

1 Project Details

1.1 Project Title. Fact and Ontology Mining for Question Answering

1.2 Project Acronym. FactMine

1.3 Summary. The goal of the proposed project is to develop unsupervised methods for the extraction of ontological information from texts. We will examine three techniques: natural language processing, pattern extraction, and clustering. We expect to employ suitable versions of these techniques for expanding the Dutch version of EuroWordNet [34] and thus create an ontology which can be used in the open domain question answering part of the IMIX Demonstrator. We will also apply these techniques to texts of a restricted domain in order to evaluate their usefulness for such domains.

This proposal fits in the paradigm of the Semantic Web, an effort for adding semantic annotation to online texts. For storing facts and relations the project will use the Resource Description Framework (RDF) and the Web Ontology Language (OWL), both developed within the Semantic Web effort.

By providing techniques for generating ontologies and other knowledge sources, the present proposal relates to three of the thematic priorities which were defined to be relevant for the IMIX Demonstrator: clarification dialogues, which require ontologies for detecting problems in the questions, architectural problems caused by slow response time as a result of large collection document search, for which we propose an alternative, and answer construction from multiple documents.

2 Applicant's Data

Prof.dr. Maarten de Rijke

Language & Information Technology Group (LIT), Informatics Institute, University of Amsterdam.

E-mail: mdr@science.uva.nl URL: <http://lit.science.uva.nl/>

3 Co-applicant(s) (factsheet IRIS)

Dr. Maarten Marx

Language & Information Technology Group (LIT), Informatics Institute, University of Amsterdam.

E-mail: marx@science.uva.nl URL: <http://lit.science.uva.nl/>

4 Previous and Future Submissions

This is a new proposal which has not been submitted elsewhere.

5 Institutional Setting

The project will be embedded within the Language & Inference Technology (LIT) Group of the University of Amsterdam. The LIT Group (with around twenty staff, post-docs, and Ph.D. students) is a major player in the experimental evaluation of Information Retrieval systems. The group has become a regular participant in the Cross-Language Evaluation Forum (CLEF); the Text REtrieval Conference (TREC); and the Initiative for the Evaluation of XML Retrieval (INEX). With its unique combination of experts in information retrieval, applied natural language processing, and knowledge representation and reasoning, the LIT Group is ideally positioned to conduct innovative and technologically relevant research in Question Answering (QA). The LIT Group has extensive experience in QA research and in building QA systems that work; it is the only Dutch team to have participated in the TREC QA track since 2001, and the only academic research team to have developed a QA system for Dutch.

6 Period of Funding

The project will have a running time of 35 months. The projected starting date is November 1, 2004.

7 Composition of the Research Team

Name	Title	Role	Expertise	Affiliation
de Rijke, M.	Prof.dr.	Applicant	language and inference technology	LIT, UvA
Marx, M.	Dr.	Co-applicant	semistructured data	LIT, UvA
Tjong Kim Sang, E.F.	Dr.ir.	Postdoc	machine learning and language technology	To be funded by NWO
Buitelaar, P.	Dr.	Advisor	ontologies / language technology for the semantic web	DFKI (Germany)
Daelemans, W.M.P.	Prof.dr.	Advisor	machine learning and language technology	UvT / UA (Belgium)
Mons, B.	Dr.	Advisor	information extraction and retrieval	Collexis B.V. / Bio- semantics, Erasmus MC

8 Thematic Priorities

The proposed project deals with a combination of thematic priorities of the IMIX project. First, as we argue below, the proposed information extraction methods and activities provide direct support the *Clarification dialogues for open domain QA* theme. Second, to facilitate dialogues with QA systems, the proposal suggests so-called offline answering strategies; these strategies are directly relevant for the topic *Construction of answers by combining information from multiple documents*. Finally, IE is a key component of any QA system, whether open domain or restricted domain; by developing dedicated IE methods for Dutch and by proposing to generate extensive knowledge sources (which will facilitate “responsive” interactive QA), the proposed project also addresses the topic *Issues related to the architecture of systems supporting multimodal interaction*.

9 Description of the Proposed Research Project

9.1 Scientific Problem and Aim. Given a collection of documents and a question, question answering (QA) systems have to find an answer to the question in the collection. At a sufficiently abstract level it is natural to view the QA process as a two-step process: first locate documents that are likely to contain the answer(s), and then extract the answers from those documents. In a slogan, QA is information retrieval (IR) plus information extraction (IE): “QA = IR + IE.” In this proposal we focus on information extraction aspects of QA, that is, on turning mostly unstructured text into structured and semistructured knowledge sources. We are especially interested in constructing two kinds of semistructured knowledge sources: (natural language) *ontologies* and *fact bases*. For us, an ontology is a knowledge source that provides a conceptualization of a particular domain, and a fact base provides a (large) collection of instances of relations and concepts in a particular domain.

