3

Crterios para a elaborao e avaliao de ontologias

A dificuldade em se chegar a um consenso sobre o que seja uma ontologia tem reflexos tanto no que concerne aos crterios metodolgicos que devem ser observados no momento da sua elaborao, quanto aos crterios de avaliao e comparao das ontologias construdas.

Ontologias so objeto de diferentes reas e, com isso, apresentam, freqentemente, diferentes objetivos. E, como afirma Gruber, o importante para que a ontologia serve – o que importa sua funo, e, portanto, diferentes diretrizes iro guiar sua elaborao.

Nesta seo, descrevo duas propostas gerais para a elaborao de ontologias, que correspondem a duas vises distintas sobre o problema: a proposta de metodologia de Hovy (2005), compatvel com o que chamei no captulo 2 de viso tradicional de ontologias e significados, e a proposta de metodologia de Brewster e Wilks (2004), compatvel com o que chamei de perspectiva relativista. Em ambas as propostas, trata-se de metodologias teoricamente motivadas que devem direcionar o “construtor de ontologias” e facilitar a consistncia e a exatido, em todas as etapas da elaborao.

3.1.1. Crterios para a elaborao de ontologias “tradicionais”

Com base na experincia prpria de coordenador de projetos relacionados ontologias, Hovy (2005) distingue cinco diferentes motivaes que iro se refletir em diferentes abordagens para a elaborao de ontologias: abordagem da filosofia, da cincia cognitiva, da lingstica, da inteligncia artificial e da computao. Assumindo ser a principal tarefa do “ontologista” a discusso sobre a criao de um novo termo e sua localizao na ontologia com relao aos outros termos j existentes, para uma posterior especificao adicional e definio, Hovy descreve como todo este processo decisrio poderia ocorrer, levando em considerao a

“personalidade” do ontologista – se filósofo, se cientista cognitivo, se lingüista, etc.

Especificamente, a proposta de metodologia de Hovy (2005), que pode ser chamada de *refinamento gradual contínuo*, é composta por 7 etapas:

- 1) Determinação das características gerais da ontologia que se quer construir: domínio, objetivo, nível desejado de granularidade, antecedentes teóricos e conceituais, público-alvo etc;
- 2) Coleta de todas as fontes de conhecimento adicionais, incluindo ontologias anteriores, estruturas de alto nível, glossários de termos, algoritmos e ferramentas, descrições teóricas já existentes, etc;
- 3) Delimitação do principal fenômeno sob consideração: identificação dos conceitos nucleares, tipos ou características permitidos, etc. Começar com uma ontologia de alto nível, já existente, pode ser útil;
- 4) Listagem de todos os termos/conceitos importantes para a tarefa. Os termos podem ser derivados de algum modelo de (meta)dados, de algum algoritmo do sistema, de relatórios de especialistas, etc;
- 5) Registro explícito, para cada conceito, dos princípios e fatores que justificam a sua criação (ainda que a definição seja incompleta ou informal, desde que contenha as principais características de interesse). Identificação e definição das relações e dos conceitos;
- 6) Inspeção da ontologia inicial, buscando corrigir irregularidades, desequilíbrios, etc;
- 7) caracterização da ontologia quando pronta, registrando seus parâmetros essenciais (cf. descrição detalhada em Hovy, 2002).

Como o próprio Hovy assume, a utilização dessa metodologia leva tempo e exige esforço. Nem todos os aspectos se aplicam a todos os casos, e nem a todos os domínios ou “estilos de ontologização”.

3.1.2. Critérios para a elaboração de ontologias baseadas em corpus

Outra sugestão de metodologia para a elaboração de ontologias é apresentada em Brewster e Wilks (2004). A principal diferença desta abordagem é o reconhecimento de que “*given the 'info-smog' we live in, hand-crafting is impractical and undesirable*” (Brewster e Wilks, 2004:4). Os autores admitem, no entanto, que embora a construção totalmente automática de ontologias ainda seja um desafio, as atuais ferramentas de PLN possibilitam em grande parte a automação da tarefa, reduzindo significativamente o trabalho manual.

Brewster e Wilks (2004) apresentam uma série de critérios metodológicos com o objetivo de ajudar na escolha das ferramentas adequadas para a construção de taxonomias/ontologias¹². São eles:

1) *Coerência*: A taxonomia deve ser, para o usuário, uma organização coerente, de bom-senso, dos conceitos ou termos. A coerência depende de os termos e as associações entre eles integrarem a “conceitualização compartilhada” a que Gruber (1993) se refere. A noção de coerência é dependente da aplicação.

