A Methodology for Ontology Learning: Deriving Ontology Schema Components from Unstructured Text

Mihaela Vela
Language Technology Lab, DFKI GmbH
Saarbrücken, Germany
email Mihaela.Vela@dfki.de

Thierry Declerck
Language Technology Lab, DFKI GmbH
Saarbrücken, Germany
email Thierry.Declerck@dfki.de

ABSTRACT

In this paper we present on-going work on the derivation of candidate components of ontology schema (so-called T-Box) from the shallow analysis of unstructured text. We discuss here examples dealing with German text in two domains: Economics and Radiology.

INTRODUCTION

In this short paper we briefly describe a rule-based methodology for both ontology extraction/learning and extension from the shallow analysis of unstructured text[1]. We prefer in fact the use of the term "derivation of T-Box components from unstructured text" for describing our investigation: The term "ontology learning" is very often understood in combination with the use of machine learning techniques, which are not considered in our experiment. The term Ontology Extraction is often linked to Ontology-Based Information Extraction (OBIE) and thus ontology population. What we have in mind here is the possibility of detecting from the text analysis candidate T-Box elements, supporting the semi-automatic building or extension of ontologies from scratch of from already existing terminologies.[2]

A MULTI-LAYERED APPROACH TO THE DETECTION OF T-BOX ELEMENTS

In one experiment we investigate a multi-layered text analysis approach for ontology extraction in the finance and economic domain. We define here three processing levels, within which candidate T-Box components can be derived: 1) The detection and analysis of compound words in text as a first base for suggesting candidate ontology classes and relations; 2) Detection and analysis of paraphrases of such compounds in text, in order to filter and validate the list of candidate classes and relations resulting from the first step and; 3) Analysis of syntactic patterns of the sentences containing the candidate ontology classes and relations, which are resulting from the two former processing steps.

A reason for implementing this multi-layered approach is: The first step requires only linguistic knowledge (what is a compound word?) but not the use of full natural language tools, a fact which saves computation time. First within the second step, some (shallow) natural language processing techniques are required, but the application of such techniques is restricted to the parts of the documents that contain paraphrases of compound words. More complex natural language processing techniques, like syntax parsing, are then applied in the third step only to sentences containing the compounds and/or their paraphrases. This way not the whole document is submitted to natural language processing, and this saves a substantial amount of processing time.

Detection and Analysis of Compound Words

In our experiment we are dealing with German texts and we are taking into account a specificity of this language: Its heavy use of compound words (but this aspect is also shared among other languages, like Dutch, Finnish, Hungarian, etc.). Compound words, in their productive use, consist in the merging of two or more lexical items, whereas the meaning of the whole compound can mostly be computed on the base of the meaning of the parts of the compound. The composition of the meaning is not always following the syntactic composition. So for example in German "Schweineschnitzel" (*Escalope from the Pork),* is a piece of meal from the pork, whereas a "Kinderschnitzel" (*a smaller piece of escalope for children)* is a small portion of meat.

A first intuition guiding our analysis of compounds is the fact that those words are good indicators for the expression of relations between entities expressed by the elements of the compound words, since the parts of the compounds can be considered as possible classes (or instances) of a potential ontology. The main condition is that both parts of the

[1] This work is done in the context of the European R&D MUSING project (See www.musing.eu for more details) and in the context of the THESEUS Program in the MEDICO Project, which is funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016. The responsibility for this publication lies with the authors.

[2] This being said, we will continue using the terms ontology extraction or ontology learning as short forms for the expression "Derivation of T-Box components from unstructured text".

compounds (limiting ourselves for the time being to the study of binary compound words) are nominal items, which are most of the time referential expressions.

