A Cross-Language Document Retrieval System Based on Semantic Annotation

Bogdan Sacaleanu
DFKI GmbH
bogdan@dfki.de

Paul Buitelaar
DFKI GmbH
paulb@dfki.de

Martin Volk
EIT AG
volk@eurospider.com

Abstract

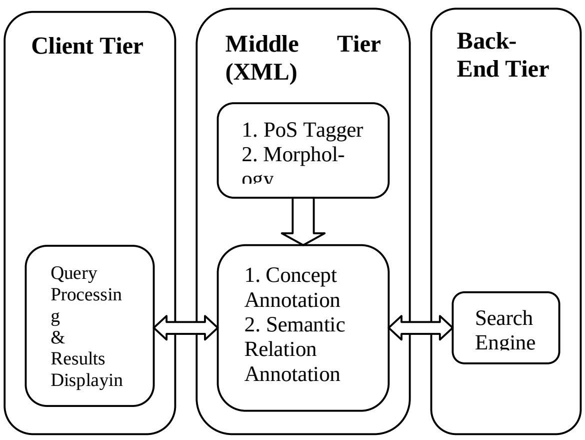
The paper describes a cross-lingual document retrieval system in the medical domain that employs a controlled vocabulary (UMLS[1]) in constructing an XML-based intermediary representation into which queries as well as documents are mapped. The system assists in the retrieval of English and German medical scientific abstracts relevant to a German query document (electronic patient record). The modularity of the system allows for deployment in other domains, given appropriate linguistic and semantic resources.

1 Introduction

The task of a cross-language information retrieval (CLIR) system is to match user queries specified in one language against documents written in a different language. In recent years, three approaches to the CLIR problem have been investigated: query translation, document translation and the use of an interlingua as specified in thesauri and similar semantic resources. The system[2] we describe here (MuchMore*) approaches the CLIR task by automatically mapping both the queries and documents into an intermediary XML-based representation by means of a multilingual medical thesaurus. The controlled vocabulary used, the Metathesaurus (or rather the MeSH[3] part of this), is one of the three knowledge sources developed within the UMLS containing semantic information about biomedical concepts, their various names and the specific relationships among them (i.e. broader_term, narrower_term, etc.). In addition we used the UMLS Semantic Network as a further knowledge source, which provides a categorization of the Metathesaurus concepts in semantic types and provides links between these types through relationships that are important for the biomedical domain (i.e. location_of, leads_to, etc.).

2 The MuchMore* Platform

At its core, MuchMore* is a multitier application configured as a client tier to provide a user interface, a middle tier annotation module that generates the intermediary data representation, and a



[1] The Unified Medical Language System (http://umls.nlm.nih.gov) integrates information from multiple machine-readable biomedical information sources.

[2] The system described here emerged in the context of the MuchMore project in close cooperation between two project partners. It is an integral part of the MuchMore prototype, which integrates additional CLIR approaches by other partners.

[3] MeSH: Medical Subject Headings (http://www.nlm.nih.gov/mesh/meshhome.html)

back-end tier consisting of a search engine system to provide the retrieval technology.

2.1 Query and Document Annotation

The middle tier annotation module consists of more subtiers representing an advanced annotation system that automatically identifies a number of relevant linguistic and semantic features. Components for part-of-speech tagging, morphological analysis, phrase tagging (chunking), concept and semantic relations annotation are being loosely integrated and generate an intermediary XML representation of the input data (an example is shown below). The loose integration of these components, through their ability to both produce and consume XML data, is an extremely flexible way for reuse. Through substitution or further chaining of such components the annotation can be extended to embrace a diverse set of domains beside the medical one.

2.2 Query Processing

The entry point to the MuchMore system is a query-processing interface that provides a user interface for completing or refining query construction (see Figure 1. below). For this purpose, the following information is displayed:

- the text of the query, serving as reference context for any further refinements
- a list of automatically extracted medical concepts along with their frequency and the semantic relations holding among the concepts
- a browsing option that helps the user to navigate through the concept space (MeSH) and include more general or more specific concepts in the constructed query