Brewster e Wilks (2004) ressaltam que, com essa escolha, torna-se muito difícil avaliar qualitativamente uma taxonomia ou hierarquia, pois não haveria um critério totalmente estabelecido capaz de decidir se uma taxonomia é correta, ou se uma é melhor que outra. Diferentemente de áreas como a recuperação de informação, em que as medidas de precisão e recuperação são amplamente aceitas, não há medidas equivalentes para ontologias, já que o conhecimento não é uma entidade quantitativa.

2) *Herança múltipla*: Esta característica, já apontada por Noy e McGuiness (2001), diz respeito à localização de um termo em múltiplas posições na taxonomia. Frequentemente, um único termo apresenta uma variedade de facetas que justifica sua localização múltipla em uma taxonomia ou ontologia. O termo

¹² Embora os autores apresentem os critérios como norteadores para a elaboração de ontologias, tais critérios também podem ser utilizados em um momento posterior, de avaliação das ontologias.

“tuberculose pulmonar”, por exemplo, é um distúrbio respiratório e, igualmente, um “distúrbio infeccioso”.

3) *Facilidade na computação*: Como um dos problemas das bases de conhecimento em geral é a manutenção/ atualização, é importante que o método escolhido não apresente uma grande complexidade computacional, de modo a não tornar sua atualização muito custosa.

4) *Rótulos únicos*: É importante que todos os nós tenham rótulos únicos, ainda que sejam compostos por mais de uma palavra. Grupos de palavras caracterizados por uma única etiqueta são mais facilmente compreendidos pelo usuário. Os autores citam o exemplo dos rótulos múltiplos de Sanderson e Croft (1999), em que um grupo de documentos é caracterizado pelo seguinte conjunto de termos: *bateria Califórnia tecnologia milha estado recarga impacto oficial custo hora governo*, o que dificulta sua categorização / identificação.

5) *Fonte de dados*: Brewster e Wilks sugerem que os dados para a construção da ontologia devam vir de duas fontes: a) documentos (fontes primárias); b) dados provenientes de uma taxonomia já existente (“seed taxonomy”), que funcionariam como uma estrutura de dados para revisar ou para servir de esboço inicial, quando necessário.

3.2. Formas de avaliação de ontologias

Boa parte dos trabalhos em PLN utiliza como forma de avaliação as medidas de precisão e abrangência. Tais medidas são tradicionalmente empregadas na área de Recuperação de Informação, e baseiam-se na noção de relevância. A precisão é a proporção de documentos recuperados em uma busca que são relevantes para o usuário. A abrangência corresponde à proporção de documentos relevantes que foram recuperados¹³. Para a realização desses cálculos,

¹³ Precisão = $\frac{\text{número de documentos relevantes identificados}}{\text{número total de documentos identificados}}$
 Abrangência = $\frac{\text{número de documentos relevantes identificados}}{\text{número de documentos relevantes na coleção}}$

é fundamental a existência de material já corretamente identificado, isto é, é fundamental a existência de um “gabarito”, com os quais os novos resultados da identificação automática possam ser comparados – esse “gabarito” pode ser um corpus etiquetado, no caso de tarefas de PLN, ou , no caso da extração de informações, uma coleção de consultas, documentos e julgamentos de relevância conhecidos.

Na avaliação de ontologias (e taxonomias e léxicos semânticos), porém, as noções de precisão e abrangência não são facilmente aplicáveis, pois a natureza da tarefa é diferente. A medida de precisão poderia refletir a quantidade de conhecimento corretamente identificado na ontologia, com relação a todo o conhecimento disponível na ontologia. A abrangência poderia refletir a quantidade de conhecimento corretamente identificado com relação a todo o conhecimento que deveria ser identificado. O problema está em como definir “o conhecimento que deve ser adquirido”, já que o mesmo conjunto de fatos pode levar a diferentes interpretações e, conseqüentemente, a diferentes tipos de “conhecimento” (Brewster et al., 2004).

De fato, a avaliação de ontologias e taxonomias vem sendo bastante discutida e ainda não existem abordagens abrangentes e gerais para o problema. Como afirmam Brewster et al.(2004), “*There are inherent problems in trying to evaluate an ontology as it is not clear what exactly one is trying to evaluate*” (2004:1). De uma maneira geral, para os autores, boas ontologias são aquelas que servem aos seus propósitos.

Alguns trabalhos sugerem que a avaliação de ontologias deve ser feita numa comparação com um modelo ideal (*golden model*), como é o caso de Hovy (2002) e Maedche e Staab (2002). Apesar de trazer a vantagem de possibilitar a utilização de medidas conhecidas como precisão e abrangência, este tipo de comparação, porém, não será abordado neste trabalho porque não temos um modelo ideal de ontologias para o português (e sua elaboração é o processo custoso que justamente se quer evitar), e porque a própria idéia de “modelo ideal” parece não se ajustar ao modelo apresentado aqui. Além disso, quando os resultados diferem dos do modelo ideal, é difícil detectar a origem do problema: se o corpus é inapropriado, se a metodologia é inadequada ou se há uma diferença entre o conhecimento presente no corpus e o modelo ideal (Brewster et al., 2004). Ainda na linha “modelo ideal”, uma possibilidade de avaliação é comparar os

resultados obtidos com as relações presentes na WordNet (Fellbaum, 1998), como fazem alguns trabalhos (Snow et al., 2005; Widdows, 2003; Lin e Pantel, 2002). Porém, considerando os limites da Wordnet com relação a termos específicos de domínio, esta proposta também parece inadequada.

