Semantic Relations in Concept-Based Cross-Language Medical Information Retrieval

Špela Vintar, Paul Buitelaar
DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany
{vintar, paulb}@dfki.de

Martin Volk
University of Stockholm
Department of Linguistics
Universitetsvägen 10C
S-10691 Stockholm, Sweden
volk@ling.su.se

Abstract

We explore and evaluate the usefulness of semantic annotation, particularly semantic relations, in cross-language information retrieval in the medical domain. As the baseline for automatic semantic annotation we use UMLS, which specifies semantic relations between medical concepts. We developed two methods to improve the accuracy and yield of relations in CLIR: a method for relation filtering and a method to discover new relation instances. Both techniques were applied to a corpus of English and German medical abstracts and evaluated for their efficiency in CLIR. Results show that filtering reduces recall without significant increase in precision, while discovery of new relation instances indeed proved a successful method to improve retrieval.

1 Introduction

The aim of Cross-Language Information Retrieval (CLIR) is to find documents in a large, possibly multilingual, collection that are most relevant for a given query, where the language of the query may be different from the language of the documents retrieved. Methods typically used to overcome this language barrier may be divided into: approaches based on bilingual dictionary look-up or Machine Translation (MT) (Hull and Grefenstette, 1996; Kraaij and Hiemstra, 1998; Oard, 1998); corpus-based approaches utilizing a range of IR-specific statistical measures (Carbonell et al., 1997; Qui, 1995); and concept-driven approaches, which exploit semantic resources (thesauri) to bridge the gap between surface linguistic form and meaning (see Section 6).

The appeal of concept-based approaches is that, in contrast with translation or corpus-based methods, they use linguistic processing and semantic resources to arrive at a language-independent representation of meaning, thus focusing on the logical content of an information search rather than its form. This is especially significant for highly specialized domains such as medicine, on the other hand this approach presupposes the existence of large domain-specific thesauri.

The identification of terms and their mapping to concepts is the first stage of semantic analysis and its efficiency largely depends on the quality of linguistic processing on the one hand and the quality and coverage of the thesaurus on the other. Semantic relations between concepts represent another layer of information, which have the potential of making the document search even more detailed and specific, and possibly interactive by allowing the user to control the directions in which a query is expanded.

We report on a series of experiments performed to test and evaluate the role of semantic relations in CLIR. The work we describe was performed within a project on the systematic comparison of concept-based and corpus-based methods in cross-language medical information retrieval. We use the Unified Medical Language System (UMLS) as the primary semantic resource and a corpus of English and German medical abstracts for development and evaluation of methods and tools. The paper focuses on semantic relations, which are a crucial element of medical knowledge representation. The basis of our experiments are the semantic relations as specified in the UMLS Semantic Network, which we seek to modify and expand for CLIR purposes. We describe a method for selecting relevant relations from those proposed by UMLS and a method for extracting new instances of relations based on statistical and NLP techniques.

2 Semantic Annotation for Concept-Based CLIR

2.1 UMLS and Semantic Relations

The essential part of any concept-based CLIR system is the identification of terms and their mapping to a language-independent conceptual level. Our basic resource for semantic annotation is UMLS, which is organized in three parts.

The **Specialist Lexicon** provides lexical information: a listing of word forms and their lemmas, part-of-speech and morphological information.

Second, the **Metathesaurus** is the core vocabulary component, which unites several medical thesauri and classifications into a complex database of concepts covering terms from 9 languages. Each term is assigned a unique string identifier, which is then mapped to a unique concept identifier (CUI). For example, the entry for the term *HIV pneumonia* in the Metathesaurus main termbank (MRCON) includes its CUI, language identifier, term status and finally the term string :

```
C0744975|ENG|P||HIV pneumonia|
```

The UMLS 2001 version includes 1.7 million terms mapped to 797,359 concepts, of which 1.4 million entries are English and only 66,381 German. Only the MeSH (Medical Subject Heading) part of the Metathesaurus covers both German and English, therefore we only use MeSH terms (564,011 term entries for English and 49,256 for German) for corpus annotation.

The third part is the **Semantic Network**, which provides a grouping of concepts according to their meaning into 134 semantic types (TUI). The concept above would be assigned to the class *T047, Disease or Syndrome*. The Semantic Network then specifies potential relations between those semantic types. There are 54 hierarchically organized domain-specific relations, such as *affects, causes, location_of* etc. All of them are binary relations (A is related to B).