The concept list consists of preferred names of the matched terminology, as found in the controlled vocabulary. Furthermore, on clicking the frequency number associated with a concept, all its instances in the query are highlighted. Thereby the user is not only presented with a normalized medical terminology, according to the controlled vocabulary, but he can also inspect which terms in the query document are instances of which concepts. A list of semantic relations that hold between co-occurring concepts is displayed for each concept. When the user clicks on a listed relation, the context of the relation and

```
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.

<text>
  <token id="w1" pos="NN">Balint</token>
  <token id="w2" pos="NN">syndrom</token>
  <token id="w3" pos="VBZ" lemma="be">is</token>
  <token id="w4" pos="DT" lemma="a">a</token>
  <token id="w5" pos="NN" lemma="combination">combination</token>
  <token id="w6" pos="IN" lemma="of">of</token>
  <token id="w7" pos="NNS" lemma="symptom">symptoms</token>
  ...
  <token id="w20" pos="JJ" lemma="spatial">spatial</token>
  <token id="w21" pos="NN" lemma="perception">perception</token>
  <token id="w22" pos="CC" lemma="and">and</token>
  <token id="w23" pos="NN" lemma="representation">representation</token>
  ...
</text>

<umlsterm id="t7" from="w20" to="w21">
   <concept id="t7.1" cui="C0037744" preferred="Space Perception" 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"/>
```

the its concepts are highlighted, helping the user to make an informed choice on the relevance of the automatically extracted relation.

For query expansion we provide a browse-able contextual view of a concept according to the MeSH hierarchy. By selecting any concept in the generated list an overview is given of ancestor, sibling and child concepts. By double-clicking any of these, the query can be extended in a way that is relevant to the user needs, with the added concepts shown in a text area below the original concept list. The text area can be directly edited to append new terms to the query, which the user considers relevant but were neither automatically extracted nor available through MeSH browsing.

Once the query has been refined according to the user needs, the underlying information about tokens, lemmas, concept codes and their relations is sent to the retrieval engine.

2.3 Indexing and Retrieval

The back-end tier of the system is a retrieval engine with XML-based indexing support. It allows to index any linguistic or semantic feature from the intermediary XML document representation. Information relevant to a user query is being retrieved through a vector space similarity match between concepts and semantic relations on the query and document side. Evidence from multiple indexing features are automatically combined into the computation of the relevancy value for each document.

The result page displays a list of relevant documents in a descending order and a list of concepts and semantic relations that the query consists of (see Figure 2. below).

For viewing the content of any retrieved document, a user interface similar to the query processing's view has been implemented, whereby the matched concepts and relations are being highlighted.

