a particular world view.

1. Introdução

Ferramentas capazes de organizar o crescente volume de informação textual são cada vez mais necessárias. Ontologias aparecem como instrumentos importantes para o desenvolvimento de técnicas de recuperação e extração de informação, pois oferecem a possibilidade de representar de maneira estruturada um dado conhecimento; estruturas hierárquicas são uma forma relativamente simples de organização – e compreensão – da informação.

Com relação à recuperação e extração de informação e de documentos, ontologias permitem buscas expandidas por meio da realização de inferências e permitem a busca por uma determinada categoria semântica. Por exemplo, sobre uma determinada base de dados do domínio economia, seria possível uma busca como "liste todos os bancos". De modo contrário, sem uma representação subjacente do conhecimento, uma busca como "liste todos os presidentes do Brasil" dificilmente teria sucesso segundo a maioria das ferramentas de busca atuais, a não ser que existisse um documento que apresentasse a seqüência de palavras "liste todos os presidentes do Brasil" e que, além disso, também listasse os presidentes do Brasil.

Porém, a elaboração de ontologias é, em geral, um processo lento de construção de consenso (Velardi et *al*. 2005): envolve discussões e tomadas de decisão entre especialistas e, freqüentemente, o uso de uma vasta mão de obra. Por isso, tem-se investido recentemente em formas de automação desse processo, ancorando na informação contida em textos o conhecimento a ser representado em uma ontologia (Buitelaar et *al*.2005).

Neste trabalho, propomos a elaboração automática de ontologia a partir da identificação de padrões léxico-sintáticos em um corpus. Nosso objetivo é duplo: tanto apresentar a metodologia e resultados quanto argumentar em favor da idéia de que um método semanticamente "superficial", lingüisticamente motivado, não é apenas prático, mas apresenta a vantagem de lidar com as imprevisibilidades das línguas naturais – compatibilizando a técnica de extração automática de informação a partir de córpus com uma perspectiva relativista do significado.

O restante do artigo está organizado da seguinte maneira: na seção 2, apresentamos a visão tradicional que sustenta a maioria das investigações sobre ontologias; na seção 3, apresentamos uma outra perspectiva – relativista – e sugerimos como ela pode auxiliar a elaboração automática de ontologias; na seção 4,

[1] claudiaf@let.puc-rio.br
[2] violetaq@let.puc-rio.br

descrevemos brevemente uma metodologia para a elaboração automática de ontologias; na seção 5 apresentamos os resultados obtidos; por fim, na seção 6, tecemos algumas considerações finais.

2. Ontologias e significados – uma visão tradicional

O estudo das ontologias, embora desperte grande interesse no campo da Inteligência Artificial (IA), remonta às origens da filosofia, há cerca de 25 séculos. Mas esta longa tradição não significa que existam respostas satisfatórias aos problemas inicialmente apresentados. O termo, originalmente, designa o estudo do ser, considerado independentemente de suas determinações particulares e naquilo que constitui sua inteligibilidade própria. Trata-se da teoria do ser em geral, da essência do real (Japiassú e Marcondes, 1989). Enquanto teoria do ser, uma ontologia busca descrever as categorias mais básicas da realidade - entidades, tipos de entidades e o relacionamento entre esses elementos.

A investigação sobre as categorias que compõem a realidade começa a receber um tratamento sistematizado com Aristóteles, que apresenta 10 categorias básicas que classificariam tudo o que pode ser dito ou predicado sobre qualquer coisa: substância, quantidade, qualidade, relação, lugar, tempo, posição, estado, atividade e passividade. Deste modo, as categorias expressas pela realidade descreveriam o real – assume-se a existência de um mundo externo à linguagem, passível de descrição. Ontologias devem, portanto, ser gerais e independentes de língua, pois descrevem a realidade, que, por sua vez, é a mesma para todos – e por isso os conceitos são gerais, independentes de língua. Ou seja, nessa abordagem, aos conceitos das ontologias são atribuídos rótulos – as palavras – , que serão dependentes de língua. (De fato, essa é a perspectiva que norteia, até hoje, redes lexicais como as wordnets (Fellbaum,1998; Vossen 1998), que freqüentemente são utilizadas como ontologias):

> In principle, the separation between ontology and lexicon is as follows: “language-neutral" meanings are stored in the former; language-specific information in the latter. (Viegas et al. 1999)

Assume-se, portanto, que o significado – ou conceito – é uma entidade extralingüística; que há uma estabilidade entre significado e (i) representações mentais (em uma perspectiva mentalista); ou (ii) realidade (em uma perspectiva realista). Os estudos na área de semântica lexical, principalmente, irão se embasar nessa visão: os significados das palavras podem ser expressos por formalismos como traços semânticos universais, e dificuldades quanto à delimitação dos significados de uma palavra não são postas em discussão.