For the detection of compounds we implemented a pattern-based approach and applied it to our corpus (a collection of texts from the German weekly newspaper "Wirtschaftswoche"). We first search for potential nominal items in the corpus (the pattern for German: a string starting with a capital letter between blanks or between a blank and a punctuation sign). We then search for larger strings (also starting with a capital letter and between blanks or between a blank and a punctuation sign), which contain the potential nominal items detected by the previous pattern search. The larger strings are considered to be compounds[3]. Since the detected compounds also start with a capital letter, we can assume that (most of) them are nominal items, such as Aktiengesellschaft (*stock company*), Bankensystem (*banking system*), Kursverfall (*slump in prices*), Notenbanken (*central banks*), Bankvertereter (*representative of a/the bank*), Datenbanken (*databases*), and can as such be considered as a natural language realization of a potential ontology class. As the (nominal) compounds are establishing a relationship between two nominal items, we can assume that they describe a relation between two potential ontology classes. We classify for the time being the relations expressed by compounds as being either of a structural type or expressing an object property. As the result of our very basic extraction of simple and local linguistic units[4], we suggest the rules in Figure 1 for deriving potential T-Box elements

compound[suggestedClass + suffix]
➔ objectProperty(suggestedClass, suffix)
compound[prefix + suggestedClass]
➔ subClassOf(compound, suggestedClass)

Figure 1: Ontology extraction rules for the string-based ontology extraction

"suggestedClass" stays for the nominal item we detect in the first pattern search. Both the prefix and suffix are the additional string context of the nominal item within the compound. Prefix and suffix are often nominal items - and thus potential classes - but this is not necessarily the case. We can have combination of adjectives and nouns, etc.

The first rule in Figure 1 states that between a possible class and its nominal suffix in the compound we may have an objectProperty-relation[5] and for the compound "Bankvertreter" we can derive the realtion: objectProperty(Bank, Vertreter). With the second rule we can derive from the same compound the relation: subClassOf(Bankvertreter, Vertreter).

Obviously, the (naive) processing strategy presented above is over-generating. So for example in the case of Aktiengesellschaft, we can correctly derive subClassOf(Aktiengesellschaft, Gesellschaft), but we would also incorrectly derive objectProperty(Aktien, Gesellschaft)[6]. We need here to formulate semantic constraints on the domain and range of the possible relations. It is not enough to have the fact that both parts of the compound are nominal items. We need to ensure for example that for the objectProperty the suggestedClass is denoting a human (or a living entity) or an institution and that for the subClassOf-relation the suggestedClass is denoting for example an institution (to be implemented and verified).

Searching for Paraphrases of Compounds in the Corpus

In the second processing step we look for paraphrases of the compounds in the corpus, since this helps in validating the role of the compounds for the extraction of classes and relations and allows to precise the type of relation marked by the compounds. The patterns for this are described in Figure 3. Table 1 further below lists ten types of paraphrase patterns as they were extracted from the corpus.

concept + [at most three words] + suffix
prefix + [at most three words] + concept

Figure 3: Patterns for finding paraphrases of compounds

We can observe the paraphrases of the compounds detected by the patterns can be divided into two categories: one for which the meaning of the paraphrase corresponds to the meaning of the original compound and one for which this is not the case. The decision whether a paraphrase does semantically correspond to the original compound (i.e. is valid for our task) is for the time being to be taken by an ontology engineer who can tell whether relevant ontology information can be extracted from the paraphrase. We are working on adding part-of-speech information and lexical semantics to the words occurring in the defined search windows in order to solve this task automatically as well. This for sure requires the use of language processing tools, but since the range of application of such tools is by now limited to the found sentences containing the paraphrase, the

[3] We do for sure filter out all possible flectional endings. And with this simple strategy, we do not detect all possible compounds in German, since certain words change their surface realization when integrated in compounds.

[4] By this we mean that there is at this stage no textual and linguistic context involved for the interpretation of the compounds.

[5] We could be more precise here and specify that the objectProperty is in fact a has-relation, but this is still too premature.

[6] In English: subClassOf(stock company, company) and objectProperty(stocks, company)

gain of accuracy is not seriously hampered by the decrease of performance due to the use of linguistic tools.

