Evaluation Resources for Concept-based Cross-Lingual Information Retrieval in the Medical Domain

Paul Buitelaar[§], Diana Steffen[§], Martin Volk[*], Dominic Widdows[♦],

Bogdan Sacaleanu[§], Špela Vintar[§], Stanley Peters[♦], Hans Uszkoreit[§]

[§] DFKI GmbH
Stuhlsatzenhausweg 3,
66123 Saarbrücken, Germany
{paulb, bogdan, uszkoreit}@dfki.de

[*] (previously at) Eurospider Information Technology AG
Schaffhauserstrasse 18
CH-8006 Zürich, Switzerland
volk@ling.su.se

[♦] Stanford University, CSLI
220 Panama Street
Stanford, CA 94305-4115, USA
{dwiddows, peters}@csli.stanford.edu

Abstract

The paper describes evaluation resources for concept-based, cross-lingual information retrieval in the medical domain. All resources were constructed in the context of the MuchMore project and are freely available through the project website. Available resources include: a bilingual, parallel document collection of German and English medical scientific abstracts, a set of queries and corresponding relevance assessments, two manually disambiguated test sets for semantic annotation (sense disambiguation), two evaluation lists for German morphological decomposition of medical terms.

MuchMore

The evaluation resources described in this paper were all constructed in the context of the MuchMore project[1] on concept-based, cross-lingual information retrieval (CLIR). The project provided a framework for integrating and refining existing technologies and developing new approaches to CLIR for the medical domain. For this purpose, the project pursued the following aims:

- Integrated and effective combination of different approaches and heterogeneous resources for cross-lingual information access and management, including performance and user evaluation for realistic information access tasks.
- Automated acquisition of domain-specific linguistic resources and effective use of multilingual concept hierarchies.
- Demonstration of a cross-lingual information access prototype system for the medical domain, that provides access to multilingual information on the basis of a combined use of corpus analysis and (domain-specific) ontologies and thesauri.

The MuchMore Prototype

The MuchMore project developed a prototype cross-lingual document retrieval system that enables users to retrieve documents (in English and/or German) that are relevant to a given query document (in English or German), see e.g. (Sacaleanu et al., 2003). In the current version of the system[2], query documents are assumed to be German electronic patient records and documents to be retrieved are medical scientific abstracts in both German and English. The cross-lingual information retrieval task has been approached with a combination of concept-based and corpus-based methods: semantic annotation, similarity thesaurus, example-based translation, pseudo-relevance feedback and concept-space model[3].

Evaluation Efforts and Constructed Evaluation Resources

One of the primary goals of the MuchMore project has been to develop and evaluate methods for the effective use of multilingual ontologies and thesauri in the semantic annotation of English and German medical texts and subsequently to evaluate and compare the impact of such semantic information on the CLIR task. In particular, work on semantic annotation with domain-specific (medical) and general semantic resources has been central in this approach.

In order to evaluate performance gain in the CLIR task, several experiments were carried out with a query set and relevance assessments defined by medical experts for a document collection that was gathered and prepared specifically for the MuchMore project goals.

Overall results of the evaluation effort with this document collection and query test set show that best performance may be obtained by a combination of corpus-

[1] http://muchmore.dfki.de

[2] http://muchmore.dfki.de/demo1.htm

[3] For a more detailed overview see the MuchMore final report: http://muchmore.dfki.de/pubs/D0.6.final.pdf

based and concept-based information, i.e. using a combination of manually constructed and automatically extracted (semantic) resources. Adding manually constructed knowledge (through semantic annotation or classification) improves performance, although disambiguation has not been shown to further improve performance significantly. For more information on the evaluation results, see e.g. (Volk et al., 2002). In this paper we rather concentrate on the structure and content of the evaluation resources used in these experiments.

