Propostas de Tese
para o Mestrado em Engenharia Informática (MEI)

Grupo de Processamento de Linguagens
DI/CCTC
Universidade do Minho

(Pedro Rangel Henriques)
ano lectivo 13/14

1 API REST para serviços de PLN

Supervisor: Alberto Simões
Resumo:
Pretende-se estudar a possibilidade de disponibilizar pequenos serviços de processamento de linguagem natural usando uma API REST. Para além da implementação de um conjunto de serviços, deverão também ser analisadas as questões de tempo de processamento de alguns dos serviços, bem como as questões de gestão de carga do servidor.

Pretende-se o desenvolvimento de uma API com a filosofia REST, devidamente documentada, para um conjunto de aplicações relacionadas com o processamento de linguagem natural.

Existe um protótipo em funcionamento, em http://api.natura.di.uminho.pt/, que deverá servir de base neste trabalho.

O desenvolvimento deverá ser em Perl, usando a framework Dancer2. Serão interligadas diversas ferramentas, algumas delas em C, outras em Perl, outras em Java. É importante que o sistema seja dinâmico, ou seja, que a introdução de novos serviços possa ser descrita usando uma Domain Specific Language.

Sempre que possível devem ser usados formatos standard para disponibilizar resultados, e ser o mais coerente possível principalmente quando se usem ferramentas diferentes.

2 An Engine for Gathering and Managing Facts

Supervisor: Alberto Simões
Resumo:
The goal of this work is to devise and implement a customizable engine for gathering and maintaining facts (in the form of triples) from a set of pre-processed texts. The pre-processing implies some text annotations and extra information (e.g., Named Entity Recognition (NER), word lemmatization and disambiguation, morphological tagging, are available).

This work implies a state-of-the-art review of already available approaches, and to devise a better approach if possible, or improve an already existing one. Many available tools expect (or are tinkered) for specific domains, or expect specific textual formats, this engine should be as domain/format agnostic as possible.

The working plan also includes an evaluation of the devised system, and a comparison with other available techniques when possible.

3 Alargamento da Wordnet PortuGal.Net

Supervisor: Alberto Simões
Resumo:

O conceito de WordNet surgiu na universidade de Princeton, com a WordNet inglesa. Embora a princípio todos lhe apontassem defeitos, cedo se tornou um trabalho de referência, que muitos investigadores usam para diversos fins, desde tradução à recolha de informação. Muitos outros investigadores têm vindo a tentar copiar para as suas línguas.

Uma WordNet não é mais que um grafo de synsets, em que cada synset é um conjunto de palavras que são sinónimas entre si. Os arcos entre os synset representam relações entre os conceitos representados pelos synsets. Estas relações podem ser taxonómicas (hiperonímia/meronímia) ou outras. Ou seja, uma WordNet também pode ser vista como uma Ontologia.

Existem várias iniciativas para a construção de uma WordNet portuguesa. Com esta dissertação pretende-se que sejam analisadas as diferentes WordNet portuguesas disponíveis, que sejam estudadas as suas potencialidades e os seus problemas. Posteriormente, pretende-se tirar partido das WordNets e recursos similares disponíveis com licenças abertas para o alargamento da WordNet em desenvolvimento a par das WordNets das várias línguas espanholas (MCR).

4 Visualização de redes sociais dinâmicas

Supervisor: Pedro Rangel Henriques, Alda Lopes Gançarski

Resumo:

O aparecimento do Web 2.0 traduz-se por um conjunto de aplicações baseadas na Internet permitindo aos utilizadores o intercâmbio de recursos, opiniões e experiências, criando assim autênticas redes sociais (RS). Estas aplicações, cada vez mais utilizadas, variam em função das suas funcionalidades e finalidades. Pode-se distinguir os blogs (p. ex. Twitter), os wikis (p. ex. Wikipédia), as redes sociais de contacto (p. ex. Facebook), os sites de partilha de recursos (p. ex. Youtube para vídeos, Flickr para fotografias), os sites para partilha de links (p. ex. Delicious).