IMIX foresees the development of a QA demonstrator that must be able to operate in two modes: open domain QA and interactive QA in a restricted domain. Natural language ontologies such as WordNet [21] contain lexical relations between concepts. In applications of natural language processing, such ontologies provide the required background knowledge. In particular, in QA they are widely used in many stages of the typical QA pipeline: e.g., in query formulation, in question classification, in answer justification, and in answer type checking filtering. The importance of natural language ontologies like WordNet has been stressed again and again by numerous TREC participants [33]. For Dutch only a relatively small WordNet exists [6, 34]; it is much smaller than the English WordNet, and we have found it to be too sparsely populated to be really useful in the setting of open domain QA right now [15]. The availability of an extended Dutch WordNet will directly benefit QA research in the Netherlands in general, and in the IMIX programme in particular.

Within the IMIX programme there are additional reasons that make the creation of WordNet-like resources essential. Clarification dialogues are collections of utterances with the purpose of clarifying or elaborating on an earlier utterance. In human-machine interaction, a clarification dialogue can be used for obtaining more information related to utterances from the human or those from the machine.

The dialogue can be invoked by misunderstandings, for example in case of speech input, detected inconsistencies or a lack of information for a successful processing of the previous utterance. One of the core motivations of this project is to facilitate clarification dialogues started by a machine in order to obtain more information regarding a human utterance especially in the case of observed inconsistencies and information gaps.

It is challenging for a QA system to detect question inconsistencies and missing information if it only has access to text documents. The best method for finding information gaps is to look for the question words and their relations in an *ontology* that conceptualizes the domain covered by the QA system. If there is no relation between the question words, the question may contain an inconsistency. In case the ontology suggests that there might be multiple relevant answers, or that certain key concepts in the question are not sufficiently specific, the system might benefit from additional information from the user. The following example illustrates this scenario:

Human: Wie is de president van Korea?
(English: *Who is the president of Korea?*)
System: Bedoelt u Noord-Korea of Zuid-Korea?
(English: *Do you mean North Korea or South Korea?*)
Human: Zuid-Korea
(English: *South Korea*)
System: Roh Moo-hyun

In sum, we propose to use the following approach: use ontologies for detecting inconsistencies and information gaps in user utterances which might give rise to a clarification dialogue. This, obviously, requires the availability of broad coverage ontologies of the kind described before.

As to the need for broad coverage *fact bases*, that is, collections of “ontology instances,” it is important to notice the following. The amount of processing needed to adequately provide the levels of exactness required by QA systems is such that QA systems that perform “live” retrieval and extraction will simply be too slow and impractical for ordinary users. This is particularly relevant for the IMIX programme, as it aims to have an *interactive* dialogue-based demonstrator, where, we take it, responsiveness is a key issue. In order to meet this challenge we propose a strategy in which information is extracted automatically from electronic texts offline, and stored for quick and easy access [7, 13].

In the open domain QA case the core idea is this. We determine what the most important (or most frequently asked) question types are, and for each of the important question types we build a knowledge source from which questions of that type are to be answered. In previous TREC activities we hand-crafted a small number of surface patterns for extracting information for the most important question types. For instance, the “Location” category concerns geographic information of the following type “Amu Darya, river, Turkmenistan, XIE19990811.0277,” where the first field indicates a location, the second its type, the third a country or region in which it is located, and the fourth the identifier for the document from which it was extracted. Following very much the same ideas, in our earlier TREC and CLEF participations we also extracted more restricted, and specialized, biographical information and made these available for rapid access in fact bases for online interaction.

Within the IMIX programme, and specifically, to facilitate an interactive demonstrator, we propose this offline strategy, both for the open domain and for the planned restricted (medical/RSI) domain. However, instead of building the required fact bases with hand-crafted extraction patterns, we believe that the only scalable and “maintainable” route here is to deploy automatic extraction methods.

All in all, then, the aim of this project is three-fold:

- First, to develop robust algorithms for knowledge extraction, and in particular, for ontologies and for fact bases.
- Second, to understand how our algorithms deal with the following scenarios: extending the Dutch WordNet, open domain vs. restricted domain (medical/RSI within IMIX), and relatively clean newspaper data vs. noisy data.
- Third, to deliver, distribute, and help integrate the ontologies and fact bases created within the proposed project within the planned IMIX demonstrators.

9.2 Scientific Setting and Method. The methods we propose for knowledge extraction from texts are *unsupervised*. We plan to examine three techniques: natural language processing, pattern detection and clustering; see Figure 1. These techniques will be used for expanding the Dutch version of EuroWordNet [34] and thus create ontology and fact bases which can be used in the open domain QA part of the IMIX Demonstrator. We will also apply these techniques to texts of a restricted domain in order to evaluate their usefulness for such domains. Let us take a look at the details.

