A Multi-Layered, XML-Based Approach to the Integration of Linguistic and Semantic Annotations

Paul Buitelaar*, Thierry Declerck§, Bogdan Sacaleanu*, Špela Vintar*, Diana Raileanu*, Claudia Crispi§

*Language Technology Lab
DFKI GmbH
{paulb|bogdan|vintar|raileanu}@dfki.de

§Department of Computational Linguistics
University of Saarland
{declerck|crispi}@dfki.de

Abstract

In this paper we present a multi-layered approach to document annotation that allows for the structural integration of linguistic and semantic annotations produced by various language technology tools and using knowledge encoded in different domain ontologies as needed for semantic web applications.

1 Introduction

Establishing the semantic web on a large scale implies the widespread annotation of web documents with ontology-based knowledge markup. For this purpose, tools have been developed that allow for semi-automatic annotation of web documents with ontology-based metadata[1]. However, given that a large number of web documents consist either fully or at least partially of free text, language technology tools will be needed to support this authoring process by providing an automatic analysis of the semantic structure of textual documents. In this way, free text documents will become available as semi-structured documents, from which meaningful units can be extracted automatically (*information extraction*) and organized through clustering or classification (*text mining*).

There are many language technology tools available that may be used in document annotation for the semantic web. Typically, such tools will cover several linguistic analysis steps, such as morphological analysis, part-of-speech tagging, or phrase recognition. Some tools will cover additional analysis levels, such as dependency structure analysis, semantic tagging, discourse analysis, etc. However, there are no tools that fully cover *all possible* linguistic and semantic analysis steps. Therefore, there is a need to integrate the results from different language technology tools as well as from different linguistic and semantic analysis steps into one coherent representation. This involves the following interrelated aspects:

- Integration of linguistic (PoS, lemma, phrase, etc.) and semantic information (concept, relation, event, etc.)
- Integration of semantic information from different semantic resources (ontologies)
- Integration of various XML formats, at both structural and semantic level (XML Schema)

In this paper we present a multi-layered approach to document annotation that allows for the structural integration of linguistic and semantic annotations produced by various language technology tools and using knowledge encoded in different domain ontologies as needed for semantic web applications. We will focus on the following linguistic analysis steps: morphological analysis,

[1] http://Annotation.SemanticWeb.org

```
Balint syndrom is a combination of symptoms including simultanagnosia, a
disorder of spatial and object-based attention, disturbed spatial percep-
tion and representation, and optic ataxia resulting from bilateral pa-
rieto-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="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>

<chunks>
      <chunk id="c1" from="w1" to="w2" type="NP"/>
      <chunk id="c7" from="w20" to="w23" type="NP"/>
</chunks>
```

Figure 1: Linguistic annotation in MuchMore (PoS, Morphology, Chunks)

part-of-speech tagging, chunking, dependency structure analysis; and the following semantic analysis steps: term tagging, relation tagging, event recognition. Examples for each are given in the context of two projects that use linguistic and semantic annotation for the purpose of cross-lingual information retrieval (MuchMore[2]) and content-based multimedia access (MUMIS[3]).

2 Linguistic Annotation

2.1 MuchMore: PoS, Morphological Analysis, Chunks

In this section we present an example of linguistic annotation for English and German texts as used in the MuchMore project on concept-based, cross-lingual information retrieval. The MuchMore annotation format integrates multiple levels of linguistic analysis in a multi-layered XML-based DTD, which organizes each level as a separate track with options of reference between them via indices [1]. Linguistic annotation in MuchMore is based on *ShProT*, a shallow processing tool that consists of a tokenizer [2], a part-of-speech tagger [3], a morphological analyser [4] and a chunk parser [5]. In addition, MuchMore also covers semantic tagging of terms and relations, using EuroWordNet [6] and UMLS (Unified Medical Language System)[4] as primary semantic resources.

Part-of-Speech

Part-of-Speech (PoS) tagging is performed by TnT [3], which is an HMM based part-of-speech tagger trained on general language corpora (the NEGRA corpus for German[5], the SUSANNE corpus for English[6]). In order to perform in an optimal way, TnT needs to be adapted to a specific domain. Two approaches may be considered: retrain TnT on an annotated domain-specific corpus or update the underlying TnT lexicon.

