

Event-coreference across Multiple Multi-lingual Sources in the MUMIS Project*

**Jan Kuper[§], Horacio Saggion[†],
Hamish Cunningham[†], Thierry Declerck[‡],
Ed Hoenkamp[¶], Marco Puts[¶],
Franciska de Jong[§], Yorick Wilks[†], Peter Wittenburg[#]**

[§]Department of Computer Science, University of Twente, The Netherlands

[†]Department of Computer Science, University of Sheffield, UK

[‡]DFKI GmbH, Saarbruecken, Germany

[¶]NICI, University of Nijmegen, The Netherlands

[#]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

email: jankuper@cs.utwente.nl

This paper describes work done in the context of the MUMIS project, especially on multi-document information extraction. The MUMIS project aims at searching and indexing video material on the domain of soccer, based on information extracted from textual documents in various languages.

As a case study, the project examined the European Soccer Championship, which took place in 2000.

The main line of the process is as follows:

- *Information extraction* from individual texts in various languages (English, German, Dutch) by information extraction systems developed in Saarbruecken, Sheffield, Enschede. This information extraction is based on an ontology and lexicon created for the soccer domain, and delivers key events (such as corner, shot-on-goal, free-kick) together with the names of the players involved in the event, and the time the event took place during a soccer match.
- *Merging* the results from the various information extraction systems into a combined result. Information extracted from single documents turns out to be often incomplete, or even incorrect. The merging part of the project uses domain knowledge, expressed in constraints, to combine partial information elements into more complete knowledge.

The novelty of the project is the merging component, and the presentation will emphasise on this component. In some greater detail, the merging component consists of the following steps:

* *MULTI-Media Indexing and Searching* (parlevink.cs.utwente.nl/projects/mumis),
Funded by EC's 5th Framework HLT Programme under grant number IST-1999-10651

- *Two document alignment* to decide which (partial) events extracted from two documents do stem from the same event in reality,
- *Multi-document alignment* combines these two-document alignments into larger structures of corresponding events, where each structure contains (partial) events which are assumed to belong together, i.e., which are about the same event in reality,
- *Unification* of events based on domain knowledge expressed in constraint rules. Erroneous elements of partial events are corrected, and empty slots are filled in. There are various types of rules used in this step: event internal rules, rules expressing possible combinations of events, rules based on the role of the teams during the corresponding moment in the match,
- *Ordering* events in the right temporal order. For example, a shot-on-goal and a goal will typically occur in that order, and not in the opposite order.

Evaluation results show a substantial increase of the quality of the extracted information.

References

This paper is a summary of papers published a.o. at EACL'03 and IJCAI'03:

- H. Saggion, J. Kuper, H. Cunningham, T. Declerck, P. Wittenburg, M. Puts, E. Hoenkamp, F. de Jong, Y. Wilks, *Event-coreference across Multiple Multi-lingual Sources in the MUMIS project*, in: Conference Companion to the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03), Budapest, April 12–17, 2003, 239–242
- J. Kuper, H. Saggion, H. Cunningham, T. Declerck, F. de Jong, D. Reidsma, Y. Wilks, P. Wittenburg, *Intelligent Multimedia Indexing and Retrieval through Multi-source Information Extraction and Merging*, in: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03), Acapulco, August 9–15, 2003 409–414