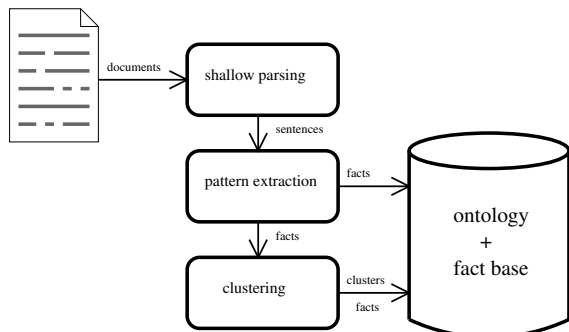


Figure 1: Schematic overview of the proposed approach. After parsing the documents, pattern extraction identifies facts which are converted to clusters by the clustering algorithm.

lemmatization, part-of-speech tagging, named entity tagging, chunking, and relation finding. Most of these tasks have been implemented with training material from Corpus Gesproken Nederlands (Corpus Spoken Dutch) [25]. For the purpose of ontological information extraction, the relation finding task is the most important. Here is an example sentence with roles marked up explicitly:

(Subject De Tweede Kamer Subject) (Modal wil Modal) (Object het hoogspanningsnet Object) (Location in Nederland Location) (Verb nationaliseren Verb)
 (English: The Dutch parliament wants to nationalize the high-voltage network in The Netherlands.)

This sentence contains one main verb (*nationalize*) which has a subject and an object. In more complex sentences, phrases can be linked to different verbs. The present version of the shallow parser does not attempt to solve the prepositional phrase attachment problem: should prepositional phrases be attached to the verb or to the previous noun phrase? However, the modular architecture of the shallow parser makes the addition of a prepositional phrase attachment module easy, although, of course, the implementation of a good attachment module remains nontrivial.

After syntactically analyzing documents the next step is to extract patterns from the documents. LIT has experience with producing string-based patterns for QA, like *Europees parlement in Straatsburg* for the question *In welke stad is het Europees Parlement?* (English: In what city is the European Parliament?) [14]. These patterns are useful for ontology and fact base building as well. This project will re-use the available LIT technology for extracting facts with these patterns.

The linguistic preprocessing step applied to the input texts allows us to extract facts based on previously unavailable semantic patterns. From the parsed example sentence, we can extract the following facts: `subject(nationaliseren,Kamer)`, `object(nationaliseren,hoogspanningsnet)`, and `sv0(Kamer,nationaliseren,hoogspanningsnet)`. In the facts we summarize the noun phrases by using their head, the main word of the phrase. This makes it easier to perform generalization over the phrases used in the facts.

In the next step, sets of semantically related words (clusters) can be extracted from the facts. For example, in the facts of the format `subject(nationaliseren,X)`, most of X will be government organizations and persons. In the format `object(nationaliseren,Y)`, Y will be companies or resources. Sets can be related as the following medical text example from Reinberger et al. [28] shows:

objects of acquire: hepatitis infection virus disease
objects of transmit: hepatitis infection disease

The lists of objects of the two verbs are almost identical. From these lists we can infer a new fact that was not found in the text documents: `object(transmit,virus)`. The approach we want to apply for detecting such new facts is clustering. This method combines sets when certain criteria are satisfied. In this particular example, it could have decided to merge the two sets. This would not have had an effect on the objects of *acquire* but it would add one element to the objects of *transmit*, thereby producing the new fact. Apart from generating new facts, clustering should also produce interesting sets of concepts. The concept sets would be most useful for a QA ontology if they were labeled, something which they are not. However, it is imaginable that a QA system can benefit even from unlabeled concepts when it is capable of some interaction with the user:

Human: Noem drie huisdieren
(English: *Give me three examples of pets*)
System: Ik weet niet wat een huisdier is. Kunt u een voorbeeld geven?
(English: *I do not know what a pet is. Can you give me an example?*)
Human: Hond
(English: *Dog*)
System: Drie voorbeelden van huisdier zijn hond, kat en goudvis.
(English: *Three examples of pets are dog, cat and gold fish.*)

Let us turn to some practical issues now. First, the extracted facts and concept sets will be stored in RDF/XML [20]. This is a standard method for encoding semantic information in the World Wide Web. It is widely used for storing ontologies, for example in the context of the Semantic Web. We foresee an interest in the usage of the extracted facts and concepts from both IMIX partners and people and groups outside the project. When the data is available, we are planning to give access to it to the Dutch and international research community via online tools.