As part-of-speech annotated medical corpora are difficult to obtain, we decided in the context of MuchMore to extend the existing TnT lexicon with information from a medical lexicon. Because of a similar syntax for general language and the medical language of the MuchMore cor-

[2] http://muchmore.dfki.de
[3] http://parlevink.cs.utwente.nl/projects/mumis/
[4] http://umls.nlm.nih.gov
[5] http://www.coli.uni-sb.de/sfb378/negra-corpus/
[6] http://www.cogs.susx.ac.uk/users/geoffs/Rsue.html

pus of scientific abstracts, we obtained good results without retraining.

Morphological Analysis

Morphological analysis for both German and English is based on a full-form lexicon generated by Mmorph [4]. Each token is looked up for a matching entry that will provide its morphological information. If no valid word form has been matched, the token is analysed as a potential compound.

Initial decompounding experiments produced poor results. However, after adapting the existing Mmorph lexicon to the medical domain results rather improved. Adaptation proceeded according to the following two steps. First, the Mmorph lexicon was updated with additional morphological information from medical lexicons for both German and English. This enabled us to avoid incorrect decompositions like:

```
Zoonoses   → zoo + nose
Epicillin  → epic + ill + in
Endoral    → end + oral
```

Secondly, general-language word forms that function as prefixes in the medical domain (e.g. `auto`) were removed from the Mmorph lexicon, avoiding incorrect decompositions like:

```
Autoimmune →
     auto + immune
Postinflammatory →
     post + inflammatory
Radiogram →
     radio + gram
```

Chunks

For chunk analysis we use Chunkie [5], which is an HMM-based partial parser that recognizes boundaries as well as the internal structure of simple and complex phrases. On the basis of the PoS and morphological information, Chunkie is able to determine noun phrases (NP), adjectival phrases (AP) and prepositional phrases (PP). As with TnT, the performance of Chunkie can be improved by adaptation to a specific domain, if a manually annotated domain-specific training corpus is available. Unfortunately, as with part-of-speech tagging this was not the case.

Example

Linguistic annotation of PoS, morphology and chunking in MuchMore is illustrated by the example in Figure 1. In the MuchMore annotation format, each sentence contains a block that holds tokens with lemma and part-of-speech information. Each phrase is annotated by use of indices over tokens. In the example an NP is found that covers the tokens *`w1-2`* (`Balint syndrom`) and a more complex NP that covers the tokens *`w20-23`* (`spatial perception and representation`).

2.2 MUMIS: Named Entities, Full Parsing, Dependency Structure and Grammatical Functions

In this section we present linguistic annotation of PoS, morphology, named entities, chunks, dependency structure (including Grammatical Functions) as performed in the MUMIS project on content-based multimedia access, dealing with the soccer domain [7].

We present here only the work done on German texts, which is based on an integrated set of linguistic tools – SCHUG: Shallow and Chunk based Unification Grammar, which implements a rule-based system of cascades [8]. The application defined by the MUMIS scenario implies annotation of named entities, head-modifier structure and grammatical functions in addition to the shallow processing information also covered by MuchMore. SCHUG has adopted the annotation schema developed in MuchMore, which allows for a smooth integration of the various annotation layers provided by these two systems. Therefore, it is possible to include dependency structure information in this annotation format and also to add a new annotation layer that provides information on grammatical functions, even if this information is provided by tools not included in the MuchMore system architecture.

SCHUG implements a modular strategy for the recognition of domain specific named entities. For MUMIS in particular, the task is to detect soccer relevant named entities (time expressions, players, teams, etc.). This information will be

```
[NP Industrie, Handel und Dienstleistungen] [VG werden] [PP in der er-
sten Liste] [VG aufgeführt], wobei [NP die in Klammern gesetzten Zahlen]
[PP auf die Mutterfirmen] [VG hinweisen].

<chunks>
      <chunk id="c1" from="w1" to="w5" type="NP" head="w1,w3,w5"/>
      <chunk id="c2" from="w6" to="w6" type="VG"/>
      <chunk id="c3" from="w7" to="w10" type="PP" head="w7"
                  complement="w8,w9,w10"/>
   <chunk id="c4" from="w11" to="w1" type="VG"/>
      ….
</chunks>

<clauses>
<clause id="cl1" from="c1" to="c4" pred_struct="c2 c4"
            GF_Subj="c1"/>
<clause id="cl2" from="c6" to="c9" pred_struct="c9"
      GF_Subj="c6"/>
</clauses>
```

Figure 2: Linguistic Annotation in MUMIS (Chunks, Clauses)

encoded on the chunk annotation level, with an index pointing to the distinct tokens that correspond to each individual named entity.