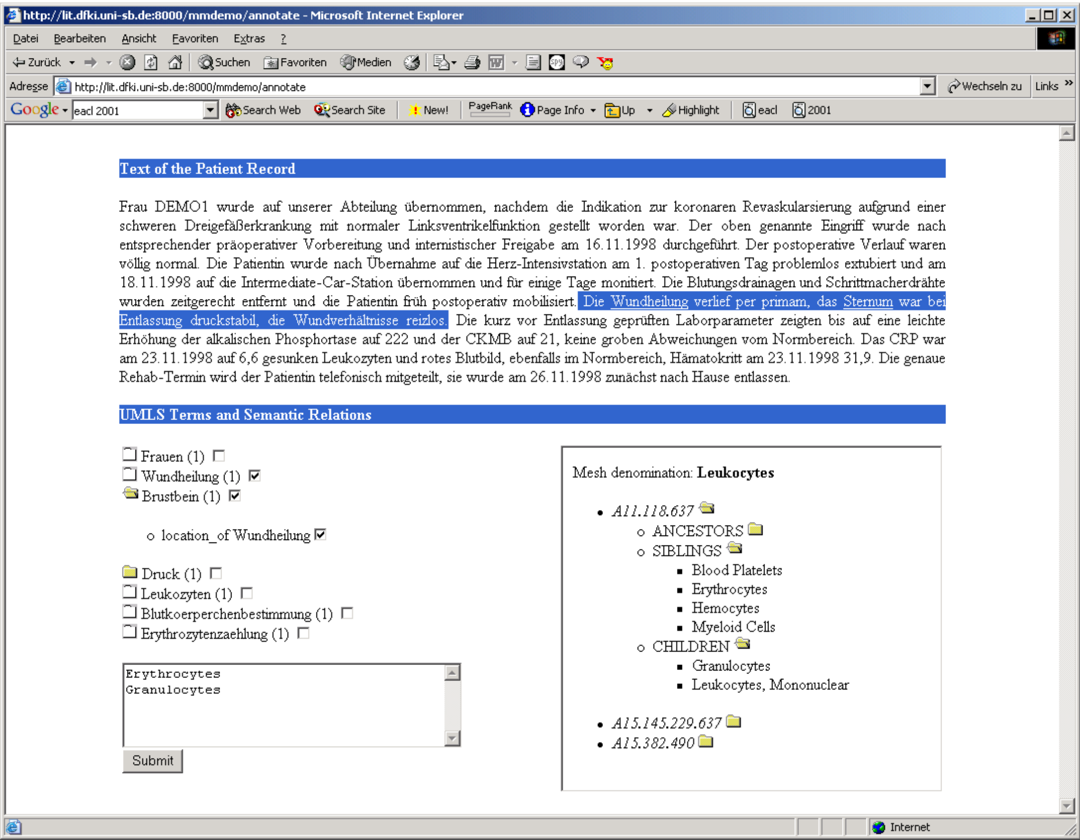


Figure 1: Query Processing Interface

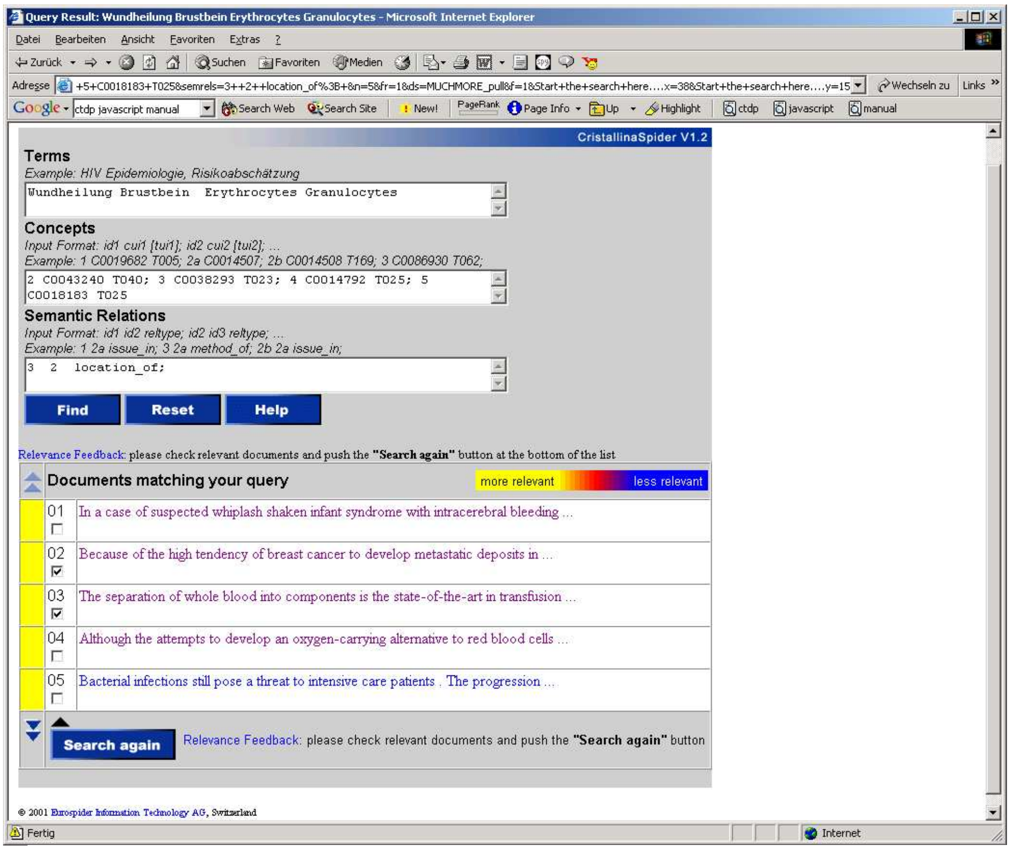


Figure 2: Retrieval Result Page

3 Future Work

The next version of the system (which will be available at the time of the conference) will integrate the following additional features:

- Sense Disambiguation
- Relation Filtering
- Results Merging

Sense Disambiguation Ambiguity is one of the inherent problems to deal with in the context of semantic annotation. The problem is that a word or even a complex term may have different meanings, i.e. concepts to be annotated with. The system will therefore be extended with a sense disambiguation component in the middle tier to tackle this problem. This component will choose, the most appropriate UMLS concept for a term according to the context.

Relation Filtering Given the UMLS Semantic Network, relations can also be ambiguous. That is, two concepts can be related in several ways as illustrated by the following example:

Diagnostic Procedure	*analyzes*	Antibiotic
Diagnostic Procedure	*measures*	Antibiotic
Diagnostic Procedure	*uses*	Antibiotic

For this purpose, a relation-filtering component will be added that selects the correct relation by means of lexical markers, such as verbs, and by a measure of context relevancy.

Results Merging As one of the goals of the project is to compare the performance of different document retrieval methods, we will integrate further retrieval systems developed in the context of the project by other partners and merge different scoring schemes so that the end user gets one unified result list with the most relevant documents ranked topmost.