Next, the approach sketched here relies heavily on the availability of text. Currently, we have access to two types of corpora we believe are useful to this project. First, there is a Dutch encyclopaedia by Het Spectrum (about 7 million words). Second, there is the Twente Nieuws Corpus (300 million words). The two corpora nicely complement each other, with one containing more structured general information, while the other is a large collection of unstructured news data. The Corpus Gesproken Nederlands [25] will be used as training material for further development of the shallow parser.

As to evaluation, the quality of the facts and concept sets that will be generated for the open domain ontology and fact bases, can be assessed by the proposed postdoc; the resulting knowledge sources will be evaluated as components of the overall QA task. Evaluating the quality of the data extracted in the restricted (medical) domain cannot be done by the postdoc (who lacks domain expertise). Here, we will rely on the expertise and advice of one of our team members, Dr. B. Mons.

9.3 Scientific Significance. Recently, and motivated mainly by performance issues, offline strategies for QA have witnessed a considerable increase in interest, although QA systems that use offline mining strategies have been online at least since 1993 [16]. To a large extent QA systems relying on offline strategies build on a body of work on extracting semantic information using lexical patterns. Hearst [10] explored the use of lexical patterns for extracting hyponym relations, with patterns such as “such as.” Berland and Charniak [2] extract “part-of” relations. Mann [19] describes a method for extracting instances from text that takes advantage of part-of-speech patterns involving proper nouns.

The use of lexical patterns in the setting of QA received lots of attention after a team taking part in one the earlier QA Tracks at TREC showed that the approach was competitive at that stage [26, 30]. Hermjakob et al. [11] showed that answer retrieval could be improved by searching for predefined paraphrases of frequent questions. Ravichandran et al. [27] collect surface text patterns automatically in an unsupervised fashion using a collection of trivia question and answer pairs as seeds. These patterns are then used to generate features for a statistical QA system. Jijkoun et al. [12, 15] combine the extraction of surface text patterns with WordNet-based filtering of name-apposition pairs to increase

precision, but found that it hurt recall more than it helped precision. Fleischman et al. [7] focus on the precision of the information extracted using simple part-of-speech patterns. They present an effective machine learning method for filtering noise from the collected data.

Information extraction for ontology building is a new and growing field. Gomez-Pérez and Monzano-Macho [8] list 18 approaches and 18 tools in a recent survey, nearly all of them referring to work published in 2000 or later. Ten of these use some kind of parsing for structuring the information in the source documents. The goal of most of the approaches is ontology expansion, which means that they rely on the existence of an initial ontology [1, 9, 17, 22, 23, 29]. Xu et al. [35] apply an unsupervised method for ontology learning with linguistic analysis limited to the chunk level. Thompson and Mooney [31] produce a more elaborate linguistic analysis of sentences and use a combination of heuristics and logical rules for generating facts to be added to a semantic lexicon. Oliveira et al. [24] parse texts and generate concept trees in an interactive way. The techniques used by Bisson et al. [3] are very close to those we propose. Their linguistic analysis converts the texts to triplets containing a verb, one of its roles, and the frequency of the pair, e.g., `object(cause,decrease,29)`. Phrases appearing in the same role for the same verb form sets which are combined with clustering techniques.

Maedche and Staab [18] apply shallow parsing to German text of a restricted domain and extract concepts which are presented to humans for validation. Reinberger et al. [28] employ shallow parsing for extracting noun-verb-noun relations from medical text and apply several clustering techniques to the sets of verbs and nouns thus obtained. They observe that even when an example ontology is available, the evaluation of the results is difficult. The clusters that are produced are unlabeled and generating labels for them is one of the future goals of the authors.

The most important quality of our ontology learning approach when compared with the systems described in this section is that it is *unsupervised*. Our approach does not rely on the existence of an initial ontology. This is important because, as stated by Bisson et al. [3], the existence of a general ontology might not be sufficient for constructing an ontology for a restricted domain because of gaps in the general lexicon. We believe that our shallow parsing technology is superior to most of the linguistic tools employed by the systems listed here. We expect that because of this technology our unsupervised approach will be robust and will be better equipped for extracting ontological information even for new restricted domains with their own specific vocabularies.

9.4 Contributions towards the IMIX Goals. The IMIX Programme “aims at funding research in [...] interactive multimodal information processing, resulting in operational software that can be integrated into the demonstrator system [...]” As we have argued in Section 9.1, the kind of knowledge sources that we aim to build, and for which we plan to develop extraction tools, are essential for operational QA systems in general, and for the envisaged interactive demonstrator in particular. The proposed project also contributes towards some of the “small programs” that have already started up within IMIX. Specifically, the Rolaquad project will benefit from the knowledge bases we aim to build (these are sources against questions can be asked), and the “Question Answering for Dutch using Dependency Relations” will benefit from the general ontologies we will create—no QA system can do without. Finally, the proposed project offers a unique opportunity to bring the extensive QA expertise of the LIT group to bear on the IMIX Programme and its demonstrators.