The chunking procedure of SCHUG consists of a rule-based sequence of cascades, which produces a richer linguistic representation than the MuchMore tools. As shown in Figure 2, also verb groups, which can be complex, are annotated (chunks are put into square brackets). The MuchMore annotation format allows for a straightforward mapping of this information in the chunk layer. Note that also information on heads, modifiers and complements are represented. In the case of a PP, the head is always the preposition and its complement is always an NP. The internal structure of the complement NP is not given here.

Next, also grammatical functions are annotated. In order to detect these accurately, an analysis of the *clauses* of a sentence is required. Clauses are the subparts of a sentence that correspond to a (possibly complex) semantic unit, each of which contains a main verb with its complements (grammatical functions) and possibly other chunks (modifiers).

The example (in Figure 2.) has two clauses, each with a span of several chunks with information on its predicative structure (`pred_struct`) and grammatical function (e.g. `GF_Subj`). Predicative structure can be complex, as with the first clause, where the predicate corresponds to a discontinuous verb group with an auxiliary verb - `werden (to be)` - and a main verb - `aufführen (perform)`.

3 Semantic Annotation

3.1 MuchMore: Terms and Relations

In addition to the annotation of corpora with shallow processing information as discussed in section 2.1, also semantic information is annotated in the context of the MuchMore project. This includes semantic tagging with EuroWordNet synsets as well as semantic tagging with UMLS concepts and relations (and MeSH[7] concepts).

UMLS

A major objective of the MuchMore project is to explore techniques for enhancing cross-lingual information retrieval (CLIR) through automatic semantic annotation of domain-specific terms and relations. For this purpose, the publicly available medical language resource UMLS is used. At the level of terms, the following semantic information is used in annotation:

- **Concept Unique Identifier** (CUI): maps a term to a concept

[7] MeSH (Medical Subject Headings) is one of the medical terminologies that are part of UMLS http://www.nlm.nih.gov/mesh/meshhome.html

```
<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>
<umlsterm id="t8" from="w26" to="w26">
      <concept id="t8.1" cui="C0029144" preferred="Optics" tui="T090">
            <msh code="H1.671.606"/>
      </concept>
</umlsterm>

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

<ewnterm id="e2" from="w21" to="w21">
      <sense offset="0487490"/>
      <sense offset="3955418"/>
      <sense offset="4002483"/>
</ewnterm>
```

Figure 3: Semantic Annotation in MuchMore (Terms, Relations)

- **Type Unique Identifier** (TUI): maps a concept to one or more semantic types
- **Medical Subject Headings** (MeSH id): maps a CUI to one or more MeSH codes
- **Preferred Term**: a term that is marked as preferred for a given set of terms and a corresponding concept

Semantic relations are currently annotated between semantic types (TUIs) that co-occur within a sentence. This means that we can only annotate relations between terms that were mapped to concepts. The `semrel` element thus refers to the level of UMLS terms by specifying the pair of terms and the type of relation found. Due to the generic nature of semantic types, the number of occurring semantic relations between terms in a given text can be considerable. However, through term disambiguation and relevance-based selection of relations it is possible to prune them.

EuroWordNet

In addition to UMLS, terms are annotated with EuroWordNet 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 XML element `sense` with the attribute `offset` (EWN code of the sense). For the purpose of cross-lingual information retrieval we limit the annotation with EWN to senses for nouns only.

Example

For the example sentence of section 2.1 (see Figure 3. below), the words *`w20-21`* (`spatial perception`) point to the concept `Space Perception` in UMLS, which corresponds to the CUI code `C0037744` and TUI code `T041` (`Mental Process`) and to two MeSH codes. Word `w26` (`optic`) triggered the concept `Optics` with corresponding CUI, TUI and MeSH codes.

