An Efficient and Flexible Format for Linguistic and Semantic Annotation

Špela Vintar[§], Paul Buitelaar[§], Bärbel Ripplinger[*],

Bogdan Sacaleanu[§], Diana Raileanu[§], Detlef Prescher[§]

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

[*] Eurospider Information Technology AG
Schaffhauserstrasse 18
CH-8006 Zürich, Switzerland
ripplinger@eurospider.ch

Abstract

The paper describes an XML annotation format and tool developed within the MUCHMORE project. The annotation scheme was designed specifically for the purposes of Cross-Lingual Information Retrieval in the medical domain so as to allow both efficient and flexible access to layers of information. We use a parallel English-German corpus of medical abstracts and annotate it with linguistic information (tokenisation, part-of-speech tagging, lemmatisation and decomposition, phrase recognition, grammatical functions) as well as semantic information from various sources. The annotation of medical terms/concepts, semantic types and semantic relations is based on the Unified Medical Language System (UMLS). Additionally, we use EuroWordNet as a general-language resource in annotating word senses and to compare domain-specific and general language use. A major aim of the project is also to complement existing ontological resources by extracting new terms and new semantic relations. We present the annotation scheme, which is conceptually related to stand-off annotation, and describe our tool for automatic semantic annotation.

1. Introduction

This paper describes the XML-based annotation format (DTD) that was developed according to the aims and needs of the MUCHMORE[1] project on Cross-lingual Information Retrieval (CLIR) in the medical domain. Our approach to CLIR can be described as concept-based or semantics-driven, where the main research goal of the project is to exploit multiple levels of semantic annotation from different sources in order to enhance document retrieval in a domain-specific, multilingual context. This task invariably includes linguistic pre-processing steps such as tokenisation, and part-of-speech tagging, morphological analysis (lemmatization, decompounding), and syntactic analysis (phrase recognition, grammatical functions).

Semantic annotation is performed on the basis of a publicly available medical language resource UMLS[2] (Unified Medical Language System), which 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. Next to UMLS in the medical domain, we also use EuroWordnet[3] as a general language semantic resourced.

Although the medical domain provides extensive online semantic resources, the project additionally seeks to complement existing sources by extracting novel terms and novel semantic relations from parallel (and comparable) corpora, which are subsequently also integrated in the annotation described here.

In a domain with a highly developed and complex scientific language, described by several – not necessarily compatible – conceptual hierarchies, which contain information that may or may not be relevant for a specific project task or for the final task of document retrieval, combining all these layers of information is not trivial. It was necessary to develop an encoding scheme that would offer efficient access to and indexing of individual data tracks, while at the same time allow flexible combinations and interactions between layers. Furthermore the format needs to be adaptable to specific project tasks, e.g. indexing for retrieval purposes, word sense disambiguation, term and relation extraction, evaluation etc.

The following section briefly outlines the corpus selection and pre-processing steps. The third section describes the resources used for linguistic and semantic annotation, while the fourth section presents the actual annotation format. We conclude by comparing and justifying our approach in relation to some other well-known annotation projects and by presenting some of our further work that is planned within the MUCHMORE project.

2. Corpus Selection and Preparation

The corpus used in the development of the annotation format is a parallel corpus of English-German scientific medical abstracts obtained from the Springer Link web

1 http://muchmore.dfki.de
2 http://umls.nlm.nih.gov
3 http://www.hum.uva.nl/~ewn/

site[4]. The corpus consists approximately of 1 million tokens for each language. Abstracts are from 41 medical journals, each of which constitutes a relatively homogeneous medical sub-domain (e.g. Neurology, Radiology, etc.).

In a preparation phase, we normalized the downloaded HTML documents in various ways, in order to produce a clean, plain text version, consisting of a title, abstract and keywords. Additionally, the corpus was aligned on the sentence level.

3. Annotation Resources and Tools

3.1. Linguistic Annotation

The corpus is linguistically analyzed 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.

On top of shallow analysis, also grammatical functions such as subject, object and indirect object are annotated, using a tool that is currently under development at DFKI.