É nesse contexto que a IA se apropria do termo ontologia: o crescente reconhecimento de que fontes computacionais devem ser as mais gerais possíveis, reutilizáveis e compartilháveis entre a comunidade de IA foi o primeiro passo para considerar o valor das questões tradicionais da filosofia: o estudo da realidade e seus objetos, independentemente do nosso conhecimento sobre eles, e a busca por uma natureza a priori das coisas (Bateman, 1995). Para Guarino (1995), uma base de conhecimento que se aproximasse à noção filosófica clássica de verdade facilitaria não apenas a interação e comunicação entre diferentes agentes, mas também o compartilhamento e reaproveitamento da própria base. Ainda segundo Bateman (1995), há apenas uma aparente confluência de interesses entre filosofia e IA, no que tange às ontologias. Na IA, o uso do termo remeteria à construção de *frameworks* para “conhecimento” que permitam a sistemas computacionais lidar com problemas tais como processamento de linguagem natural e “*real world reasoning*”. De acordo com essa perspectiva, um sistema deve ser capaz de realizar deduções com relação a algum corpo de informação, e os componentes organizacionais mais gerais desta informação são chamados coletivamente de ontologia. Já Guarino (1995) defende a introdução sistemática de princípios de ontologia formal na engenharia de conhecimento, a fim de explorar as várias relações entre ontologia e representação de conhecimento. Para a área de sistemas de informação, uma ontologia seria uma linguagem formal elaborada para representar um domínio particular de conhecimento, cujo objetivo é, essencialmente, funcional (Zúñiga 2001). Em última análise, a própria discussão sobre o que venha a ser uma ontologia é ilustrativa da dificuldade de se estabelecerem definições e conceitos comuns e compartilháveis entre domínios. Ou seja, a dificuldade em se chegar a um acordo sobre o que são ontologias põe em xeque a própria existência de ontologias nos moldes propostos – uma ontologia geral, multilingüe e, algumas vezes, independente de domínio.

De fato, a elaboração de ontologias sustentadas por representações de conhecimento gerais, independentes de língua, parece ser problemática. O projeto de construção de uma única ontologia, que pudesse ser ao mesmo tempo não-trivial e adaptável para diferentes comunidades de sistemas de informação,

foi em grande parte abandonado; a tarefa se mostrou muito mais difícil que o previsto inicialmente, confirmando os problemas já enfrentados por filósofos há 2000 anos (Smith 2001).

O declínio da crença (e desapontamento) na construção de ontologias gerais, levou, por sua vez, ao investimento em ontologias específicas de um domínio. Neste contexto, uma das definições de ontologia mais difundidas é a de Gruber (1993), segundo a qual uma ontologia é "uma especificação formal explícita de uma conceitualização compartilhada". Como esse compartilhamento é também duvidoso, investe-se hoje em modelos de mapeamento de informações entre ontologias, como forma de reduzir a heterogeneidade entre as diversas representações de domínio e as diversas linguagens utilizadas para essa representação (Kalfoglou e Schorlemmer, 2003).

No âmbito da pesquisa em PLN, ontologias podem ser vistas como "modelos de domínios específicos", que têm como objetivo facilitar buscas semânticas (Brewster e Wilks, 2004).

3. Ontologias e significados – uma visão relativista

Paralelamente à visão tradicional, desenvolve-se, na filosofia, uma outra abordagem, relativista, não-representacionista, cujo embrião pode ser encontrado já no pensamento sofista, e que sustenta não existir uma realidade independente e exterior à linguagem e, portanto, passível de descrição. Segundo essa perspectiva pragmática radical, a própria empreitada ontológica perde o sentido – isto é, não se trata de uma tarefa difícil, mas de uma tarefa sem sentido: não há conceitos independentes de língua que descrevem o universo (ou parte dele) – em última análise, não há universo a ser descrito independente de língua. O estabelecimento de verdades universalmente válidas, autônomas com relação às circunstâncias concretas é impossível (Martins, 2004). Somos constituídos pela linguagem, o que impossibilita a realização de julgamentos sobre ela. Ontologias gerais, aproximações às noções de verdade, não são questões que devem ser consideradas.

Mas, se não há "entidades mentais" ou realidade às quais as palavras se "colam", e que corresponderiam ao significado das palavras, o que é o significado então? A visão não-representacionista de Wittgenstein, expressa principalmente nas *Investigações Filosóficas* (1953), é de grande valia para lidar com o terreno movediço do significado – intimamente relacionado à questão da elaboração de ontologias. Os significados correspondem aos usos culturalmente determinados que fazemos das palavras – o significado não é uma entidade, ele está no uso (Martins, 2004).

Nessa perspectiva, a dificuldade em se responder à pergunta *o que é o significado* se deve à natureza equivocada da pergunta. A conexão – pretensamente estável – entre linguagem e realidade é forjada, na medida em que a própria linguagem constitui a realidade (Hacker e Backer, 1984). A linguagem diz as opiniões dos homens, e por isso sua imprevisibilidade não é um desvio, mas conseqüência dessas opiniões ou impressões, que são naturalmente contraditórias (Martins, 2004).

Porém, assumir a inadequação da questão *o que é significado* não significa a defesa de uma posição reducionista segundo a qual significados não existem. Eles existem, mas não como entidades autônomas, e não com a precisão ou os limites definidos, necessários à formalização que sempre se buscou fazer. O significado é flexível e maleável, não cabe no molde fixo que lhe desejam impor. Esta recusa dos significados a uma formalização exaustiva pode ser uma forte limitação para as semânticas formais, mas, por outro lado, pode representar uma motivação para outras formas de lidar com o significado, como sugere Wittgenstein. O significado não é uma propriedade imanente à palavra, mas uma função que expressões lingüísticas exercem em um contexto específico e com objetivos específicos (Marcondes, 2005). Com isso, o significado de uma palavra pode variar conforme o contexto em que é utilizado, conforme o objetivo desse uso.

Se não há uma essência única e fixa do significado, como lidar com as definições? Dicionários não só existem como são úteis. Negar esse fato parece um contra-senso. Porém, o que Wittgenstein enfatiza é o caráter parcial e incompleto das definições – que nem por isso as torna menos úteis. Desse modo, se, em uma perspectiva essencialista, esbarraríamos, em algum momento, nos "indefiníveis" – traços ou universais como "humano" ou "masculino", que compreendemos sem dificuldade – Wittgenstein argumenta que as definições são sempre fundamentadas em um conhecimento prévio, derivado do uso (do contexto, da situação de explicação, de inúmeros outros fatores). Isto é, definições, embora úteis nos contextos em que são utilizadas, são sempre parciais.