Table 1: Validation of compounds by reformulating the compounds

Compound	Paraphrase of compound
Bankexperten	Experten *der* Bank
Expertenschaetzungen	Schaetzungen *von* Experten
Buerofachmesse	Fachmesse *fuer* Buero
Westloehne	Loehne *im* Westen
Auslaenderhass	Hasses *gegen* Auslaender
Partnersuche	Suche *nach* einem neuen Partner
Designchef	Chef *ueber* deutsches Design
Einkommensteuerveranlagung	Veranlagung *zur* Einkommensteuer
Teilverkauf	Verkauf *zu* drei gleichen Teilen
Umweltvetraeglichekeit	Vertraeglichkeit *mit* der natuerlichen Umwelt

In Table 1, the reader can see how the compounds are split in different parts, and how those parts are linked to each other either by a determiner, indicating mostly a possession or a part-of relation, or by a preposition. The semantic interpretation of the preposition is also giving some hint on how to interpret the relation existing between parts of the original compound.

By using information about Part-of-Speech (PoS) and lexical semantics we can propose a refinement of the objectProperty and the subClassOf relation suggested by the first processing step along the line of the two basic types of paraphrases. The first type is the paraphrase in genitive case, which introduces a has-relation between the concept and the affix. The extended rule is depicted in Figure 5:

```
suggestedClass + art[genitive] + modifier? + suffix
➔ hasSuffix(suggestedClass, suffix)
prefix + art[genitive] + modifier? + suggestedClass
➔ hasSuggestedClass(prefix, suggestedClass)
```

Figure 5: Rule for genitive paraphrase of compounds

The second type of paraphrase pattern found concerns the reformulations with prepositions occurring between the two parts of the original compound. In this case the generic objectProperty is replaced by a new relation reflecting the semantics of the preposition in the paraphrase. Figure 6 contains the generic rules for this kind of reformulations.

```
suggestedClass +
prep[von|fuer|in|gegen|nach|ueber|zu|mit]
+ modifier? + suffix
➔ prepRelation(suggestedClass, suffix)
prefix + prep[von|fuer|in|gegen|nach|ueber|zu|mit]
+ modifier? + suggestedClass
➔ prepRelation(prefix, suggestedClass)
```

Figure 6: Rule-pattern for deriving classes and relations from reformulations of compounds using prepositions as links between the original segments of the compounds.

Phrase Structure and Syntactic Information

In the paraphrases we described in the former section, we can then still extract more relevant information for suggested T-Box elements. This is valid for the type of semantic relation that can be extracted from the structure modifier-nominal head, such as “jährliche Bilanz” (*annual balance*) that can appear in a paraphrase: Here we can extract the information that the class “balance” has a periodic time associated with it. In order to be able to achieve this result, we need to consider beyond phrase structure information (“jährliche Bilanz” is a nominal phrase, or NP) also a lexical semantic point of view. We apply for example to adjectives the semantic classification by Lee (1994) and to adverbs the classification by Lobeck (2000).

For the time being we identify seven linguistic phenomena on which the heuristics for semantic relation extraction can be applied. One example is shown in Figure 7.

```
Premodification
np[np_spec? np_mod np_head]:
if np_mod(introduces some_rel)
==> np[np_modn np_head] relation_introduced_by_np_mod [np_head]
```

Figure 7: Ontology derivation rule pattern for pre-modification of nominal heads

This rule is for phrases with one pre-modifier. Depending on the class of the modifier, a specific semantic relation is introduced. The presence of the determiner (NP SPEC) in the NP is optional, but the occurrence of exactly one modifier (NP MOD) and of the head (NP HEAD) is obligatory. In the example depicted in Figure 7, the phrase *deutsche Tochterfirmen* introduces the relation hasNationality(Tochterfirmen, Deutsch) according to the classification of “deutsch” as an adjective related to nationality or origin.

For reason of space we do not list here all the examples of the identified linguistic phenomena, as they also work using a similar heuristic. But to close this section, we would like to give some statistics about the corpus we are using, and the related analysis steps we described above. The corpus consists of 200107 tokens. If in the beginning we had 19767 potential concepts to be used, in the compound selection process we had only 3088 relevant concepts, which make 15.6% from the initial number of concepts.