9.5 Innovative Aspects. Although application of natural language processing (NLP) technology for building ontologies seems obvious, most information retrieval systems rely on statistical methods based on a bag of words model [33]. Reasons for this are the lack of either coverage or analysis depth of former generation NLP tools. With the proposed shallow parsing approach we put forward an NLP method that is both robust and has a sufficient level of sentence analysis depth for making possible interesting approaches for knowledge extraction. Another innovative aspect of the proposed project is the suggested application of language technology for fact extraction from documents. As far as we know, it the first of such an application for Dutch open domain question answering. Additionally, as pointed out before, we believe that the most important quality of our ontology learning approach when compared with systems described in the literature is that it is *unsupervised*. One further innovative aspect is the integration of Semantic Web technology within the QA setting; we believe that ontology-based approaches and tools will provide a significant increase in the quality of information retrieval

methods in general, and that they are indispensable in the QA setting.

9.6 Societal, Cultural, and Technological Relevance. The significance of the project lies in several areas, both societal and scientific. With its ultimate aim of laying the groundwork for intelligent ways of accessing Dutch language information sources that will enable users to obtain real answers to real questions, the long-term societal benefits of the proposal are obvious. To this end, ontologies and fact bases to be created with the IMIX partners, and to the extent allowed by the licenses governing the corpora from which the information has been extracted, with other research teams in the Netherlands and elsewhere. The main scientific interest of the proposal is its development of robust and linguistically informed methods for knowledge extraction. Extraction methods that combine robustness with “intelligence” are of tremendous importance — meeting these two, often conflicting requirements is far from trivial and has been one of the “holy grails” in artificial intelligence. The unsupervised methods we propose try to meet this challenge head on.

9.7 References

- [1] E. Alfonseca and P. Rodríguez. Automatically generating hypermedia documents depending on user goals. In *Workshop on Document Compression and Synthesis in Adaptive Hypermedia Systems (AH-2002)*. Málaga, Spain, 2002.
- [2] M. Berland and E. Charniak. Finding parts in very large corpora. In *Proc. 37th Annual Meeting of the Association for Computational Linguistics*, 1999.
- [3] G. Bisson, C. Nédellec, and D. Cañamero. Designing clustering methods for ontology building - the mo'k workbench. In *Proc. of the workshop on Ontology Learning at ECAI-2000*, pages 13–19. Berlin, Germany, 2000.
- [4] S. Buchholz. *Memory-Based Grammatical Relation Finding*. PhD Thesis, University of Tilburg, 2002.
- [5] M. Craven, D. DiPasquo, D. Freitag, A. K. McCallum, T. M. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the World Wide Web. *Artificial Intelligence*, 118:69–113, 2000.
- [6] EuroWordNet. Building a multilingual database with wordnets for several European languages. URL: <http://www.illc.uva.nl/EuroWordNet/>.
- [7] M. Fleischman, E. Hovy, and A. Echihiabi. Offline strategies for online question answering: answering questions before they are asked. In *Proc. ACL 2003*, 2003.
- [8] A. Gomez-Pérez and D. Monzano-Macho. *A survey of ontology learning methods and techniques*. OntoWeb technical report, deliverable 1.5, 2003.
- [9] U. Hahn and K. Markó. Joint knowledge capture for grammars and ontologies. In *First International Conference on Knowledge Capture (K-Cap 2001)*. Victoria, B.C., Canada, 2001.
- [10] M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proc. Fourteenth International Conference on Computational Linguistics*, 1992.
- [11] U. Hermjakob, A. Echihiabi, and D. Marcu. Natural language based reformulation resource and web exploitation for question answering. In *Proc. TREC 2002*, 2003.
- [12] V. Jijkoun, G. Mishne, and M. de Rijke. Preprocessing Documents to Answer Dutch Questions. In *Proc. BNAIC'03*, 2003.
- [13] V. Jijkoun, G. Mishne, C. Monz, M. de Rijke, S. Schlobach, and O. Tsur. The University of Amsterdam at the TREC 2003 Question Answering Track. In *Proc. TREC 2003*, 2004.
- [14] V. Jijkoun, G. Mishne, and M. de Rijke. Building infrastructure for Dutch question answering. In A. de Vries, editor, *Proc. DIR 2003*, 2003.
- [15] V. Jijkoun, G. Mishne, and M. de Rijke. How frogs built the Berlin Wall. In *Proc. CLEF 2003*, LNCS. Springer, 2004.
- [16] B. Katz. Annotating the World Wide Web Using Natural Language. In *Proc. RIAO '97*, 1997.
- [17] D. Lonsdale, Y. Ding, D. Embley, and A. Melby. Peppering knowledge sources with salt; boosting conceptual content for ontology generation. In *Proc. AAAI Workshop on Semantic Web Meets Language Resources*, 2002.
- [18] A. Maedche and S. Staab. Discovering conceptual relations from text. In *Proc. ECAI 2000*, pages 321–325, 2000.
- [19] G. Mann. Fine-grained proper noun ontologies for question answering. In *SemaNet'02: Building and Using Semantic Networks*, 2002.
- [20] F. Manola and E. Miller. *RDF Primer*. World Wide Web Consortium, 2003.
- [21] G. A. Miller. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312, 1990.
- [22] M. Missikoff, R. Navigli, and P. Velardi. An integrated approach for web ontology learning and engineering. *IEEE Computer*, November:60–63, 2002.
- [23] D. Moldovan, R. Girju, and V. Rus. Domain-specific knowledge acquisition from text. In *Proc. ANLP-2000*, 2000.
- [24] A. Oliveira, F. C. Pereira, and A. Cardoso. Automatic reading and learning from text. In *Proc. ISAI 2001*, 2001.
- [25] N. Oostdijk and D. Broeder. The Spoken Dutch Corpus and its exploitation environment. In *Proc. LINC-03*, 2003.
- [26] D. Ravichandran and E. Hovy. Learning surface text patterns for a question answering system. In *Proc. ACL 2002*, 2002.
- [27] D. Ravichandran, A. Ittycheriah, and S. Roukos. Automatic derivation of surface text patterns for a maximum entropy based question answering system. In *Proc. HLT-NAACL 2003*, Edmonton, Canada, 2003.
- [28] M.-L. Reinberger, P. Spyns, W. Daelemans, and R. Meersman. Mining for lexons: applying unsupervised learning methods to create ontology bases. In *Proc. of ODBase'03*. Catania, Italy, Springer-Verlag, 2003.

