Calidoscópio
Vol. 3, n.1, p. 47-52, jan/abr 2005

Rove Luiza de Oliveira Chishman
rove@icaro.unisinos.br
Isa Mara da Rosa Alves
isamralves@gmail.com

Extração de informações e *web* semântica: a importância da semântica verbal

RESUMO - O objetivo deste trabalho é apresentar um estudo descritivo de verbos do domínio universidade com vistas à construção de uma ontologia que contribua para o aperfeiçoamento de sistemas de Processamento Automático de Língua Natural (PLN), em especial para a construção de um Sistema de Busca e Extração de Informações na *Web*. Com o auxílio de ferramentas de extração, analisamos os verbos com base nas noções de valência e papéis temáticos propostas por Borba (1996) e chegamos a uma tipologia para os verbos de estado do referido domínio.

Palavras-chave: semântica verbal, papéis temáticos, *web* semântica, ontologia, Sistemas de Busca Extração de Informações

ABSTRACT - The objective of this paper is to present a descriptive study of verbs of the university domain aiming at the construction of an ontology which may contribute towards the enhancement of Natural Language Processing (NLP) systems, particularly to the construction of an on-line Information Retrieval System. Using a special tool for extraction pourposes, we analized verb valence and thematic roles according to the aproach proposed by Borba (1996) and we got a typology to describe stative verbs of the domain.

Key words: verbal semantics, thematic roles, semantic web, ontology, Information Retrieval Systems.

Introdução

Este trabalho está vinculado a um campo de investigação interdisciplinar. Situado na área da semântica lexical computacional, explora a interface da Semântica com a área da Computação dedicado ao Processamento da Linguagem Natural (PLN), subárea da Lingüística Computacional e, por sua vez, da Inteligência Artificial (IA).

Este artigo, em especial, traz algumas reflexões empreendidas no âmbito do projeto ONTOVERB, cujo objetivo prático é a construção de uma ontologia do domínio universidade a partir da descrição da semântica verbal. Com o estudo dos nominais já bem encaminhado, ocupamo-nos aqui em relatar a fase do projeto voltada para o estudo das abordagens semânticas que se prestam à representação do conteúdo veiculado pelos verbos. Interessa-nos essencialmente responder às seguintes questões:

a) Em que medida as informações semânticas relacionadas aos verbos podem fazer parte de uma ontologia, de maneira a possibilitar uma melhoria do desempenho do sistema de busca e extração de informações?
b) Que abordagens semânticas se adequam a essa aplicação?
c) Como podemos armazenar esses dados de forma a permitir a sua usabilidade não apenas pelo usuário, mas também pelo sistema?

As questões (a) e (b) dizem respeito à fase lingüística do projeto, que prevê o estudo exploratório de experiências bem sucedidas de construção de léxicos computacionais ou ontologias e a conseqüente definição dos recursos lingüísticos que se ajustam aos nossos interesses. Compreendem o trabalho de caracterização detalhada do significado dos verbos que compõem o domínio. A questão (c) vincula-se à fase computacional da pesquisa, que abarca a definição dos critérios de formalização da descrição lingüística. Realiza-se, nessa fase, o estudo do editor de ontologias Protégé.

A extração de informações e a *Web* Semântica

A recuperação de informações é uma atividade que pode ser realizada tanto de forma manual como com o auxílio de sistemas computacionais específicos. A recuperação manual de informações se torna muitas vezes uma atividade impraticável devido à variedade de materiais a ser analisado. Como exemplo, podemos ter a busca de um livro em uma biblioteca que trate de um tema específico. Manualmente, o pesquisador demoraria muito tempo – e,

talvez nunca terminaria – até localizar todos os livros que tratam do assunto específico de seu interesse. A partir da criação da *Internet*, em 1989, a quantidade de materiais disponíveis sobre os mais diversos assuntos aumentou extraordinariamente ainda mais a necessidade de sistemas que auxiliem o homem em nessa tarefa.

