AQUISIÇÃO SEMI-AUTOMÁTICA DE INFORMAÇÕES LINGÜÍSTICAS: EXPRESSÕES MULTIPALAVRAS E FRASEOLOGIAS

Daniel Emilio Beck, Aline Villavicencio
Instituto de Informática
Universidade Federal do Rio Grande do Sul
{debeck, avillavicencio}@inf.ufrgs.br

Introdução

O Processamento da Linguagem Natural (PLN) é área que trata da interpretação e manipulação automática de palavras e textos escritos na língua natural. É uma área multidisciplinar, tendo relações com a Inteligência Artificial e com a Linguística, por exemplo, e possui aplicações em áreas como Recuperação de Informação, Processamento da Fala e Educação a Distância (EAD).

Dentro dos Grandes Desafios propostos pela SBC no período de 2006 a 2016 [3], está o acesso participativo e universal do cidadão brasileiro ao conhecimento, que está diretamente relacionado à EAD. Nesse enfoque, desenvolver a área de PLN possui grande importância, pois possibilita um maior acesso à informação contida na web utilizando ferramentas como tradução automática, pesquisas por conteúdo, etc.

Recursos lingüísticos como dicionários e ontologias são essenciais para o desenvolvimento e utilização dessas ferramentas [1]. Nesse âmbito, a detecção de expressões multipalavras através de corpora é fundamental para garantir que a abrangência desses recursos seja a mais completa e correta possível, pois estima-se que o total delas seja equivalente ao total de palavras simples de uma língua (no caso do WordNet, um dicionário para o inglês, 41% das entradas são de expressões compostas [2]).

Objetivos

A idéia é criar um ambiente na forma de uma página web que forneça ferramentas de PLN, com enfoque na extração de expressões compostas. A página seria utilizada inicialmente para a realização de trabalhos nas disciplinas de Tópicos Especiais em Computação: Processamento de Linguagens Naturais

(oferecida no Instituto de Informática) e Sintaxe do Texto (oferecida no Instituto de Letras), mas poderia ser expandida para outras disciplinas (como Linguagens Formais, do Instituto de Informática) à medida que sejam acrescentadas novas ferramentas com o passar do tempo.

Com esse ambiente, os alunos das disciplinas poderão acessar a página e utilizar as ferramentas disponibilizadas para realizar seus próprios trabalhos práticos de criação de recursos lingüísticos. Além disso, ela também poderá ser usada como forma de comunicação entre os alunos e os professores, permitindo a submissão de trabalhos, por exemplo.

Metodologia

Para a criação de página, será dada preferência para a utilização de uma plataforma de EAD já pronta (como o Moodle[4], por exemplo). Essas plataformas geralmente fornecem um esqueleto básico para a criação do ambiente de EAD, tornando mais dinâmico o processo de elaboração da página. Além disso, a plataforma utilizada será de código aberto, pois caso alguma funcionalidade necessária não seja fornecida, ela poderá ser criada e acoplada ao esqueleto.

As ferramentas serão disponibilizadas para serem utilizadas diretamente na página ou para download, sendo acessadas através de uma conta criada para cada aluno. Essa mesma conta também será utilizada para a submissão dos trabalhos das disciplinas. Também os alunos poderão contribuir com a página melhorando as ferramentas já existentes ou implementando outras.

Uma lista preliminar das ferramentas a serem colocadas a disposição é relatada a seguir:

- Corpora
- Pesquisa de multipalavras na web
- Pluralizador de substantivos (inglês)
- Part-of-speech tagger
- Ferramentas para análise estatística dos dados

Ressalta-se que essas ferramentas serão disponibilizadas de forma a serem o mais amigável possível para o usuário, sendo que aquelas a serem

utilizadas offline serão acompanhadas de uma GUI (Graphical User Interface). Elas terão ainda instruções detalhadas de utilização e exemplos disponíveis no site.

Resultados

Espera-se que o projeto facilite o aprendizado da disciplina, disponibilizando um ambiente onde os alunos podem experimentar na prática os conteúdos teóricos apresentados em aula. Este projeto ainda promove um ambiente centralizado com as ferramentas necessárias para realizar os trabalhos práticos dessas disciplinas. Também espera-se que os professores sejam beneficiados, pois facilitará a correção dos trabalhos dos alunos, já que estes estarão disponíveis neste ambiente.

Um outro aspecto importante é que nem todas as ferramentas da página serão dependentes de linguagem. Isso possibilita que sejam feitos trabalhos para a língua portuguesa, facilitando a criação de novos recursos para essa língua e a ampliação dos já existentes. No futuro, espera-se que a página forneça mais ferramentas para o português, à medida que essas sejam implementadas (o que poderia ser feito inclusive pelos alunos das disciplinas).

Conclusões

O projeto, junto com as disciplinas, ajudará a divulgar, explicar e expandir conhecimentos da área de PLN, que, como foi vista, é crucial para o desenvolvimento de melhorias no acesso à informação e ao conhecimento e, portanto, ajudará a gerar novos conhecimentos e tecnologias, além de promover a cidadania, a educação e a inclusão digital.

Também o trabalho trará contribuições para a EAD, pois para a concretização do projeto, poderá ser necessária a criação de melhorias ou modificações de alguma plataforma específica dessa área. Além disto todos os recursos e tecnologias desenvolvidos ficarão disponíveis para a comunidade em geral, aumentando a variedade e foco do material de EAD. Dessa forma, espera-se também que o projeto traga experiências novas que ajudem a criar novas implementações e melhorar as já existentes.

Palavras-chave

processamento da linguagem natural, expressões multipalavras, lingüística, aquisição de palavras, dicionários, ontologias

Bibliografia

[1]Strube de Lima, V. L., Nunes, M.G.V., Vieira, R. (2007) Desafios do Processamento de Línguas Naturais. In: Anais do XXVII Congresso da SBC, pp. 2202-2216

[2]Sag, I. A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. (2002) Multiword Expressions: A Pain in the Neck for NLP. Computational Linguistics and Intelligent Text Processing, pp. 189-206

[3]Grandes Desafios da Pesquisa em Computação no Brasil – 2006 – 2016. Relatório sobre o Seminário realizado em 8 e 9 de maio de 2006.

[4]http://moodle.org – Página oficial da ferramenta Moodle.