- [29] C. Roux, D. Proux, F. Rechermann, and L. Julliard. An ontology enrichment method for a pragmatic information extraction system gathering data on genetic interactions. In *Proc. ECAI 2000 Workshop on Ontology Learning*, 2000.
- [30] M. Soubbotin and S. Soubbotin. Patterns of potential answer expressions as clues to the right answer. In *Proc. TREC-10*, pages 134–143, 2001.
- [31] C. A. Thompson and R. J. Mooney. Semantic lexicon acquisition for learning natural language interfaces. Technical Report AI98-273, 1, 1998.
- [32] E. F. Tjong Kim Sang, W. Daelemans, and A. Hothker. Reduction of dutch sentences for automatic subtitling, 2004. To be submitted to CLIN-2003.
- [33] E. M. Voorhees. Natural language processing and information retrieval. In *School on Information Extraction, SCIE99*, pages 32–48, 1999.
- [34] P. Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer, 1998.
- [35] F. Xu, D. Kurz, J. Piskorski, and S. Schmeier. A domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping. In *Proc. LREC 2002*. Las Palmas, Canary Island, Spain, 2002.

10 Word Count

Section 9 contains 3348 words.

11 International Perspective

In Section 9.3 we detailed how the proposed project is situated amongst other international developments in the area. Foreign partners that will play an “official” role in the proposed project have been listed in Section 7; we will interact with Buitelaar on issues related to language technology and the Semantic Web, with Daelemans on machine learning and natural language, and with Mons on information extraction, especially in restricted domains. Furthermore, we maintain close contacts with a number of colleagues world-wide, both concerning QA and concerning knowledge extraction methods. As to the QA contacts, the LIT group maintains active research collaborations with groups across Europe—Alicante (Peñas), Edinburgh (Webber), Saarbrücken (Neumann, Uszkoreit), Sheffield (Gaizauskas), Trento (Magnini)—and beyond—CMU, MIT, and UMD. The QA aspects of our proposed research are related to the START system [16], but we go beyond it in the extraction methods that we propose and in the coverage that we foresee. More generally, the LIT group plays an active and highly visible role in QA, both in terms of publications and in terms of organizational involvement in QA-related events (EACL 2003 Workshop on NLP for QA, ACL 2004 Workshop on QA in Restricted Domains, SIGIR 2004 Workshop on IR for QA). Moreover, the LIT group is the coordinator of the Dutch task at the CLEF QA track. Finally, together with researchers at CWI, TNO, and in Twente, the group will host the 2007 edition of SIGIR, the leading annual event on Information Retrieval.

As to knowledge extraction methods, our proposed research is related to ongoing work in Antwerp (Daelemans, Reinberger), Paris (Zweigenbaum), Rotterdam (Mons), Saarbrücken (Buitelaar), and Sheffield (Gaizauskas); we are in close contact with all of the groups involved. Furthermore, our work is related to the Web→KB project [5] but goes beyond it by using different extraction methods and by integrating the resulting knowledge sources in a QA setting.

