Área: CV ( ) CHSA ( ) ECET (X)

IDENTIFICAÇÃO DE RELAÇÕES SEMÂNTICAS EM DESCRIÇÕES TEXTUAIS DE REQUISITOS DE SOFTWARE

Rogério Figueredo de Sousa (Bolsista PIBIC/UFPI), Raimundo Santos Moura (Orientador, Departamento de Informática e Estatística/UFPI)

Introdução

A Engenharia de Requisitos é a fase inicial do processo de desenvolvimento de *software, e* é considerada por inúmeros autores como a fase mais crítica do processo de construção de software. É nessa fase que ocorre a maior proximidade entre os profissionais responsáveis pela construção desses sistemas e os profissionais da área de conhecimento do problema.

Segundo KOTONYA, G. & SOMMERVILLE, I. (1998) a engenharia de requisitos engloba todas as atividades envolvidas na descoberta, documentação e manutenção de um conjunto de requisitos para um sistema computacional. Durante esta fase, são gerados os insumos e coletadas as informações necessárias para a continuidade do processo de desenvolvimento, e tais informações são condensadas em um documento conhecido por Especificação de Requisitos.

De acordo com uma pesquisa online realizada pela Universidade de Trento na Itália (BECK, K. et al., 2001), 79% dos documentos de requisitos são escritos em linguagem natural comum e dentro desse contexto ANCHIÊTA R. T. et al. (2010) implementou uma ferramenta para identificar elementos da UML, tais como classes, atributos e métodos, usando apenas análise sintática das palavras de uma determinada descrição informal.

A identificação de elementos UML utilizando apenas a análise sintática das palavras possui algumas limitações como: geração excessiva de elementos, ausência do uso do significado das palavras, duplicação de palavras que exprimem a mesma ideia.

Conhecendo tais limitações, o presente projeto incorpora um novo nível na extração de informações de descrições textuais da língua portuguesa para a geração de diagramas UML, incluindo assim, as relações semânticas entre os principais termos nas descrições e o uso de técnicas estatísticas para essa análise. A ideia é ao analisar uma descrição textual de requisitos, coletar as frequências de palavras simples e de grupos de palavras e as relações de sinonímia entre todas as palavras da descrição; em seguida, usar tais informações como pré-processamento da geração dos elementos UML possíveis na Especificação de Requisitos. Essas características foram incorporadas a ferramenta desenvolvida por ANCHIÊTA R. T. et al. (2010), gerando assim um novo protótipo.

Metodologia

Para o desenvolvimento do protótipo, inicialmente, identificou-se os padrões semânticos entre as palavras, que foram utilizados pela ferramenta, chegando a cinco padrões recorrentes: sinonímia, hiperonímia/hiponímia e holonímia/meronímia, porém utilizamos na ferramenta apenas a relação de sinonímia, pois essa relação é mais comumente encontrada em comparação com as outras, tratar essas ocorrências nos fornece informações mais relevantes sobre o documento.

Tendo definidos os padrões, na fase seguinte, por meio de pesquisa, foram identificadas as ferramentas utilizadas para a etiquetagem de textos, tais etiquetas definem as classes gramaticais das palavras pertencentes ao texto analisado.

Para efetivamente iniciar a construção da ferramenta foram definidos alguns requisitos e funcionalidades que a ferramenta deveria contemplar, e nessa fase surgiu à necessidade do uso de uma ferramenta para a obtenção dos sinônimos das palavras das descrições. Além dos sinônimos, foi definido que seriam tratados apenas os radicais das palavras, eliminando, portanto, o problema das flexões das palavras.

Após a definição dos requisitos e funcionalidades, a etapa seguinte consistiu na implementação do protótipo computacional e, por fim, foram realizados experimentos com a Especificação de Requisitos do Sistema Merci, que é utilizado como documento modelo na disciplina de Engenharia de Software I do curso de Ciência da Computação da UFPI.

Resultados e Discussão

Para o tratamento das relações entre as palavras de um texto, inicialmente adquiriu-se informações relevantes sobre as mesmas, tais como a classe gramatical de cada palavra e as flexões de gênero, número e pessoa.

Neste projeto, utilizou-se o glossário que é resultado do projeto “Dicionário Eletrônico da Língua Portuguesa” do Laboratório EaSII e o etiquetador probabilístico *TreeTagger* como etiquetadores textuais, o uso de dois etiquetadores se justifica pelo fato de existirem palavras que não pertencem ao banco de dados do Glossário Easy. O etiquetador TreeTagger sempre infere uma classe gramatical para qualquer que seja a palavra.

É importante destacar que a maior dificuldade encontrada no processo de etiquetagem de textos em Português é a existência de acentuação, para minimizar esse fato, utilizamos a ferramenta ANTLR (PARR, T., 2007) para a separação do texto em *tokens* e a posterior etiquetagem dos mesmos. A ferramenta ANTLR (*Another Tool for Language Recognition*) é um gerador de *parser* que automatiza a construção de reconhecedores para DSLs (Linguagens Específicas de Domínio).