Um outro tipo de avaliação é apresentado por Velardi et al. (2005), que sugerem a comparação entre os resultados da ontologia automática e a avaliação humana. Para os autores, um dos objetivos da avaliação das ontologias geradas automaticamente não é apenas a comparação das diferentes abordagens, mas a verificação da capacidade de um dado processo automático de competir com o processo tipicamente humano de conceitualização de um determinado domínio. Para Verlardi et al., portanto, as questões fundamentais que se apresentam são (i) poderia um método automático simular esse processo humano? e (ii) é possível oferecer a especialistas formas de mensurar a adequação de um conjunto de conceitos como modelo de um domínio?

Além disso, como frequentemente especialistas têm dificuldades para avaliar o conteúdo formal de uma ontologia computacional, sua função é comparar a intuição que têm do domínio com a descrição deste fornecida pela ontologia.

Tentando minimizar estas dificuldades, Velardi et al. desenvolveram um método para a geração automática de glosas a partir das relações entre os conceitos da ontologia. Com isso, eliminam a dificuldade dos especialistas de lidar com os aspectos formais do conteúdo – glosas oferecem uma descrição, em linguagem natural, das especificações formais atribuídas aos conceitos da ontologia¹⁴. A comparação entre intuição e glosas é, sem dúvida, um trabalho mais simples.

De qualquer maneira, contudo, a tarefa de avaliação é manual, um trabalho custoso e subjetivo, dependente da intuição do especialista – o rigor da avaliação é de difícil verificação. Porém, a idéia de criação automática de glosas a partir das definições parece bastante interessante não apenas para a avaliação por humanos, mas em outros contextos que envolvam geração automática de textos.

¹⁴ Exemplo de glosa gerada, no domínio turismo:

termo: *affiliated_hotel*

glosa: “*a kind of hotel, a building where travelers can pay for lodging and meals and other services, being joined in close association*”.

O modelo de avaliação de Brewster et al. (2004) é o que mais combina com a proposta apresentada neste trabalho. Compartilhando os pressupostos de que ontologias devem ser baseadas em corpus, Brewster et al. propõem uma avaliação direcionada aos dados (*data-driven*). Sugerem, especificamente, uma medida para a adequação entre a ontologia e o domínio de conhecimento (*ontological fit*).

Deste modo, se corpora devem ser a fonte mais efetiva de informação para a construção de ontologias (ou de uma grande porção da ontologia, como é sugerido em Brewster et al., 2001), a avaliação deve consistir na identificação da adequação da ontologia com o corpus. Para tanto, uma sugestão simples de Brewster et al. (2004) é a extração automática de termos do corpus e a contagem do número de termos que estão, simultaneamente, no corpus e na ontologia. A ontologia seria “penalizada” para termos presentes no corpus e ausentes na ontologia, e para termos presentes na ontologia e ausentes no corpus. A proposta dos autores, no entanto, é um pouco mais sofisticada: uma metodologia em 3 etapas, que dificilmente poderíamos aplicar por ser dependente de WordNet e de um corpus semanticamente anotado, ferramentas das quais não dispomos em português.

Por fim, descrevo brevemente a proposta de Etzioni et al. (2005) que, embora não seja aplicada a ontologias, diz respeito à avaliação de extração de relações de hiperonímia em um corpus não-anotado – prescindindo, portanto, das tradicionais medidas de precisão e abrangência.

Etzioni et al. apresentam um sistema – KnowItAll – de aquisição automática de instâncias (nomes próprios) e de classes dos domínios Turismo/Geografia, Filmes e Cientistas, que toma como fonte a Web.

Para avaliar a extração das instâncias, um módulo do sistema, chamado “avaliador”, calcula a PMI (*Pointwise Mutual Information*) entre a instância extraída e “sintagmas discriminadores” associados à classe em questão. Esses sintagmas discriminadores podem ser “X é um Y”, ou “Y X” em que X é a instância e Y a classe-alvo. Por exemplo, para as instâncias “Brasil” e “O Invasor” (X) e as classes-alvo “país” e “filme” (Y), seriam sintagmas discriminadores construções como

“*Brasil é um país*” ou “*o país Brasil*”;

“*O Invasor é um filme*” ou “*o filme O Invasor*”

O escore PMI calculado é o número de respostas (hits) para uma busca que combina o sintagma discriminador e a instância, dividido pelo número de hits para a instância sozinha:

$$\text{PMI} = \frac{|\text{Hits (Brasil é um país)}|}{|\text{Hits (Brasil)}|}$$

Porém, como destacam os autores, o score PMI bruto é, tipicamente, uma fração muito pequena, mesmo para instâncias positivas. Além disso, o cálculo não dá a probabilidade de uma instância ser membro da classe, apenas a probabilidade de ver o sintagma discriminador em *webpages* que contêm a instância.