Definições analíticas, que analisam termos com base em uma conjunção de marcas características, deixam de ser encaradas como "as" definições por excelência: trata-se apenas de mais uma forma de definição, dentre outras possíveis (Glock, 1997).

Nesse sentido, a incompletude inerente às definições é uma faceta da ausência de um ideal de exatidão, tanto no que se refere às explicações, quanto aos limites entre os diferentes significados – diferentes usos – de um termo. Precisão e exatidão, novamente, são relativos. Não há um padrão único de exatidão; a precisão é uma questão de adequação às circunstâncias e aos propósitos.

> "É inexato se eu não indicar a distância que nos separa até o sol até exatamente 1 m? E se eu não indicar ao marceneiro a largura da mesa até 0,001 mm?
> **Um** ideal de exatidão não está previsto; não sabemos o que devemos nos representar com isso – a menos que você mesmo estabeleça o que deve ser assim chamado. Mas ser lhe á difícil encontrar tal determinação; uma que o satisfaça." (Investigações Filosóficas, § 88)

A língua é naturalmente vaga, imprevisível e ambígua, e grande parte de sua robustez se deve justamente a isso. Nem todos os conceitos, porém, são realmente vagos, e, embora a maior parte dos conceitos empíricos admita casos fronteiriços, nem por isso se tornam inúteis (Glock, 1997).

> "It is precisely the lack of clarity in our use of the word culture which makes it such a handy word to have at one's disposal." (Stock 1983, apud Kilgarriff 1997: 39)

E no que as considerações de Wittgenstein podem ser úteis à semântica computacional, à elaboração de ontologias?

Na IA, como já mencionado, a ambição inicial de ontologias gerais foi substituída pela idéia de ontologias de domínio (o que não deixa de ser um exemplo curioso de como os limites entre conceitos são imprecisos, mas nem por isso pouco úteis). Além da redução no escopo da tarefa, a constatação de que a elaboração de ontologias exige um processo longo de construção de concordância entre um número grande de especialistas, levou à pesquisa sobre formas de automação desse processo, tomando como princípio que o conhecimento a ser representado na ontologia deve ser a informação contida em textos (Buitelaar et al., 2005). Adotando uma perspectiva relativista, na qual a linguagem constitui a realidade, é difícil pensar em ontologias baseadas em conceitos pré-definidos. Por outro lado, é igualmente difícil transpor a "linguagem enquanto prática de vida" para um ambiente computacional. Diante desse impasse, o que propomos é a substituição (grosseira, é verdade) de "práticas de vida" pelo córpus – assumimos que o conhecimento disponível em textos, expresso em linguagem natural, pode funcionar como uma fonte confiável para a busca de informações e categorizações.

Conseqüentemente, a principal característica da ontologia proposta é a ausência de categorias pré-definidas. Categorias em uma taxonomia são construtos humanos, abstrações que refletem uma perspectiva particular do mundo (Kilgarriff 2003, 1997; Brewster e Wilks, 2004). A idéia de sustentar a ontologia em córpus busca deslocar o espaço de discussão sobre quais seriam as categorias relevantes de um domínio: mudamos o foco do desejado consenso entre especialistas para as categorias que emergem do córpus, que, por sua vez, refletiriam o conhecimento implícito do domínio em questão.

4. A elaboração da ontologia

Os principais obstáculos na construção de ontologias referem-se a (i) como definir o conhecimento a ser codificado, isto é, quais informações e/ou categorias devem estar presentes na sua aquisição e (ii) como manter e atualizar este conhecimento, tendo em vista que ontologias são úteis na medida em que contêm informação suficiente e relevante para representar o domínio em questão, e que este conhecimento, por princípio, está em constante desenvolvimento, mesmo que esse desenvolvimento se dê apenas por novas formas de análise do domínio.

Com o objetivo de enfrentar estas dificuldades, e sustentadas por uma visão relativista do significado, apresentamos aqui alguns resultados decorrentes de um extrator automático de relações de hiponímia, baseado em expressões regulares e em padrões léxico-sintáticos. A metodologia não é nova – a extração de relações taxonômicas a partir de padrões lexicais foi proposta originalmente em Hearst (1992), e desde então vem sendo amplamente utilizada. Nossa contribuição está no desenvolvimento de alguns ajustes para sua aplicação para a língua portuguesa, e, principalmente, na incorporação de novos padrões de expressão de hiperonímia e na possibilidade de cruzamento dos dados obtidos para a realização de inferências.

4.1. Metodologia:

Utilizamos um córpus de 1.846.502 palavras, composto por textos da área de saúde pública disponíveis na Internet. Para a aplicação dos algoritmos de identificação de padrões sobre o córpus, é necessário que ele já tenha passado pelas seguintes etapas:

a) Etiquetagem morfossintática: é fundamental que o córpus contenha etiquetas de classes gramaticais (POS tags). Para isso, o córpus foi processado pelo etiquetador automático do parser PALAVRAS (Bick, 2000)

b) Etiquetagem de Sintagmas Nominais: já com as etiquetas de classes gramaticais, o córpus foi analisado por um etiquetador automático de Sintagmas Nominais (Santos & Oliveira, 2005)

c) Revisão manual, a fim de minimizar, principalmente, erros decorrentes da identificação / segmentação de nomes próprios e do reconhecimento de expressões multivocabulares verbais e nominais.

