

DERIVING KNOWLEDGE FROM DUTCH MEDICAL TEXT

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

General

- **Framework:** The NWO-sponsored project FactMine (number 264-70-050) is part of the research programme Interactive Multimodal Information Extraction (IMIX).
- **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 October 2005

- **Database contents:** medical definitions and disorder information on cause, diagnosis, prevention, symptoms and treatment have been extracted from two encyclopedias.
- **Intrinsic evaluation:** the 4718 derived facts have an estimated correctness of 99%.
- **Extrinsic evaluation:** the first answer quality of a QA system measured on 50 RSI questions was: precision 46% and recall 6%.
- **Demo:** the derived information tables are accessible via a question answering interface at <http://staff.science.uva.nl/~erikt/factmine/>

Question examples

Wat is slaapziekte?

infectieziekte die wordt veroorzaakt door protozoon (eencelligen) uit het geslacht Trypanosoma.

Hoe weet je dat je het hebt?

Door middel van microscopisch onderzoek van bloed, vocht uit de lymfeklieren, hersenvocht (liquor) of beenmerg kunnen de trypanosomen aangetoond worden...

Wat kan je ertegen doen?

Doden van de trypanosomen met middelen als suramine en pentamidine...

References

- Erik Tjong Kim Sang, Gosse Bouma and Maarten de Rijke, Developing Offline Strategies for Answering Medical Questions. In: *Proceedings of the AAAI-05 Workshop on Question Answering in Restricted Domains*, Pittsburgh, USA, 2005, pp. 41-45.

