Universidade Federal do Rio Grande do Sul
Instituto de Informática

Processamento de Linguagem Natural: Extração de Expressões Multipalavra

Bolsista: Vítor Bujés Ubatuba De Araújo
Orientador: Prof. Dr. Edson Prestes e Silva Jr.
Coordenadores: Profª Drª Aline Villavicencio, Carlos Eduardo Ramisch

Expressões multipalavra (EMs) são combinações de palavras que apresentam idiossincrasias lingüísticas ou estatísticas. EMs são uma característica importante das línguas humanas, e incluem fenômenos tais como verbos frasais (*carry up, consist of*), verbos de suporte (*tomar um banho, dar uma caminhada*), compostos (*carro de polícia, bode expiatório*) e expressões idiomáticas.

O presente trabalho consiste na extensão de uma ferramenta automatizada para identificação e extração de EMs a partir de corpora, o mwetoolkit. O mwetoolkit funciona primeiramente extraindo candidatos a EMs baseando-se em padrões morfossintáticos, e em seguida filtrando os resultados usando medidas estatísticas. Buscou-se melhorar tanto a eficiência da ferramenta em termos de consumo tempo e recursos computacionais, quanto a qualidade dos resultados obtidos na extração.

Para tal fim, foram realizadas uma série de modificações no toolkit. O mecanismo para casamento de padrões morfossintáticos foi alterado para utilizar expressões regulares extendidas, o que resultou tanto em uma maior flexibilidade na especificação dos padrões e uma maior cobertura na identificação de EMs, quanto a um aumento de eficiência, devido ao suporte nativo da linguagem de programação utilizada a expressões regulares. A rotina de indexação de palavras e expressões foi reescrita, o que levou a uma redução significativa de consumo de tempo e memória. Adicionou-se ainda a possibilidade de descartar palavras entre componentes de uma EM, permitindo a identificação de EMs não-contíguas (*ficar totalmente de escanteio*), entre outras. Finalmente, adicionou-se suporte a padrões com dependências sintáticas, o que permite uma identificação mais precisa de EMs. A adição dessas funcionalidades possibilita a análise de corpora muito maiores (e.g., British National Corpus, com 110 milhões de palavras) de maneira eficiente, com maior cobertura e mais flexível. Os resultados desse trabalho encontram-se descritos em [1].

Referências

[1] V. de Araújo, C. Ramisch and A. Villavicencio (2011). Fast and Flexible MWE Candidate Generation with the mwetoolkit. In *Proceedings of ACL 2011 Workshop on Multiword Expressions: from Parsing and Generation to the real world*.