Generic NLP Tools for Supporting Shallow Ontology Building

Thierry Declerck[1], Mihaela Vela[2]

[1]DFKI GmbH, Language Technology Lab Affiliation
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
declerck@dfki.de
[2]Universität des Saarlandes
Saarbrücken, Germany
m.vela@mx.uni-saarland.de

Abstract

In this paper we present on-going investigations on how complex syntactic annotation, combined with linguistic semantics, can possibly help in supporting the semi-automatic building of (shallow) ontologies from text by proposing an automated extraction of (possibly underspecified) semantic relations from linguistically annotated text.

1. Introduction

The Semantic Web is marking a new stage in advanced automated textual analysis, ontologies becoming a key instrument in the development of applications requiring semantic resources, like for example information extraction (IE), knowledge acquisition (KA) and text-based knowledge discovery (KD). Some research projects have already investigated the combination of Semantic web technologies and natural language processing (see for example Ontoweb, Esperonto or Sekt[1]), whereas our work is based mainly on results of (Esperonto), and it seems that today the two communities are agreeing on the basic principles for the integration of both kind of technologies for the purpose of semantic annotation (or Knowledge Mark-Up) of textual documents (see for example Workshop, 2003).

But it remains the problem that domain specific ontologies are still missing (or are incomplete) for deployment in real world applications and the design and construction of (domain specific) ontologies is itself a time consuming task, which requires many human resources. Therefore there is need for investigating methods and tools for supporting (supervised) automated ontology building, also from textual documents, as this has already been mentioned in (Maedche & Staab, 2002). In (Esperonto) and (Declerck & Vela, 2005) a methodology for extracting generic ontology relevant semantic relations from linguistically annotated text has been discussed. The linguistic annotation suggested there consists in a combination of structural information and linguistic dependencies, within linguistic fragments (e.g. head-modifier structures within NPs) and between linguistic fragments (Subject, Object etc.). For the time being we identify several linguistic phenomena on which the heuristics for semantic relation extraction can apply. Those phenomena and associated (heuristic) rules will be present in some details later in this paper, after a short presentation of the NLP tool, which is delivering the linguistic annotation for our experiment.

2. The NLP tools supporting the Extraction of Semantic Relation

In this section we describe briefly the linguistic tools, called SCHUG (Shallow and Chunk-based Unification Grammar Tools) that have been partially implemented in (Esperonto), and more precisely the recent extensions of those tools that are supporting the shallow ontology building on the basis of the automatically extracted semantic relation.

The NLP tools implement a cascaded chunk parsing up to the level of clausal analysis, including annotation of dependencies. The semantic extraction itself is based on a set of rules that apply to various linguistic phenomena, like adjectival pre-modification, post-nominal genetive and prepositional modification, NP and clausal coordination, etc. Those rules apply bottom up, which means that the relation extraction starts within the phrases and extends then to the relations between linguistic fragments.

The set of rules has been implemented as a Perl module that has been added to the processing chain of SCHUG. This module delivers the extracted semantic relations in the form of a graph. Graphs resulting from various documents can then be merged and so propose a unified structure for the semantic relations extracted from a larger set of linguistically annotated documents.

Below an example of such a graph is shown (figure 2), together with the (bracketed) sentence (figure 1) out of whose subject it has been generated:

```
[Synovial inflammation in rheumatoid
arthritis] [is] [closely related] [to
the formation of ectopic lymphoid
microstructures] [.]
```

Figure 1: The shallow constituency bracketing above shows a NP, followed by a VG (VerbGroup), an Adjective Phrase and a PP.

In Figure 1, we do not display the dependencies within and between linguistic fragments, but the reader can see in Figure 2 the kind of semantic relations we can automatically extract from those dependencies.

[1] URLs of those projects are given in the references.

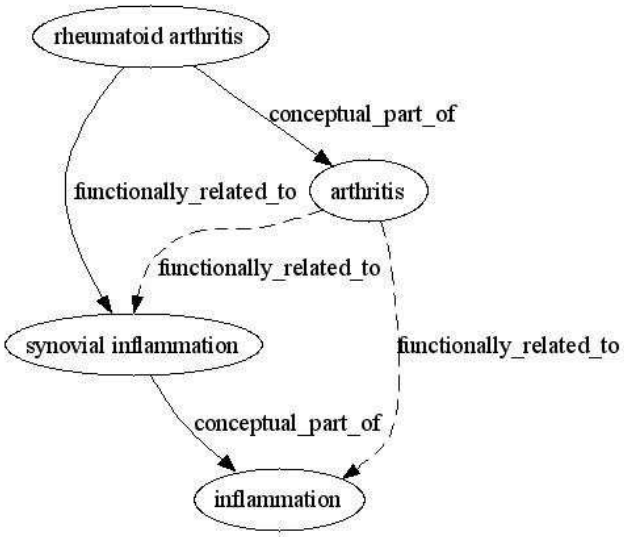


Figure 2: In this graph the reader can see that the head-modifier relation between "synovial" and "inflammation" leads to a kind of "sub-type" relation between the head noun and the pre-modifying adjectives. And she can also see that a semantic relation can be established between the subject and its predication.