12 Work Programme

- Yr 1**
 - Adaption of the Dutch shallow parser for unsupervised information extraction.
 - Evaluation of pattern formats for generating facts.
 - Testing and evaluation of clustering techniques for building clusters and extracting facts.
- Yr 2**
 - Testing and evaluation of clustering techniques for building clusters and extracting facts.
 - Generate and evaluate a prototype ontology for open domain question answering.
 - Create an online demo for accessing the generated fact bases.
- Yr 3**
 - Generate and evaluate a prototype ontology for restricted domain question answering.
 - Create an online browsing tool for ontologies.
 - Integration of the ontologies in IMIX Demonstrator.

13 Planned Deliverables and Knowledge Dissemination

- Yr 1**
 - Deliverable 1: Dutch shallow parser adapted for unsupervised information extraction.
 - Deliverable 2: fact base generated from patterns.

- Report 1: report describing deliverable 1.
 - Report 2: report describing deliverable 2.
- Yr 2**
- Deliverable 3: clustering techniques for building clusters and extracting facts.
 - Deliverable 4: prototype ontology for open domain question answering.
 - Deliverable 5: online demo for accessing fact bases
 - Report 3: report describing deliverable 3.
 - Report 4: report describing deliverable 4.
 - Report 5: report describing deliverable 5.
- Yr 3**
- Deliverable 6: prototype ontology for restricted domain question answering.
 - Deliverable 7: online browsing tool for generated ontologies.
 - Report 6: report describing deliverable 6.
 - Report 7: report describing deliverable 7.
 - Report 8: report describing integration of ontology in IMIX Demonstrator.

14 Short CV Principal Applicant(s) and Candidate Postdoc

Maarten de Rijke (Vlissingen, 1961) holds MSc degrees in Philosophy and Mathematics, and a PhD in Theoretical Computer Science. He worked as a postdoc at CWI, before becoming a Warwick Research Fellow at the University of Warwick, UK. He joined the University of Amsterdam in 1998, initially as assistant professor, then as associate professor (2001), and recently he was appointed full professor (Internet and Information Processing). He currently holds one of the prestigious Pionier grants, and has published over 200 papers. He founded the Language & Inference Technology Group in 2001, which has rapidly become one of the leading IR groups in Europe.

Erik F. Tjong Kim Sang (Utrecht, 1966) studied Electrical Engineering at the Delft University of Technology, The Netherlands (1984–1988). He was a PhD Student at the Computational Linguistics Group of the University of Groningen, The Netherlands (1990–1995, PhD 1998) and a lecturer at the Linguistics department of Uppsala University in Sweden (1995–1998). Currently, he is employed at the CNTS - Language Technology Group at the University of Antwerp, Belgium. He was appointed in the European TMR project Learning Computational Grammars (1998–2001) and his current project is Automatic Transcription and Normalisation of Speech (2001–2004).

15 Literature

Ten most significant publications by members of the research team:

1. C.C. van der Eijk, E.M. van Mulligen, J.A. Kors, B. Mons, and J. van den Berg. Constructing an Associative Concept Space for Literature-based Discovery. *J. of the American Society for Information Science and Technology*, to appear.
2. V. Hollink, J. Kamps, C. Monz, and M. de Rijke. Monolingual Document Retrieval for European Languages. *Information Retrieval*, 7:33-52, 2004.
3. V. Jijkoun, G. Mishne, and M. de Rijke. How frogs built the Berlin Wall. In *Proceedings CLEF 2003*, LNCS. Springer, 2004.
4. V. Jijkoun and M. de Rijke. Answer Selection in a Multi-Stream Open Domain Question Answering System. In: *Proceedings 26th European Conference on Information Retrieval (ECIR'04)*, LNCS, Springer, 2004.
5. J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. XML Retrieval: What to Retrieve? In *Proceedings SIGIR 2003*, 2003.
6. M. Marx. XCPATH, the First-order Complete Fragment of XPath. In *Proceedings PODS 2004*, 2004.
7. M.-L. Reinberger, P. Spyns, W. Daelemans, and R. Meersman. Mining for lexons: applying unsupervised learning methods to create ontology bases. In *Proceeding of ODBase'03*, LNCS, Springer, 2003
8. E.F. Tjong Kim Sang. Memory-Based Shallow Parsing. *J. of Machine Learning Research*, 2:559–594, 2002
9. E.F. Tjong Kim Sang. Memory-Based Named Entity Recognition. In *Proceedings of CoNLL-2002*, 2002.
10. E.F. Tjong Kim Sang, W. Daelemans, and A. Hothker. Reduction of Dutch sentences for automatic subtitling, 2004. Submitted to CLIN-2003.