Podemos encontrar na *web* atual diversos *sites* de busca (Google, Yahoo, Cadê, etc.), ou seja, sistemas dedicados à Recuperação de Informações. Atualmente, a recuperação/busca de informações é realizada em geral através de palavras-chave, o que acaba por gerar uma grande quantidade de lixo no retorno da busca. Os sistemas atuais, em sua grande maioria, não estão aptos a realizar inferências a partir do conteúdo dos documentos a fim de estabelecer relações de sentido entre eles. O Google, por exemplo, é um *site* de busca que utiliza uma tecnologia capaz de rastrear os *links* internos dos documentos a fim de verificar quando um *site* remete a outro, o que é um indicativo de que há alguma relação entre o conteúdo dos documentos. Ressaltamos, no entanto, que, mesmo com essa tecnologia, as informações semânticas permanecem disponíveis apenas para os humanos e não para a máquina.

É nesse cenário de *World Wide Web*, onde há uma vasta quantidade de informações disponibilizadas sem qualquer organização semântica, que surge a *Web Semântica.* Essa nova concepção de internet propõe a substituição dos tradicionais sistemas de busca de textos baseados simplesmente em palavras-chave por avançados recursos capazes de manipular e realizar inferências reconhecendo relações entre o conteúdo dos textos. É sob essa nova concepção de *web* que surgem os Sistemas de Extração de Informações, os quais não apenas localizam informações para o usuário, mas são capazes de interagir em linguagem natural.

Os sistemas de extração de informações não são capazes apenas de encontrar documentos relacionados a determinado tema, mas de analisar o conteúdo de tais documentos e interagir com o usuário em linguagem natural. Lima e Vieira (2000) afirmam que tais sistemas são capazes de alterar a forma de apresentação das informações relevantes contidas em determinados segmentos do texto e apresentá-las em formato coerente com a questão solicitada na busca. A busca por informações em sistemas de extração de informações pode ser realizada através de palavras-chave ou questões (expressão, frase ou pergunta).

O grande desafio para qualquer sistema de extração de informações é estabelecer relações entre os textos a fim de retornar ao usuário um maior número de documentos relevantes com base na solicitação de busca. A seleção de documentos relevante exige a realização de processos de inferências. Para um humano, é tarefa relativamente simples, apesar de demorada; no entanto, para que um sistema computacional seja capaz de simular tal processo mental, é necessária a construção de regras que possibilitem quantificar as decisões sobre a relevância. Como se pode imaginar, não é tarefa simples. Mas, sabe-se atualmente que a forma de fazer isso é através da implementação dos sistemas a partir de *ontologias*.

No contexto de nosso projeto, uma *ontologia* é um documento ou arquivo composto por termos referentes a um domínio organizado por seu conteúdo semântico formalmente definido. Entre as principais vantagens em usar uma ontologia, destacamos a possibilidade de tornar acessíveis para a máquina os conteúdos semânticos e de reutilização desse conhecimento por meio do intercâmbio com outras ontologias através da unificação de códigos e conceitos empregados. Isso só é possível graças ao novo padrão de Internet criado pela W3C, conhecido como *Semantic Web*, que permite que informações semânticas sejam estruturadas de maneira explícita. De acordo com essa nova tecnologia, no lugar da conhecida linguagem padrão HTML (HyperText Markup Language), teremos a OWL (Ontology Web Language), que permite a modelação do conteúdo de textos.

A ontologia que estamos construindo será aplicada a um sistema de extração de informações, mais precisamente um sistema de busca na Internet, tal como Google, Altavista, Yahoo. Entre as características que objetivamos contemplar, destacamos: (a) sua capacidade de representar informações lingüísticas e de efetuar inferências a partir desse conhecimento; (b) a possibilidade de interagir com o usuário; (c) a capacidade de processar documentos; (d) a capacidade de inserir novas regras a partir de documentos classificados.

Dedicamo-nos, na seção seguinte, a refletir sobre os recursos descritivos que se prestam à representação da semântica verbal. Dentre as abordagens estudadas, uma, em especial, é analisada neste trabalho: a semântica de situações e papéis temáticos. Na seqüência, apresentamos também a análise dos verbos extraídos das páginas do *site* da universidade.