3.2. Semantic Annotation

3.2.1. Terms, Concepts and Semantic Types

A major objective of the MUCHMORE project is to explore techniques for enhancing cross-lingual information retrieval through automatic semantic annotation of domain-specific terms and relations. For this purpose, the publicly available medical language resource UMLS (Unified Medical Language System) is used. As mentioned above, UMLS organizes linguistic, terminological and semantic information in three interrelated parts: Specialist Lexicon, Metathesaurus and Semantic Network. At the level of terms, the following semantic information is used in annotation:

- Concept Unique Identifier (CUI)

  maps a term to a concept in the Metathesaurus

- Type Unique Identifier (TUI)

  maps a concept to one or more semantic types in the Semantic Network

- Medical Subject Headings ID (MeSH is one of the medical thesauri underlying the MetaThesaurus)

  maps a CUI to one or more MeSH[5] codes

- Preferred Term

  a term that is marked as preferred for a given set of terms and a corresponding concept

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

[5] http://www.nlm.nih.gov/mesh/meshhome.html

The decision to use MeSH codes in addition to CUIs was based on our observation, confirmed by medical experts, that the UMLS Semantic Network, especially the semantic types and relations, does not always adequately – or even accurately – represent the domain-specific relationships that we intend to exploit for CLIR purposes.

MeSH codes on the other hand have a transparent structure, from which both the semantic class of a concept and its depth in the hierarchy can be inferred. For example, the term *infarction* C23.550.717.489 and *myocardial infarction* C14.907.553.470.500 both belong to the group of diseases (C), but the node of the first term lies higher in the hierarchy as its code has fewer fields.

There are several possible levels of ambiguity at the level of terms and concepts: a single term may be assigned several CUIs, and a single CUI may be mapped to several MeSH codes. Since UMLS is designed to become the ultimate ontological resource for the medical domain, which unifies all previously existing conceptual hierarchies, the MeSH hierarchy is viewed as a subset of the UMLS Metathesaurus. This relationship is reflected in our annotation scheme in the following way: We treat terms with several CUIs as ambiguities, i.e. different possible *readings* of a term, and therefore annotate each reading as a separate element with its corresponding information. However, if a CUI can be mapped to several MeSH codes, we annotate those as possible *mappings* subordinate to the CUI.

Another ambiguity occurs on the level of semantic types - TUI's. Thus, for example, the term *Type I Collagen* with C0041455 can have the semantic type T116 or T123, meaning *Amino Acid, Peptide or Protein* or *Biologically Active Substance* respectively. But since in this case we are still dealing with a single concept, which can be viewed from different perspectives depending on the context, we do not consider multiple TUIs to represent real ambiguities and thus do not treat them as different readings of a term.

Example:

```
<umlsterm id="t5" from="w23">
    <cui code="C0078414"
    preferred="cisplatin/etoposide
    protocol" tui="T061"/>
    <cui code="C0031618"
    preferred="Phosphatidylethanolamines"
    tui="T119">
      <msh code="D10. ... .400.840"/>
    </cui>
</umlsterm>
```

In a similar way we integrate and annotate new terms extracted by the bilingual term extraction tool developed at Xerox Research Center Europe (XRCE), one of the partners in the MUCHMORE project.

3.2.2. Semantic relations

Semantic relations are currently annotated between semantic types (TUIs) that co-occur within a sentence. This means that we can only annotate relations between items that were previously identified as terms. The semrel element thus refers to the level of UMLS or novel (XRCE) terms by specifying the pair of terms and the type of relation found.

Example:

```
<semrel id="r5" term1="t5" term2="t3"
reltype="affects"/>
```

Due to the generic nature of semantic types, the number of possible semantic relations specified between them in UMLS can be considerable, while their actual usefulness for document retrieval seems questionable.

However, through term disambiguation and relevance-based selection of relations it is already possible to prune them. Further, we also intend to use MeSH codes to distinguish between higher (UMLS Semantic Network) and lower (MeSH) level relations, i.e. generic vs. specific ones. For this purpose, work is underway to identify novel semantic relations, based in particular on co-occurrence analysis and clustering of MeSH concepts, identification of verbal and other lexical patterns, and on grammatical function analysis.

3.2.3. EuroWordNet Senses

In addition to UMLS, terms are annotated with EuroWordNet (Vossen, 1997) to compare domain-specific and general language use. We annotate both single- and multi-word EWN terms, whereby each possible sense of a term represents a separate element `sense` with the attribute `offset` giving the EWN code of the sense. For practical reasons we limit the annotation of EWN senses to nouns only.

Example:

```
<ewnterm id="e1" from="w5">
   <sense offset="4690182"/>
   <sense offset="8542711"/>
</ewnterm>
```

3.2.4. Annotation Tool

The tool for automatic annotation of the semantic levels described above is written in Perl and was developed at DFKI. It takes the linguistically processed files produced by ShProt as input and adds all the missing semantic levels step by step.