We consider such semantic relations as a possible starting point for the support of text-based (supervised) ontology building. In our experiment the extracted generic semantic rules are mapped onto UMLS[2] relations, on the base of the lexical semantic properties of selected words belonging to distinct POS. The lexical semantic information is taken from both UMLS and EuroWordNet (for English and German). We used for example 24 different classes of adjectives that introduce different relations,

On the base of a first (very limited) evaluation (see Declerck & Vela, 2005), it seems to be that we can claim that linguistic dependencies might really offer appropriate means for extracting shallow semantic relation for supporting the semi-automatic building of ontologies, and that NLP tools when applied to a specialised set of documents can really help in automatizing the procedure of knowledge extraction from text.

3. The basic underlying linguistic phenomena

In the following we list the linguistic phenomena we have been dealing with, with concrete examples, and show the kind of semantic relation we could associate automatically with them, using relation names as stated in UMLS.

[2] UMLS (Unified Medical Language System, see http://www.nlm.nih.gov/research/umls/) is a set of semantic resources for the medical domain, containing a Metathesaurus, a Semantic Network, a specialised lexicon and lexical programs. See. The mapping to UMLS relation names will allow future extensive comparisons of our work with the UMLS resources and documents manually indexed/annotated with those resources.

3.1. Apposition and Paranthesis (1)

"The effects of rheumatoid arthritis on bone include structural joint damage (erosions) and osteoporosis"

Linguistic Structure: [[The effects of rheumatoid arthritis] [on bone]] [include] [[structural *joint* damage (*erosions*)] [and] [osteoporosis]]

=> The Apposition (2 syntactic heads "joint" and "erosions" in one NP) including a parenthesis construction suggests a synonymy relation or a definition. Heuristic: Establishing Semantic Relations on the top of linguistic "head-modifiers" constructions

3.2. Apposition with Paranthesis (2)

"For symptoms of rheumatoid arthritis (pain, joint stiffness), the reference treatment is a nonsteroidal antiinflammatory drug (NSAID) such as diclofenac or ibuprofen."

Linguistic Structure: [For *symptoms* of rheumatoid arthritis (*pain , joint stiffness*)] , [the reference treatment] [is] [a nonsteroidal antiinflammatory drug (NSAID)]

⇒Suggesting a semantic relation between („pain" and „joint stiffness")

⇒Classify „pain" and „joint stiffness" as symptom of RA. The word „symptom" is linguistically annotated as the head of the Compl-NP of the PP starting with „For".

3.3. Apposition with commas

"Etoricoxib, a selective COX2 inhibitor, has been shown to be as effective as non-selective non-steroidal anti-inflammatory drugs in the management of chronic pain in rheumatoid arthritis and osteoarthritis, …"

Linguistic Structure: [Etoricoxib, a selective COX2 inhibitor,] [has been shown]…

Similar hypothesis as in the former examples: a semantic relation between "Etoricoxib" and "selective COX2 inhibitor". Probably a "isa" relation

3.4. Phrase Internal Coordination (1)

"The effects of rheumatoid arthritis on bone include structural joint damage (erosions) and structural joint damage "

Linguistic Structure: [[The effects of rheumatoid arthritis] [on bone]] [include] [[structural *joint* damage (*erosions*)] [and] [osteoporosis]]

⇒RA causes structural joint damage AND structural joint damage (interpreting the head noun "effects" as a causation).

⇒Hypothesis: The two heads of an NP coordination are somehow related.

3.5. Phrase Internal Coordination (2)

"A study was conducted to determine the incidence of ulnar and peripheral neuropathy "

Linguistic Structure:

… [The incidence of [[ulnar and peripheral] neuropathy]]

⇒ The AP "ulnar and peripheral" AP modifies the head noun "neuropathy". The AP is a coordinated one, having two Adjectival heads.

⇒ Hypothesis: They correspond to two types of neuropathy

3.6. Nominal phrases with (exact) one modifier

"chronic inflammation"

=> in dependency of the semantic class of the adjective, introduces a sub-type relation (here on the base of the lexical semantic information encoded in UMLS.