A semântica de situações e papéis temáticos

Ainda que o propósito deste trabalho seja refletir sobre um tipo específico de abordagem semântica, a que se apóia na descrição das situações e de seus participantes, convém referir que há diversas formas de descrever a semântica dos verbos. Sem perder de vista a compatibilidade de tais recursos com a construção de um léxico computacional, podemos citar, pelo menos, três recursos: (a) os modelos relacionais; (b) os modelos baseados em *frames*; (c) os modelos baseados na noção de predicação e papéis temáticos.

Os modelos relacionais do léxico têm servido como ponto de partida para a organização de extensas bases de dados lexicais. As *wordnets* (de Princeton e a

EuroWordNet) são certamente bons exemplos. Ao organizarem suas bases lexicais a partir de relações semânticas ou conceituais, e não seguindo uma ordenação alfabética, apresentam o léxico em uma organização inspirada no que seria a de um léxico mental. Sinonímia, antonímia, hiponímia e meronímia estão entre as relações contempladas pelas *wordnets*. Vale considerar, contudo, que, em se tratando da inclusão de verbos em uma ontologia, informações de tipo relacional não são mais importantes, ao contrário do que ocorre com a organização dos nominais.

Os modelos baseados em *frames*, por sua vez, apresentam representações esquemáticas das situações envolvendo participantes, propriedades e outros papéis conceituais que constituem um *frame*. Com certeza, o FrameNet, de responsabilidade de Charles Fillmore, é o projeto mais representativo dessa forma de organizar o conhecimento. O objetivo dessa base de dados é documentar possibilidades combinatórias semânticas e sintáticas (valências) de cada palavra predicativa (nominais, verbos e adjetivos) em cada um de seus sentidos.

Os modelos baseados em situações e participantes exploram a interface entre sintaxe e semântica ao se afastarem do nível relacional de análise e ao explorarem o nível sentencial. Trata-se de um recurso de descrição especialmente rico, se considerarmos que os verbos são as entidades que, por excelência, organizam as sentenças. Pensando nas vantagens que esse recurso pode trazer para a representação do conhecimento em uma ontologia, podemos apontar a possibilidade de descrever o papel que cada participante tem uma situação específica.

Saeed (1997) emprega o termo *situação* para referir-se à relação sintático-semântica estabelecida entre um verbo (enquanto elemento predicador) e seus argumentos externo e interno, ou seja, os participantes da situação.

Há, na literatura, diferentes formas de descrição das situações. Certamente a classificação aspectual apresentada por Vendler (1969) – verbos de atividade, estado, *accomplishment* e *achievement* – serviu de inspiração para muitos trabalhos posteriores. Dowty (1979), Van Valin (1997) e Pustejovsky (1995), sem dúvida, têm a influência de Vendler.

Ainda que tenhamos à disposição diferentes abordagens para o estudo dos verbos sob essa ótica, havendo notável distinção terminológica, todas compartilham de um mesmo pressuposto teórico: a centralidade no verbos. Eles sustentam que é o verbo que dita a presença e a natureza do nome, e não o contrário.

Para a análise dos verbos de nosso *corpus*, optamos por adotar a classificação proposta por Francisco Borba, na obra *Uma gramática de valências para o português* (Borba, 1996) e no *Dicionário de usos do português do Brasil* (Borba, 2002).

Segundo Borba (1996), da associação entre um verbo e um nome, resulta um caso/papel para o nome e uma classe para o verbo. Para o autor, os verbos podem ser organizados em quatro classes semânticas: *verbos de ação*, *de processo*, *de ação-processo* e *de estado*.

Os verbos de ação se caracterizam por não apresentarem nenhuma mudança de estado, físico ou moral, de condição, de posicionamento no tempo ou no espaço, e expressam uma atividade realizada por um sujeito *agente*. Tomando como exemplo em verbo extraído de nosso *corpus*, temos:

> (1) Ludger Teodoro Herzog atua na Unisinos desde 1975.