Para a elaboração de uma ontologia, a primeira etapa é a indicação das relações semânticas desejadas. Como, até o momento, só há extração de relação de hiponímia, o resultado final tem a forma de uma taxonomia – taxonomias, por sua vez, podem ser compreendidas como porções de uma ontologia. Uma das grandes vantagens da identificação das relações de hiponímia é a sua característica transitiva: se A é maior que B, e B é maior que C, então A é maior que C. Isto é, é a transitividade que permite a realização de inferências, que, por sua vez, acarretam a produção de conhecimento novo, derivado do corpus. Assim, se extraímos do córpus duas relações distintas: (i) um cachorro é um animal ; (ii) Rex é um cachorro, conseguimos, por meio de inferência, obter (iii) Rex é um animal, informação que não estava explicitada no córpus.

Na área de processamento automático de linguagem natural, o trabalho de Hearst (1992) é pioneiro na identificação de padrões lexicais capazes de expressar, sistematicamente, determinadas relações semânticas. Especificamente, Hearst (1992, 1998) apresenta seis pistas textuais para a extração da relação de hiponímia/hiperonímia:

(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 é o sintagma nominal (SN) hiperônimo e NP_1, NP_2, NPi são os SN hipônimos, isto é, SNo>SNi

Desde então, a maioria dos trabalhos continua a explorar essas pistas, acrescentando novas técnicas na maneira de utilizá-los – principalmente técnicas de aprendizado de máquina. Nenhum destes trabalhos, porém, está voltado para o português. Para a elaboração da ontologia, utilizamos as pistas (iii) e (iv) de Hearst, que podem ser transformadas em apenas uma, acoplando-se as relações de conjunção e disjunção:

SN_1 { ,SN_i } * { , } e|ou outros(as) SN_0

E a pista (i)

SN_0 tais como SN_1 { , SN_2 ... ,} (e | ou) SN_i

Porém, a observação do córpus em português mostrou que, para que o padrão (i) revele uma quantidade significativa de relações de hiperonímia no português, é preciso considerar tanto “tais como” quanto a sua variante “como”. Isto é, “such as” pode ser literalmente traduzido para “tais como” mas, na língua portuguesa, freqüentemente apenas o “como” é utilizado neste tipo de construção:

A tentativa posterior de clonar outros mamíferos **tais como** camundongos, porcos, bezerros, ...
A tentativa posterior de clonar outros mamíferos **como** camundongos, porcos, bezerros,...

Contudo, essa inclusão é, ao mesmo tempo, um complicador: "como" é uma palavra que se enquadra em diferentes classes gramaticais, dificultando o trabalho dos etiquetadores automáticos e, conseqüentemente, acarretando problemas na identificação do padrão desejado. O "como" capaz de expressar a relação de hiponímia é um "como" que pode ser acrescido de "por exemplo", e cuja classe gramatical é "palavra denotativa" (Oliveira e Freitas, 2006):

A tentativa posterior de clonar outros mamíferos **como por exemplo** camundongos, porcos , bezerros...

De fato, o córpus utilizado contém cerca de 2700 ocorrências de "como" palavra denotativa, contra apenas 232 ocorrências de "tais como", o que mostra o ganho na alteração do padrão original.

Além disso, acrescentamos mais dois novos padrões, que também expressam sistematicamente uma relação de inclusão de classe:

tipos de SN_0: SN_1 { , SN_2 ... ,} (e | ou) SNi
SN_0 **chamado/s/a/as** (de **)** SN_1

5. Resultados

Uma vez assumida a perspectiva relativista para tratar do significado e das relações de significado entre as palavras, a questão mais pertinente diz respeito a como proceder a uma análise dos resultados ou, dada a metodologia proposta, o que deve ser considerado erro. Isto porque partimos do pressuposto de que os padrões investigados expressam, de fato, relações de hiponímia, ainda que tais relações não sejam relações "convencionais" de um ponto de vista lexicográfico. Desse modo, consideramos corretas relações extraídas do corpus a partir dessas regras, como

sensibilidade<condição
socos<traumas
reforma de um jardim<trabalhos voluntários

Consideramos erros casos em que: a relação extraída não estava correta devido à ambigüidade do sintagma preposicionado (a); presença de uma estrutura adverbial deslocada da ordem direta ou encaixada (b) e (c); elipse de algum termo (d); presença de uma oração no interior do sintagma hiperônimo ou hipônimo(e).

(a) (...) *transmissão de o HBV vivo* ***e outros*** *patógenos de transmissão sangüínea*

(b) (...) mesmo em *países de prevalência relativamente baixa* **como** *a China, as taxas em algumas cidades* chegam a quase 20%.

(c) *as inundações aumentam os riscos de aquisição de doenças infecciosas transmitidas por água contaminada, através de contato ou ingestão,* ***como*** *leptospirose, hepatite A, hepatite E, ...*

(d)(...) *e atender a um amplo número de indivíduos ,* ***como*** *grupos comunitários e de trabalhadores , estudantes , grupos étnicos isolados ou centros religiosos*

(e)...*preparações que não atinjam esta concentração energética mínima,* ***tais como*** *sopas e mingaus...*

Ou seja, nessa etapa, consideramos erros principalmente os padrões extraídos que correspondem a uma estrutura sintática diferente da estrutura-alvo ou em que peculiaridades sintáticas contribuem para um desvio do padrão-alvo, já que, em termos semânticos, assumimos que os padrões expressam as relações desejadas, ainda que de uma forma pouco convencional. Com esses critérios, fizemos uma avaliação manual dos resultados (tabela 1).

