Colaboração em projecto do L2F do INESC

Francisco Dias

(Relatório de Actividade)

Resumo— Foi desenvolvida uma aplicação para geração de regras para identificação de expressões fixas da Língua Portuguesa para serem usadas num Sistema de Processamento de Língua Natural. A geração de regras possibilitou uma taxa de identificação de 93,7% para este tipo de expressões. Dado os bons resultados obtidos, fui convidado a escrever um artigo científico que descreve os resultados obtidos.

Palavras Chave—Portfólio, língua natural, expressões fixas.

1 Introdução

Processamento de Língua Natural é uma área multidisciplinar que une a Inteligência Artificial à Linguística. Nesta área do conhecimento são necessários muitos recursos humanos para se poder criar aplicações, bases de dados, anotações de textos para aprendizagem automática.

O Laboratório de Língua Falada (L2F) no Instituto Nacional de Engenharia de Sistemas e Computação (INESC-ID) é uma organização onde se desenvolvem importantes contribuições para o processamento computacional da Língua Portuguesa. Sediado junto ao Instituto Superior Técnico (IST), entrei em contacto com o L2F no início do semestre para manifestar o meu interesse em participar num projecto deste laboratório como actividade para a disciplina de Portfólio IV.

Este relatório apresenta a minha actividade de desenvolvimento de uma aplicação na área de Língua Natural.

2 Objectivo

Este projecto teve como objectivo a construção de uma aplicação de computador que gera automaticamente regras de identificação de expressões fixas da Língua Portuguesa para serem utilizadas na cadeia Statistical and Rule Based Natural Language Processing Chain (STRING), um sistema de processamento de língua natural desenvolvido no L2F no INESC-ID.

Uma expressão fixa é uma construção idiomática que tem um significado próprio, que não deve ser levado à letra, e que se apresenta sempre com uma forma fixa. Alguns exemplos deste tipo de expressões são "dançar a mesma dança", "moer o juízo", "vender gato por lebre".

A aplicação desenvolvida gera regras a partir de uma representação sintáctica das expressões para de seguida serem adicionadas ao sistema XEROX Incremental Parser (XIP), o parser utilizado pela cadeia STRING. Adicionadas as regras, os textos processados pela STRING vêm anotados com as expressões fixas existentes.

A motivação para a identificação destas expressões é aumentar a eficácia do reconhecimento dos significados de um texto e, dessa forma, poder aumentar o desempenho do sistema.

3 Actividade Desenvolvida

3.1 Planeamento

Este projecto foi planeado para ter uma duração de 12 semanas, durante as quais houveram reuniões semanais às sexta-feiras para

• *Francisco Dias, nº. 47619,*
E-mail: francisco.m.c.dias@tecnico.ulisboa.pt,
é aluno do curso de Mestrado em Engenharia Informática e Computadores,
Instituto Superior Técnico, Universidade de Lisboa.

Manuscrito entregue em 30 de Maio de 2014.

(1.0) Excelent	ACTIVITY					DOCUMENT						
(0.8) Very Good	Objectives x2	Options x1	Execution x4	S+C x1	SCORE	Structure x0.25	Ortogr. x0.25	Gramm. x0,.25	Format x0.25	Title x0.5	Filename x0.5	SCORE
(0.6) Good												
(0.4) Fair	1.6	0.7	3.2	0.6	6.1	0.2	0.25	0.2	0.25	0.5	0.5	1.9
(0.2) Weak												

apresentação dos resultados obtidos durante a semana de trabalho anterior.

A primeira reunião ocorreu a 28 de Fevereiro onde me foi apresentado o grupo de trabalho, o ambiente de desenvolvimento, as ferramentas a usar e o objectivo do projecto. O grupo de trabalho é formado por um investigador na área da computação e por um linguista.

Em cada uma das reuniões seguintes, eu tirei as minhas dúvidas acerca do funcionamento sistema e de interpretações linguísticas, e apresentei propostas de desenvolvimento que foram discutidas com o investigador e com o linguista, de forma a achar um consenso.

3.2 Material Inicial

O projecto foi iniciado tendo como base um dicionário electrónico com cerca de 2500 expressões fixas [1] no formato de matrizes, desenvolvido durante um projecto anterior, em que cada matriz contém uma classe de expressões.

O material para início de desenvolvimento consistiu em 3 matrizes com cerca de 20 expressões cada uma.

3.3 Modelação

A partir do estudo do material disponível e das anotações feitas durante as reuniões, comecei a definir, num papel, um esquema para o desenvolvimento da aplicação. A partir deste esquema foi possível encontrar os caminhos de execução semelhantes nos vários tipos de matrizes, e assim modelar uma sequência de algoritmos que fosse comum entre todas elas.