Outro tipo de verbo não-estativo são os *verbos de processo*. Estes expressam um evento ou uma sucessão de eventos que afetam um sujeito *paciente* ou *experimentador*. Construções com essa classe semântica traduzem um acontecer ou um experimentar, isto é, algo que se passa com o sujeito ou que ele experimenta. Borba (1996) explica que o sujeito é afetado por aquilo que o verbo indica ou é um experimentador. Observemos o exemplo a seguir:

> (2) A deliberação final ocorre nas próprias câmaras ou no Colegiado Pleno.

Já os *verbos de ação-processo* exprimem uma ação realizada por um sujeito *agente* ou uma causação levada a efeito por um sujeito *causativo* afetando o complemento. Esses verbos sempre atingem uma mudança de estado, de condição ou de posição.

> (3) A Dicom coordena a organização e o funcionamento das ações desenvolvidas pelo centro.

A quarta classe sintático-semântica proposta por Borba (1996) diz respeito aos *estativos*. Estes, segundo o autor, expressam uma propriedade – estado, condição, situação – localizada no sujeito, que é mero suporte dessa propriedade ou então seu experimentador ou beneficiário. Entre as características desses verbos, está a obrigatoriedade de se ter um argumento inativo. Observemos o exemplo a seguir:

> (4) A Unisinos possui um Programa de Bolsas de Iniciação Científica.

O autor indica que os *verbos de estado* que apenas compõem predicados estativos são chamados copulativos, por ligarem o núcleo do predicado ao sujeito. É, sem dúvida, a classificação das orações estativas em três subconjuntos que expressa o diferencial dessa abordagem. O primeiro tipo refere-se às construções que contêm verbos de estado de peso semântico específico. Incluem-se nesse grupo verbos como *amar*, *crer*, *saber*, *compreender*, *duvidar*. O segundo grupo refere-se às sentenças contendo predicado de

existência. *Existir*, *haver*, *ser* são verbos que se aplicam a essa tipologia. As orações estativas que têm como núcleo do predicado um adjetivo ou um nome formam o terceiro grupo. Esse terceiro grupo, tendo o predicado introduzido por um verbo copulativo da classe de *ser*, *estar*, *parecer*, divide-se em *equativas*, *atributivas*, *locativas* e *possessivas*.

Exemplificando a tipologia proposta por Borba para as orações estativas, temos as seguintes construções extraídas de nosso corpus:

(5) A reitoria compreende três instâncias.
(6) Há lugares para estudar.
(7) a. A Unicon Empresa Júnior é uma associação civil.
b. A resposta foi positiva.
c. O atendimento é das 8 às 22h.
d. A Biblioteca tem livros sobre os jesuítas.

A identificação dos papéis semânticos parte da descrição semântica do verbo a que o argumento se refere. Borba (1996), ao propor uma gramática de valências para o Português, é uma referência importante para o estudo dos papéis semânticos. O autor defende que uma análise valencial deve tanto identificar matrizes valenciais ou descrever a estrutura externa dos constituintes quanto determinar as relações sintático-semânticas ou temáticas que constituem a estrutura conceitual dos itens lexicais.

A noção de matriz valencial é definida pelo autor como o esquema que explicita a valência do verbo, que pode ser em três níveis: valência quantitativa, valência lógica ou lógico-semântica; valência quantitativa, valência sintática ou morfossintática; valência semântica, incluindo traços que compõem cada categoria, das funções temáticas, das restrições selecionais.