2.2 Linguistic and Semantic Annotation

The foundation of our CLIR setting is the automatic linguistic and semantic annotation of the document collection, in our case a parallel corpus of about 9000 English and German medical abstracts, obtained from the Springer web site[1]. For linguistic processing we are using ShProT, a shallow processing tool that consists of four integrated components: the SPPC tokenizer (Piskorski and Neumann, 2000), TnT (Brants, 2000) for part-of-speech tagging, Mmorph (Petitpierre and Russell, 1995) for morphological analysis and Chunkie (Skut and Brants, 1998) for phrase recognition.

The next stage is the annotation of various semantic information. At the level of terms, the following information is used:

- Concept Unique Identifier (CUI)
- Type Unique Identifier (TUI)
- Medical Subject Headings ID - an alternative code to the CUIs
- Preferred Term - a term that is marked as the preferred name for a particular concept

The identification of UMLS terms in the documents is based on morphological processing of both the term bank and the document, so that term lemmas are matched rather than word forms. The annotation tool matches terms of lengths 1 to 3 tokens, based on lemmas if available and word forms otherwise. Term matching on the sub-token level is also implemented to ensure the identification of terms that are a part of a more complex compound, which is crucial for German.

[1] http://link.springer.de

In addition to concept identifiers (CUIs) we also annotate the codes of MeSH tree nodes. The decision to do so was based on our observation that the UMLS Semantic Network, especially the semantic types and relations, does not always adequately represent the domain-specific relationships. MeSH on the other hand has a transparent tree structure, from which both the semantic class of a concept and its depth in the tree can be inferred. For example, the terms *infarction* (C23.550.717.489) and *myocardial infarction* (C14.907.553.470.500) both belong to the group of diseases, but the node of the first term lies higher in the hierarchy as its code has fewer fields.

The term inventory in our documents is further expanded by integrating newly extracted terms provided by our project partners (cf. (Gaussier, 1998)). This slightly improves term-based retrieval, but since the new terms cannot be assigned a semantic type nor a MeSH code, they have no effect upon semantic relations.

Semantic relations are annotated on the basis of the UMLS Semantic Network, which defines binary relations between semantic types (TUIs) in the form of triplets, for example *T195 - T151 - T042* meaning *Antibiotic - affects - Organ or Tissue Function*. We search for all pairs of semantic types that co-occur within a sentence, which means that we can only annotate relations between items that were previously identified as UMLS terms. According to the Semantic Network relations can be ambiguous, meaning that two concepts may be related in several ways. For example:

Diagnostic Procedure	assesses_effect_of	Antibiotic
Diagnostic Procedure	analyzes	Antibiotic
Diagnostic Procedure	measures	Antibiotic
Diagnostic Procedure	uses	Antibiotic

Since the semantic types are rather general (e.g. *Pharmacological Substance, Patient or Group*), the relations are often found to be vague or even incorrect when they are mapped to a document. If for example the Semantic Network defines the relation *Therapeutic Procedure – method_of – Occupation or Discipline*, this may not hold true for all combinations of members of those two semantic classes, as seen in **discectomy – method_of – history*. Given the ambiguity of relations and their generic nature, the number of potential relations found in a sentence can be high, which makes their usefulness questionable. A manual evaluation of automatic relation tagging in a small sample by medical experts showed that only about 17% of relations were correct, of which only 38% were perceived as significant in the context of information retrieval.

On the other hand, many relations undoubtedly present in our texts are not identified by automatic relation tagging. One possible reason for this may be the incompleteness of the Semantic Network, but a more accurate explanation is that relationships are constantly being woven between concepts occurring together in a specific context, thus creating novel or unexpected links that would not exist between concepts in isolation.

For the above reasons we developed methods to deal with each of the problems described, relation filtering and relation extraction.

3 Extending Existing Resources: Relation Filtering and Relation Extraction

3.1 Relation Filtering

The first task was to tackle relation ambiguity, i.e. to select correct and significant relations from the ones proposed by automatic UMLS lookup; a procedure we refer to as *relation filtering*. The method is composed of two steps following two initial hypotheses:

- Interesting relations will occur between interesting concepts.
- Relations are expressed by typical lexical markers, such as verbs.

