Departamento de Letras

RELAÇÕES SEMÂNTICAS ENTRE ENTIDADES MENCIONADAS

Aluna: Jaqueline Xavier da Silva
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 projeto de que participei inicialmente como bolsista, intitulado Elaboração de Dicionário Eletrônico, tinha como um de seus objetivos a análise das expressões prepositivas do português que podem ser consideradas expressões multivocabulares com grau de fixidez suficiente para serem incorporadas a um léxico eletrônico. Esse projeto foi desenvolvido em parceria com o Projeto VISL (Visual Interactive Syntax Learning), da Southern Denmark University, através do Prof. Eckhard Bick, e seus resultados devem ser incorporados ao dicionário do parser Palavras (Bick, 2000).

Expressões multivocabulares (EMVs) são formadas por mais de um item lexical (N+PREP ou N+PREP+N) cuja co-ocorrência é significativa em um corpus. A decisão de análise de expressões prepositivas como EMVs é realizada a partir de coleta dessas expressões em seus contextos de ocorrência e análise linguística e estatística de sua relevância como multivocábulo. Quando submetidas a um parsing, são analisadas como um vocábulo único, por economia de processamento do parser e por corresponder mais adequadamente ao fenômeno da composição lexical.

Em relação a esse projeto, de janeiro a abril de 2009, meu trabalho foi o de compatibilizar várias listas de buscas por "concordância" (palavras em contexto) de expressões preposicionadas multivocabulares, realizadas pelo bolsista anterior, e organizá-las por função sintática. Para isso, foi necessária a leitura de alguns textos básicos sobre o tratamento computacional do léxico e sobre expressões multivocabulares. Os resultados da pesquisa estão sendo organizados pela Profa. Orientadora para posterior dicionarização.

Em meados de abril de 2009, findo o projeto "Elaboração de Dicionário Eletrônico", foi solicitada sua substituição pelo projeto "Relações semânticas entre entidades mencionadas", em andamento, abordando questões de natureza semântica relacionadas ao léxico.

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 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 de busca pode procurar 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 classificação e identificação de EM apresenta dificuldades expressivas. A entidade "Brasil", por exemplo, pode significar o nome de país, de equipe esportiva, de povo, de instituição governamental. 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, em textos do inglês.

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, em textos escritos em português (Santos & Cardoso, 2008). Dois anos depois foi apresentada, junto com o Segundo HAREM, uma nova tarefa (ou sistema), o ReRelEM (Reconhecimento de Relações entre Entidades Mencionadas), com o objetivo de identificar as relações semânticas entre as EM reconhecidas no HAREM (Freitas et al., 2008).

Objetivo

O projeto atual tem como objetivo validar e, se possível, ampliar o material anotado no ReRelEM (Freitas et al, 2008), permitindo uma avaliação mais consistente das relações semânticas inicialmente propostas entre as entidades mencionadas. Pretende-se, inicialmente, rever as anotações da coleção de textos usada no ReRelEM e posteriormente, se 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 torna-se-ia maior e, possivelmente novas relações poderiam ser adicionadas ao conjunto já identificado.

Metodologia

Em relação ao primeiro momento de pesquisa como bolsista de iniciação científica, ainda no projeto Elaboração de Dicionário Eletrônico, foram lidos textos sobre identificação e uso de combinações multivocabulares (Garrão, 2006), extração de sintagmas nominais a partir de corpus (Oliveira, C. et al, 2006; Qliveira & Quental, s/d), regras para extração de sintagmas nominais (Oliveira et al, 2006), reconhecimento de locuções prepositivas (Oliveira, Garrão e Amaral, 2003), dentre outros. Foram organizados 1310 arquivos de locuções preposicionadas, eliminando e consolidando os arquivos repetidos. Por fim, foi elaborada uma lista de expressões multivocabulares retiradas de textos selecionados pela orientadora.

A partir da mudança de projeto de pesquisa, o foco do trabalho passou a ser o estudo das relações semânticas entre entidades mencionadas. Até o momento, foram 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;

---

[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/

- familiarização com o formato de anotação em linguagem XML.

Para tanto, estou reanotando 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 serão revistos durante o próximo semestre. Além desses textos, prevê-se também a anotação de outros textos da Coleção Dourada do HAREM.

Conclusão

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, é possível que surjam outras relações, que serão avaliadas e discutidas com a organização do HAREM/ ReRelEM. No primeiro contato com alguns textos da CD do ReRelEM, percebemos a existência de uma relação entre entidades que não foi classificada: a relação *idade_de*, entre as categorias PESSOA e VALOR-QUANTIDADE .

Espera-se que a continuação do trabalho possibilite mais generalizações e a legitimação das opções tomadas, assim como uma anotação mais precisa e segura.

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.