O autor define papéis semânticos como noções relacionais que se apresentam como configurações estruturais, com estatuto comparável aos das noções de sujeito e objeto em muitas teorias gramaticais. Como noções relacionais, as relações temáticas (ou semânticas) representam um sistema de casos ou gramática de casos, sendo caso definido como a atuação dos argumentos na predicação. Seguindo basicamente a proposta de Fillmore (1977) para a identificação dos papéis semânticos, Borba (1996) enumera um conjunto de doze casos e procura caracterizá-los a partir de traços semânticos. Abaixo os casos adotados na proposta de Borba:

- Agentivo: o que age ou faz; desencadeia uma atividade, sendo origem dela e seu controlador.
- Beneficiário: o que se beneficia de.
- Locativo: o que localiza.
- Experimentador: traduz uma experiência ou uma disposição mental.
- Objetivo: mais neutro, é o afetado por aquilo que o verbo indica.
- Instrumental: exprime uma causa indireta.
- Causativo: provoca um efeito ou desencadeia algo; expressa uma atividade ligada a um estímulo.
- Meta: contém os traços *afetado* e *transição* e expressa o ponto de partida.
- Origem: contém os traços *afetado* e *transição* e expressa o ponto de chegada.
- Resultativo: é um efetuado, liga-se a verbos de existência.
- Temporal: indica localização no tempo.
- Comitativo: indica associação, é sempre um afetado.

Antes de apresentarmos a análise semântica dos verbos do *corpus*, convém ressaltar que não há um conjunto fechado e estático de papéis semânticos ideal para descrever as relações de sentido expressas entre o predicador, seus argumentos e componentes situacionais. A teoria dos papéis semânticos tem sido constantemente reformulada desde seu princípio. Kearns (2000) comenta que alguns dos papéis semânticos tradicionais mais antigos têm sido abandonados, enquanto outros estão sendo subdivididos em subtipos. Isso justifica a dificuldade de aplicar em sua totalidade a classificação proposta por Borba e de identificar com segurança muitos dos casos de nosso *corpus*, o que indica a necessidade de nos valermos de traços mais definitórios para definir os papéis.

A descrição semântica dos verbos do *corpus*

O *corpus* da pesquisa é constituído de páginas do *site* da universidade. A fim de identificar as entidades verbais do domínio, foi realizada a extração automática através da ferramenta EXTRACTOR, que possibilitou a geração de duas listas: uma com verbos na forma original e outra com verbos na forma canônica. Utilizamos também a ferramenta CONCORDANCE, que nos forneceu, além da lista dos verbos, a estatística de ocorrências e os co-textos de cada verbo.

A partir desse procedimento inicial de extração e organização, passamos à análise com base na classificação proposta por Borba (1996). Nosso objetivo é descrever a semântica com base na tipologia apresentada acima, o que significa aplicar as características de cada um dos quatro subconjuntos ao conjunto de 1043 verbos resultantes da extração e construir uma matriz valencial explicitando, em especial, as propriedades semânticas dos verbos e seus argumentos.

A primeira análise empreendida segue os critérios que Borba (1996) utiliza para o estudo da valência se-

mântica. Restringindo o estudo da valência verbal aos *verbos plenos*, isto é, verbos que semanticamente têm significado lexical e sintaticamente ocupam o núcleo do predicado num sintagma verbal, o autor exclui os verbos funcionais, os modais e os substitutos. Considera funcionais os auxiliares e os verbos-suporte. Como exemplos de expressões de modalidade por verbos, temos a forma imperativa e as formas de possiblilidade (dever, poder, querer + inf.), obrigatoriedade (dever, ter que) e permissão (deixar). A substituição ocorre quando se usa uma unidade para referir-se anafórica ou cataforicamente a outra.

Com base nesse critério, excluímos os verbos que pertencem a essas categorias, consideradas pelo autor como casos residuais. Sem aprofundar o estudo desses casos que não formam predicado, mas que também têm suas peculiaridades semânticas, submetemos o *corpus* a uma primeira filtragem: excluir os casos residuais. Chegamos, então, ao seguinte resultado:

- Número total de ocorrências verbais: 1043.
- Número de ocorrências com verbos plenos: 630.
- Número de ocorrências de casos residuais: 255.
- Outros casos: 158.

Além dos verbos que Borba denomina como casos residuais, encontramos outros tipos de resíduo: são casos de ambigüidade categorial, como a forma *aprovada*, empregada como adjetivo e não como verbo, e a forma *são*, empregada como adjetivo e não como verbo. O número 158 refere-se a tais casos.