Embora satisfeitos com suas medidas de avaliação, Etzoini et al. alertam para os seguintes problemas:

- dados esparsos (mesmo usando a web como corpus);
- polissemia: *Botafogo* pode ser um bairro e um clube de futebol. Quando o sentido em que se está interessado é o menos freqüente, os scores tendem a ser baixos.

Como, neste trabalho, utilizo um corpus de dimensões mínimas, se comparado à web (cf. capítulo 5), fica claro que o problema da escassez de dados irá se repetir. Por outro lado, a utilização de “sintagmas discriminadores” parece uma boa idéia para testar a validade das relações extraídas automaticamente do corpus. Assim, para a frase (1) abaixo

(1) ..., *transtorno obsessivo-compulsivo, distímia, transtorno afetivo bipolar, abuso de álcool e outras substâncias*

em que são extraídas as relações

transtorno obsessivo-compulsivo < substância;
 distímia < substância;
 transtorno afetivo bipolar < substância;
 abuso de álcool < substância,

uma forma de avaliação seria a realização de uma pesquisa na internet utilizando como expressão de busca as expressões “*abuso de álcool é uma substância*”; “*distímia é uma substância*”, ou “*substância abuso de álcool*” ou “*substância*

distímia”. Como em todas as expressões o resultado da busca é zero, as relações são descartadas.

Porém, nem sempre os resultados dos padrões discriminadores serão confiáveis. Na frase (2)

(2) *O grupo étnico materno foi definido por branco, pardo e negro, considerando critérios como cor da pele, textura do cabelo e formato do nariz.*

são extraídas as relações

cor da pele < critérios
 textura do cabelo < critérios
 formato do nariz < critérios

Em seguida, uma pesquisa na internet usando os “padrões discriminadores” obtém os seguintes resultados (tabela 1):

Padrão discriminador	Qtde de documentos recuperados
"cor da pele é um critério"	1
"critério cor"	89
"critério cor da pele"	2
"textura do cabelo é um critério"	Zero
"textura é um critério"	Zero
"critério textura"	Zero
"critério formato"	3
"critério formato do nariz"	Zero
"formato é um critério"	Zero
"formato do nariz é um critério"	Zero

Tabela 1: Resultados de busca na Internet por padrão discriminador

Outro problema na utilização dos “padrões discriminadores” está na estrutura morfossintática do sintagma utilizado na busca. O fato de mecanismos de busca não considerarem acentos é mais um complicador para a avaliação em língua portuguesa, pois, a uma busca como “*textura é um critério*”, também correspondem resultados como “*textura e um critério*”. Outro problema é que é preciso fazer a concordância de gênero (“*é um*” / “*é uma*”), complicador que não aparece para a língua inglesa (“*is a*”).

Neste capítulo, foram apresentados critérios para a elaboração e avaliação de ontologias. Quanto à elaboração, existem basicamente duas maneiras de enfrentar o problema:

- assumir que ontologias devem ser elaboradas sobretudo de forma manual, em um trabalho que visa, principalmente, a inserção de conceitos relevantes do domínio em questão em uma estrutura formal;
- utilizar critérios mais gerais de elaboração, que levem em conta simultaneamente características desejáveis do ponto de vista da caracterização do domínio e da automação do processo.

Em ambas as propostas está prevista a intervenção humana, seja de maneira explícita, em todas as etapas do processo, como sugere Hovy (2005), seja de maneira implícita e minimizada, como na incorporação de uma taxonomia pré-existente (“seed taxonomy”), que pode funcionar como um esboço de ontologia, como sugerem Brewster e Wilks (2004).

Uma consequência da dificuldade em se atingir um consenso sobre metodologia para a elaboração de ontologias é a ausência de critérios sistemáticos para sua avaliação.

Em geral, as avaliações são sustentadas por dois diferentes paradigmas:

- A comparação com algum modelo ideal, que pode ser uma outra ontologia ou a WordNet;
- A avaliação manual por especialistas.

Porém, ao se considerar especificamente a elaboração de ontologias a partir de corpus, evidencia-se a necessidade de uma outra forma de avaliação – uma avaliação que meça a adequação entre a ontologia e o corpus, como sugerem Brewster et al. (2004).