Usando o FrameNet para a descrição semântica: um experimento de anotação de corpus

Rove Chishman (UNISINOS)

Anderson Bertoldi (UNISINOS/ PG CAPES)

João Gabriel Padilha (UNISINOS/IC FAPERGS)

Apresentamos aqui os resultados de um trabalho de anotação semântica realizado no âmbito do projeto FrameCorp, cujo principal propósito é aplicar o conceito de *frame* semântico, conforme proposto pelo projeto FrameNet (FILLMORE, et al., 2003), na tarefa de anotação manual de *corpus*. Nosso objetivo aqui é relatar o processo de anotação de *frames* realizado no *corpus* Summ-it, desenvolvido inicialmente com o objetivo de embasar pesquisas envolvendo relações anafóricas e retóricas e a sumarização automática (ALUÍSIO et al., 2003). Os poucos recursos lingüísticos com anotação semântica para a língua portuguesa motivaram esta pesquisa. Nosso trabalho envolveu três principais recursos: (1) a base de dados XML do projeto FrameNet, (2) a ferramenta de anotação SALTO (BURCHARDT et al., 2003) e (3) os corpora Summ-it e PLN-Br. O *frame* é a caracterização de uma pequena "cena" ou "situação" abstrata, bem como dos participantes dessa cena. A atividade de anotação foi parcialmente inspirada na metodologia do FrameNet. A descrição de um *frame* semântico implica a associação de uma unidade lexical a um *frame* específico, que, por sua vez, é constituído de diversos papéis conceituais, ou *elementos frame*. A tarefa de anotação consiste em: (i) identificar o elemento evocador de *frame* na sentença, (ii) identificar um equivalente de tradução na base de dados FrameNet, (iii) identificar o *frame* associado ao item lexical do inglês e (iv) anotar a sentença com os elementos *frame* apropriados. Tal tarefa não é trivial, pois envolve a utilização de um recurso lexical da língua inglesa para a anotação semântica do português. A anotação do Summ-it foi dividida em três momentos distintos. Primeiramente, nós anotamos o *frame Statement*, seguindo uma abordagem lexicográfica de anotação, procurando por determinadas unidades lexicais no *corpus*. Assim, optamos por iniciar a anotação pela categoria semântica mais recorrente. Das 774 sentenças que compõem os textos do *corpus*, 135 delas foram anotados com o *frame Statement*. Em um segundo momento, utilizamos uma abordagem *running-text*, anotando todos os verbos evocadores de *frames*. A terceira etapa envolveu o estudo e anotação dos casos residuais, considerados em nossa atividade de anotação

como aqueles verbos que não são tipicamente predicadores, como os modais e os suportes. Uma das maiores dificuldades desta etapa foi diferenciar o *frame Statement* de *frames* semanticamente relacionados, como *Telling*, *Quarreling*, *Adducing* e *Summarizing*, pois todos esses *frames* podem ser evocados por unidades lexicais em discurso indireto, assim como o *frame Statement*. A anotação foi realizada por dois anotadores, as anotações confrontadas e os casos de divergência foram solucionados por um terceiro anotador. Foram anotadas 512 sentenças: os anotadores concordaram em 337 sentenças e divergiram em 91 sentenças. A anotação pôde ser dividida em casos simples, em que há paralelismo entre o inglês e o português, de forma que se torna fácil identificar um equivalente de tradução. Já os casos de discordância são provocados pelo não-paralelismo, vaguidade ou polissemia. O verbo *ter*, por exemplo, ilustra um caso de vaguidade, podendo evocar os frames *Possession*, *Have-associated* e *Inclusion*. Na terceira etapa, ocupamo-nos da análise semântica das construções com verbos suporte e modalizadores, totalizando 262 sentenças do *corpus*. No que tange às construções com verbos suporte, o desafio consistiu em lidar com diferentes estruturas predicativas e graus de composicionalidade variados. As construções com verbos modais, por sua vez, apresentaram desafios de outra ordem. Uma das dificuldades consistiu em identificar na base de dados os frames correspondentes às diferentes categorias de modalização, como epistêmicos, deônticos e dinâmicos. A outra dificuldade consistiu em avaliar se as categorias semânticas destes frames são descritivamente adequadas para expressar a semântica da modalidade. Vale, por fim, destacar que este trabalho de anotação nos mostrou a utilidade das descrições baseadas em frames, se compararmos com outros modelos semânticos utilizados para o mesmo fim, como é o caso dos papéis temáticos adotados pelo projeto PropBank. Como limitação, destacam-se as dificuldades naturais de uma anotação seguindo o modo *running-text*, haja vista a diversidade de problemas teóricos a serem enfrentados.

Referências

ALUÍSIO, S. et al. The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. In: Macnery, T. et al. (eds.), CORPUS LINGUISTICS 2003, Lancaster. *Proceedings of the Corpus Linguistics 2003*, UCREL Technical Papers, v.16. p.14-21, 2003.

BURCHARDT, et al. SALTO - A Versatile Multi-Level Annotation Tool. *Proceedings of LREC 2006*, Genoa, Italy, 2006.

FILLMORE, C. J. Frame Semantics. The Linguistic Society of Korea. *Linguistic in the Morning Calm,* Seoul, Hansinh Publishing Co., 111-137. 1982.

FILLMORE, C. J.; JOHNSON, C. R.; PETRUCK, M. Background to FrameNet. *International Journal of Lexicography.* Vol.16, no.3, p.235-250, 2003.