Evaluation Resources for Cross-Lingual Information Retrieval in the Medical Domain

MuchMore Document Collection

The MuchMore document collection is a parallel corpus of English and German scientific medical abstracts obtained from the Springer LINK web site[4]. The corpus consists of approximately 9000 documents with a total of one million tokens for each language. Abstracts are taken from 41 medical journals, e.g. "Der Nervenarzt", "Der Radiologe", etc., each of which constitutes a homogeneous medical sub-domain, e.g. Neurology, Radiology, etc. The corpus of downloaded HTML documents has been normalized in various ways, in order to produce a clean, plain text version. Additionally, the corpus has been aligned on the sentence level. An example abstract document is given in Figure 1., below:

Balint sydrom and associated disorders. Anamnesis, diagnostic and treatment approaches

Balint syndrom is a combination of symptoms including simultanagnosia, a disorder of spatial and object-based attention, disturbed spatial perception and representation, and optic ataxia resulting from bilateral parieto-occipital lesions. Fixation and ocular exploration of space are severely impaired, as are reading, writing, drawing and orientation as well as movement in space. Low-level visual impairments may be associated and difficult to evaluate but are not a necessary element of Balint syndrom. This review summarizes the relevant facts of the etiology, localization of lesions, and the core features as well as frequently associated disorders of the syndrom. Instructions for the examination of patients are given and approaches for the management and treatment outlined. Models of attention and space representation are described for the explanation of this multifaceted and exciting syndrome. Finally the key features of the syndrom are summarized in a table at the end of the paper.

Balint syndrome - Optic ataxia - Space representation - Parietal lesion - Attention - Reading

Figure 1: Example abstract document from the MuchMore corpus

Each abstract document consists of a Title (if available), the Abstract itself and (if available) one or more Keywords. English abstracts are translations of the German originals, which influences their quality as most translations seem to have been done by the authors themselves.

Automatic Annotation

The corpus is automatically annotated using "ShProT", a shallow processing tool that consists of three integrated components: "TnT" (Brants, 2000) for part-of-speech tagging, "Mmorph" (based on Petitpierre and Russell, 1995) for morphological analysis and "Chunkie" (Skut and Brants, 1998) for phrase recognition. Both TnT and Mmorph were adapted to the medical domain by updating the lexicon with information from English and German medical dictionaries.

Semantic annotation is performed on the basis of UMLS[5] (Unified Medical Language System), a publicly available semantic resource in the medical domain that consists of an English medical lexicon (Specialist Lexicon), a multilingual terminology database (MetaThesaurus) that links several standard medical thesauri and a Semantic Network of relations between concepts in the MetaThesaurus. MeSH (Medical Subject Headings), which is one of the medical thesauri that are contained in UMLS, is used also separately in semantic annotation. Finally, also EuroWordNet[6], as a general language semantic resource, is used in the semantic annotation of the MuchMore parallel corpus.

Automatically annotated and plain versions of the corpus are available from the project web site[7]. An example section of an annotated abstract document is given in Figure 2., below.

```
… spatial and object-based attention, disturbed
spatial perception and representation ….

<text>
  ...
  <token id="w20" pos="JJ" lemma="spatial">
   spatial </token>
  <token id="w21" pos="NN" lemma="perception">
    perception </token>
  ...
</text>

<umlsterm id="t7" from="w20" to="w21">
  <concept id="t7.1" cui="C0037744" tui="T041">
    <msh code="F2.463.593.778"/>
    <msh code="F2.463.593.932.869"/>
  </concept>
</umlsterm>

<semrel id="r7" term1="t7.1" term2="t8.1"
  reltype="issue_in"/>
```

Figure 2: Example section of an annotated abstract document from the MuchMore corpus

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

[5] http://umls.nlm.nih.gov

[6] http://www.hum.uva.nl/~ewn/

[7] http://muchmore.dfki.de/resources1.htm

Annotation Format and Viewer

The annotation format (as illustrated in Figure 2.) has been developed specifically for the purposes of the MuchMore project. It combines multiple levels of linguistic and semantic information that are interrelated in various ways. Our aim was to design an annotation format that would encompass all of these layers and adequately represent the relationships between them, while at the same time remaining logical and readable, efficient for parsing and indexing as well as flexible for future additions and adjustments. For more information on the annotation format see (Vintar at al., 2002), (Buitelaar et al., 2003).

An online demo is available that allows for automatic annotation of small documents with PoS, morphology, chunks and semantic annotation as discussed above. Documents that are annotated in this format can be displayed by use of the "MuchMore Viewer", which allows also for interactive correction of annotated documents. The annotation demo and viewer are both available from the MuchMore web site[8].