3.7. Nominal phrases with multiple pre-modifying adjectives

"chronic synovial inflammation" vs. "severe, destructive, juvenile rheumatoid arthritis"

- ⇨ the non enumerative sequence of adjectives suggest that "chronic" is a sub-type of "synovial inflammation"
- ⇨ the enumerative sequence of adjectives suggests that each adjective is introducing a sub-type relation to the head noun, and only to this.

3.8. PP postmodifcation of a head-noun

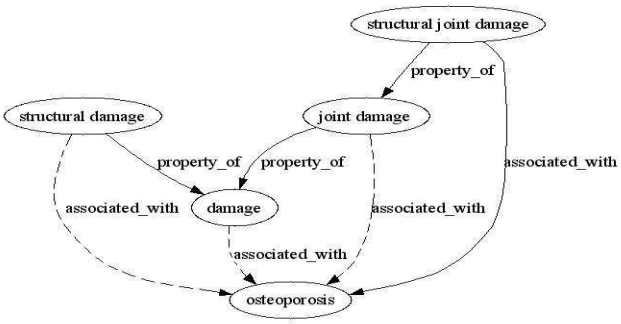
An example is given above in figure 1, together with the graph displayed in figure 2.

3.9. Phrasal Coordination

[structural joint damage] and [osteoporosis]]

- ⇨ suggests an association between "structural joint damage" and "osteoporosis". But still unclear if the relation holds between the whole NP or only between the head nouns. Further investigation is needed here.

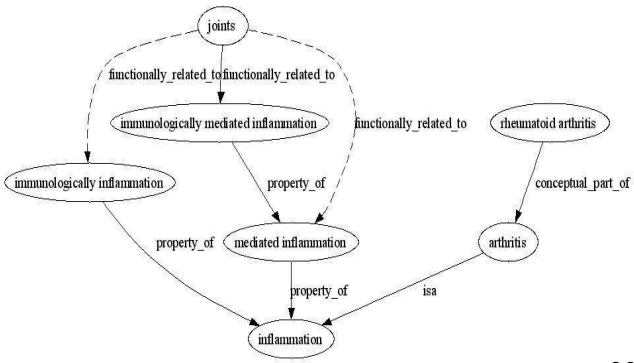
We display the graph below:



3.10. Predication

[Rheumatoid arthritis] [is] [an immunologically mediated inflammation of joints of unknown aetiology] [and] [often] [leads] [to disability]

- ⇨ introduces a "isa" relation between RA and "inflammation", with the additional qualification of "immunologically mediated", as the graph below shows:



3.11. Other verbs (causation)

Intransitive vs. transitive verbs like in "[Rheumatoid arthritis] [often] [leads] [to disability]" vs. "[Rheumatoid arthritis] [causes] [pain]"

=> in both cases the linguistic analysis infers a cause relation, having the information that in one case the verb "leads" need a PP initiated by "to". See the graph from the former section to see the output of the system applied to "lead to".

4. Conclusions and further Work

On the base of a first (very limited) evaluation (see Declerck & Vela, 2005), it seems to be that we can claim that linguistic dependencies might really offer appropriate means for extracting shallow semantic relation for supporting on building, and that NLP tools when applied to a specialised set of documents can really help in automatizing the procedure of knowledge extraction from text. In future studies we will extend our approach for literature-based scientific discovery from text. We will also have to look for a standardisation of the linguistic output of SCHUG for ensuring large-scale experiments on all kind of know extraction from text.

We will also investigate the use of such semantic extraction methods for supporting semantic based text summarization. In fact, we noticed that our incremental graph building from a corpus dedicated to symptoms of RA allow a very compact representation of the knowledge deposited in various scientific texts. One can see the global graph resulting from the analysis of ten sentences, as displayed in the next page.

5. References

Baud R. , Patrick Ruch: The Future of Natural Language Processing for Biomedical Applications. International Journal of Medical Informatics on NLPBA 67(1-3) (2002) 75-83

Benslimane D., Ahmed Arara, Kokou Yetongnon, Faiez Gargouri, Hanene Ben Abdallah: Two Approaches for Ontologies Building: From-scratch and From Existing Data Sources. Proceedings of The 2003 International Conference on Information Systems and Engineering (ISE) (2003)

Bodenreider O., Serguei V. Pakhomov: Exploring Adjectival Modification in Biomedical Discourse Across Two Genres. Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine (2003) 105-112

Bodenreider O., Thomas Rindflesch, Anita Burgun: Unsupervised, corpus-based method for extending a biomedical terminology. Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain (2002) 53-60

Bowers N.: Graph-ReadWrite-2.00 2005

Declerck T. (2002). A set of tools for integrating linguistic and non-linguistic information. *In Proceedings of SAAKM (ECAI Workshop)*, 2002.