UMLS further defines that `Space Perception` is an issue in `Optics` (expressed by the relation: `issue_in`), which is encoded as a relation between term/concept `t7.1` and term/concept `t8.1`.

In EuroWordNet annotation, word *`w21`* (`perception`) has 3 senses, corresponding to the following synsets:

```
0487490:   perceiving
           perception
```

```
Ein Freistoss von Christian Ziege aus 25 Metern geht über das Tor.

(A 25-meter free-kick by Christian Ziege goes over the goal.)

<chunks>
      <chunk id="c1" from="w1" to="w5" type="NP" head=”w2”
            pp_modifier=”w3 w4 w5”/>
      <chunk id="c2" from="w6" to="8" type="PP" head=”w6”
            complement=”w7 w8”/>
      <chunk id="c3" from="w9" to="9" type="VG"/>
      <chunk id="c4" from="w10" to="w12" type="PP" head=”w10”
            complement=”w11 w12”/>
</chunks>

<clauses>
<clause id="cls1" from="c1" to="c4" pred_struct="c3" GF_Subj="c1"/>
</clauses>

<events>
      <event id="e1" clause=”cls1” event-name=”free-kick”>
            <arguments>
               <argument id="arg1" name="player” value=”w4, w5”/>
   <argument id="arg2" name="location” value=”25-meter”/>
               <argument id="arg3" name="time” value=”07:00”/>
            </arguments>
      </event>
      <event id="e2" clause=”cls1” event-name=”goal-scene-fail”>
            <arguments>
               <argument id="arg1" name="player” value=”w4, w5”/>
               <argument id="arg2" name="location” value=”25-meter”/>
               <argument id="arg3" name="time” value=”07:00”/>
            </arguments>
      </event>
</events>
```

Figure 4: Semantic Annotation in MUMIS (Relations, Events)

```
           sensing
3955418:   perception
4002483:   percept
           perception
           perceptual experience
```

We are currently working on a word sense disambiguation module to cut down on ambiguities concerning EuroWordNet senses, based on unsupervised training as described in [9] [10].

3.2 MUMIS: Relations and Events

The primary goal of the MUMIS project is to generate formal annotation from heterogeneous textual sources that allows for content-based indexing of soccer videos. The project provides highly structured domain-specific annotation of relevant entities, relations and events that can be extracted from transcribed audio and video broadcasts of specific soccer games and corresponding on-line textual documents.

The linguistic annotations of these documents as provided by the SCHUG tools are enriched with domain-specific information that is encoded in an ontology of the soccer domain. The ontology represents in a hierarchical fashion the main events, relations and entities relevant to the domain. The nodes of the ontology are associated with multilingual terms (acting as instances of the concepts), encoded in the XML-compliant, TMX-format (Translation Memory eXchange format)[8].

[8] For more information on the TMX format: http://www.lisa.org/tmx/tmx.htm

Example

One of the main text types that MUMIS is dealing with is the *on-line ticker* (short descriptions of interesting soccer events, each event indicated by a time-code), as shown in the example in Figure 4. Here, various relations (`player, location,…`) and events (`free-kick, goal-scene-fail`) that are relevant to the soccer domain are recognized

Some relations are not explicitly mentioned, but can still be inferred by the MUMIS system. For example, the `team` for which `Ziege` is playing can be inferred from the ontological information that a `player` is part of a `team` and the instance of this particular `team` can be extracted from additional texts or metadata. In this way, information not present in the text directly can be added by additional information extraction and reasoning (see also [11]).

4 Conclusions

We presented an approach that allows for the integration of linguistic annotations, using different language technology tools, and domain-specific knowledge markup (semantic annotations), using information from various semantic networks or ontologies. In this way, the approach allows for different semantic views on a document (or set of documents), each corresponding to specific domain ontologies.

Both annotation examples are based on a common XML Scheme we are developing for integration of various resources in XML format. W3C XML Schema borrows a number of concepts from object-oriented programming (abstract types, type substitution, polymorphism) that allow development of schemas, which define generic base types and extend these types without affecting the original schema. Our approach is driven by the idea of a linguistic core as base type that consists of some “basic units” (token, part-of-speech, morphology and chunk information) to which further linguistic and semantic annotations can be added by extension.