MuchMore Query Set and Relevance Assessments

In our CLIR experiments, we used relevance assessments based on 25 queries provided by medical experts. We obtained relevance assessments for the German documents as well as for the English documents from two teams of experts.

Unfortunately however there was a large discrepancy in their assessments. The German language team judged 959 documents to be relevant, the English language team 500 documents. The main reason for this discrepancy was the different types of experts doing the assessments (professionals vs. students). The overlap was 382 documents while 118 were only deemed relevant by the English language judges and 577 were only relevant for the German language judges.

Because the document collection is parallel, we decided to use only the German relevance assessments for our experiments in order to get comparable data. In these assessments the number of relevant documents per query varies between 7 and 104.

Queries are short and usually consist of a complex noun phrase extended by attributes (including prepositional phrases) and coordination. Figure 3. gives some examples.

arthroscopic treatment of cruciate ligament injuries Arthroskopische Behandlung bei Kreuzbandverletzungen
indication for implantable cardioverter defibrillator (ICD) Indikation für einen implantierbaren Kardioverter-Defibrillator (ICD)
Therapy in chronic lower back pain Therapie beim chronischen Kreuzschmerz

Figure 3: Example Queries

[8] http://muchmore.dfki.de/demo2.htm

Queries (in German and English) and relevance assessments (from the German and the English language teams) are available from the MuchMore web site[9].

Evaluation Resources for Sense Disambiguation in the Medical Domain

One of the research areas that the MuchMore project focused on was sense disambiguation, which is an enabling task in concept-based, cross-lingual information access. Unfortunately, there is a lack of test sets for sense disambiguation evaluation, specifically for languages other than English and even more so for specific domains like medicine. Given that MuchMore had a focus on English as well as German in the medical domain, the project developed its own evaluation sets in order to test different disambiguation methods (Raileanu et al., 2002). The sets consist of disambiguated instances that were selected from the German part of the MuchMore corpus annotated with GermaNet (Hamp and Feldweg, 1997) and UMLS (German). Both sets are available from the project web site[10]. For more information on some of the sense disambiguation evaluation results see (Widdows et al., 2003).

Inter-Annotator Agreement

In order to assess the difficulty of the disambiguation task, inter-annotator agreement scores were computed. Here, "annotators" are the domain experts who manually disambiguated selected ambiguous word occurrences.

The agreement scores for the GermaNet evaluation corpus have been discussed in (Raileanu et al., 2002). The agreement scores for UMLS vary from very low to very high, with an average of 65%. In some cases, the UMLS definitions were insufficient to give a clear distinction between concepts, especially when the concepts came from different original thesauri. This allowed the decision of whether a particular definition made "sense" to be more or less subjective.

Approximately half of the disagreements between annotators occurred with terms where inter-annotator agreement was less than 10%, which is evidence that a significant amount of the disagreement between annotators was on the *type* level rather than the *token* level. In other cases, it is possible that there was insufficient contextual information provided for annotators to agree. If one of the annotators was unable to choose any of the senses and declared an instance to be 'unspecified', this also counted against inter-annotator agreement.

Evaluation Resource for Morphological Analysis in the Medical Domain

In many languages other than English the morphological system is very rich and enables the construction of semantically complex compound words. For instance the German word *Kreuzbandverletzung* corresponds in English with three words: *cruciate ligament injury*. In a task like CLIR it is therefore important to know the morphological structure of (domain-specific) words, in order to allow for an appropriate cross-lingual match.