3.1.1 Relation Filtering with IDF

Following our first hypothesis we expect interesting and true relations to occur between items that are specific rather than general, and thus not too frequent. To measure this specificity we use the *inverse document frequency (IDF)* of the concept's code (CUI), which assigns a higher weight to concepts occurring only in a subset of documents in the collection. We thus take N_t to be the number of documents containing the CUI t and N the number of all documents.

$$IDF_t = log_2 \frac{N_t}{N}$$

We decided to use IDF instead of the generally used TF-IDF, because term frequency (TF), if multiplied with IDF, will assign a higher score to frequent terms like *patient, therapy, disease*. Relations between items with the IDF weight below a certain value are removed; the threshold value was set experimentally to 2.7.

Consider the following example where two instances of the relation diagnoses are found in a sentence:

Diagnostic – diagnoses – Disease
Diagnostic – diagnoses – Lyme Disease

As the IDF weights of both *Diagnostic* and *Disease* are below the set threshold the first relation instance is removed.

3.1.2 Relation Filtering with Verbal Markers

Relations that are semantic links between (mainly) nominal items may be represented by various linguistic means or *lexical markers*. In a rule-based approach such markers would be specified manually, however we chose to use a co-occurrence matrix of *lexical verbs* and automatically tagged relations. This is based on the assumption that some verbs are more likely to signify a certain relation than others. The co-occurrences are normalized and non-lexical verbs filtered out, so that for each lexical verb we get a list of relations it most likely occurs with. This information is then used to remove relations that occur with an untypical verb.

Below are the frequencies of five relations that are assigned to the verb *activate*:

interacts_with (197)
produces (83)
affects (52)
disrupts (32)
result_of (29)

Table 1 shows the number of relation instances using UMLS *[umls]* and after each filtering step *[umls_idf_filt, umls_idf_vb_filt]*.

3.2 Extraction of New Relation Instances

The identification of new instances of relations was based on observed co-occurrences of concepts, where instead of the semantic types (TUI) from the Metathesaurus we use MeSH classes. This gives us flexibility in choosing the number of semantic classes, depending on the level in the hierarchy.

The MeSH tree is organized into 15 top tree nodes, each of which is marked with a letter and subdivided into further branches. These top nodes are the following:

Anatomy [A]
Organisms [B]
Diseases [C]
Chemicals and Drugs [D]
Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
Psychiatry and Psychology [F]
Biological Sciences [G]
Physical Sciences [H]
Anthropology, Education, Sociology and Social Phenomena [I]
Technology and Food and Beverages [J]
Humanities [K]
Information Science [L]
Persons [M]
Health Care [N]
Geographic Locations [Z]

We use co-occurrences on the second level, meaning that we strip full MeSH codes assigned to each concept to only the top node letter and first order children. Looking at the structure of MeSH, this leaves us with 114 semantic classes, though some of them do not occur in our corpus. An example of a co-occurring MeSH pair is *D3 [Heterocyclic Compounds] + C10 [Nervous System Diseases]*.

For each UMLS semantic relation we then compute a list of typical MeSH pairs, for example *treats: D27|C23, D3|C23, E7|C23, E7|C2, ...*. Once these patterns of correspondence between pairs and relations are established, we may extract new instances of relations on the basis of co-occurring MeSH codes within the sentence.

Since the Semantic Network defines as many as 54 relations, of which several are very generic (e.g. *associated_with*) and some very rare, we chose to limit the extraction procedure to 15 most frequent and at the same time most specific relations. These are: *result_of, location_of, interacts_with, produces, degree_of, issue_in, uses, performs, treats, measures, causes, disrupts, diagnoses, analyzes*.

Table 1 shows the number of relation instances in the corpus if we use only UMLS *[umls]*, if we add new ones to those that may be found in UMLS *[umls_new]*, the number of instances if we first perform the filtering and then add new ones *[umls_idf_vb_new]*, and the number of instances we find using only our extraction method*[only_new]*.

Corpus version	Relation instances
umls	702,449
umls_idf_filt	405,844
umls_idf_vb_filt	290,250
umls_idf_vb_new	461,823
umls_new	819,202
0nly_new	1,009,847

Table 1: Number of relation instances in the corpus using UMLS, filtering and relation extraction

The extraction method can be tuned in terms of precision and recall by setting the MeSH-pair frequency threshold. For our current document collection and CLIR purposes this was set to 150, however other applications utilizing relation extraction, such as ontology building, might require a higher threshold.

4 Evaluation

The main goal of the experiments that we describe was to evaluate the usefulness of semantic relations in CLIR, where we explore the possibilities of modifying and expanding existing semantic resources, i.e. UMLS. The baseline of retrieval experiments is therefore to use UMLS as it is, and then compare performance achieved with pruning and expansion techniques.