First it identifies the UMLS terms, whereby the term matching is performed for uni-, bi- and trigrams based on word stems, if available, and on word forms otherwise. Term lookup is also done for individual lemmas in the case of analyzed compounds as well as for word parts in the case of non-analyzed hyphenated tokens. Part-of-speech filters, case normalization and word order inversion are also implemented to improve term matching.

Once a string has been identified as a UMLS or EWN term, corresponding databases are consulted to provide mappings to concept codes and semantic types or senses respectively.

Semantic relations are identified on the basis of combinations of semantic types (TUIs) found within the sentence, which are then matched against the database of semantic relations in the UMLS Semantic Network. Additional filters are used to remove relations occurring between different senses of the same term.

The output of the tool is an XML document corresponding to the DTD described below.

4. Annotation Format

The annotation task involves combining 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.

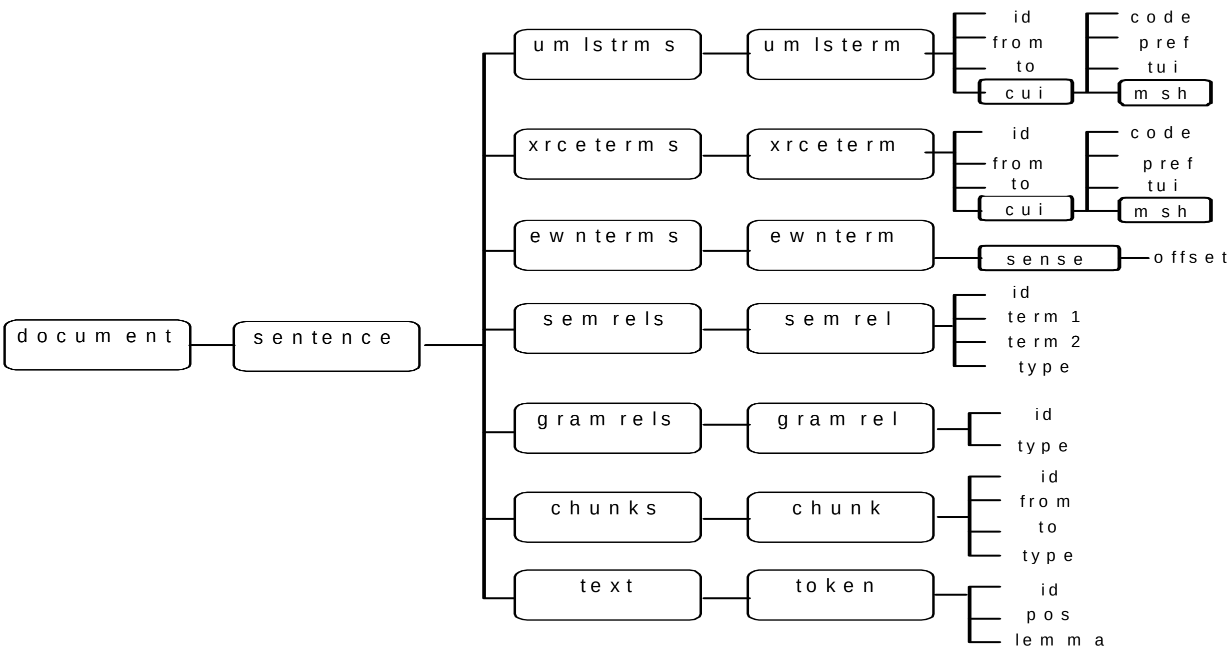

Figure 1: Annotation scheme

A document consists of a `title` (optional), any number of sentences and a set of keywords (also optional). The division into different layers introduces elements for UMLS terms, novel XRCE terms, EuroWordNet terms, UMLS semantic relations, grammatical relations, chunks and for the text itself (tokens). The elements `umlsterm`, `xrceterm`, `ewnterm`, `chunk`, and `gramrel` refer to the `text` level through indices on tokens, whereby we use two attributes (`from` and `to`) to mark the beginning and end token of the referring element. The `semrel` element refers to the `umlsterms` level through indices on the pair of terms between which the relation was identified (see Figure 1).

As explained above, the treatment of ambiguities and alternative or parallel concept mappings depends on the type of ambiguity and its relevance for IR in the medical domain. We therefore treat terms with multiple CUIs as "real" ambiguities that need to be annotated in separate elements, whereas ambiguity of semantic type (TUI) represents the same concept viewed from different perspectives. A similar approach was taken for novel XRCE terms.

The annotation of semantic information is of particular importance for our purpose of creating an efficient concept-based framework for Cross-Language Information Retrieval. The format must allow for individual layers being indexed separately or used in various combinations. New annotation levels can be added simply by referring to existing indices (tokens, terms).