Padrão	Quantidade de Relações	Acertos
como/tais como	2428	1824 (75%)
e outros	394	321 (81.4%)
tipos de	21	18 (85%)
chamado	89	78 (87.6%)
TOTAL	2932	2241 (76.4%)

Tabela 1: avaliação manual dos resultados considerando o critério sintático

Porém, embora coerente com o ponto de vista teórico assumido, o critério de erro utilizado é pouco útil em dois aspectos importantes:

a)comparação de resultados: não há como comparar nossos resultados com os apresentados em outros trabalhos (Hearst 1998; Widdows e Dorow 2003; Snow et *al.* 2005);

b)funcionalidade das relações extraídas: concordamos com outros autores que uma relação como *doença<fator*, embora correta, pode ser pouco significativa e pode, portanto, ser eliminada sem prejuízo (ou com um prejuízo mínimo) de informação.

Assim, com o objetivo de tornar nossos resultados "mais comparáveis" e "mais significativos", passamos à segunda etapa da avaliação, que consistiu na verificação de uma amostra dos resultados considerados "corretos" por avaliadores humanos.

5.1 Validação

Das 2241 relações corretamente extraídas – assumindo o critério puramente sintático –, selecionamos uma amostra de 436 relações para avaliação humana. Numa pequena adaptação dos processos de validação utilizados por Hearst (1998) e Cederberg & Widdows (2003), foi pedido aos avaliadores que pontuassem as relações obedecendo aos seguintes critérios:

3	a relação está correta da forma como foi extraída;
2	a relação está "um pouco" correta, isto é, o substantivo núcleo está correto, mas preposições, adjetivos etc que o acompanham deixam a relação estranha;.
1	a relação está correta em termos gerais; isto é, é muito geral ou muito específica para ser útil;
0	a relação está errada.

Porém, os critérios acima, se por um lado pretendem oferecer alguma objetividade à tarefa de avaliação, por outro não têm como assegurar a objetividade pretendida. Freqüentemente é difícil distinguir entre uma "*relação correta*" (3) e uma "*muito específica para ser útil*" (1). De fato, grande parte da dificuldade da tarefa está justamente em determinar o que é o "ser útil". Relações como as abaixo, estão corretas ou são muito específicas – e pouco úteis?

Superposição de tarefas<características de a organização do trabalho
Reavaliação do uso de anti-retrovirais<formas de recaptação do paciente

Além disso, no momento da validação, freqüentemente o senso comum difere do conhecimento enciclopédico, e então ocorrem divergências entre os avaliadores.

Por exemplo, do ponto de vista do senso comum, "cereais" podem ser um grupo alimentar; porém, do ponto de vista enciclopédico, isto é, do conhecimento científico, "fibras" são um grupo alimentar, e não cereais. Qual deve ser o critério? A instrução dada aos avaliadores para que determinada relação fosse considerada correta é que a relação fosse verdadeira em algum mundo possível, isto é, existe pelo menos 1 circunstância em que a relação pode ser verdadeira. Com isso, "cereais" foi aceito como "grupo alimentar".

Os resultados da avaliação humana estão na tabela 2, abaixo:

Classificação	Qtd de relações	Exemplos
3 -	(73.4%)	superóxido dismutase<enzimas suco<bebidas
2 -	(3.4%)	sofrimento<sentimentos inerentes à condição psicólogos<agentes da equipe
1 -	(16%)	proteção<valores queima de neurônios<comprometimentos
0 -	(7.1%)	setor público<serviços soco<traumas

Tabela 2: Resultados da avaliação humana

Os resultados da avaliação indicam que a maior parte dos erros está na categoria 1, sendo decorrência de definições gerais demais ou específicas demais – e, conseqüentemente, pouco úteis. É o caso de relações cujo hiperônimo é um substantivo do tipo "fator", "termo" "elemento", "questão", aspecto", entre outros. Tais hiperônimos se enquadram na lista de substantivos genéricos descritos em Marques (1995), e de substantivos-suporte descritos em Oliveira (2006): trata-se de substantivos com um alto grau de generalidade ou falta de especificidade, independentes de contexto temático.

De modo a eliminar tais relações gerais demais e pouco informativas, aplicamos um filtro para eliminar as relações cujo hiperônimo era um substantivo genérico – suporte.

Para diminuirmos os erros da categoria 2, relativos principalmente à "dependência contextual" de algumas relações, aplicamos mais dois outros filtros: um para eliminação de pronomes dêiticos e outro para eliminação de alguns adjetivos que Hearst chama de adjetivos comparativos, como *importante* e *menor*.

Após a aplicação dos filtros, o número de relações extraídas caiu de 2241 para 1937. Dessas, separamos 430 para serem avaliadas manualmente. Os novos resultados estão na tabela 3.

A comparação dos resultados antes e depois da aplicação de filtros (tabela 3) mostra que a eliminação dos substantivos e adjetivos genéricos aumentou em 7% a precisão dos resultados da categoria 3 (corretos da maneira como foram extraídos) – e um conseqüente declínio das relações classificadas como 1, isto é, muito gerais ou mito específicas. Houve também uma pequena melhora nas relações classificadas como 2 (núcleo do SN correto).

Classificação	Qtd de relações **com** filtro	Qtd de relações **sem** filtro
3 -	**349 (81%)**	320 (73.4%)
2 -	**28 (6.5%)**	15 (3.4%)
1 -	**20 (4.6%)**	70 (16%)
0 -	**33 (7.6%)**	31 (7.1%)

Tabela 3: Resultados da avaliação humana após a aplicação de filtros

A comparação com os resultados obtidos em outros trabalhos demonstra que a metodologia empregada, lingüisticamente motivada, embora simples, foi bastante eficaz. Porém, é importante ter em mente que o alto grau de subjetividade da tarefa de avaliação pode comprometer um pouco a comparação.