The detection of paraphrases for the compounds allows to reduce the set of candidate concepts to 206 (1% from the initial number of potential nominal items, and 6.6% from the number of candidate concepts being part of a compound), which we can consider as being now serious candidates. From 17704 compounds found in the corpus only

284 have indeed paraphrases. So we can conclude that in our entire corpus we found 206 concepts which appear as part of a compound and the compound has a reformulation.

.

POSSIBLE EVALUATION STRATEGIES

Concerning the evaluation we intend to apply two approaches. First, based on the chi-square calculus, we intend to measure to what extent the extracted triples are relevant for the finance domain. The second evaluation concerns the comparison of the ontology constructed on the base of the rules presented here with the manual built ontologies in MUSING.

AN ADDITIONAL SCENARIO: ANALYSIS OF RADIOLOGY REPORTS

In the context of the Medico project we started a similar experiment. The documents here are radiology reports in German language. Those documents are very special and do not contain for example verbs and real sentences. Most of the text consists in "nominal phrases" in telegraphic style, with a lot of abbreviations. And the lexical category mostly represented is the noun. This is thus a very good type of documents for testing our shallow approach. But for the time being we do not have at our disposal a large corpus, and so we can not yet apply the processing step consisting in searching for paraphrases of the compounds we can detect in the reports.

But since we have at our disposal a first version of a terminology for Radiology (RadLex) in German, which is also integrated in a beta version of an ontology (See more information about RadLex at http://www.radlex.org/), we decided to start our work with the terms, as they are given in the ontology. We then apply the same step as described in the section "Detection and Analysis of Compound Words" above. This allows for example to detect "Lebervene", the combination of "Leber" (*Liver*) and "Vene" (*Vein*) in the reports, which is not listed in the terminology. Interesting is but that both terms are occurring in the terminology. What we can not do yet is to specify automatically the type of relation between the two terms, since our corpus is not large enough for allowing the search of paraphrase. Searching the Web for possible paraphrases shows us that there are a not significant numbers of paraphrases and that "Lebervene" is probably not a productive compound but rather a term (or class) per se, thus not expressing any relation.

We also got the other way round: We wrote patterns that allow finding in the reports parts of what are probably compounds in the terminology. So for example: "Gallenblase" (*gallbladder*) is in the terminology and is a class in the ontology, but not "Blasé". So we suggest this term to the specialist (but there might be very good reasons for omitting this term in the ontology for Radiology). At the same time we can detect then in the report "Gallenstauug" (*gallcongestion?),* which is not in the terminology. From the ending of the compound in "ung", we know that we deal with a nominalization, and thus that the compound can not be classified as an anatomic concept.

With the help of domain specialists, our work is being currently evaluated, and we will soon get information about the relevance of our work for the terminology and ontology building. It will be interesting to know if it is worth to include all possible compounds in the terminology/ontology or to rather go for a mix of basic classes and association rules that correspond to what the compounds are expressing.

The next step will also consist in analysing the head modifier structures in the "nominal phrases" of the reports. But this will require a preliminary adaptation of our NP detector, since the authors of the reports very often use post nominal modification in term of predicative use of adjectives.

CONCLUSION

We have described on-going work on extracting T-Box elements of ontologies on the base of a multi-layered linguistic approach, taking into account as well performance issues. While an evaluation has till to be performed, comparing in the financial use our suggested T-Box schema to existing ontologies, In the Radiology use case, the evaluation is being done on the base of relevance appreciation delivered by domain experts.

REFERENCES

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In Proceedings of 14th International Conference on Computational Linguistics (COLING-92), pages 539{545, Nantes, France, 2002.

Sun-Muk Lee. Untersuchungen zur Valenz des Adjektivs in der deutschen Gegenwartssprache. Lang, 1994. URL http://www.sfs.uni-tuebingen.de/ GermaNet/.

Anne Lobeck. Discovering Grammar: An Introduction to English Sentence Structure. Oxford University Press, 2000. URL: http://www.ac.wwu.edu/annelob/TESOL402assignment3.htm.

RadLex: http://www.radlex.org/