Ten key publications in the area:

1. M. Banko, E. Brill, S. Dumais, and J. Lin. AskMSR: Question Answering Using the Worldwide Web. In *Proceedings of 2002 AAAI Spring Symposium in Mining Answers from Texts and Knowledge Bases*, 2002
2. G. Bisson, C. Nédellec, and D. Cañamero. Designing clustering methods for ontology building - The Mo'K workbench. In *Proceedings of the workshop on Ontology Learning at ECAI-2000*. pages 13–19, 2000.

3. M. Fleischman, E. Hovy, and A. Echihiabi. Offline Strategies for Online Question Answering: Answering Questions Before They Are Asked. In *Proceedings of ACL-03*, 2003.
4. A. Gomez-Pérez and D. Monzano-Macho. A survey of ontology learning methods and techniques. OntoWeb technical report, deliverable 1.5, 2003. URL: <http://ontoweb.aifb.uni-karlsruhe.de/About/Deliverables/OverviewProjectPhase4>
5. V. Jijkoun and M. de Rijke. Answer Selection in a Multi-Stream Open Domain Question Answering System. In: *Proceedings 26th European Conference on Information Retrieval (ECIR'04)*, LNCS, Springer, 2004.
6. A. Maedche and S. Staab. Discovering Conceptual Relations from Text. In *Proceedings of ECAI 2000*, pages 321–325, 2000.
7. B. Magnini, M. Negri, R. Prevete, and H. Tanev. Is it the Right Answer? Exploiting Web Redundancy for Answer Validation. In *Proceedings of the 40st Annual Meeting of the Association for Computational Linguistics*, 2002
8. D. Moldovan, S. Harabagiu, R. Girju, P. Morarescu, F. Lacatusu, A. Novischi, A. Badulescu, and O. Bolohan. LCC Tools for Question Answering. In *Proceedings of TREC 2002*, 2003
9. M.-L. Reinberger, P. Spyns, W. Daelemans, and R. Meersman. Mining for lexons: applying unsupervised learning methods to create ontology bases. In *Proceeding of ODBase'03*, LNCS, Springer, 2003.
10. F. Xu, D. Kurz, J. Piskorski. and S. Schmeier. A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping. In *Proceedings of LREC 2002*, 2002.

16 Summary for Non-Specialists

Het doel van dit project is het ontwikkelen van automatische technieken voor het afleiden van kennis uit tekst. Deze kennis zal worden opgeslagen in een ontologie (kennisdatabank) die als hulp zal dienen bij het beantwoorden van vragen. We willen met behulp van deze technieken het digitale Nederlandse woordenboek EuroWordNet uitbreiden. Daarnaast willen we de technieken toepassen bij het afleiden van “feitenbanken” en specialistische ontologieën, bijvoorbeeld op basis van medische teksten.

Dit voorstel past in het thema Semantisch Web, een wereldwijd project waarin men probeert de kennis die algemeen toegankelijk is op Internet zodanig te formuleren dat hij ook door machines kan worden begrepen. In dit wereldwijde project worden ook ontologieën gemaakt. Wij willen twee van hun definitietalen (RDF en OWL) in ons project gaan gebruiken.

Het voorgestelde project is nuttig voor de automatische vragenbeantwoorder van IMIX. Immers zoals de plannen nu zijn, worden vragen in IMIX beantwoord door te zoeken in teksten. Dit kost erg veel tijd is derhalve problematisch voor interactief gebruik. Daarnaast is het op die manier erg lastig om fouten in de vragen te herkennen. Een antwoordsysteem dat beschikking heeft over een ontologie en uitgebreide feitenbanken is sneller en kan eenvoudiger fouten in vragen herkennen. Bij het afleiden van de ontologie en de feitenbanken maken we bovendien gebruik van verschillende documenten waardoor de kwaliteit van de antwoorden verbetert.

17 Research Budget

The total budget of the project is 168.259 Euros. This includes salary, project equipment (data and web server), and the standard benchfee. In accordance with the requirements, the proposed post-doc will spend at least 10% of his time on activities aimed at knowledge transfer. These activities will include (i) building and maintaining an online QA system that answers questions from the fact bases produced by this project, (ii) integration of tools and resources in the wider IMIX programme, (iii) active interaction with other members of the IMIX team, through site visits and demos.

(1) post-doc		
a) aanstelling postdoc	1 fte, 35 months	= Euro 157.721
b) benchfee	1 fte	= Euro 4.538
c) projectgebonden apparatuur/software		= Euro 6.000
Subtotaal postdoc		= Euro 168.259
(2) Overig (technisch personeel/programmeurs)		
		= Euro 0
(3) Investerings (bedragen incl. BTW)		
		= Euro 0
Totale kosten (1) postdoc + (2) Overig		= Euro 168.259
Totaal gevraagde subsidie voor (3) investeringen		= Euro 0
Totaal van deze aanvraag		= Euro 168.259