- **LINGÜÍSTICA COMPUTACIONAL III**
  Coordenador(a): *Bento Carlos Dias da Silva*

A SEMÂNTICA VERBAL: CONTRIBUIÇÕES PARA O APERFEIÇOAMENTO DE SISTEMAS DE EXTRAÇÃO DE INFORMAÇÃO NA ERA DA WEB SEMÂNTICA

Isa Mara da Rosa Alves (UNESP), Rove Luiza de Oliveira Chishman

Sob a perspectiva da Semântica Lexical Computacional, este trabalho apresenta alguns dos resultados de um estudo descritivo de verbos do domínio jurídico realizado com o intuito de construir uma ontologia aplicável ao aperfeiçoamento de sistemas de Processamento Automático da Língua Natural (PLN), em especial, à ferramenta de busca e extração de informações na Web da Procuradoria Geral da República de Portugal (PGR-PT). Tendo em vista a interdisciplinaridade das investigações feitas, a pesquisa como um todo foi organizado em três fases: (i) fase teórica lingüístico-computacional; (ii) fase lingüística teórico-aplicada; e (iii) fase computacional. Este artigo focalizará a fase lingüística teórico-aplicada da pesquisa; porém, não deixaremos de mencionar as demais. A partir de um estudo de aplicações semelhantes com o intuito de identificar que tipo de informação lingüística se presta para a construção de uma ontologia de

domínio jurídico - fase (i) -, o objetivo deste artigo é apresentar uma proposta de descrição ontológica dos verbos do domínio jurídico com base na análise de córpus - fase (ii). O resultado desse estudo serviu para - na fase (iii) - estabelecer critérios para a formalização das descrições semânticas em uma ontologia com o auxílio do editor Protégé. Esta ferramenta possibilita a conversão dos dados para a linguagem padrão da Web Semântica, a Ontology Web Language (OWL). Das abordagens discutidas neste trabalho, as que mais se prestaram para a descrição da semântica verbal com os fins aqui desejados foram as relações lógico-semânticas, os papéis semânticos e os frames. Com o uso da ontologia proposta aqui, o sistema da PGR-PT, de simples busca por palavras-chave, estará apto a interagir com o usuário em língua natural através de pergunta e resposta de maneira eficiente.

A SUMARIZAÇÃO TEXTUAL COM BASE EM CONHECIMENTO DISCURSIVO PARA PRESERVAÇÃO DA COERÊNCIA

Eloize Rossi Marques Seno (UFSCAR), Lucia Helena Machado Rino (UFSCAR), Thiago Alexandre Salgueiro Pardo (USP)

Apresenta-se, neste resumo, o RHeSumaRST, um sumarizador automático de textos baseado na organização do discurso, visando à preservação da coerência dos sumários.

A sumarização de um texto é realizada pela aplicação de heurísticas definidas manualmente a partir do estudo de um corpus produzido para este fim (Pardo e Rino, 2003). Essas heurísticas podam a estrutura discursiva dos textos, que são construídas, neste trabalho, com base na Rhetorical Structure Theory - RST (Mann and Thompson, 1987). Nesta teoria, o discurso é representado por uma estrutura em que as proposições do texto são relacionadas por relações retóricas, diferenciando-se as proposições de acordo com sua importância.

As heurísticas visam, principalmente, à preservar a coerência dos sumários. Coerência, neste caso, é limitada à verificação das cadeias de co-referências, que, sabidamente, são um de seus aspectos principais. Para isso, as heurísticas incorporam elementos da Veins Theory (Cristea et al., 1998), que delimita domínios de acessibilidade referencial para cada proposição do discurso com base na estruturação RST.

Para a realização da sumarização automática, a estrutura discursiva de um texto deve ser produzida anteriormente à aplicação das heurísticas, o que é feito pelo DiZer (Pardo et al., 2004), um analisador retórico automático para o português do Brasil. Além disso, os textos devem ser anotados com suas cadeias de co-referências. Neste trabalho, utilizou-se a ferramenta MMAX (Müller and Strube, 2001) como auxílio a esta anotação.