A visualização é uma forma importante de ajudar os utilizadores das RS a perceber o seu funcionamento e evolução, bem como a existência de comunidades latentes. Devido à grande escala de utilização de RS, a visualização de um grande número de utilizadores e suas interacções torna-se difícil [1]. Um aspecto importante a ter em conta para a visualização é a dimensão temporal. De facto, o dinamismo das RS implica o aumento e a diminuição frequente de utilizadores, assim como das suas inter-conexões [2]. Outro aspecto importante que pode ajudar a uma melhor compreensão do funcionamento duma RS é a localização geográfica dos utilizadores no mundo [3].

Por exemplo, o Facebook Stats (aplicação desenvolvida no âmbito de um projeto de 3° ano de LCC (2013-2014)) é uma aplicação que tem o intuito de permitir ao utilizador diversas visualizações numa base temporal sobre a utilização do Facebook por parte desse utilizador e a sua interação com os seus amigos (incluindo geo-localização).

Pretende-se nesta tese propor a visualização de uma RS não apenas do ponto de vista de um utilizador específico (como no Facebook Stats), mas também como um todo, i.e. tendo em conta o conjunto dos utilizadores e suas ações e interações. Os mecanismos de visualização propostos permitirão estudar o comportamento da RS em função de diferentes parâmetros, como a geo-localização dos utilizadores, a dimensão temporal (p. ex. período escolar vs férias, dia de trabalho vs fim de semana, etc) ou outros parâmetros de consulta considerados importantes. O sistema de visualização deve ser desenvolvido de forma adequada para diferentes tipos de RS, conduzindo a um estudo comparativo. Bibliografia citada: [1] Visualizing Overlapping Latent Communities Using POI-Based Visualisations, P. Dudas, J. Ahn, M. Jongh, P. Brusilovsky, iConference 2013. [2] Visualizing the Evolution of Communities Structures in Dynamic Social Networks, K. Reda, C. Tantipathananandh, A. Johnson, J. Leigh, T. Berger-Wolf, IEEE Symposium on Visualization 2011. [3] Visualization of social media data : mapping changing social networks, Faculty of Geo-information Science and Earth Observation of the University of Twente, The Netherlands, Ding Ma, 2012.

5 Procura de informação na Web Social

Supervisor: Pedro Rangel Henriques, Alda Lopes Gançarski

Resumo:

O aparecimento da Web 2.0 traduz-se por um conjunto de aplicações baseadas na Internet permitindo aos utilizadores o intercâmbio de recursos, opiniões e experiências, constituindo os media sociais. Com o enorme crescimento, nos últimos anos, deste tipo de aplicações, foram desenvolvidas ferramentas de procura de informação na Web social em que os algoritmos de procura tomam em conta como os grupos sociais influenciam e melhoram a capacidade de encontrar informação interessante, sendo os resultados classificados em função do grafo social.

O objectivo desta tese é propor um sistema de procura de informação na Web social, tendo em conta o conteúdo da

pesquisa e a relação social entre o utilizador que lança a procura e os proprietários dos recursos encontrados.

6 Legibilidade de Código

Supervisor: Pedro Rangel Henriques + Maria João Varanda Pereira
Resumo:
A legibilidade de código é um factor muito importante na compreensão de programas. As palavras usadas como identificadores, a formatação do código e a organização/modularização do código são parâmetros de avaliação passíveis de serem medidos e que influenciam fortemente a percepção do domínio do problema envolvido em cada codificação.
Existindo muito trabalho já desenvolvido nesta área, propõem-se como tema desta tese a recolha exaustiva de métodos de avaliação de legibilidade de código, assim como de abordagens/ferramentas para otimização (automática) dessa mesma legibilidade. Pretende-se também estudar a sua possível aplicação a diferentes tipos de linguagens de programação.