Similarly, for project tasks that do not require all information, levels can be removed through simple reformatting without corrupting the document's consistency.

5. Related Work

Especially over the last decade, the NLP community has put considerable efforts in development of standards and conventions for text encoding, out of which TEI (Sperberg-McQueen and Burnard, 2002) and CES/XCES (Ide et al., 2000) are probably the most widely used.

The annotation scheme developed in the MUCHMORE project is related to the concept of stand-off annotation (e.g. McKelvie et al., 1997; Thompson and McKelvie, 1997), which is recommended also by the TEI Consortium. Various levels of information are encoded in separate annotation layers, although we still keep them within the same document. At present, our annotation framework is designed primarily to serve project-internal purposes, however state-of-the-art XSLT-based tools allow conversion into TEI/XCES-compliant format in case a broader dissemination of our corpora is required.

In addition, several other projects[6] are related to our work described here, e.g. the ATLAS architecture, which builds on the notion of annotation graphs (Bird and Liberman, 2001). Within the MATE project (Dybkjær et al., 1998) a TEI/CES conformant annotation scheme implementing stand-off annotation was designed, providing for complex multilevel encoding of speech and including solutions for overlapping or alternative annotations. The project also developed an annotation workbench facilitating manual tagging and validation.

The main reason for developing a project-internal format was that none of the formats proposed in any of the above projects fulfills the specific requirements of a concept-based CLIR setting. Most annotation tools are tailored to the needs of speech annotation, where many of the tasks are still performed manually, where – in the case of dialogue – utterances may overlap and the text stream is not linear, and where many layers, related for instance to prosody, pragmatics of communication, or non-linguistic elements are a matter of subjective judgment.

All of these issues are very distinct from the annotation task in the context of cross-lingual medical IR, where both document and query processing must be performed fully automatically and the main challenge lies in issues like effective term matching, concept mapping, word sense disambiguation and relation extraction, all of which in turn have an impact on further important issues related to the indexing and weighting of individual layers of data.

6. Conclusions

We described the annotation format, tools and resources used in the MUCHMORE project on Concept-Based Cross-Lingual Information Retrieval. Our annotation setting incorporates several linguistic levels (tokenisation, lemmatisation, POS-tagging, chunking, grammatical functions) and semantic levels, using different semantic resources (UMLS terms, semantic types and relations; novel extracted terms and relations; EuroWordNet senses). For specific project tasks and for the document retrieval itself, these data structures are used in various combinations. Therefore our annotation format is organized around several layers of information, all referring to the basic text level via indices.

We conducted some initial experiments in evaluating the benefits of semantic annotation within a CLIR setting (Ripplinger et al, 2002). Most of the monolingual test runs do not yet achieve the performance we aim for, partly due to incomplete semantic resources and partly also originating out of errors in morphological analysis (both especially for German). Nevertheless, a clear increase in both precision and recall when using semantic data was already observed in cross-lingual runs, and we expect to obtain substantially better results after the project tasks of disambiguation, (novel) term extraction and (novel) relation extraction have been completed and fully integrated.

Acknowledgements

This research has in part been supported by EC/NSF grant *IST-1999-11438* for the MUCHMORE project.

References

Bird S. and Liberman, M., 2001. *A Formal Framework for Linguistic Annotation.* Speech Communication 33 (1,2), 23-60.

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

Dybkjær, L., Bernsen, N.O., Dybkjær, H., McKelvie, D. and Mengel, A. 1998. The MATE Markup Framework.

[6] For a comprehensive list see http://www.ldc.upenn.edu/annotation/

MATE Deliverable D1.2, November 1998. http://mate.nis.sdu.dk/information/d12/

Ide, N., Bonhomme, P., Romary, L. 2000. XCES: An XML-based Standard for Linguistic Corpora.. Proceedings of the Second Language Resources and Evaluation Conference (LREC), Athens, Greece, 825-30.

McKelvie D., Brew C. and Thompson H. 1997. *Using SGML as a Basis for Data-Intensive NLP.* In Proceedings of ANLP97, Washington, DC.

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

Ripplinger B., Vintar S. and Buitelaar P. 2002. *Cross-Lingual Medical Information Retrieval through Semantic Annotation.* Proceedings of the EFMI Workshop on Natural Language Processing in Biomedical Applications, EFMI, ??.

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

Thompson H. and McKelvie D. 1997. *Hyperlink semantics for standoff markup of read-only documents.* In Proceedings of SGML Europe 97, Barcelona.

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