IDENTIFICAÇÃO DE RELAÇÕES SEMÂNTICAS ENTRE ENTIDADES MENCIONADAS

Aluno: Andrea da Fonseca Barreto
Orientador: Violeta Quental

Introdução

O reconhecimento e categorização semântica de nomes próprios, entidades mencionadas em textos (EM), é fundamental para a análise automática de textos, por sua carga informacional elevada e devido à grande frequência com que ocorrem na língua. Sua identificação e classificação, além de possibilitar uma maior riqueza e especificidade quando da realização da análise semântica de textos, auxiliam o desempenho de uma série de tarefas (ou sistemas) que lidam com o processamento automático de uma língua, especialmente no que diz respeito à busca e recuperação de informação.

Nomes próprios não fazem parte de dicionários e sua listagem seria ilimitada, o que leva à necessidade de usar estratégias de outra ordem para seu reconhecimento – a convenção de iniciais maiúsculas, por exemplo – e classificação semântica – a observação dos ambientes sintático-semânticos em que ocorrem, o uso de métodos de aprendizagem de máquina a partir de corpora anotados manualmente.

Pretendendo incentivar o desenvolvimento de pesquisas relativas ao reconhecimento e classificação de EM na língua portuguesa, a Linguateca (http://www.linguateca.pt) propôs o evento de avaliação denominado HAREM (Avaliação e Reconhecimento de Entidades Mencionadas) [3] e, junto com a segunda versão desse evento, uma nova tarefa que foi chamada de ReRelEM (Reconhecimento de Relações entre EM) [1], com o objetivo de aprofundar a compreensão de textos em língua portuguesa, ampliando seu objeto de estudo. A avaliação não mais se limitaria à identificação e classificação de EM, mas abarcaria, também, a identificação das relações semânticas entre tais entidades. As relações semânticas que unem as Entidades Mencionadas em um texto são, de forma geral, a de identidade entre referentes e diversas nuances de relações metonímicas. Assim, reconhecida a EM "Brasil" na frase "*Apesar da mudança de discurso, Brasil deixou a Copa na mesma fase de 2006*", anotada como uma EM do tipo **pessoa/grupo** (uma equipe), reconhece-se que **inclui** a **pessoa/indivíduo** "Kaká" e, inversamente, que esse jogador é parte desse **grupo**.

Objetivos

Nossos objetivos com esse projeto foram estudar as formas de reconhecimento de Entidades Mencionadas [2] e as relações semânticas existentes entre EM; a validação e ampliação do material já anotado no evento ReRelEM, para uma melhor compreensão do relacionamento semântico entre EM em textos. A partir desse estudo, pretendia-se contribuir para a disponibilização à comunidade científica de mais material anotado, que pudesse ser útil para o treino de sistemas e a investigação de outros fenômenos que envolvam também relações semânticas entre entidades de um texto. Muitos casos de relações entre EM foram inicialmente classificados como "outras relações", pelos participantes do ReRelEM, pois não foi possível chegar a um consenso sobre a natureza dessas relações.

Metodologia

A metodologia de trabalho proposta foi, além da leitura da literatura sobre semântica lexical e sobre Reconhecimento de Entidades mencionadas (especificamente sobre essa tarefa no que se refere à língua portuguesa, conforme referências a seguir), a análise da anotação de

EM da Coleção Dourada do Segundo HAREM, disponível no endereço do evento. Pretendíamos, como contribuição nova ao material já disponibilizado, analisar as relações lexicais tratadas em conjunto como "outras" relações, propondo uma classificação apropriada a casos considerados de difícil categorizacão. Para a etiquetagem das novas relações do corpus, contamos com a utilização da ferramenta Etiquet(H)AREM, desenvolvida pela equipe do Polo de Coimbra da Linguateca. A anotação do corpus com essas etiquetas seria então avaliada pela organização do evento ReRelEM.

Conclusão

Consideramos que os objetivos propostos foram parcialmente atingidos, já que não foram propostas novas categorizações para as relações entre EM classificadas inicialmente como "outras relações", que permanecem ainda em discussão. Consideramos, no entanto, a pesquisa produtiva, tendo proporcionado uma melhor compreensão dos fenômenos semânticos analisados e das formas de tratamento computacional de nomes próprios.

Referências:

1. FREITAS, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho e Cristina Mota. «Relações semânticas do ReRelEM: além das entidades no Segundo HAREM», p. 77-96. In: Cristina Mota & Diana Santos (eds.). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008. http://www.linguateca.pt/LivroSegundoHAREM/. (ISBN: 978-989-20-1656-6)

2. MOTA, Cristina & Diana Santos (eds.). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008. http://www.linguateca.pt/LivroSegundoHAREM/. (ISBN: 978-989-20-1656-6)

3. SANTOS, Diana & Nuno Cardoso (eds.) *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 2007. ISBN: 978-989-20-0731-1 (versão electrónica); ISBN 978-989-20-1297-1 (versão em papel, Lisboa, 2008)