5 Related and Future Work

The work presented here uses ontologies (EuroWordNet, UMLS, MUMIS ontology) that are not yet (fully) represented in an RDF/S format. Therefore, also the representation of concepts from these ontologies in the MUMIS and MuchMore semantic annotation has not yet been designed to work with an RDF/S format.

There are two ways of achieving this: first by substituting the current XML-based semantic annotation by one based on RDF/S, which would in effect could be reduced to the copying of relevant ontology items into the annotation; secondly, and following from this, we maintain the XML-based semantic annotation as it currently is but substituting concept ids or codes with pointers into RDF/S-based ontologies. In ongoing work we are focusing on this aspect, more specifically within the EU-funded *Esperonto* project[9] and the Special Interest Group *OntoWeb-lt* on Language Technology in Ontology Development and Use[10] within the EU-funded Thematic Network *OntoWeb*[11].

Future work will be concerned also with a closer compliance with emerging standards, such as those of the ISO working group on language resources (ISO/TC37/SC4)[12], which addresses exactly those issues in document annotation presented in this paper: multi-layer annotation for the exchange of information between language technology modules.

Finally, future work will take into account results from related initiatives, projects and tools, such as: the GDA[13] initiative for linguistics-based semantic web annotation, the document annotation tool MELITA[14] (based on information extraction), and the ContentWeb project [12].

Acknowledgements

This research has in part been supported by EC/NSF grant IST-1999-11438 for the MuchMore project, EC grant IST-1999-10651 for the MUMIS project, EC grant IST- 2000-29243 for the OntoWeb project and EC grant IST-2001-34373 for the Esperonto Project.

---

[9] http://www.esperonto.net
[10] http://ontoweb-lt.dfki.de
[11] http://www.ontoweb.org
[12] http://www.iso.ch/iso/en/stdsdevelopment/tc/tclist/TechnicalCommitteeDetailPage.TechnicalCommitteeDetail?COMMID=5393
[13] http://www.i-content.com/GDA
[14] http://www.dcs.shef.ac.uk/%7Ealexiei/Melita.htm

References

[1] Vintar Š., Buitelaar P., Ripplinger B., Sacaleanu B., Raileanu D., Prescher D. *An Efficient and Flexible Format for Linguistic and Semantic Annotation* In: Proceedings of LREC2002 , Las Palmas, Canary Islands - Spain, May 29-31, 2002.

[2] Piskorski J., G. Neumann. *An Intelligent Text Extraction and Navigation System.* Proceedings of the 6th International Conference on Computer-Assisted Information Retrieval (RIAO). 2000.

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

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

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

[6] 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.

[7] Declerck T., Wittenburg P., Cunningham H. *The Automatic Generation of Formal Annotations in a Multimedia Indexing and Searching Environment*. Proceedings of the Workshop on Human Language Technology and Knowledge Management, ACL-2001.

[8] Declerck T. *A set of tools for integrating linguistic and non-linguistic information*. Proceedings of SAAKM 2002, ECAI 2002, Lyon.

[9] Buitelaar P., Alexandersson J., Jaeger T., Lesch S., Pfleger N., Raileanu D., von den Berg T., Klöckner K., Neis H., Schlarb H. *An Unsupervised Semantic Tagger Applied to German*. In: Proceedings of Recent Advances in NLP (RANLP) **,** Tzigov Chark, Bulgaria. 2001.

[10] Buitelaar P., Sacaleanu B. *Ranking and Selecting Synsets by Domain Relevance*. In: Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations. NAACL 2001 Workshop, Carnegie Mellon University, Pittsburgh. 2001.

[11] Saggion H., Kuper J., Declerck T., Reidsma D., Cunningham H. *Intelligent Multimedia Indexing and Retrieval through Multi-source Information Extraction and Merging.* Technical Report, University of Sheffield.

[12] Guadeloupe Aguado-de-Cea, Inmaculada Alvarez-de-Mon, Antonio Pareja-Lora and Rosario Plaza-Arteche: *RDF(S)/XML Linguistic Annotation of Semantic Web Page,* In Proceedings of NLPXML 2002, COLING, Taipei. 2002.