[9] http://muchmore.dfki.de/resources2.htm

[10] http://muchmore.dfki.de/resources3.htm

Therefore, to evaluate the performance of the annotation system on morphological decomposition of German medical terms, evaluation lists were developed of manually decomposed terms. Two lists are available[11], compiled for different sub-corpora of the MuchMore corpus. An example section is given in Figure 4., below:

Angiofibroms > Angiofibrom
Angstsymptomatik > Angst + Symptomatik
Anitbiotikasensitivität > Antibiotika + Sensitivität
Ann-Arbor-Klassifikation > Ann-Arbor + Klassifikation
Anpassungsfähigkeit > Anpassung + Fähigkeit
Antimyosin-Szintigraphie > Antimyosin + Szintigraphie
Antiphlogistika > Antiphlogistika
Antiphlogistikum > Antiphlogistikum
Antirefluxchirurgie > antireflux + Chirurgie
Anäthesieeinleitung > Anästhesie + Einleitung
Aortenaneurysmen > Aorta + Aneurysma
Aortenbogen > Aorta + Bogen
Aortendissektion > Aorta + Dissektion
Argonlaser > Argon + Laser
Arrhythmie > Arrhythmie
Arrhythmieform > Arrhythmie + Form
Artikel > Artikel
Arzneimittel > Arzneimittel
Aspirationspneumonie > Aspiration + Pneumonie
Atemartefakte > Atem + Artefakt
Atemdepression > Atem + Depression
Atemluft > Atem + Luft
Atemphase > Atem + Phase
Atemsystem > Atem + System
Atherectomy > Atherectomy
Atherektomie > Atherektomie
Bewegungsstörungen > Bewegung + Störung

Figure 4: Examples of the evaluation list for morphological decomposition of German medical terms

Conclusions

We described a number of evaluation resources for concept-based, cross-lingual information retrieval in the medical domain that have been constructed in the context of the MuchMore project and which are freely available from the project web site[12]. We hope that their availability will allow for their further use in other projects and applications and that they will proof to be a benchmark for the comparison of evaluation results in CLIR for the medical domain.

Acknowledgements

This research has been supported by EC/NSF grant IST-1999-11438 for the MuchMore project.

Special thanks to our colleagues at DFKI (Daniel Olejnik, Sarah Schmitt, Markus Endres, Sebastian Thull) ZInfo (Jörg Bay, Oktavian Weiser) and CMU (Ralf Brown) for their contributions in corpus collection and preparation, query construction, relevance assessment and sense annotation.

[11] http://muchmore.dfki.de/resources6.htm
[12] http://muchmore.dfki.de

References

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

Buitelaar P., Declerck Th., Sacaleanu B., Vintar Š., Raileanu D., Crispi C. A Multi-Layered, XML-Based Approach to the Integration of Linguistic and Semantic Annotations. In: Proceedings of EACL 2003 Workshop on Language Technology and the Semantic Web (NLPXML'03), Budapest, Hungary, April 2003.

Hamp, B. and Feldweg, H. GermaNet: a Lexical-Semantic Net for German. In: Proceedings of the ACL/EACL97 workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid,1997.

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

Raileanu D., Buitelaar P., Bay J., Vintar, S. An Evaluation Corpus for Sense Disambiguation in the Medical Domain In: Proceedings of LREC2002, Las Palmas, Canary Islands - Spain, May 29-31, 2002.

Sacaleanu B., Volk M., Buitelaar P. A Cross-Language Document Retrieval System Based on Semantic Annotation. In: Proceedings of the EACL 2003 Demo Session, Budapest, Hungary, April 2003.

Skut W. and Brants T. A Maximum Entropy partial parser for unrestricted text. In Proceedings of the 6th ACL Workshop on Very Large Corpora, Montreal, 1998.

Vintar, S., Buitelaar P., Ripplinger B., Sacaleanu B., Raileanu D. and Prescher D. An Efficient and Flexible Format for Linguistic and Semantic Annotation. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), May 29-31, Las Palmas, Canary Islands, Spain.

Volk M., Ripplinger B., Vintar Š., Buitelaar P., Raileanu D., Sacaleanu B. Semantic Annotation for Concept-Based Cross-Language Medical Information Retrieval In: International Journal of Medical Informatics, Volume 67:1-3. 2002.

Vossen, P. EuroWordNet: a multilingual database for information retrieval. In: Proceedings of the DELOS workshop on Cross-language Information Retrieval. Zurich, March 5-7, 1997.

Widdows, D., Peters, S., Cederberg, S., Chan, C.-K., Steffen, D., Buitelaar, P. Unsupervised monolingual and bilingual disambiguation of medical documents using umls. In: ACL workshop on Natural Language Processing in Biomedicine. Sapporo, Japan, July 2003.