Enhancing Semantic Annotation through Coreference Chaining: An Ontology-based Approach

Till Christopher Lech
CognIT as, Oslo
till.christopher.lech@cognit.no

Koenraad de Smedt
University of Bergen
desmedt@uib.no

Abstract. Semantic annotation of natural language text requires a certain degree of understanding of the document in question. Especially the resolution of unclear reference is a major challenge when detecting relevant information units in a document. The ongoing KunDoc project examines how domain specific ontologies can support the task of Coreference chaining in order to enhance applications such as automatic annotation, information extraction or automatic summarization. In this paper, we present a robust methodology for acquisition of semantic contexts that does not depend on a thorough syntactic parsing as necessary tools often are unavailable for "smaller" languages. Based on a shallow corpus-analysis, verb-subject relations constitute the framework for the extraction of semantic contexts. Our approach either adds the semantic contexts to concepts and instances in an existing ontology or builds up the domain knowledge necessary for coreference chaining from scratch.

Introduction

Automatic semantic annotation of natural language text such as web documents requires a certain degree of text understanding. An important task in order to constitute a coherent semantic representation of a document is the resolution of anaphoric expressions and coreference chains. In the current NLP landscape there are numerous approaches for anaphora resolution based on either heuristics, such as (Mitkov 1998) and (Stuckardt 2000), or statistics, such as (Soon, Ng et al. 2001) or (Ng and Cardie 2003). Only few efforts have been made so far to explore background knowledge stored in ontologies in order to resolve unclear reference. The aim of the ongoing KunDoc Project[1] is to examine how ontologies can be acquired, enhanced and reused for detecting coreference chains in natural language text. In this paper, we will describe a methodology for learning and use of domain specific ontologies in order to support the coreference chaining task. We present a ontology-based methodology for use of semantic contexts for coreference chaining, followed the acquisition of these semantic contexts that can either be added to existing ontologies or constitute a starting point for ontology engineering, based on verb-subject relations extracted from a domain-specific text corpus.

[1] http://kundoc.net

The KunDoc Methodology

Most of the heuristics-based methods for Coreference chaining rely mainly on morpho-syntactic features such as number and gender agreement, syntactic function, topicalisation etc. Only few efforts have been made – besides the knowledge-based methods in the early days of AI, such as frames or scripts – in order to utilize semantic cues in order to resolve anaphoric expressions. In the KunDoc project, these semantic cues are retrieved from domain-specific ontologies. The semantic features used in KunDoc consist of:

- Class/subclass relations according to the taxonomy used (e.g. person, organization, etc.)
- Verbs or adjectives with which the concept in question frequently co-occurs

As a methodological framework for extraction of possible referents and their antecedents we used the CORPORUM system (Engels and Lech 2003), a toolkit for semantic analysis of natural language text, which extracts the most relevant concepts and proper nouns as well as associations between these concepts from text.

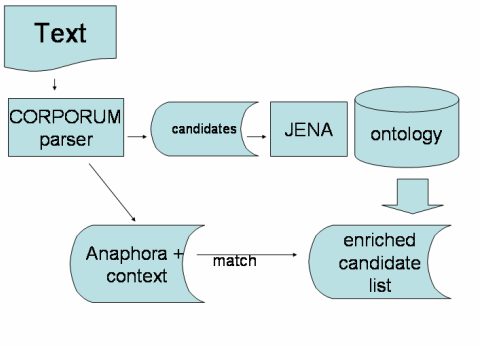

Fig. 1. The KunDoc Architecture

During the text analysis, possible antecedents are collected and stored in a candidate list. These candidates are enriched with their semantic features that are retrieved from the ontology using the JENA[2] interface. As soon potential anaphoric expressions are encountered (definite noun phrases or pronouns), their features are matched with the candidates in order to find the correct antecedent, as shown in Figure 1. Whereas class/subclass relations are inherent in existing taxonomies, the semantic contexts for coreference chaining may have to be added. The following section presents a methodology for generating these contexts.

