Uma Abordagem ao PÁGICO baseada no Processamento e Análise de Sintagmas dos Tópicos

Ricardo Rodrigues
CISUC, Universidade de Coimbra
rmanuel@dei.uc.pt

Hugo Gonçalo Oliveira
CISUC, Universidade de Coimbra
hroliv@dei.uc.pt

Paulo Gomes
CISUC, Universidade de Coimbra
pgomes@dei.uc.pt

University of Coimbra
Faculty of Sciences and Technology
Department of Informatics Engineering

Knowledge and Intelligent Systems Laboratory
Cognitive and Media Systems Group
Centre of Informatics and Systems of the University of Coimbra

Plano

- Introdução
- Abordagem
- Processamento dos Tópicos
- Descrição das Corridas
- Resultados
- Conclusões

Introdução

- PÁGICO: "[...] *avaliação conjunta* [...] *que tem por objectivo avaliar sistemas que encontrem respostas não triviais a necessidades de informação complexas* [...]"
- Conjugação de trabalhos de doutoramento
  - RAPPORT: Sistema de Resposta Automática a Perguntas para o Português
  - ONTO.PT: Ontologia Lexical para o Português

Abordagem

- Abordagem dividida em quatro partes
  - **Indexação** dos artigos
  - **Análise** e **processamento** dos tópicos
  - **Pesquisa** sobre os conteúdos dos artigos
  - **Tratamento** das respostas

Abordagem

- Indexação
  - **Conteúdos** de artigos **radicalizados** através do *portuguese analyzer* do Lucene (Hatcher e Gospodnetic, 2004)
    - Ignorados números, géneros e tempos verbais
    - Aumentada a abrangência de eventuais *queries*
  - Criado **índice** utilizando o **Lucene**

Abordagem

- Processamento e Análise
  - Frases dos **tópicos** segmentadas em **sintagmas**
  - Sintagmas usados como base para a criação de ***queries***

Abordagem

- Pesquisa
  - ***Queries*** usadas para realização de pesquisas sobre os **conteúdos** dos artigos
  - **Resultados ordenados** em função da pontuação atribuída pelo Lucene
  - Um valor máximo para o número de **resultados** devolvidos (**25**)

Abordagem

- Tratamento
  - **Excluídos** os resultados referentes à **estrutura** da Wikipédia ou demasiado **genéricos**:
    - Páginas começadas por Wikipédia, Portal, Lista ou Anexo
    - Páginas de desambiguação
    - Artigos começados por dígitos
    - Artigos referentes a disciplinas (e.g., Economia, História, etc.) ou começados por “ismos” (e.g., Anarquismo, Academicismo, Abolicionismo, etc.)

Processamento dos Tópicos

- Identificação dos Sintagmas
  - Tópicos: Sintagmas Nominais (SNs) e Sintagmas Verbais (SVs) – essencialmente os verbos
  - Duas etapas:
    - Etiquetagem gramatical (*PoS tagging*) – OpenNLP (http://incubator.apache.org/opennlp/)
    - Identificação dos Sintagmas (*chunking*) – regras extraídas do Bosque (Freitas, Rocha e Bick, 2008)

Processamento dos Tópicos

- Categoria da Resposta
  - Primeiro nome do primeiro SN: alvo do tópico ou categoria a que as respostas obedecem
  - Padrão **`<hipónimo> é um <hiperónimo>`** para identificação de frases referentes a definições

Processamento dos Tópicos

- Expansão de Sinónimos
  - Verbos nos SVs expandidos de forma a aumentar a abrangência das *queries*
  - Usados *synsets* do ONTO.PT com um de dois métodos de desambiguação:
    - *Bag of Words*
    - *Personalized PageRank*
  - Apenas considerados sinónimos com uma frequência acima de determinado patamar, de acordo com listas de frequências disponibilizadas pela Linguateca

Processamento dos Tópicos

- Expansão de Nacionalidades e de Países
  - Boa parte dos tópicos referentes a países ou nacionalidades lusófonos
  - Expressões como **`país lusófono`** substituídas por **`(país lusófono) OR Portugal OR Brasil OR Angola OR Moçambique OR (Cabo Verde) OR (Guiné Bissau) OR (São Tomé e Príncipe) OR Timor`**.
  - Expressões como **`futebol brasileiro`** substituídas por **`(futebol brasileiro) OR (futebol AND Brasil)`**.

Descrição das Corridas

- Três corridas oficiais:
  1. Versão base, apenas expansão de sinónimos
  2. Expansão de sinónimos dos SVs com apenas um verbo, utilizando o método *Bag of Words*
  3. Expansão de sinónimos dos SVs com apenas um verbo, utilizando o método *Personalized PageRank*
- Algumas corridas não oficiais, para testar alguns aspectos (e.g., a expansão de sinónimos de SNs)

Resultados

Corrida	Limite	# Submetidas	# Respostas	Precisão	Pseudo-abrangência	Pontuação	# Tópicos
1	5	512	86	*0,1680*	0,0383	14,4453	47
	10	918	122	0,1329	0,0543	16,2135	51
	15	1275	147	0,1153	0,0654	16,9482	54
	20	1577	164	0,1040	0,0730	17,0551	56
	25	**1718**	***181***	**0,1054**	***0,0805***	***19,0693***	***59***
2	5	516	90	*0,1744*	0,0400	15,6977	50
	10	927	132	0,1424	0,0587	18,7961	53
	15	1289	164	0,1272	0,0730	18,3986	58
	20	1591	184	0,1157	0,0819	21,2797	58
	25	**1736**	***203***	**0,1169**	***0,0903***	***23,7379***	***59***
3	5	518	92	*0,1776*	0,0409	16,3398	48
	10	940	135	0,1436	0,0601	19,3883	53
	15	1305	166	0,1272	0,0738	21,1157	57
	20	1601	188	0,1174	0,0836	22,0762	58
	25	**1730**	***208***	**0,1202**	***0,0925***	***25,0081***	***59***

Resultados

- Corrida com melhores resultados (a terceira):
  - Pontuação: 25,0081
  - Precisão: 0,1202
  - Pseudo-abrangência: 0,0925
  - Respostas correctas: 208
  - Tópicos com respostas correctas: 59

Conclusões

- Expansão de sinónimos com bons resultados, bem como a segmentação prévia dos tópicos em sintagmas – principalmente nos SVs
- Ideias para melhorar a abrangência das *queries* e os resultados:
  - Expansão de termos nos seus hipónimos (para além dos sinónimos)
  - Melhorar a lista de artigos a excluir dos resultados
  - Determinar o valor ideal de resultados a devolver (talvez flexibilizando-o de tópico para tópico)

Obrigado!

Uma Abordagem ao PÁGICO baseada no Processamento e Análise de Sintagmas dos Tópicos