7 Visualização ontológica de Programas

Supervisor: Pedro Rangel Henriques + Maria João Varanda Pereira
Resumo:
Para apoio à complexa atividade de Compreensão de Programas (área de investigação conhecida por *PC – program comprehension*, na qual o nosso grupo de investigação vem trabalhando há anos), propõe-se um tema de tese que consiste em criar visualizações de um programa (o programa em análise) baseadas na apresentação da ontologia do Domínio da Linguagem (a qual deveria ser inferida automaticamente a partir de respetiva gramática) e na apresentação da respetiva ontologia populada (extraída de cada programa).
Note-se que já em trabalho de mestrado anterior resolvemos, com sucesso e interessantes resultados, o problema de derivar uma gramática para uma linguagem concreta a partir da ontologia do Domínio do Problema. Agora neste trabalho pretende-se investigar o oposto: saber se é possível extrair a ontologia a partir da gramática e como o fazer sistematicamente. Será ainda tema de investigação a procura de soluções de visualização que realmente ajudem a compreender o código em análise.

8 Reflection em Java

Supervisor: Pedro Rangel Henriques + Nuno Oliveira
Resumo:
Nesta proposta de tema de tese de mestrado pretende-se que sejam estudadas várias packages para Java que sejam específicas para ajudar a trabalhar com reflection, de modo a que seja possível fazer um estudo comparativo, baseado em casos de estudo, e tirar conclusões sobre as vantagens e limitações desses pacotes de software livre disponíveis.
Com base nesse estudo, pretende-se que a seguir seja desenvolvido um pacote para trabalhar com Reflection que permita trabalhar com tipos genéricos e ultrapassar as limitações encontradas.

9 Técnicas de Análise de Código para Otimização de Chamadas a Funções

Supervisor: Pedro Rangel Henriques + Maria João Varanda Pereira
Área:

Resumo:
Usando boas-práticas há muito advogadas no âmbito da programação imperativa (incluindo a programação orientada-a-objetos) o programador é levado a organizar o seu código-fonte em muitas pequenas funções[1] com uma semântica muito específica e bem definida. Essas funções são depois invocadas dentro de outras funções ou do programa-principal. Esta técnica é muito importante em termos de modularidade e de clareza (o código-fonte torna-se muito mais legível e

[1] Aqui o termo *função* é usado, *à la C*, para designar genericamente quaisquer subprogramas, quer funções, quer procedimentos, ou mesmo no âmbito da POO, para designar métodos.

fácil de manter), mas introduz um considerável atraso em termos de tempo de execução (como é sabido, o mecanismo de invocação de funções é computacionalmente pesado devido à necessidade de criar um novo *activation-record*, passar os valores dos parâmetros reais para os parâmetros formais, fazer a passagem de controlo e depois retomar o controlo após ter recuperado o *activation-record* inicial.

O ideal é permitir que o código-fonte seja mantido nesta organização funcional, mas o mesmo seja pré-processado para substituir as chamadas a funções por *código in-line* sempre que se reúnam condições para que tal seja possível sem alterar a semântica do programa original.

O objetivo desta proposta consiste em recorrer a técnica de análise de código-fonte (como, por exemplo, as que se empregam no âmbito da Compreensão de Programas) para identificar todas as situações de invocação que possam ser substituídas pela inserção do código da função no ponto da chamada, produzindo um relatório (visual) dessa análise. O trabalho terá de começar pela escolha de uma linguagem de programação que será alvo do estudo e pela identificação das condições que uma função deve reunir para que tal substituição seja possível. Após esta fase, será necessário definir os esquemas gerais para realizar a substituição com total preservação da semântica.

10 Implementação de um editor dirigido pela sintaxe para logoLISS com compilação incremental

Supervisor: Pedro Rangel Henriques + Daniela da Cruz

Resumo:
A linguagem LISS é uma linguagem de programação imperativa—que permite a manipulação de inteiros, sequências dinâmicas de inteiros e conjuntos de inteiros (sets) definidos em compreensão—que vem sendo usada na UM, no seio do nosso grupo de investigação (gEPL) desde há vários anos para testar diferentes questões relacionadas com a compilação e o uso de geradores automáticos de compiladores.