To retrieve documents from the collection we are using a set of 25 medical queries, for which relevance assessments were provided by medical experts. Those queries are available in English and German, and for the majority of previous CLIR evaluation tasks performed within our project we used German queries over the English document collection, in accordance with the envisaged user requirements. Unfortunately, due to low term coverage for German, only very few semantic relations were found on the query side, and it was therefore impossible to assess their value. For this reason we opted for using English queries over the English document collection, however without indexing tokens and lemmas but relying solely on semantic information. We believe that this – though still monolingual – setting allows us to generalise our observations for CLIR, because we are using concepts and relations as the interlingua.

All experiments were carried out using the Rotondo [2] retrieval system, which indexes all semantic information provided in the XML annotated documents as separate categories: UMLS terms, MeSH terms, XRCE terms, semantic relations. The system uses a straight *lnu.ltn* weighting scheme. In the tables below we present the retrieval results in four columns: mean average precision **(mAP)**, absolute number of relevant documents retrieved **(RD)**, average precision at 0.1 recall **(AP01)** and precision for the top 10 documents retrieved **(P10)** (These metrics are also used in TREC experiments; cf. (Gaussier et al., 1998)). The total number of relevant documents for the 25 queries is 956.

4.1 Evaluation of relation filtering and relation extraction

Previous experiments within our project have shown that on the level of concepts MeSH codes achieve a higher precision than CUI's from the UMLS (Volk et al., 2002). We therefore choose MeSH codes as the primary semantic category on the level of concepts (mesh), and to this we wish to compare retrieval results achieved by using semantic relations together with MeSH codes (mesh_semrel) as well as the results of using semantic relations only (semrel).

Table 2 gives the results obtained by using UMLS-based semantic annotations, to be consid-

[2] A retrieval system from Eurospider Information Technology AG

	mAP	RD	AP01	P10
umls_mesh	0.311	541	0.659	0.536
umls_mesh_semrel	0.302	542	0.644	0.544
umls_semrel	0.146	253	0.384	0.340

Table 2: Results of using UMLS-based semantic annotations

	mAP	RD	AP01	P10
umls_idf_filt	0.309	541	0.651	0.540
umls_idf_vb_filt	0.306	541	0.661	0.524
umls_idf_vb_new	0.305	541	0.665	0.520
umls_new	0.300	542	0.647	0.532
only_new	0.298	543	0.637	0.508

Table 3: Results of relation filtering and extraction indexing MeSH concepts and relations

	mAP	RD	AP01	P10
umls_idf_filt	0.126	203	0.315	0.280
umls_idf_vb_filt	0.107	175	0.282	0.264
umls_idf_vb_new	0.124	197	0.336	0.308
umls_new	**0.153**	**259**	**0.419**	**0.344**
only_new	0.116	213	0.363	0.280

Table 4: Results of relation filtering and extraction indexing relations only

ered our baseline. We see that the average precision of using only concepts (mAP = 0.311) decreases slightly if we introduce semantic relations, however with an equally slight increase in recall and precision at top 10 documents. Semantic relations are always based on prior identification of two concepts, they are thus very specific and inevitably produce low recall if used alone (mAP = 0.146). We nevertheless consider this information useful for assessing the impact of relation filtering and expansion.

To this baseline we now compare five versions of our document collection, each annotated with a different set of semantic relations. The first contains UMLS-based relations filtered with the IDF method (umls_idf_filt), the second was additionally filtered with the verb method (umls_idf_vb_filt). We then introduce newly extracted relation instances, first to the filtered version of the corpus (umls_idf_vb_new), then to the baseline UMLS-annotated version (umls_new) and finally, we annotate relations using only our method for extracting new relation instances (only_new). For each corpus version we use queries that were processed identically to the document collection.

Table 3 gives the results for the combination of MeSH codes and semantic relations, and Table 4 shows the results for semantic relations only.

If we use semantic relations on top of MeSH concept codes, almost no difference can be observed, except perhaps that filtering with IDF seems to have a positive effect on high-end precision and that adding new relations slightly increases recall. However if we look at the results obtained by using only semantic relations, the differences between approaches become more apparent. It seems that each filtering step significantly decreases both recall and precision, while adding new relations – as we would expect – works well. The highest precision and recall were achieved with a combination of UMLS annotation and new relations, and this combination also outperforms the baseline.

