Compressão Automática de Frases

Proposta de Projecto

Orientador: João Paulo da Costa Cordeiro

Ano Lectivo de 2011/2012

1 Objectivos

A compressão de uma frase consiste em remover partes desta, de modo a satisfazer dois grandes objectivos: a redução do tamanho e a manutenção da informação relevante. Estes dois objectivos apontam, naturalmente, em direcções opostas e um bom compressor é aquele que consegue um bom compromisso entre ambos. Um grande número de frases, das mais variadas fontes e géneros, são susceptíveis de serem comprimidas, atendendo ao elevado número de irrelevâncias que as compõe, com evidentes vantagens para um utilizador. Por exemplo, na seguinte frase:

> *Foi sem surpresa que* ***o*** *Banco Central Europeu (****BCE****)* ***anunciou a manutenção da taxa de juro*** *de referência para a Zona Euro* ***em 1%****.* ***Da reunião*** *de hoje* ***não eram esperadas*** *quaisquer* ***novidades****, nem no nível de juros, nem em termos de medidas não convencionais.*

os segmentos a negrito seria preservados, dando origem à seguinte versão reduzida, com uma taxa de compressão de 37.78% ($\frac{17}{45}$)[1]:

> ***O BCE anunciou a manutenção da taxa de juro em 1%. Da reunião não eram esperadas novidades****.*

Nos últimos anos tem havido várias abordagens a esta problemática que continua sendo um tópico muito activo, no domínio da investigação, com importantes contributos para a a área da *Sumarização Automática de Texto*.

Posto isto, este projecto pretende servir uma abordagem ao problema enunciado, baseada em *Aprendizagem Supervisionada* (Supervised Learning). O trabalho será desenvolvido no Centro de Tecnologia da Linguagem Humana e Bioinformática da UBI (`http://hultig.di.ubi.pt`).

2 Plano de Trabalho

O desenvolvimento deste projecto deve seguir a seguinte ordem de trabalho:

- **T1**: Criação de uma aplicação com interface gráfica para anotação de texto, que permita ao utilizador assinalar segmentos de texto a eliminar. A aplicação permitirá criar um conjunto de exemplos de treino para ser utilizado na aprendizagem supervisionada. (4 semanas)
- **T2**: Criação dos exemplos de treino, com a ferramenta desenvolvida em **T1**. Pretende-se um conjunto significativo e de elevada qualidade, com cerca de 10000 casos. (4 semanas)

[1] Número de palavras que se mantém, sobre o número de palavras na frase original.

- **T3**: Definição de características consideradas relevantes e aplicação de um algoritmo de aprendizagem supervisionada, que será definido pelo docente. (3 semanas)
- **T4**: Definição de um conjunto de regras de compressão de frases, com vista à comparação com as regras induzidas. (2 semanas)
- **T5**: Escrita do relatório de projecto (3 semanas).

3 Requisitos Académicos

O aluno deve possuir boas classificações/competências em domínios fundamentais, tais como *Programação*, *Programação e Algoritmos*, e *Programação Orientada a Objectos*. O aluno deve compreender bem o mecanismo da Aprendizagem Supervisionada, do domínio da Inteligência Artificial. Deve também estar preparado para trabalhar em linguagem Java, a um nível avançado.

4 Grau de Dificuldade

Difícil.

5 Resultados Esperados

Os seguintes resultados são esperados:

- Uma aplicação, em linguagem Java, com interface gráfica, para uma cómoda anotação de texto.
- Um conjunto de cerca de 10000 frases com segmentos de eliminação anotadas, armazenado em XML.
- Conjunto de regras de compressão de frases, definidas pelo aluno.
- Um relatório de projecto.

6 Contactos

João Paulo Cordeiro (`jpaulo@di.ubi.pt`)
Departamento de Informática, Gabinete 4.3

Referências

[1] Knight K., Marcu D. (2002). *Summarization beyond sentence extraction: A probabilistic approach to sentence compression.*. Journal of Artificial Intelligence, Volume 139:91–107.

[2] Tom Mitchell (1997). *Machine Learning*. McGraw Hill