Esta modelação é importante, tanto para um desenvolvimento metódico como para no futuro ser mais fácil a modificação e reaproveitamento do código para outras tarefas.

3.4 Documentação

Devido à complexidade da aplicação, revelou-se necessário desde o início proceder à sua documentação.

A documentação de aplicações na área de Língua Natural mostrou-se diferente dos outros tipos de aplicações. Para além do código dever se encontrar bem comentado devido à sua sofisticação, necessita também que seja agregado um conjunto de exemplos de frases em Língua Natural que justifiquem cada regra e cada rotina da aplicação.

Estas frases de exemplo tornam o código mais compreensível por justificar a razão da sua inclusão.

3.5 Pro-actividade

Para poder medir o desempenho da aplicação passei a executar a aplicação sobre as matrizes de expressões fixas para gerar as suas regras e de seguida validei-as usando o avaliador de regras. Verifiquei então que algumas expressões não eram aceites porque não se encontravam bem definidas na matriz.

Por isso, procedi a uma alteração e correcção das matrizes, tarefa que durou três dias. Essa tarefa consistiu em verificar as expressões uma por uma, analisar os erros que ocorriam e fazer a correcção sempre que possível. Quando a correcção não era possível, ou houve dúvidas sobre a sua correcção, marquei essas expressões para as enviar para o linguista.

Embora a tarefa de correcção não se encontrasse no âmbito inicial deste projecto, esta iria ser necessária aquando da importação das regras para o sistema e tendo-se revelado de grande utilidade para demonstrar a correcção da aplicação e possibilitando que o sistema, desde já, apresente resultados a partir das regras por ela geradas.

3.6 Finalização

A apresentação final ocorreu no dia 23 de Maio em reunião no INESC-ID. Nela foram definidas as últimas afinações ao sistema

No final fui agradecido pela qualidade do trabalho desenvolvido, pelo empenho e pela minha pro-actividade.

4 RESULTADOS

As regras geradas foram importadas para o sistema XIP e testadas com um conjunto de frases de exemplo para cada uma das expressões. Estas regras permitiram que a cadeia STRING passasse a identificar 93,7% das expressões fixas identificadas inicialmente, resultados que foram considerados muito bons.

As matrizes com as expressões fixas que foram alteradas e corrigidas por mim passarão a ser a base para o futuro desenvolvimento, e os seus resultados servirão para melhorar o desempenho do sistema.

5 Artigo Científico

Devido aos resultados obtidos terem sido considerados muito bons fui convidado para a co-autoria de um artigo científico que descreve a solução encontrada para a identificação.

A realização deste artigo está fora do âmbito da actividade de Portfólio IV.

6 Conclusão

O desenvolvimento desta aplicação foi um sucesso e os seus resultados foram muito além dos esperados inicialmente.

Fui um projecto muito interessante para mim, e será certamente importante para o meu futuro e para o meu projecto de mestrado na área de Língua Natural.

Agradecimentos

O autor agradece à equipa do L2F pelo bom ambiente encontrado e pela sua disponibilidade e confiança para me receber neste projecto.

Referências

[1] J. Baptista et al., 2007, Frozen Sentences of Portuguese: Formal Descriptions for NLP, *Proceedings of the Workshop on Multiword Expressions*, Association for Computational Linguistics, pags. 72-74.

Francisco Dias Depois de frequentar os três anos de Engenharia Aeroespacial no IST e de ter trabalhado em várias áreas da ciência da computação, tomei como desafio voltar ao IST e fazer um mestrado em Robótica e Inteligência Artificial, com ênfase na Interacção Humano-Máquina e nos Sistemas de Processamento de Língua Natural.

APÊNDICE
DEMONSTRAÇÃO DA APLICAÇÃO

A demonstração do reconhecimento de dependências e de expressões fixas da cadeia STRING está disponível ao público no seguinte endereço:

`https://string.l2f.inesc-id.pt/demo/dependencyextraction.pl`

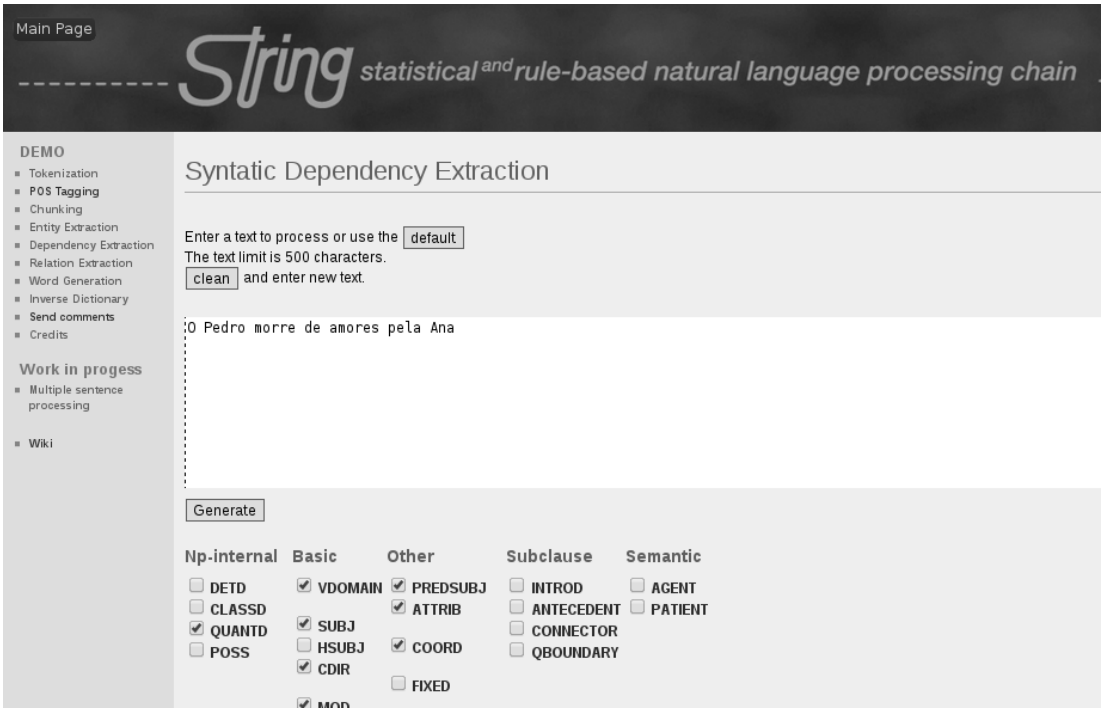
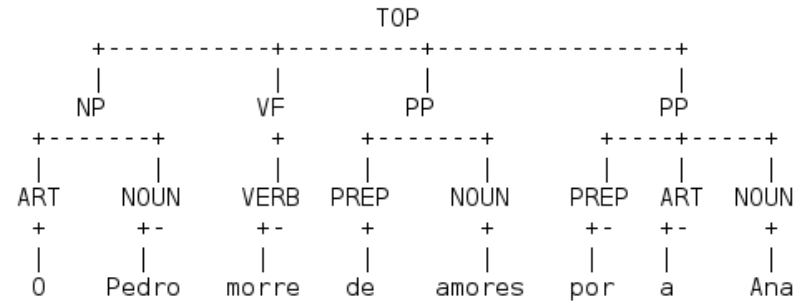


Figura 1. Interface da cadeia STRING

APÊNDICE
EXEMPLO DA SAÍDA DA APLICAÇÃO




```
MAIN(morre)
DETD(Pedro,O)
DETD(Ana,a)
VDOMAIN(morre,morre)
MOD_POST(morre,amores)
MOD_POST(morre,Ana)
SUBJ_PRE(morre,Pedro)
FIXED(morre,amores)
NE_PEOPLE_INDIVIDUAL(Pedro)
NE_INSTITUTION_COLLECTIVE(Ana)
0>TOP{NP{O Pedro} VF{morre} PP{de amores} PP{por a Ana}}
```

Figura 2. Execução com a frase *O Pedro morre de amores pela Ana*, identificando com sucesso uma expressão fixa como *FIXED(morre,amores)*

APÊNDICE
COMPROVATIVOS DE EXECUÇÃO

Declaração do coordenador do projecto STRING no L2F:

DECLARAÇÃO

Eu, Nuno João Neves Mamede, Professor Associado com Agregação do Departamento de Engenharia Informática (DEI) e Investigador no Laboratório de Sistemas de Língua Falada (L^2F) do INESC-ID Lisboa, declaro para efeitos de comprovativo de atividade para a cadeira de Portfólio Pessoal IV, que Francisco Manuel Carvalho Dias, aluno n.º 47619, foi responsável pela elaboração de um sistema de geração de regras para o sistema XIP que visa a identificação de expressões fixas em textos.

O trabalho foi elaborado no período compreendido entre 1 de Março de 2014 a 26 de Maio de 2014, no INESC-ID, na Rua Alves Redol, Nº 9..

O trabalho realizado foi para além do estritamente necessário e excedeu em muito o acordado inicialmente: cerca de 40 horas de trabalho. A futura integração do trabalho na cadeia de processamento de língua Natural do INESC-ID vai permitir uma melhoria do seu desempenho.

O Francisco também demostrou ter excelentes capacidades de relacionamento, o que permitiu a sua fácil integração na equipa de investigação.O entusiasmo colocado na execução das tarefas acordadas também merece um grande destaque.

Lisboa, 28 de Maio de 2014

(Nuno Mamede)