4.2 Evaluation with manually annotated queries

Relations represent a secondary, highly specific level of semantic information, which is difficult to evaluate in traditional CLIR settings. Within our project, responding to user requirements of the medical domain, we designed a retrieval prototype where semantic information can be used interactively. If semantic relations are understood as a specific point of view on top of the initial request, the user may first submit a query and then select the relations she would find useful.

In an approximation of this scenario we had our 25 queries first automatically tagged for terms and concepts, and then manually annotated for semantic relations by a medical expert. The expert was asked to use only the 15 relations listed above. Table 5 shows the retrieval results using manually annotated queries over all five versions of our corpus, where only semantic relations were indexed.

Although the overall results of this run are very

	mAP	RD	AP01	P10
umls_idf_filt	0.035	85	0.080	0.080
umls_idf_vb_filt	0.027	68	0.077	0.080
umls_idf_vb_new	0.031	77	0.080	0.080
umls_new	0.045	104	0.106	0.124
only_new	**0.085**	**154**	**0.274**	**0.232**

Table 5: Retrieval results using manually annotated queries (indexing only semantic relations)

low, which is due to the fact that manual annotation was much less 'generous' than the automatic, we see a dramatic increase in recall and precision using the corpus annotated only by our method. This indicates a high correspondence between the 'true', expert-provided information and the automatic extraction model, and thus confirms our intuitions about the relevance of MeSH co-occurrences.

5 Discussion

Although the initial motivation for this research was to enhance document retrieval by introducing semantic relations, results obtained from the set of experiments we describe above lead to other – possibly even more promising – fields of application. In a domain such as medicine where extensive semantic resources can be used for concept-based retrieval, the most influential factor in the performance of a CLIR system is concept or term coverage, while semantic relations should probably be implemented in an interactive way allowing the user to narrow the focus of an otherwise overproductive query. Another implication of the results presented is that a smaller set of relations is beneficial both for document retrieval and automatic extraction of relation instances.

In a broader context or in another domain these methods might be adapted to ontology expansion or, possibly in combination with term extraction, to ontology construction. Hand-crafted ontologies more often than not focus on concepts and hierarchical relations between them. Automatic relation extraction is an important method of revealing domain-specific, possibly even previously unknown links between concepts and is therefore an integral part of Text Mining and Knowledge Discovery.

6 Related Work

Domain-specific multilingual thesauri have been used for English-German CLIR within social science (Gey and Jiang, 1999), while (Eichmann et al., 1998) describe the use of UMLS for French and Spanish queries on the OHSUMED text collection. Both of these approaches use the thesaurus to compile a bilingual lexicon, which is then used for query translation. Mandala et al. (Mandala et al., 1999) seek to complement WordNet by using corpus-derived thesauri and report improved performance in monolingual IR, however their approach only indirectly subsumes (unlabelled) relation extraction by using term co-occurrences.

Many approaches use lexical markers for extracting relations between terms or concepts (Hearst, 1992; Davidson et al., 1998; Finkelstein-Landau and Morin 1999), some also in combination with shallow parsing, but this method is generally low in recall and therefore not suitable for retrieval purposes. We are using lexical markers as probabilistic contexts for semantic classification, an approach similar to that of (Bisson et al., 2000). As for the relevance of MeSH classes for medical semantic relations, (Cimino and Barnett, 1993) already defined some combinations of top MeSH nodes that indicate specific medical relations. In our approach these combinations are established by statistical measures applied to our semantically annotated corpus, and we use a finer grained network of semantic classes.

7 Conclusions

In this paper we focus on the role of semantic relations, as specified in the UMLS Semantic Network, in concept-based medical CLIR. Proposed are two methods for improving their use: relation filtering and relation extraction. The evaluation of these methods shows that the first does not score well in retrieval, whereas relation extraction on the basis of co-occurrences of MeSH classes looks promising for query expansion. The evaluation with a set of manually annotated queries shows that newly extracted relation instances have the highest level of correspondence with relations as identified by medical experts, which can especially be exploited in an interactive retrieval setting.

Future research will include learning semantic relations using classification techniques, where the context features of MeSH co-occurrences will be expanded from verbs to other linguistic markers including grammatical functions. For CLIR tasks it remains to be established which number of different relations works best. Although in our experiments relations do not lead to a major gain in precision and recall compared to using only concepts, the techniques we develop may find further application in related areas such as ontology construction and adaptation.