Em um estudo preliminar, verificou-se que, de fato, o RHeSumaRST preserva a coerência do sumário mais do que alguns métodos da literatura considerados importantes na área de sumarização automática.

O ESTABELECIMENTO DA CORRESPONDÊNCIA LÉXICO-CONCEITUAL ENTRE BASES RELACIONAIS DE DADOS LEXICAIS DO PORTUGUÊS E DO INGLÊS

Bento Carlos Dias da Silva (UNESP)

O objetivo deste trabalho é discutir o método de especificação da correspondência léxico-conceitual entre a base nuclear de uma rede wordnet em fase de construção para o português do Brasil, a base lexical da Wordnet.Br, e a base análoga, desenvolvida para o inglês norte-americano, a base lexical da WordNet de Princeton, e, por extensão à base lexical da EuroWordNet, que se compõe das bases construídas para um subconjunto de línguas da União Européia. Depois de breve contextualização histórica e metodológica da investigação em Tecnologias da Linguagem

Humana, contexto de pesquisa em que se insere este estudo, resumem-se os passos da compilação da rede Wordnet.Br, isto é, as etapas da montagem da base relacional de dados lexicais da rede, em que substantivos, verbos, adjetivos e advérbios são estruturados e armazenados, em módulos independentes, porém comunicantes, essencialmente em termos de uma coleção de relações paradigmáticas de significado: as relações léxico-semânticas de sinonímia e antonímia e as lógico-conceituais de hiponímia, troponímia, acarretamento, meronímia e causa. Na seqüencia, apresenta-se a estratégia de mapeamento léxico-conceitual entre as redes brasileira e norte-americana. A discussão abrange os três domínios envolvidos na pesquisa e no desenvolvimento de recursos lingüísticos para o processamento automático de línguas naturais: o lingüístico, o lingüístico-computacional e o computacional. A conclusão resume os pontos principais do trabalho e salienta a contribuição do método para a automatização do processo de compilação de recursos lexicais robustos, quer para fins computacionais, quer para fins lexicográficos.

PADRÕES DE LEXICALIZAÇÃO E A INDEXAÇÃO DA BASE DA WORDNET.BR

Ariani di Felippo (UNESP)

A WordNet, desenvolvida para o inglês americano, é uma base de dados lexicais que se estrutura em conjuntos de sinônimos, os "synsets", que visam a representar o conceito lexicalizado pelas unidades que o compõem; cada synset constitui um nó e as ligações entre os diferentes nós, representada por meio de arcos rotulados, visam a exprimir a relação léxico-semântica de antonímia e as relações lógico-conceptuais de hiponímia, troponímia, meronímia, causa e acarretamento. Diante da relevância lingüística e da potencialidade tecnológica, o desenvolvimento de wordnets para diferentes línguas frutificou-se e sofisticou-se. As wordnets em construção para as línguas dos países da União Européia, por exemplo, estão aglutinadas em uma base multilíngüe, a EuroWordNet, em que os synsets das diferentes línguas que lexicalizam um conceito comum são interligados. Em uma das etapas futuras do desenvolvimento da wordnet brasileira, a Wordnet.Br, está prevista a interligação desta a wordnet americana, resultando, assim, em uma base bilíngüe que poderá ser utilizada em pesquisas lingüísticas e do PLN. Essa indexação, em especial, requer a identificação de padrões de lexicalização de conceitos (isto é, mapeamento ou associação regular entre conceitos e itens lexicais) do inglês e do português. Motivado por essa tarefa de indexação das wordnets brasileira e americana, apresenta-se neste trabalho a proposta de estudo do mapeamento dos padrões de lexicalização de conceitos codificados nos nomes concretos do inglês e do português. Além de subsidiar parte da futura indexação da Wordnet.Br a WordNet de Princeton, o trabalho ora proposto busca contribuir para o desenvolvimento de questões como o desenvolvimento de ontologias lingüísticas e conceituais e a especificação semântica dos itens lexicais. [CNPq]