A título de ilustração dos casos residuais encontrados no *corpus*, consideremos os seguintes exemplos:

(8) Auxiliares ou verbos-suporte:

a) O diretor é escolhido pelo reitor.
b) A universidade tem obrigação de ajudar a produzir.
c) A universidade parou de oferecer o curso.

(9) Expressões de modalidade:

a) O aluno pode inscrever-se nesse programa.
b) Veja as áreas de conhecimento por centro.
c) Essas estratégias devem resultar no aprimoramento de performance dos professores e funcionários.

Passemos à análise dos 630 verbos plenos. Com o auxílio do Dicionário de Usos (Borba, 2002), chegamos à seguinte subclassificação:

- 233 verbos de estado (36,98%);
- 201 verbos de ação-processo (31,90%);
- 140 verbos de ação (22,22%);
- 56 verbos de processo (8,88%).

Considerando o número expressivo de verbos estativos, finalizamos este trabalho refletindo sobre o papel que as informações semânticas veiculadas por essa categoria podem assumir na ontologia. Dizendo de outra forma, queremos saber como as informações relacionadas ao predicado e argumento(s) dessas construções podem ser aproveitadas em um sistema de busca. Lembremos que Borba subdivide os estativos em três subtipos: um contendo verbos com peso semântico específico como predicado, outro correspondendo aos existenciais e outro descrevendo as construções com adjetivo ou nome como núcleos do predicado. Esse terceiro tipo, por sua vez, abarca orações equativas, atributivas, locativas e possessivas.

Dos três subtipos, o terceiro é o mais produtivo. Das 233 construções com estativos, 101 têm nominais ou adjetivos como núcleos. O subconjunto das equativas nos remete a elementos semanticamente permutáveis, o que, na formalização, se reverteria na clássica relação hiperônimo-hipônimo. As atributivas nos levam às propriedades das classes e as locativas trazem informações sobre local e hora dos eventos. As possessivas nos remetem a uma relação do tipo holônimo-merônimo também prevista na fase de descrição dos nominais. Para o domínio estudado – universidade –, as construções com posse atributiva servem para relacionar os diferentes lugares e seus componentes.

Considerações finais

Neste trabalho, apresentamos os resultados parciais do projeto de construção de ontologia de domínio. Detivemos-nos na classificação proposta por Borba (1996) e limitamo-nos a analisar os verbos estativos. De forma preliminar, concluímos que, apesar de tais entidades não terem o mesmo semântico das outras três categorias, já que essencialmente não funcionam como predicados, trazem informações pertinentes e relevantes para um sistema de busca na *web*.

Referências

BORBA, F. 1996. *Uma gramática de valências para o Português*. São Paulo, Editora Ática.
BORBA, F. 2002. *Dicionário de usos do Português do Brasil*. São Paulo, Editora Ática.
DOWTY, D. 1979. *Word meaning and Montague grammar*. Dordrecht, D. Reidel.
FILLMORE, C. 1977. Em favor do caso. *In*: L. LOBATO (org.), *A semântica na lingüística moderna: o léxico*. Rio de Janeiro, Francisco Alves, p. 275-365.
KEARNS, K. 2000. *Semantics*. New York, St. Martin Press.
LIMA, V. L. S. de e VIEIRA, R. 2001. *Lingüística computacional: princípios e aplicações.* Apostila de curso de extensão.
PUSTEJOVSKY, J. 1995. *The generative lexicon*. Cambridge, MIT Press.
SAEED, J. 1997. *Semantics*. Oxford, Blackwell.

VAN VALIN, R. D. 1997. *Syntax.* Cambridge, Cambridge University Press.
VENDLER, Z. 1969. *Adjetives and nominalizations*. The Hague/ Paris, Mouton.

Recebido em 16/03/2005
Aceito em 08/04/2005

Rove Luiza de Oliveira Chishman
UNISINOS
Isa Mara da Rosa Alves
UNISINOS