Recentemente e dentro da mesma linha de exploração, a linguagem foi estendida para suportar também números complexos, polinómios e polígonos. Essa versão foi produzida com o gerador LISA sendo traduzida para código Assembly da Máquina Virtual VM.

Mais recentemente foi proposto como projeto da disciplina de Processamento de Linguagens a implementação, com o gerador Yacc, de uma extensão chamada LogoLISS, em que se pedia para acrescentar à linguagem LISS os comandos típicos da linguagem LOGO para manusear a tartaruga no écran; nesta versão o compilador continuava a gerar Assembly da VM.

A proposta para este projeto de mestrado é implementar a linguagem LogoLISS em AnTLR (um dos Geradores de Compiladores mais usados atualmente, talvez o 2º, a seguir o Yacc) usando uma gramática de atributos para criar uma árvore de sintaxe abstrata (AST) e gerar código Assembly da máquina virtual do Java, JVM.

Depois usando essa AST pretende-se criar um Editor Dirigido pela Sintaxe (SDE) para a linguagem que forneça toda a ajuda típica de um editor que controla a escrita em função da gramática e que permita fazer compilação incremental (cada vez que se altera no editor o código fonte só se gera o código final correspondente à parte alterada).

Embora este projeto traga variados desafios com alguma complexidade, tudo será feito de forma progressiva e sistemática tomando como ponto de partida todo o material que já existe sobre as anteriores implementações da linguagem.

11 Sistema Inteligente de apoio à Avaliação da Qualidade de linguagens e gramáticas

Supervisor: Pedro Rangel Henriques + Daniela da Cruz
Área:

Resumo:
Nesta proposta de projeto de mestrado pretende-se criar um sistema inteligente que permita avaliar a qualidade das linguagens de programação com base nos 8 critérios definidos pelo nosso grupo de investigação, conforme documentado num relatório das provas de agregação de Pedro Rangel Henriques (abril 2012) que será disponibilizado logo de início

ao aluno.

O sistema deve apoiar o máximo possível à classificação da linguagem segundo esses parâmetros (que embora muito bem definidos, são de quantificação subjetiva) recorrendo a um sistema tipo Case-Based Reasoning (CBR) para recordar decisões passadas e auxiliar a tomar decisões no caso presente.

A ferramenta que se pede é totalmente inovadora e se for bem concebida e desenvolvida pode ser um instrumento de grande valor prático.

12 QG - Um Sistema baseado em Gramáticas de Atributos para Avaliar a Qualidade de Gramáticas

Supervisor: Pedro Rangel Henriques + Daniela da Cruz
Área:

Resumo:
Seja G uma gramática e ML a meta-linguagem em que se especifica G. Seja MG a meta-gramática que gera ML; MG é uma gramática de atributos. A qualidade da gramática G pode ser medida através do cálculo de um conjunto de métricas (conforme documentado num relatório das provas de agregação de Pedro Rangel Henriques (abril 2012) que será disponibilizado logo de início ao aluno). Ora cada uma dessas métricas pode ser formalizada através de um atributo de MG.

O que se pretende, nesta proposta de tema para tese de mestrado, é o desenvolvimento de um sistema, QG, que: + permita definir as métricas a calcular, associando atributos sintetizados aos símbolos de MG; + permita definir as regras de cálculo desses atributos, de acordo com o significado da respetiva métrica; + aceite uma dada gramática G, escrita em ML, e avalie a sua qualidade calculando os atributos; + permita manipular G, por exemplo transformando-a numa gramática equivalente com maior qualidade.

A proposta está aberta à imaginação do aluno e à evolução dos trabalhos. A ferramenta que se pede é totalmente inovadora e se for bem concebida e desenvolvida pode ser um instrumento de grande valor prático.

13 Análise de Código Máquina: abstração do significado e cálculo de métricas

Supervisor: Pedro Rangel Henriques + Daniela da Cruz
Área:

Resumo:
Muito se tem falado e avançado em termos de análise de código-fonte (análise de programas escritos em linguagens de alto-nível) e são inúmeras as aplicações destas tecnologias, desde a re-engenharia e restauro à compreensão de programas e à avaliação da qualidade).

