IDENTIFICAÇÃO DE RELAÇÕES SEMÂNTICAS ENTRE ENTIDADES MENCIONADAS

Aluna: Andrea da Fonseca Barreto
Orientadora: Prof. Dra.Violeta de San Tiago Dantas Barbosa Quental

Introdução

A área de pesquisa na qual se insere esse trabalho investiga fenômenos relacionados ao léxico do português, para aplicação em sistemas computacionais de processamento automático da língua.

O termo "entidades mencionadas" (EM), no âmbito do processamento automático de linguagem natural (PLN), é a adaptação do conceito "named entities" e pode ser compreendido como referente a entidades expressas em textos através de nomes próprios (Santos, 2008). As EM são instâncias de classes ontológicas que possuem alto poder de informação e por isso seu reconhecimento é fundamental para extração de informação em textos. Um sistema computacional de busca em princípio busca por informações específicas e não por generalidades, e muitas dessas informações são relacionadas a nomes de entidades.

A teoria linguística geralmente considera nomes próprios como um fenômeno menos importante na gramática da língua, mas sua identificação e classificação são de grande importância para sistemas que trabalham com o processamento automático de uma língua, como, por exemplo, sistemas de extração de informação ou sistemas de diálogo.

A identificação e classificação semântica de EM apresenta dificuldades expressivas, como se observa, por exemplo, com a entidade "Brasil" nos exemplos a seguir:

1.a. O Rio de Janeiro foi capital do Brasil.
1.b. O Brasil apresentou proposta conciliadora na reunião de ontem.
1.c. O Brasil jogou mal na Copa.
1.d. O Brasil não se considera racista.

Nos exemplos, o termo Brasil pode estar se referindo a nome de país, de equipe esportiva, de povo, de instituição governamental, e a classificação semântica desta entidade só pode ser definida dentro de um 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, em textos do inglês.

Em 2006 foi criado o evento de avaliação HAREM (Avaliação e Reconhecimento de Entidades Mencionadas), com o propósito de incentivar o desenvolvimento de sistemas voltados para a tarefa de identificar e classificar automaticamente nomes próprios em categorias previamente definidas, em textos escritos em português (Santos & Cardoso, 2008). Dois anos depois, foi realizado um Segundo HAREM, que manteve a filosofia do primeiro, no tocante ao modelo semântico e ao modelo geral de avaliação, mas que passou a incluir duas novas tarefas de pesquisa: i) o reconhecimento e a padronização de expressões temporais; e ii) o reconhecimento de relações semânticas entre EM (o ReRelEM).

Segundo o modelo do HAREM, a tipificação de uma EM só pode ser feita em uma situação de uso concreto da língua, ou seja, dentro de um determinado contexto, como pode ser observado a partir do exemplo acima mencionado, em que a EM 'Brasil' pode fazer referência a uma gama variada de sentidos.

Da mesma forma, o ReRelEM, como tarefa dependente e integrada do HAREM, propôs uma anotação que considera o valor semântico das relações entre EM apenas quando inseridas em um contexto. Após a análise de textos com suas EM identificadas, buscou-se verificar que relações semânticas existem entre essas EM, de modo a reconhecer a cadeia referencial de um texto. Foram estabelecidas as seguintes relações entre EM: **Identidade (ident)**, **Inclusão (inclui/incluído)**, **Localização (ocorre_em/ sede_de)**, e **Outra** (que engloba todas as relações que não corresponderam a nenhum dos tipos anteriormente citados, mas que foram consideradas relevantes e que perfazem um total de 22 'outras relações').

A relação de Identidade estabelece-se entre EM que designem a mesma entidade, ou seja, expressões textuais formalmente idênticas e que tenham a mesma classificação semântica; expressões textuais que são resultado de transformações lexicais, mas que designam a mesma entidade; e também abreviaturas, acrônimos, traduções ou 'nomes alternativos'. A título de exemplo, podemos apontar a relação de identidade existente entre as entidades *Cidade Maravilhosa / Rio de Janeiro*.

A relação de Inclusão é estabelecida entre EM quando a entidade descrita por uma EM inclui a entidade descrita por outra. Nesse caso, a relação entre essas duas EM é marcada como 'inclui'. Quando a relação for inversa (uma EM está incluída em uma entidade descrita por outra), é marcada como 'incluido'. Assim, por exemplo, Brasil_incluído_Países Emergentes / Rio de Janeiro_inclui_Gávea.

A relação de Localização indica a localização espacial de um evento ou de uma organização. Exemplo: Em 7 de setembro de 2008, foi realizado em *Aveiro* o encontro do *Segundo Harem*.

A relação Outra, que indica relações não contempladas no elenco acima, é tarefa altamente subjetiva, já que supõe conhecimento lingüístico, conhecimento enciclopédico e conhecimento de mundo. A análise das relações do tipo Outra levou a um total de 22 sub-categorias, a saber: natural_de, povo_de, residente_de, 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_de_nascimento, data_da_morte, período_de_vida, personagem_de, localizada_em, e outra_relação. Tais sub-categorias permitiram criar um recurso semântico mais rico e informativo para servir de base a outros estudos e aplicações futuras.

Outra definição feita pelo HAREM foi a proposta de assumir a existência de vagueza em algumas EM. A vagueza se caracterizaria quando uma mesma EM representar, em um mesmo contexto, mais do que uma das classes semânticas pré-definidas no modelo de classificação. Nesses casos, uma opção seria anotar mais de uma classe para uma EM.