Percebemos, por exemplo, que alguns substantivos, embora não se encaixassem classes de genéricos e/ou suporte, também deveriam ser eliminados, por seu caráter transitivo:

X < concorrente;
X < adversário
X < parceiro

Tais relações foram consideradas categoria 1, isto é, relações muito gerais para serem úteis. Já em Hearst (1998), a relação

Nippon < partner

foi considerada uma relação útil, com o que não estamos de acordo. E assim voltamos à fragilidade da forma de validação empregada, com o julgamento humano. Outras relações que apareceram no córpus também são de julgamento difícil e subjetivo, como

avião < peça feita com dobradura
alça de sutiã < lingerie,

que foram classificadas como 1 e 0, respectivamente, confirmando nossa opção por uma validação "conservadora".Por outro lado, vale ressaltar que, com o objetivo de caracterização do domínio, as classificações "útil" e "pouco útil" perdem seu valor. Pode ser interessante, sim, deixar categorias gerais como *problemas*, *fatores* etc, como níveis mais altos na ontologia.

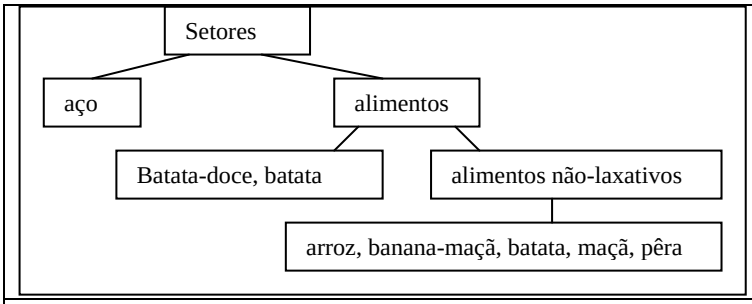
5.2. Produzindo conhecimento: a realização de inferências.

A maioria dos trabalhos que envolvem a extração de relações de hiponímia não utiliza os resultados obtidos para a realização de inferências. Acreditamos que o principal motivo é a grande quantidade de erros produzidos, principalmente quando se trata de relações extraídas de corpus gerais quanto ao domínio, como é o caso de córpus de textos jornalísticos. Kilgarriff (2003) se opõe à utilização de tesauros baseados em palavras (com relações extraídas diretamente do córpus) como ontologias na IA justamente por ser a realização de inferências baseada em conceito, e não em termo. Compartilhando uma perspectiva relativista com relação ao significado, Kilgarriff é consciente das imprecisões dos significados das palavras – dos conceitos –, e por isso argumenta que inferências são um problema para trabalhos baseados em corpus. Um exemplo: em uma ontologia baseada em córpus – e em palavras – , teríamos que *tucanos* são *aves*. Poderíamos encontrar, também, que alguns *políticos* são *tucanos*, mas não gostaríamos de inferir que alguns *políticos* são *aves*. Porém, em favor de uma ontologia baseada em palavras, argumentamos que o fato de nos apoiarmos em um córpus específico de domínio deve evitar a ocorrência de situações como essa. Para tanto, invocamos a restrição "*one sense per discourse*" (Yarowsky, 1995), segundo a qual o significado de uma dada palavra é altamente consistente em um determinado texto. Como estamos lidando com um córpus específico de domínio, esperamos que a restrição possa ser ampliada de "texto" para "domínio".

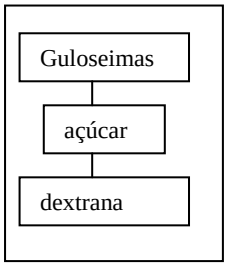
Com o cruzamento das informações obtidas na extração dos padrões léxico sintáticos, foram encontradas 420 taxonomias. Dessas, selecionamos 140 para avaliação manual. Surpreendentemente, encontramos erros em apenas 14, o que significa um total de 90% de acertos, o que contradiz a posição de Kilgarriff. Por outro lado, esse alto índice de acertos se deve, em grande parte, à utilização de um domínio restrito e técnico, o que dá pouca margem à ocorrência de variações entre os significados. De fato, como já assinala Cruse (1986), o vocabulário científico é mais preciso que o vocabulário cotidiano. Mas nem por isso deixaram de ocorrer casos em que a vagueza levou a inferências problemáticas, como mostram os quadros 1 e 2.

No exemplo do quadro 1, o problema da inferência está em *alimentos*, que pode ser visto tanto como um *setor*, assim como *roupas*, *transportes*, etc., quanto como "*substância digerível*". Em conseqüência, são realizadas as inferências "banana é um setor", "maçã é um setor" etc., que vão contra o senso-comum. É importante notar, contudo, que a tarefa de avaliação de inferências também é altamente subjetiva. Imaginemos, por exemplo, que, dentre as categorias de alimentos, aparecesse *frutas*, o que estaria correto. Com isso, a inferência produzida seria "frutas é um setor", o que, assumindo o sentido de *setor* como hiperônimo de *aço*, *vestuário* etc., estaria errado. Porém, não é difícil perceber, em um domínio de *supermercados*, por exemplo, a classificação de *frutas* como *setor* . Qual a solução? Devemos considerar a inferência correta neste caso, pois representa um significado possível na língua, ou devemos considerá-la um erro, pois não é neste sentido que o termo está sendo tratado?