Uma das formas de limitar o espaço de busca por elementos UML é usar a frequência de termos relevantes dentro das descrições textuais. Assim, durante a análise é gerada uma tabela com todas as palavras do texto e suas respectivas frequências, além da frequência de palavras simples, todas as ocorrências de conjuntos frequentes de palavras também foram contabilizados, tais conjuntos são conhecidos por Expressões Multipalavra (PERNA, C.L. et al., 2010).

Além das classes gramaticais definidas pelos etiquetadores e as frequências contabilizadas de palavras simples e expressões multipalavra, necessitamos do relacionamento de sinonímia existentes no texto, para tanto, utilizamos o Tep 2.0 que é um dicionário eletrônico de sinônimos e antônimos para o Português do Brasil (DIAS-DA-SILVA, B., et. AL, 2000).

Os métodos descritos foram aplicados sobre os radicais das palavras da especificação de requisitos do Sistema Merci. Os radicais foram obtidos com o auxílio de duas ferramentas: o Glossário Easy e o PTStemmer por meio do algoritmo de radicalização Orengo e Huick (ORENGO, V. and HUYCK, C., 2001), vez que este algoritmo foi desenvolvido exclusivamente para a língua portuguesa. O algoritmo é disponibilizado pela Linguateca, que é um centro de recursos distribuído para o processamento computacional da língua portuguesa disponível em: http://www.linguateca.pt/ferramentas.html.

O uso de dois radicalizadores também foi devido a não abrangência do Glossário Easy a todas as palavras da língua portuguesa, o que é inviável. O PTStemmer consegue estimar um possível radical de uma palavra analisada, por isso as palavras não pertencentes a base do Glossário Easy, foram analisadas pelo PTStemmer.

O módulo foi desenvolvido utilizando a linguagem de programação Java, e recebe uma descrição textual em português e após realizar todas as etapas descritas nesta sessão, retorna ao usuário todas as informações encontradas durante a análise. Tais informações podem ser utilizadas antes da busca por elementos UML, limitando a busca aos termos mais relevantes designados pelo módulo, diminuindo o tempo de busca e procurando melhorar o resultado da análise, comparado ao resultado da busca sem o módulo.

Conclusão

O presente projeto tem como contribuição final, estender a verificação de Especificações de Requisitos a fim de reconhecer e extrair informações relevantes e diminuir ao máximo a identificação de termos irrelevantes para a geração de diagramas UML, de forma a auxiliar Engenheiros de Requisitos no processo de elicitação de requisitos de software.

Fez-se uso das ferramentas ANTLR para a geração do documento a ser etiquetado pelo Glossário Easy e pelo etiquetador TreeTagger para a detecção das classes gramaticais, do radicalizador PtStemmer e do Glossário Easy para extração dos radicais dos termos encontrados no conjunto de requisitos, permitindo assim o uso flexões de número e gênero das palavras e a base de Sinônimos Tep 2.0 para a identificação de sinônimos nas especificações.

Referências

ANCHIETA, R. T., RICARTE NETO, F. A. & MOURA, R. (2010) Identificação de Elementos UML a partir de Descrições Informais de Requisitos de Softwares. In: Anais da IV ERCEMAPI, Sobral, CE, 2010.

BECK, K., BEEDLE, M., BENNEKUM, A., COCKBURN, A., CUNNINGHAM, W., FOWLER, M., GRENNING, J., HIGHSMITH, J., HUNT, A., JEFFRIES, R., KERN, J., MARICK, B., MARTIN, R., MELLOR, S., SCHWABER, K., SUTHERLAND, J. and THOMAS, D. (2001) Manifest for Agile Software Development. Disponível em: http://www.agilemanifesto.org/iso/ptbr/, 2001. Último acesso: junho, 2011.

DIAS-DA-SILVA, B., OLIVEIRA, M., MORAES, H., HASEGAWA, R., AMORIM, D., PASCHOALINO, C. and NASCIMENTO, A. (2000) Construção de um Thesaurus Eletrônico para o Português do Brasil. In: Anais do V Encontro para o processamento computacional da Língua Portuguesa Escrita e Falada. Atibaia, São Paulo, Brazil, 2000.

KOTONYA, G. & SOMMERVILLE, I. (1998) Requirements Engineering: Process and Techniques. 1998.

ORENGO, V. and HUYCK, C. (2001) A Stemming Algorithm for Portuguese Language. In Proc. of Eight Symposium on String Processing and Information Retrieval (SPIRE 2001) - Chile, 2001.

PARR, T. (2007) The Definitive ANTLR Reference: Building Domain-Specific Languages. The Pragmatic Programmers. (PDF available in http://www.praprog.com/titles/tpantlr/the-definitive-antlr-reference), 2007.

PERNA, C. L., DELGADO, H. K. and FINATTO, M. J. (2010) Linguagens Especializadas em Corpora: Modos de Dizer e Interfaces de Pesquisa. EDIPUCRS, 2010.

Palavras-chave: Engenharia de Requisitos. Processamento de Linguagem Natural. Engenharia de Software.