Declerck T. and Vela M. (2005). Linguistic Dependencies as a Basis for the Extraction of Semantic Relations. *In Proceedings of the Workshop on Biomedical Ontologies and Text Processing held at ECCB*, 2005.

Alexander Maedche and Steffen Staab (2000). Mining Ontologies from Text. In: *R.Dieng & O Corby. EKAW-2000 - 12th International Conference on Knowledge Engineering and Knowledge Management.* October 2-6, 2000, Juan-les-Pins, France. LNAI, Springer..

Esperonto (http://www.esperonto.net/), ended 2005.

Ontoweb (http://www.ontoweb.org/) and more specially the Ontoweb SIG-5 on Language Technology in Ontology Development and Use (http://ontoweb-lt.dfki.de/), ended 2004

Sekt (http://www.sekt-project.org/) ending in 2006.

Workshop (2003) Workshop on Human Language Technology for the Semantic Web and Web Services, (http://gate.ac.uk/conferences/iswc2003/), or the SemAnnot Workshop series (http://km.aifb.uni-karlsruhe.de/ws/semannot2005)

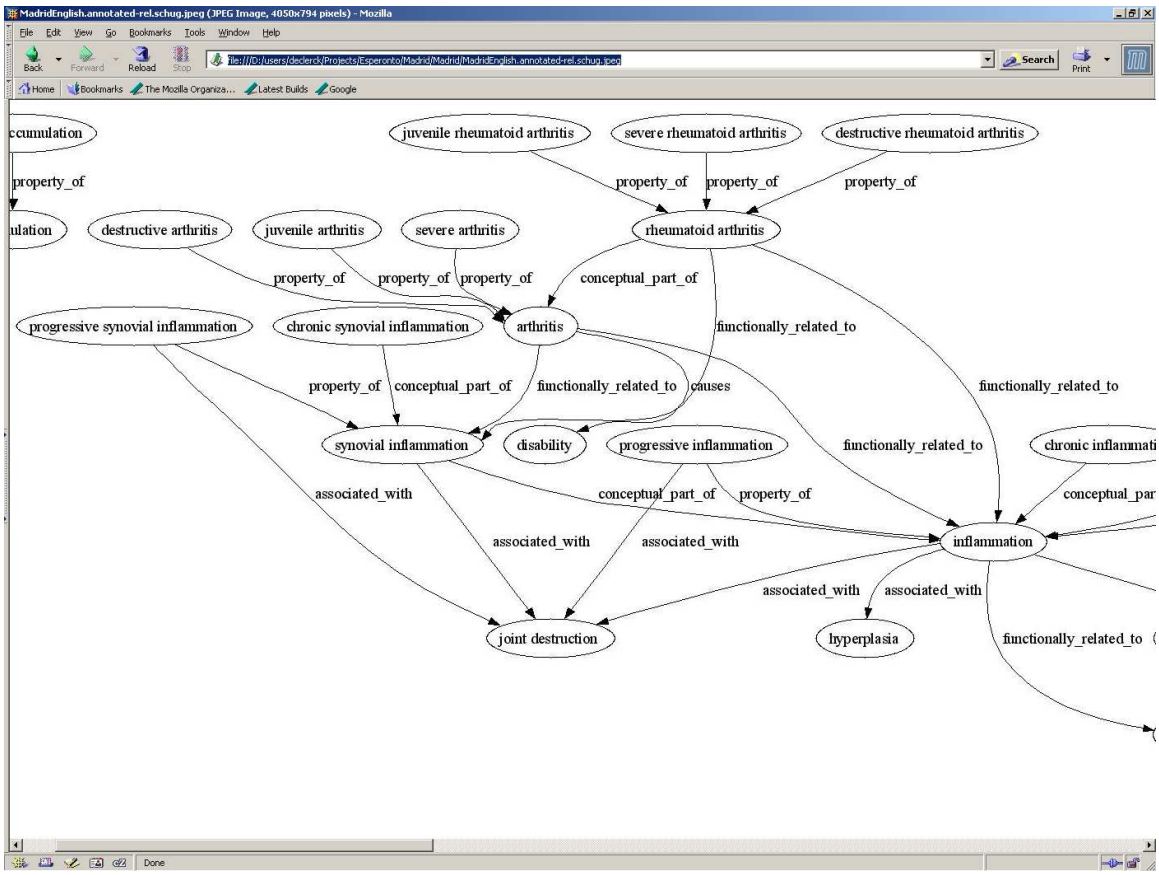


Figure 3: Screenshot of a subpart of the whole graph generated from the SCHUG tools applied to 10 sentences from a corpus on symptoms of RA