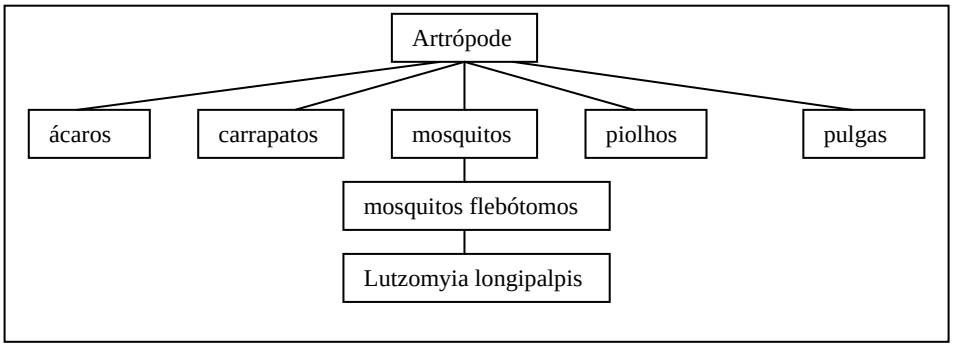
Já o exemplo do quadro 2 é uma clara evidência de diferença quanto aos registros utilizados – do lado técnico, *dextrana* é um tipo de açúcar; do lado da linguagem ordinária, açúcar é uma guloseima, mas, embora seja um açúcar, *dextrana* dificilmente seria aceito como uma *guloseima*. Embora o córpus seja de um domínio técnico, ele também possui textos de divulgação, o que justifica este tipo de ocorrência. Aliás, é justamente a presença de textos não tão técnicos no córpus que possibilita grande parte dos acertos, como mostra o exemplo do quadro 3. A relação entre *mosquitos flebótomos* e *artrópodes* dificilmente seria explicitada em algum texto, pois estão em níveis diferentes de especialidade. Já o exemplo do quadro 4 ilustra como a taxonomia produzida automaticamente a partir de córpus pode ser rica.



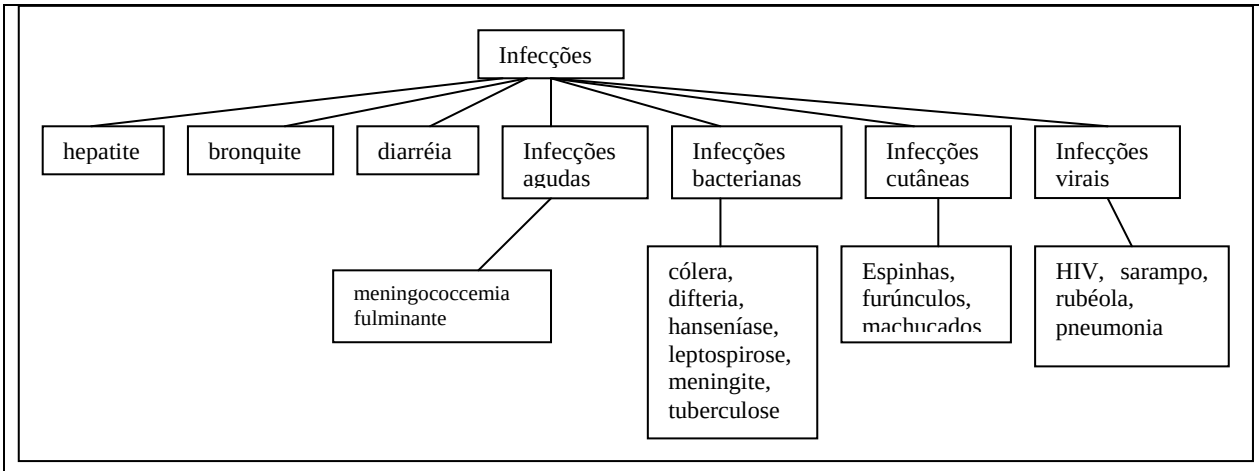
Quadro 1: Vagueza - exemplo de taxonomia para o termo “setores”



Quadro 2: Vagueza - exemplo de taxonomia para “guloseimas”

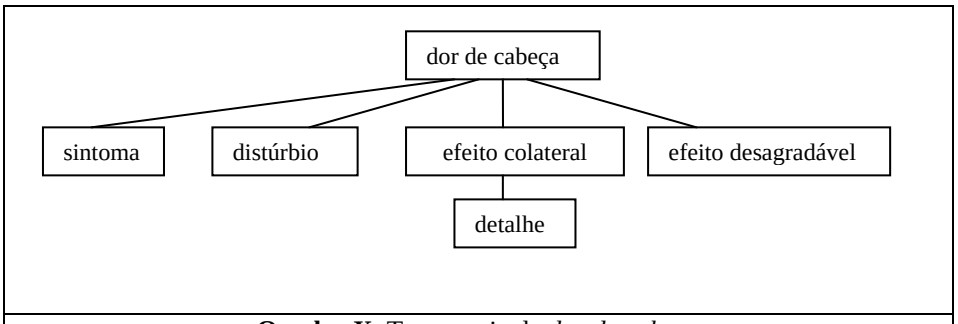


Quadro 3: Taxonomia de *artrópode*



Quadro 4: Taxonomia de *infecções*

Por fim, o cruzamento dos dados para a inferência acabou possibilitando a realização de heranças múltiplas, característica que diz respeito à localização de um termo em múltiplas posições na taxonomia. Freqüentemente, um único termo apresenta uma variedade de facetas, o que justifica sua localização múltipla. O exemplo do quadro 5 ilustra o fato com a taxonomia de *dor de cabeça*.



Quadro X: Taxonomia de *dor de cabeça*

6. Considerações Finais

Apresentamos aqui resultados de uma metodologia para a construção automática de ontologias. Acreditamos que nossa contribuição está (i) na identificação de novos padrões especificamente para o português; (ii) na possibilidade de cruzamento das informações extraídas com os padrões, gerando inferências; e (iii) na perspectiva teórica adotada, que tem como conseqüência principalmente a análise do que poderia ser considerado "erro". Uma perspectiva relativista se mostra produtiva na medida em que aceita os dados vindos do córpus e as relações de significado que nele aparecem.

