IMIX in Amsterdam: FactMine

Erik Tjong Kim Sang
Informatics Institute
University of Amsterdam

FactMine

- FactMine: Fact and Ontology Mining for Question Answering
- Running time: 1 November 2004 - 30 September 2007
- Employs one postdoc

Approach

The goal of the project is to develop unsupervised methods for extracting **facts** and **ontological information** from text.

Texts will be processed with appropriate natural language processing tools like a part-of-speech tagger and a named entity recognizer.

Hand-crafted and learned patterns will be applied to the processed text in order to extract interesting facts.

We will use clustering techniques for discovering useful relations between phrases in the texts.

Fact extraction example

As a first experiment, keywords with the first description sentence from the Spectrum encyclopedia were stored in a database.

The length of the article describing the keyword has been chosen as an indicator for its importance.

Simple questions can be answered with the resulting database, like:

```
Wat is RSI?
Waar staat de Taj Mahal?
Wie is de koningin van Nederland?
```

Clustering example

In a second experiment, all description-initial nouns associated with presumed persons were extracted from the Spectrum encyclopedia.

The resulting list was compared with a list extracted from the Dutch part of EuroWordNet.

Here are ten words missing from EuroWordNet that were most frequent in the encyclopedia: `schrijfster`, `filmregisseur`, `chemicus`, `staatsman`, `actrice`, `astronoom`, `zangeres`, `filmacteur`, `journalist` and `dichteres`.

We would like to apply these techniques for **expanding EuroWordNet**.

Plans for first project year

- Adapt existing preprocessing NLP tools
- Develop useful patterns for fact extraction
- Mine facts from encyclopedic and online sources
- Generate and make available fact bases
- Examine clustering techniques for extracting phrase relations

Situating FactMine in Imix

The project will generate useful resources for the two projects dealing with question answering (ROLAQUAD and QADR): **fact bases** and **ontological information**.

FactMine will be a subcontractor for ROLAQUAD and in order to ensure proper coordination between the two projects, the FactMine postdoc will work in Tilburg for one day in a week.

Additionally we will develop a QA engine which will answer questions by only performing table lookup.

THE END