Extração de informação de senso comum da Wikipédia

Daniel Feitosa, Thiago Alexandre Salgueiro Pardo
Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos
SCC, NILC
fdaniel@grad.icmc.usp.br, taspardo@icmc.usp.br

Resumo

Este artigo aborda a aceleração da alimentação de uma base de sensos comuns a partir da Wikipédia. Utilizando-se, para tal, de métodos de Processamento de Linguagem Natural (PLN) aplicados a Web.

1. Introdução

A utilização de avanços na área computacional para prover suporte ao processo de ensino-aprendizagem e atribuir a ele um caráter flexível e portável, em abordagens que combinam Aprendizagem Eletrônica (AE) e Educação à Distância (EaD) tem se tornado cada vez mais comum [1].

Dentre estes avanços está o projeto Open Mind Commom Sense (OMCS), originado do Media Lab (MIT), com o objetivo de obter uma grande base de senso comum. Isto alavancou o surgimento de diversos ferramentais baseados em senso comum, desde jogos educacionais a ambientes de edição do objetos de aprendizado.

Com seu sucesso surge uma extensão deste projeto no Brasil (OMCS-Br), que possui um sistema web de aprendizado por meio de interação humana, onde se escolhe um tema e o sistema traz sentenças para serem completadas. Estas sentenças serão os sensos comuns incorporados à base.

2. Objetivos

Apesar do Brasil possuir sua própria extensão para do projeto OMCS, é fato que sua base ainda possui uma quantidade de informação fora do desejado pelos colaboradores. A forma atual com que seu sistema coleta sensos comuns é eficaz porém faz com que sua base cresça a uma velocidade dependente de dois fatores principais: a quantidade de usuários e a quantidade de acessos destes.

Neste contexto, este projeto tem como meta acelerar, de forma eficiente, o crescimento da base da OMCS-Br.

3. Material e Métodos

A aceleração pode ser feita de duas formas: alimentando-se a base automaticamente ou fazendo com que, a cada acesso, o usuário consiga informar uma quantidade bem maior de sensos comuns do que a normalmente produzida.

Os dois métodos acima dependem de um corpus para extrair informações, quer sejam completas, quer sejam complementares às do usuário. E a utilização de um ou outro depende dos resultados encontrados.

A internet seria o corpus de busca, porém precisa-se restringir a textos bem escritos. Tentando alcançar os dois ideais, opta-se pela Wikipédia, que se acredita ser uma forte base de compartilhamento de informação.

Desta forma, precisa-se agora de um método eficiente para extração de informação desta base. Existem alguns métodos de exploração de conteúdo web já testados com sucesso [2], e este projeto visa justamente utilizar estes métodos ou utilizá-los como base para produção de um sistema de extração de sensos comuns da Wikipédia.

4. Trabalhos Correlatos

Com objetivo parecido com o deste projeto, Ruiz-Casado et al. [3] extraem relações semânticas da Wikipédia, porém eram para complementar a base da WordNet, uma rede de palavras e suas relações semânticas (por exemplo, sinonímia, antonímia, etc.). Após descrever o processo de extração, que utiliza templates, os autores mostram que obtiveram sucesso, levando-se a crer que tal abordagem também poderia ser útil na extração de sensos comuns.

Focando nos métodos de extração para web, o principal artigo que impulsionou a pesquisa foi o de Adafre et al. [4], onde são apresentadas duas formas de extração de informação: Link-Based e Content-Based. Dentre os dois métodos e um híbrido destes dois, o melhor foi o que apenas utiliza o Link-Based.

Estes dois trabalhos contribuíram principalmente para contrução da arquitetura do sitema e do método de extração.

5. Resultados e Experimentos

A arquitetura do sistema (Figura 1) se inicia com a interação do usuário com o projeto OMCS-Br, na qual o usuário preenche templates com informações sobre algum tema específico. A partir desse tema, executa-se o método de extração dos sensos comuns, que irá ou incrementar diretamente a base ou passará pela validação do usuário antes alimentar a base.

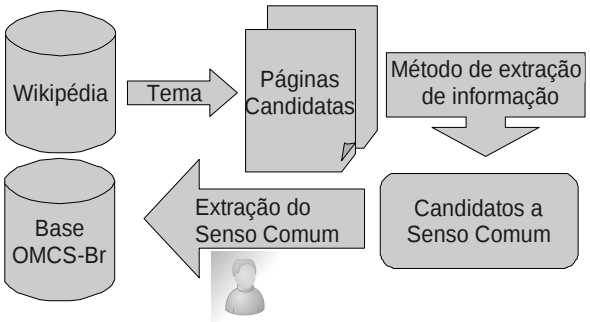

Figura 1. Arquitetura do sistema

Para acessar a Wikipédia da forma mais rápida e eficiente optou-se por manter em servidor uma cópia em formato SQL, o que causa problemas de atualização, que são menos importantes do ponto de vista do projeto.

Para o método de extração foram feitas três implemetações, que serão avaliadas na continuidade do projeto por meio de experimentos com cada implementação e versões híbridas destes, com o objetivo de encontrar o melhor método para o sistema.

6. Conclusões

Até o desenvolvimento atual, estão claras as diversas formas de extração dos candidatos a senso comum, bem como acessar a base da Wikipédia. Mesmo com a forma de acesso à base definida, este projeto pretende, até seu término, encontrar a forma mais eficiente de extrair os candidatos. Como conclusões, espera-se acelerar o crescimento da base do OMCS-Br, bem como afirmar a Wikipédia como uma base forte para obtenção de informação.

Referências

[1] Zhang D.; Zhao, J. L.; Zhou, L.; Nunamaker, J. F. (2004). Can e-learning replace classroom learning? **Communications of the ACM,** US (New York, NY): ACM Press, v. 47, n. 5, mai. 2004. p. 75-79.

[2] Jackson, P.; Moulinier, I. Natural language processing for online applications: Text retrieval, extraction, and categorization, cáp. 4 John Benjamins Publishing Co (2002).

[3] R. M. Casado, E. Alfonseca, P. Castells, *Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from Wikipedia, in Natural Language Processing and Information Systems: 10th International Conference on Applications of Natural Language to Information Systems* (2005).

[4] S. F. Adafre; V. Jijkoun; M. de Rijke; In: C. Peters et al, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval. 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Revised Selected Papers*, LNCS 4730, pages 537-540, September 2007.