Os resultados obtidos, porém, demonstram que freqüentemente nem todas as relações possíveis serão explicitadas na ontologia, indicando a necessidade de um trabalho humano complementar. Não há, por exemplo, nos nossos 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, é possível que em um córpus maior a dificuldade seja minimizada.

Acreditamos também que a construção automática a partir de grandes corpora é interessante tanto por minimizar a preocupação com o conhecimento a ser codificado, visto que esse conhecimento estaria no córpus, 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 córpus a "responsabilidade" de direcionar a construção da ontologia.

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 córpus poder ser constantemente atualizado faz com que esteja menos sujeito a falhas. Suas principais desvantagens são a alta dependência de um córpus etiquetado e a dificuldade de avaliação sistemática (e de comparação) dos resultados.

7. Referências Bibliográficas

BATEMAN, J. On the relationship between ontology construction and natural language: a socio-semiotic view. *International Journal of Human-Computer Studies*, 43, 929–944, 1995.

BICK, E. *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework*. PhD thesis, Aarhus: Aarhus University, 2000.

BREWSTER, C. e WILKS, Y. Ontologies, Taxonomies, Thesauri: Learning from Texts. In *Proceedings The Use of Computational Linguistics in the Extraction of Keyword Information from Digital Library Content Workshop*, Kings College, London, UK, 2004.

BUITELAAR, P., CIMIANO, P. e MAGNINI, B. (Eds.) *Ontology Learning from Text: Methods, Applications and Evaluation*. IOS Press, 2005.

CEDERBERG, S. e WIDDOWS, D. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Proceedings of the Seventh Conference on Natural Language Learning At HLT-NAACL 2003 - Volume 4* (Edmonton, Canada). Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, 111-118, 2003.

DOROW, D. e WIDDOWS, D. Discovering Corpus-Specific Word Senses. *EACL*, 79-82, 2003.

FELLBAUM, C. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

GLOCK, H.J. *Dicionário Wittgenstein*. Rio de Janeiro: Jorge Zahar, 1997.

GRUBER, T. A translation approach to portable ontology specifications. In *Knowledge Acquisition* (2):199-220, 1993.

GUARINO, N. Formal Ontology, Conceptual Analysis and Knowledge Representation. *International Journal of Human-Computer Studies*, 43, 625–640, 1995.

BACKER, G. e HACKER, P. *An analytical Commentary on Wittgenstein's Philosophical Investigations.* Volume 1. Oxford: Blackwell, 1984.

HEARST, M. Automated Discovery of WordNet Relations. In FELLBAUM, C (org) *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

_____. Automatic Acquisition of Hyponyms from Large Text Corpora. *Proc. of the Fourteenth International Conference on Computational Linguistics*, Nantes, France, 1992.

JAPIASSÚ, H. e MARCONDES, D. *Dicionário Básico de Filosofia*. Rio de Janeiro, Jorge Zahar, 2001.
KALFOGLOU, Y. e SCHORLEMMER, M. Ontology Mapping: The State of the Art. *The Knowledge Engineering Review Journal*, vol. 18:1, 1-31. Cambridge University Press, 2003.
KILGARRIFF, A. I don't believe in word senses. *Computers and the Humanities* 31 (2), pp 91-113, 1997.

_____. Thesauruses for Natural Language Processing. *Proceedings of NLP-KE*, Beijing, China, pp.5-1, 2003.

MARCONDES, D. *Pragmática*. Rio de Janeiro: Jorge Zahar, 2005.

MARTINS, H. Três Caminhos na Filosofia da Linguagem. In MUSSALIM, F; BENTES, A.C. (orgs.). Introdução à Lingüística. Volume III, São Paulo: Cortez Editora, p. 439-474, 2004.

OLIVEIRA, C. e FREITAS, M.C. Classes de palavras e etiquetagem na Lingüística Computacional. Artigo Submetido, 2006.
SANTOS, C.N., OLIVEIRA, C.: Aplicação de aprendizado baseado em transformações na identificação de sintagmas nominais. In: Anais do XXV Congresso da Sociedade Brasileira de Computação, Brasil, 2005.

SMITH, B. *Ontology and Information Systems*. Disponível em ontology.buffalo.edu/ontology(PIC).pdf. Acessado em 22/11/2006.

SNOW, R., JURAFSKY, D. e NG, A. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems* 17, 2004.

VELARDI, P., NAVIGLI, R., CUCHIARELLI, A. e NERI, F.. Evaluation of ontolearn, a methodology for automatic population of domain ontologies. In Buitelaar, P., Cimiano, P. e Magnini, B. (orgs). *Ontology Learning from Text: Methods, Applications and Evaluation.* IOS Press, 2005.

VIEGAS, E., MAHESH, K., NIRENBURG, S. e BEALE, S. Semantics in Action. In Saint-Dizier, P. (org.). *Predicative Forms in Natural Language and in Lexical Knowledge Bases.* Dordrecht-Boston: Kluwer, 171-203, 1999.

VOSSEN, P. Ontologies. In MITKOV, Ruslan (org). *The Oxford handbook of computational linguistic*. Oxford: Oxford University Press, 2003.

WIDDOWS. D. Unsupervised methods for developing taxonomies using syntactic and statistical information. In *HLT-NAACL*, Edmonton, Canadá, 2003.

WITTGENSTEIN, L. *Investigações Filosóficas.* Abril Cultural,São Paulo, 1953

YAROWSKY D. Unsupervised Word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA, pp 189-196, 1995.

ZÚNIGA, G.L. Ontology: Its transformation from philosophy to information systems. *Proceedings of the Second International Conference (FOIS '01).*Ogunquit, Maine, New York: ACM Press, 187-197, 2001.