References

G. Bisson, C. Nedellec, D. Canamero: Designing Clustering Methods for Ontology Building - The Mo'K Workbench. In: S. Staab, A. Maedche, C. Nedellec, P. WiemerHastings (eds.), Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artifi cial Intelligence ECAI'00, Berlin, Germany, August 20-25, 2000.

Brants T. 2000. TnT - A Statistical Part-of-Speech Tagger. In: *Proc. of the 6th ANLP Conference*, Seattle, WA.

Carbonell J., Y. Yang, R. Frederking, R. D. Brown, Y. Geng, and D. Lee. 1997. Translingual Information Retrieval: A Comparative Evaluation. In: *Proc. of the Fifteenth International Joint Conference on Artifi cial Intelligence.*

Cimino, J. and G. Barnett. Automatic knowledge acquisition from Medline. Methods of Information in Medicine, 32(2):120-130, 1993.

Davidson, L., J. Kavanagh, K. Mackintosh, I. Meyer, and D. Skuce. 1998. Semi-automatic extraction of knowledge-rich contexts from corpora. Proceedings, 1st Workshop on Computational Terminology (COMPUTERM'98), pages 50–56, Montreal.

Eichmann D., M. Ruiz, and P. Srinivasan. 1998. Cross-Language Information Retrieval with the UMLS Metathesaurus. In: *Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia.

M. Finkelstein-Landau and E. Morin. Extracting semantic relationships between terms: Supervised vs. unsupervised methods. In International Workshop on Ontological Engineering on the Global Information Infrastructure, pages 71-80, 1999.

Gaussier E., G. Grefenstette , D. A. Hull, and B. M. Schulze. 1998. Xerox TREC-6 site report: Cross language text retrieval. In: *Proc. of the Sixth TExt Retrieval Conference (TREC-6)*. National Institute of Standards Technology (NIST), Gaithersburg, MD.

Gaussier, E. 1998. Flow Network Models for Word Alingment and Terminology Extraction from Bilingual Corpora. In: *Proceedings of the 36th Annual Meeting of the association for Computational Linguistics and the 17th International Conference on Computational Linguistics, COLING-ACL'98*, Montreal, Canada.

Gey F. C., and H. Jiang. 1999. English-German Cross-Language Retrieval for the GIRT Collection - Exploiting a Multilingual Thesaurus. In: *Proc. of the Eighth Text REtrieval Conference (TREC-8)*, National Institute of Standards Technology (NIST), Gaithersburg, MD.

Gonzalo J., F. Verdejo, and I. Chugur. 1999. Using EuroWordNet in a Concept-based Approach to Cross-Language Text Retrieval, *Applied Artifi cial Intelligence:13*, 1999.

Hearst, M. Automatic Acquisition of Hyponyms from Large Text Corpora, Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France, July 1992.

Hull D. A., and G. Grefenstette. 1996. Querying Across Languages: A Dictionary based Approach to Multilingual Information Retrieval. In: *Proc. of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: ACM SIGIR*. 1996. http://www.xerox.fr/people/grenoble/hull/papers/sigir96.ps

Kraaij, W. and D. Hiemstra. 1998. TREC6 Working Notes: Baseline Tests for Cross Language Retrieval with the Twenty-One System. In: *TREC6 working notes*. National Institute of Standards and Technology (NIST), Gaithersburg, MD.

Mandala, R., T. Tokunaga and H. Tanaka. 1999. Complementing WordNet with Roget's and Corpus-based Thesauri for Information Retrieval. In: *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, Bergen, Norway.

Oard D. 1998. A comparative study of query and document translation for cross-lingual information retrieval In: *Proc. of AMTA*, Philadelphia, PA.

Petitpierre D., and G. Russell. 1995. MMORPH - The Multext Morphology Program. *Multext deliverable report for the task 2.3.1*, ISSCO, University of Geneva, Switzerland.

Piskorski, J. and G. Neumann. 2000. An intelligent text extraction and navigation system. In: *Proc. of the 6th RIAO*. Paris.

Qui, Y. 1995. Automatic Query Expansion Based on a Similarity Thesaurus. *PhD thesis*, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland.

Skut W. and T. Brants. 1998. A Maximum Entropy partial parser for unrestricted text. In: *Proc. of the 6th ACL Workshop on Very Large Corpora (WVLC)*, Montreal, Canada.

Volk M., B. Ripplinger, S. Vintar, P. Buitelaar, D. Raileanu, B. Sacaleanu. 2002. Semantic Annotation for Concept-Based Cross-Language Medical Information Retrieval. To appear in the International Journal of Medical Informatics, 2002.