Nesta proposta de projeto de mestrado pretende-se fazer algo semelhante mas trabalhando sobre código-objeto (programas escritos em linguagem de baixo-nível como seja o código-máquina). A ideia é trabalhar com o Assembly concreto de uma máquina específica real (a escolher no início do trabalho).
O projeto terá duas grandes metas. Por um lado, pretende-se criar uma abstração do código-máquina ao nível de uma linguagem algorítmica (a escolher) de modo a que se possa mostrar ao utilizador uma visão de alto-nível que facilite uma rápida compreensão do que o programa faz (de modo a entender o seu significado operacional). Por outro lado, pretende-se aferir a qualidade do código-máquina calculando uma conjunto de métricas a estabelecer, mas que incluia coisas como número de instruções, tamanho das instruções, velocidade e consumo de energia.

14 Exploração do conceito de Algorithmic Debugging na aceleração da localização do código errado

Supervisor: Pedro Rangel Henriques + Daniela da Cruz

Área:

Resumo:
Algorithmic debugging was introduced by Shapiro in [1] as an alternative to trace debuggers for the Logic Programming paradigm. The idea was afterwards employed in other declarative programming paradigms such as functional [X] and functional-logic [X] programming. When a program's execution reveals a bug, the technique can automatically isolate a buggy portion of the source code by asking a series of questions to the programmer about computations performed during this execution. The answers of the programmer are used to discard those parts of the program that executed correctly, and thus, they do not caused the bug.

In this master thesis, the idea is to study how the presence of assertions could help to reduce the number of questions and more quickly find the bug. The verification techniques developed by the group, under Gama project (Daniela's Ph.D. work), to deal efficiently with programs with assertions (written according to DbC principles), shall be used in this proposal.

Bibliographic Reference: [1] Ehud Y. Shapiro, Algorithmic Program Debugging, MIT Press Cambridge. 1983.

15 Uso de uma Ontologia sobre Métricas de Software para avaliação de programas

Supervisor: Pedro Rangel Henriques + Nuno Oliveira
Área: Ontologias, Métricas de Software

Resumo:
Pretende-se criar uma ontologia para organização de métricas de software o mais variadas possível. A ontologia a definir deve ser capaz de definir métricas específicas ao paradigma ou à linguagem de programação. As métricas a utilizar devem ser distinguíveis em vários domínios e ser padronizáveis para possibilitar a comparação entre linguagens (nos parâmetros espectáveis).

O objectivo final é produzir avaliações (nota) a partir do código fonte e das medições obtidas pelas métricas organizadas pela ontologia.

16 Avaliação Individual em Contexto de Grupo num sistema de E-learning

Supervisor: Pedro Rangel Henriques + Nuno Oliveira
Área: Aprendizagem Colaborativa, E-Learning

Resumo:
Pretende-se detectar e padronizar o trabalho efectuado por alunos dentro de um sistema de E-learning colaborativo.

O objectivo é detectar padrões de actividade de alunos por forma a identificar os que se distinguem (positiva ou negativamente) no trabalho em grupo. Deverá ainda ser creditado o esforço efectuado em cada iteração por forma a quantificar o trabalho efectivo de cada elemento do grupo. O objectivo final é creditar o esforço individual de cada aluno no contexto do trabalho em equipa.

17 Criação automática de Currículos a partir das Redes Sociais

Supervisor: Pedro Rangel Henriques + Alda Gançarski
Área: Gestão de Currículo, Integração de sistemas

Resumo:
Actualmente é comum a população activa partilhar em redes sociais especializadas os projectos em que participa. É comum as entidades associadas às publicações de conteúdos creditarem os autores nas suas plataformas online.

Pretende-se com esta proposta produzir ou integrar numa ferramenta de gestão de currículos estes conteúdos para constarem do currículo dos candidatos. Espera-se que sejam integrados conteúdos disponíveis em redes sociais como

o *Linkedin* ou *Behänce* numa plataforma unificada com suporte à exportação para os comuns formatos de Currículos.