Área: CV ( ) CHSA ( ) ECET (X)

DICIONÁRIO ELETRÔNICO DA LÍNGUA PORTUGUESA PARA APOIAR PESQUISAS DO LABORATÓRIO DE ENGENHARIA DE *SOFTWARE* E INFORMÁTICA INDUSTRIAL (EASII)

Francisco Assis Ricarte Neto (Iniciação Científica Voluntária - UFPI), Raimundo Santos Moura (Orientador, Depto. Informática e Estatística - UFPI)

Introdução

No contexto dos grandes desafios da pesquisa em computação no Brasil, o artigo **Desafios do Processamento de Línguas Naturais** [6] destaca o uso do português como facilitador do acesso à informação digital, onde os sistemas de busca e recuperação de documentos ou informações a partir de padrões textuais são bastante populares, principalmente devido a Web. Ele destaca também a extração de conhecimento de texto não estruturado e a transformação em conhecimento estruturado como foco central da área de extração da informação.

De acordo com o mapeamento da área de Processamento de Linguagem Natural (PLN) [3], o segundo maior desafio enfrentado pelos pesquisadores da área foi à ausência de recursos básicos de qualidade para o português (por exemplo, *corpus* anotado ou não, um bom *parser*, *wordnet*), perdendo apenas para a falta de financiamento de projetos.

No nosso entendimento, para alavancar pesquisas que usam o PLN como meio, sobretudo para a criação de recursos linguístico-computacionais e recuperação e extração de informação é de suma importância a criação de um dicionário eletrônico de palavras.

De acordo com M. C. Rosa, em [5], quanto à possibilidade de gerar vocabulário, as palavras podem ser classificadas em: i) **palavras de classe aberta** ou palavras de conteúdo e ii) **palavras de classe fechada** ou palavras funcionais. De um lado, as palavras de classe aberta apresentam significado lexical; essas classes, em princípio, sempre podem ser acrescentadas novas palavras. São exemplos: os substantivos, os adjetivos, os verbos, os advérbios e os numerais. Por outro lado, as palavras de classe fechada apresentam um significado gramatical e são índices de propriedades gramaticais que fazem a diferença entre as línguas. São exemplos: os artigos, as preposições, as conjunções, os pronomes e as interjeições.

Metodologia

Para realização desse trabalho as seguintes etapas foram propostas: Levantamento Bibliográfico, Levantamento de Requisitos, Implementação, e Experimentos. A etapa de Levantamento Bibliográfico foi dedicada para o estudo de *corpora* linguísticos, dicionários eletrônicos, e classificadores de palavras. A etapa do Levantamento de Requisitos foi reservada para a identificação dos requisitos necessários para o desenvolvimento do dicionário, e suas ferramentas. A fase de Implementação foi dedicada para a construção do dicionário eletrônico, do corretor ortográfico, do etiquetador de frases capaz de encontrar

elementos básicos de descrição textual, e por fim um protótipo para gerar a conjugação verbal dos principais verbos irregulares e verbos da primeira, segunda e terceira conjugação, i.e., infinitivo em AR, ER E IR, respectivamente. A etapa de Experimentos foi reservada para a realização de testes de desempenho e completude das ferramentas geradas.

Resultados e Discussão

Para a criação do dicionário eletrônico de palavras a partir de *corpora* linguísticos, foi necessário fazer uma seleção de material disponível na Linguateca que é **centro de recursos -- distribuído -- para o processamento computacional da língua portuguesa** (ver site http://www.linguateca.pt). Os *corpora* BosqueFolha e MAC-MORPHO foram selecionados, juntamente com o TeP 2.0 – ***Thesaurus Eletrônico***, que é uma base de dados lexical onde é possível consultar sinônimos e antônimos das palavras pertencentes a essa base.

A partir desses materiais criou-se um *corpus*, onde foram retiradas palavras repetidas, e adotou-se um padrão na referência às classes gramaticais das palavras, que eram diferentes entre os *corpora*. De posse desse *corpus*, o passo seguinte foi desenvolver um algoritmo capaz de capturar as palavras, e salvá-las em um banco de dados de acordo com a estrutura do dicionário (ver Figura 1). O algoritmo tem como entrada um arquivo texto (extensão ".txt"), com as palavras e as classes gramaticais, separadas pelo caractere '_ ', i.e., "*palavra_classe gramatical*"; Os passos consistem em: i) identificar as palavras; ii) separá-las em radical e sufixo; e iii) persisti-las na base de dados. A base de sufixos é pré-definida no programa, e foi criada durante fase do Levantamento de Requisitos. Para a adição dos verbos, o algoritmo possui uma variável que representa a quantidade de comparações entre o radical + terminações verbais, com as palavras do *corpus* que são da classe verbo. Essa comparação é necessária para tentar garantir que o algoritmo adicione somente verbos regulares ao banco de dados. O programa não se aplica para o caso dos verbos irregulares, pois não é possível definir um radical para todas as conjugações, vez que eles variam entre os tempos verbais. Tais verbos foram adicionados manualmente.