Objetivo

O projeto teve como objetivo validar e ampliar o material anotado no ReRelEM (Freitas et al, 2008), permitindo uma avaliação mais consistente das relações semânticas entre as entidades mencionadas inicialmente propostas pela equipe responsável. Propusemo-nos, para isso, a rever as anotações da coleção de textos usada no ReRelEM e posteriormente, se fosse possível, anotar

as relações entre EM nos textos da Coleção Dourada[1] do HAREM. Com isso, a coleção de textos anotados com relações entre entidades tornar-se-ia maior e, possivelmente, novas relações poderiam ser adicionadas ao conjunto já identificado.

Metodologia

Esse relatório apresenta inicialmente as atividades desenvolvidas pela bolsista anterior, substituída quando se formou, Jaqueline Xavier.

Foram por ela desenvolvidas as seguintes atividades:

- leitura de bibliografia relativa a relações semânticas e reconhecimento de entidades mencionadas;
- familiarização com o Etiquet(h)arem[2] – uma ferramenta de auxílio à anotação de EM e de relações semânticas entre EM;
- familiarização com as relações semânticas propostas no ReRelEM;
- familiarização com o formato de anotação em linguagem XML.

A partir daí, foram reanotados alguns textos da Coleção Dourada do ReRelEM (um subconjunto da Coleção Dourada do Segundo HAREM), tendo também em vista a possível detecção de novas relações que não fizeram parte do ReRelEM.

A Coleção Dourada do ReRelEM é composta de 12 textos, com 4417 palavras, 573 entidades mencionadas e 614 relações manualmente anotadas, que seriam revistos durante o primeiro semestre de 2010. Além desses textos, previa-se também a anotação de outros textos da Coleção Dourada do HAREM. Esse último objetivo não foi cumprido, dada a troca de bolsista, e a necessidade de retomar as etapas de familiarização com o tema e a metodologia de trabalho.

Durante o primeiro semestre de 2010, então, foi possível apenas retomar as leituras e rever o trabalho desenvolvido pela bolsista anterior.

Resultados

A Coleção do ReRelEM é um corpus pequeno para generalizações acerca das relações semânticas entre entidades mencionadas e sua ampliação é necessária. Com a revisão dos textos já anotados e com a análise e etiquetagem de mais textos, pretendia-se caracterizar outras relações, que seriam avaliadas e discutidas com a organização do HAREM/ ReRelEM. A partir da análise dos textos da CD do ReRelEM, percebemos a existência de uma relação entre entidades que não tinha sido classificada: a relação *idade_de*, entre as categorias PESSOA e VALOR-QUANTIDADE. As outras relações idenificadas não foram alteradas.

Não foi possível anotar mais textos, dada a escassez de tempo da bolsista atual.

---

[1] Coleção de textos anotados e revistos manualmente, em que estão marcadas as entidades mencionadas e as categorias semânticas a que pertencem. Essa coleção serve de base de comparação para o desempenho dos sistemas participantes do HAREM.

[2] Disponível em: http://www.linguateca.pt/HAREM/

Referências:

BICK, E. **The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework** . Dinamarca: Aarhus University Press, 2000.

GARRAO, Milena de Uzeda. **O corpus não mente jamais: sobre a identificação e duso de combinações multivocabulares do tipo verbo mais sintagma nominal** / Milena de Uzeda Garrão. Tese de Doutorado. Rio de Janeiro: PUC, Departamento de Letras, 2006.

OLIVEIRA, C. ; FREITAS, M. C. ; QUENTAL, V. ; SANTOS, C. N. ; LEME, R. ; SOUZA, L. . *A Set of NP-extraction rules for Portuguese: defining and learning*. In: **7th Workshop on Computational Processing of Written and Spoken Portuguese, 2006, Itatiaia. Computational Processing of the Portuguese Language**. Berlin: Springer, 2006. p. 150-159.

OLIVEIRA, C; GARRÃO, M.; AMARAL, L. *Recognizing Complex Preposition Prep+N+Prep as Negative Patterns in Automatic Term Extrection from Texts*. In: **Proceedings of 1 st Workshop em Tecnologia da Informação e da Linguagem Humana (TIL 2003)**. São Carlos – SP. 2003.

OLIVEIRA, C; FREITAS, C. *Classes de palavras e etiquetagem na Lingüística Computacional***.** In: **Calidoscópio,**Vol. 4, n. 3 , p. 179-188, set/dez 2006

FREITAS, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho & Cristina Mota *Relações semânticas do ReRelEM: além das entidades no Segundo HAREM*. In: MOTA, Cristina & SANTOS, Diana (eds.). **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM.** Linguateca, 2008.

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

CARVALHO, Paula; OLIVEIRA, Hugo Gonçalo. **Manual de Utilização do Etiquet(h)arem**. Disponível em: http://www.linguateca.pt/aval_conjunta/HAREM/ManualUtilEtiquetHAREM.pdf Acesso: 06/06/2009.

Coleção Dourada do Segundo HAREM/ReRelEM. Disponível em: http://www.linguateca.pt/aval_conjunta/HAREM/CDSegundoHAREM.xml. Acesso: 06/06/2009.