Introduction to FactMine

Erik F. Tjong Kim Sang, Informatics Institute, University of Amsterdam

General

- **Framework**: FactMine is part of the research programme Interactive Multimodal Information Extraction (IMIX) which is sponsored by the Netherlands Organization for Scientific Research (NWO).
- **Goal**: to develop and evaluate unsupervised methods for the extraction of **fact bases** and **ontological information** from text.
- **When**: November 2004–September 2007

Approach

- **Preprocessing**: natural language processing tools, like part-of-speech taggers and named entity recognizers, are used for automatically annotating texts.
- **Fact extraction**: hand-crafted and learned information extraction patterns are applied to the processed text in order to derive interesting facts.
- **Finding classes**: clustering techniques are employed for discovering useful concept classes and concept relations.

Evaluation

- **Intrinsic evaluation**: the developed techniques are evaluated regularly by counting the number of extracted facts and relations, and by estimating the quality of random samples.
- **Extrinsic evaluation**: the performance of a question answering system is measured, both with and without the extracted resources.

Project status

- **Database contents**: information pairs (keywords and description-initial sentences) from two encyclopedias.
- **Intrinsic evaluation**: 64674 facts were derived with an estimated correctness of 96%.
- **Extrinsic evaluation**: first-answer quality of a QA system measured on Dutch CLEF-2003: precision 72% and recall 2%.

Question examples

Wat is RSI?

(repetitive strain injury) Algemene benaming voor (blijvende) beschadigingen aan het lichaam ten gevolge van langdurige, eentonige werkzaamheden.

Wie is Cruijff?

Hendrik Johannes (Johan) Cruijff: (geb. 1947) Nederlands voetballer, kwam als voorhoedespeler 48 maal uit voor het Nederlands elftal.

Hoe heet de koningin van Nederland?

koningin Emma

References

- M. Fleischman, E. Hovy, and A. Echihabi, Offline Strategies for Online Question Answering: Answering Questions Before They Are Asked. In: *Proceedings of ACL-2003*.
- V. Jijkoun, G. Mishne, and M. de Rijke. Preprocessing Documents to Answer Dutch Questions. In: *Proceedings of BNAIC'03*.
- V. Jijkoun, M. de Rijke, and J. Mur. Information Extraction for Question Answering. In: *Proceedings of COLING 2004*.
- B. Roark and E. Charniak, Noun-Phrase Co-Occurrence Statistics for Semi-Automatic Semantic Lexicon Construction. In: *Proceedings of ACL-98*.
- L. IJzereef, *Automatische extractie van hyponiemrelaties uit grote tekstcorpora*. Masters thesis, 2004.