RELAÇÕES SEMÂNTICAS ENTRE ENTIDADES MENCIONADAS

Aluna: Jaqueline Xavier da Silva
Orientadora: Prof. Dra.Violeta de San Tiago Dantas Barbosa Quental

Introdução

A teoria linguística geralmente considera os nomes próprios como um fenômeno periférico da língua, entretanto sua classificação e identificação são de extrema importância para sistemas que trabalham com o processamento automático de uma língua, como, por exemplo, sistemas de extração de informação, de resposta automática a perguntas ou de análise de sentimento em textos (sentiment analysis).

O termo "entidades mencionadas" (EM), no âmbito do processamento automático de linguagem natural (PLN), é a adaptação do conceito "names entities" e pode ser traduzido para "entidades com nomes próprios" (Santos, 2008). As EM são instâncias de classes ontológicas, possuem alto poder de informação e por isso seu reconhecimento é fundamental para extração de informação em textos. Um sistema, ao se submeter a analisar ou compreender um texto, procura por informações específicas e não por generalidades.

A classificação e identificação de EM não é uma tarefa simples e apresenta dificuldades expressivas. Como uma das dificuldades, podemos apontar a vagueza da língua. A vagueza, na esfera da classificação de EM ocorre quando podemos atribuir mais de uma classificação a uma entidade. A entidade "Brasil", por exemplo, pode significar o nome de país, de equipe esportiva, de povo, de empresa. A classificação desta entidade só pode ser definida e identificada pelo contexto.

Reconhecer a importância de EM para sistemas de processamento automático da língua não é assunto novo. No MUC (Message Understanding Conference), criado em 1987, estudava-se o reconhecimento de EM correspondentes a três conceitos gerais: pessoas (person), organizações (organization) e locais (location) (Santos, 2008). O objetivo do MUC era reconhecer as EM e classificá-las em uma dessas três categorias, a partir de textos em inglês.

Após o MUC, os sistemas começaram a se sofisticar, propondo-se a análises mais complexas. O ACE (Automatic Content Extration), criado em 1999, procurava identificar todas as entidades, não apenas os nomes próprios (Santos, 2008). Buscava-se analisar as entidades semanticamente, a partir do conteúdo, enquanto que o MUC analisava lingüisticamente, através da forma.

Em 2006 foi criado o HAREM (Avaliação e Reconhecimento de Entidades Mencionadas), com o propósito de identificar e classificar automaticamente os nomes próprios em categorias previamente definidas, mas a partir de textos escritos em português (Santos & Cardoso, 2008). Dois anos depois foi apresentada uma nova tarefa (ou sistema), o ReRelEM (Reconhecimento de Relações entre Entidades Mencionadas), com o objetivo de identificar as relações semânticas entre as EM no HAREM (Freitas et al., 2009).

Objetivo

O projeto tem como objetivo ampliar o material anotado no ReRelEM, permitindo assim uma avaliação mais consistente das categorias inicialmente propostas para a classificação das relações semânticas entre as entidades mencionadas. Pretende-se anotar as

relações entre EM já indicadas pelo grupo organizador do HAREM/ ReRelEM nos textos da Coleção Dourada do HAREM.

Metodologia

No ReRelEM de 2008, a partir dos textos disponíveis na Coleção Dourada do HAREM, foram analisadas e anotadas algumas relações semânticas entre entidades mencionadas. As anotações foram feitas com o apoio do Etiquet(h)arem, um sistema de auxílio à etiquetagem de EM e de relações entre EM.

As relações entre EM analisadas no escopo da avaliação de 2008 foram: identidade (ident), inclusão (inclui/incluído), localização (ocorre_em/ sede_de) e "outra". A categoria "outra" abrange todas as outras relações consideradas relevantes, mas que não correspondem a nenhuma das categorias citadas anteriormente e que, portanto, não foram anotadas. Até o momento são 22 as relações classificadas como "outras": natural_de, povo_de, residente_em, vínculo_institucional, relação_profissional, relação_familiar, autor_de, produtor_de, proprietário_de, datado_de, causa_de, outra_edição, representante_de, praticado_em, participante_em, nome_de, data_nascimento, data_morte, período_vida, personagem_de, localizado_em, e outra_relação.

A proposta de pesquisa é a de, inicialmente, etiquetar os textos da Coleção Dourada do HAREM com essas relações denominadas "outras", também usando para isso o programa Etiquet(h)arem. Para isso, estamos, no momento, estudando a bibliografia relativa ao reconhecimento de entidades mencionadas, e começando a treinar a utilização do etiquetador. A etiquetagem proposta será avaliada e, a partir da revisão dos resultados, estes serão divulgados para a organização do HAREM/ ReRelEM.

Conclusão

A Coleção Dourada do ReRelEM contém um conjunto de 12 textos, 4417 palavras, 573 entidades mencionadas e 614 relações manualmente anotadas, o que ainda é, sem dúvida, um corpus pequeno para generalizações acerca das relações semânticas entre entidades mencionadas. A continuação do trabalho possibilitará mais generalizações e a legitimação das opções tomadas, assim como uma anotação mais precisa e segura.

Referências:

FREITAS, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho & Cristina Mota Relações semânticas do ReRelEM: além das entidades no Segundo HAREM. In: SANTOS, Diana e CARDOSO, Nuno, (ed.) **Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na** área. 2 ed. Linguateca, 2008.

SANTOS, Diana; CARDOSO, Nuno. Breve introdução ao HAREM. In: SANTOS, Diana e CARDOSO, Nuno, (ed.) **Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área**. 2 ed. Linguateca, 2008.

SANTOS, Diana. O modelo semântico usado no Primeiro HAREM. In: SANTOS, Diana e CARDOSO, Nuno, (ed.) **Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área**. 2 ed. Linguateca, 2008.