propor 2008

International Conference on Computational Processing of Portuguese Language

Applications of Portuguese Speech and Language Technologies

Applications of Portuguese Speech and Language Technologies - Propor 2008 Special Session

Hosted by:

Universidade de Aveiro

Promoted by:

Microsoft Language Development Center

Propor 2008 Special Session Commitee

Special Session Chair

- **António Teixeira -** DETI/IEETA, Universidade de Aveiro, Portugal

Organising Committee

- **Daniela Braga,** Microsoft Language Development Center, Portugal
- **Miguel Sales Dias,** Microsoft Language Development Center, Portugal
- **António Teixeira -** DETI/IEETA, Universidade de Aveiro, Portugal

Programme Committee

- **António Teixeira -** DETI/IEETA, Universidade de Aveiro, Portugal
- **Daniela Braga,** Microsoft Language Development Center, Portugal
- **Vera Strube de Lima,** Pontifícia Universidade Católica do Rio Grande do Sul, Brasil
- **Luís Caldas de Oliveira,** INESC-ID/IST, Portugal

Editorial Board

- **Daniela Braga,** Microsoft Language Development Center, Portugal
- **Miguel Sales Dias,** Microsoft Language Development Center, Portugal
- **Luanda Braga Batista,** Microsoft Language Development Center, Portugal

IdSay: A Question Answering system for Portuguese powered by Wikipedia

Gracinda Carvalho, David Martins de Matos, Vitor Rocio

Universidade Aberta L2F/INESC-ID Lisboa CITI – FCT/UNL,
Rua da Escola Politécnica, 147. 1269-001 Lisboa, Portugal
{ gracindac, vjr}@ niv-ab.pt
Instituto Superior Técnico/UTL
Rua Alves Redol 9,1000-029 Lisboa, Portugal
david.matos@inesc-id.pt

Extended Abstract

IdSay is a system for answering open domain questions in the Portuguese language that are liable to be answered in an encyclopaedic repository. The system indexes a collection of texts and then it is ready to answer questions. The knowledge base of the current demonstration is the Portuguese version of Wikipedia from November 2006.

The system starts by analyzing the question, to determine what kind of information is required, and it extracts the entities from the question (set of words that occur together). Afterwards it looks for documents related to the question, that is to say documents that contain the words of the question together with the entities found, if any. If there are no documents retrieved, the process is repeated without the most frequent word. Once we have the list of documents, we analyze it to extract the passages related to the question and from them we produce an answer to the question, depending on the nature of the question. Finally the system returns answers ordered by a frequency based scoring mechanism.

IdSay was built in order to minimize the response time. The indexing of the HTML version of Wikipedia, which consists of 602,002 files in 20,048 directories with a total of 7,13 GB, took about 4 hours, producing an index file with 204,161 documents, 726,659 distinct words, 60,397,743 total words, 215,757 entities and 451 MB. The index takes about 1 minute to load into memory, and a question is generally answered in less than 1 second.

The question types treated can be the search for simple facts, for example:
Q: Qual a altitude da Serra do Buçaco? [What is the altitude of the "Serra do Buçaco"?]
A: 549 m

The other type of question treated is definitions, for instance:
Q: O que é a Curia? [What is "Curia"?]
A: A Curia é uma localidade Portuguesa, localizada na freguesia de Tamengos (concelho de Anadia), onde se podem encontrar as famosas termas com o mesmo nome [Curia is a Portuguese location, in Tamengos (Anadia), where the famous baths with the same name can be found]

Though the nature of the data is unstructured, natural language text, we try to take advantage of the structure of our data source. For instance in the case of definitions, and if there is a page with the name of the entity we are searching for in Wikipedia, we use the first phrase of that page as definition, as was the case of the last question. However this is not a limitation because we can ask for definitions of entities that do not have a Wikipedia page with that name, as the following question:
Q: O que é a UNICEF? [What is the "UNICEF"?]
A: fundo das nações unidas para a infância [The United Nations Children's Fund]

The main entities in the question should be written in upper case, or enclosed in the following three punctuation marks: " ", ' ' or « », otherwise the concept can pass unnoticed.

The translation of titles of famous art works from other languages, that have in several cases more then one translation for Portuguese, sometimes different translations for the Portuguese market and the Brazilian market. In this case it may be worth searching for the original name.

Here is an example, using the name of an American film:
Q: Quem é o realizador do filme "The African Queen"? [Who is the director of the film "The African Queen"?]

A: John Huston (which is the correct answer)

Q: Quem é o realizador do filme "A Rainha Africana"? [Name of the film for Portugal]
Q: Quem é o realizador do filme "Uma Aventura na África"? [Name of the film for Brazil]
Both cases do not succeed in finding the correct answer.

IdSay was submitted for evaluation for the first time at the Portuguese monolingual task of QA@CLEF 2008. More information in Working Notes paper of the CLEF Workshop.
We invite you to test the system in the current demonstration session and let us know your comments and suggestions.