Com o protótipo do dicionário construído, foi possível criar as ferramentas de conjugação verbal, etiquetador de frases, e uma interface onde usuários podem utilizar essas ferramentas e colaborar com o projeto, adicionando novos vocábulos. Para a criação desta interface foi utilizado VRaptor (http://vraptor.caelum.com.br/), que é um framework de desenvolvimento Web ágil, que segue os padrões MVC, e que é compatível com a linguagem de programação Java. Para persistência de dados, utilizou-se o framework Hibernate (http://www.hibernate.org/), aliado ao banco de dados MySQL (http://www.mysql.com/). Todas essas ferramentas são *open source*, ou seja, são ferramentas liberadas gratuitamente para o uso, e algumas delas disponibilizam até seu código fonte.

Área: CV ( ) CHSA ( ) ECET (X)

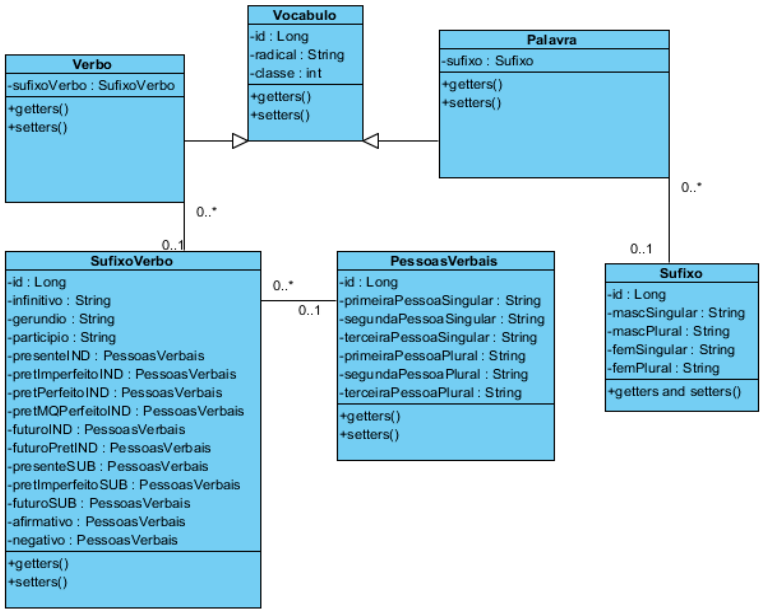

Figura 1. Diagrama UML das classes do Dicionário

Conclusão

É importante ressaltar a grandeza desse trabalho tanto para projetos futuros que serão realizados no Laboratório EaSII, como para a comunidade de linguística da UFPI e de outras instituições brasileiras. Para realização de outros trabalhos na área de PLN, como por exemplo, um tradutor automático e ferramentas de auxílio à escrita, é imprescindível um dicionário o mais completo possível. Atualmente o dicionário consta com aproximadamente 9.000 raízes de palavras, que deve passar por um processo de revisão para verificar se estão corretas.

Como trabalhos futuros destacam-se a inclusão de novas categorias associadas às palavras do dicionário. A partir dessas categorias, será possível verificar a característica semântica de uma palavra de acordo com o contexto que ela estiver inserida. Essas novas categorias poderão ser também referências às palavras de outros idiomas.

Referências Bibliográficas

[1] AIRES, Rachel Virgínia Xavier. **Implementação, Adaptação, Combinação e Avaliação de Etiquetadores para o Português do Brasil**. Dissertação de Mestrado – USP – São Carlos. São Carlos – SP, 2000.

[2] MOURA, R. S.; LINS, R. D.; CAMELO, H. A. L. **Um SOS para a Língua Portuguesa**. In: IV PROPOR, 1999, Évora - Portugal, 1999.

[3] PRADO, T.A.S.; CASELI, H.M. & NUNES, M.G.V. **Mapeamento da Comunidade Brasileira de Processamento de Línguas Naturais**.

[4] RATNAPARKHI, A. **A Maximum Entropy Part-Of-Speech Tagger**. In: Proc. Of the Empirical Methods in Naturalk Language Processing Conference, University of Pennsylvania, 1996.

[5] ROSA, Maria Carlota. **Introdução a Morfologia**. 3ª Edição. São Paulo : Contexto, 2003.

[6] STRUBE DE LIMA, V.L.; NUNES, M.G.V. & VIEIRA, R. **Desafios do Processamento de Línguas Naturais**. SEMISH 2007.

Palavras-chave: Processamento de Linguagem Natural. Dicionário Eletrônico de Palavras. Corpus.