[2] http://jena.sourceforge.net/

Acquisition of Semantic Contexts

The idea of deriving semantic classes from noun phrase/verb co-occurrences is not new in itself. Most of the work in this area is based on the distributional hypothesis, i.e. that nouns are similar to the extent that they share context. We assume that certain actions – denoted by verbs – are typically performed by a semantically restricted set of entities. Notable work in this area has been done by both Hindle (Hindle 1990), as well as by Nédellec and Faure (Faure and Nédellec 1998) Starting point for Hindle's approach is the pointwise mutual information of verb-object and verb-subject co-occurrences. In order to calculate a weighting for each verb-subject pair, Hindle (1) derives a score from the observed frequencies of verb-subject co-occurrences,

$$C_{subj}(n\ v) = \log_2 \frac{\frac{f(n\ v)}{N}}{\frac{f(n)}{N}\frac{f(v)}{N}} \quad (\mathbf{1})$$

where $f(n\ v)$ is the frequency of a noun n occurring as subject of verb v, and N is the total of all verb-subject pairs in the data set.

The extraction of explicit Predicate-Argument Structures requires somewhat accurate parses of the sentences in the corpus, which can be problematic, due to the availability of the necessary tools for various languages. Especially for the "smaller" languages, such as Norwegian, this is a well known obstacle for the development of tools for information extraction or annotation. As an alternative – and more robust – approach, a shallow parsing of the text was chosen by using the Oslo-Bergen Tagger (OBT), a PoS-Tagger developed within a cooperation between the Universities of Oslo and Bergen, Norway. The OBT consists of a morphologic tagger and a CG-based module for disambiguation of tags. The CG component gives all found options that cannot be excluded. This gives a fair recall, but low precision. As the analysis of the data set will show, some of these mistakes will be filtered out as noise, whereas others may obscure the results.

The data set was extracted from a corpus of newspaper articles about a murder case in the village of Førde, Norway. All 94 texts were published in the Norwegian online newspaper VG Nett (www.vg.no), yielding a total of 1619 subject-verb structures. In order to provide a basic benchmark for semantic classification all subjects were grouped manually into 6 conceptual classes:

- politi (police)
- offer (victim)
- etterforskning (investigation)
- spor (trace)
- pårørende (relatives)
- gjerningsmann (perpetrator)

We assume that subjects in most cases denote the agent of the action described by the respective verbs. Therefore, in our first experiments, a co-occurrence score is calculated for subject-verb pairs only. An alternative approach to Hindle's similarity measure – also based on the distributional hypothesis – is presented in (Cimiano, Staab et al. 2003). Inspired by this work, we test the *cosine* similarity. In our approach we compute the cosine similarity of the the co-occurrence weighting $C_{subj}(v, n)$ of the VSS, computed as mentioned above through the mutual information:

$$SIM = \frac{\left(\sum_{v \in A(n_1) \cap A(n_2)} C_{subj}(v\ n_1) * C_{subj}(v\ n_2)\right)}{\sqrt{\sum_{v \in A(n_1)} C_{subj}(v\ n_1)^2 * \sum_{v \in A(n_1)} C_{subj}(v\ n_2)^2}}$$

where for each subject *n*, *A* is the set of verbs *v* that share a subject-verb structures with *n*. The results for the concept "police" are depicted in Table 1.

politi	politi (police)	1
politi	etterforsker (detective)	0,280248
politi	lensmann (seargeant)	0,184824
politi	Politiet (police, definite form)	0,180558
politi	Broberg (person name)	0,174987
politi	tekniker (technician)	0,158986
politi	VG (name of newspaper)	0,157164
politi	mannskap (squad)	0,15334
politi	Fonn (person name)	0,151243
politi	vitne (witness)	0,150081
politi	Borlaug (person name)	0,149111
politi	Naustdal (person name)	0,147527

Table 1. Concepts similar to „police"

This approach seems promising as there is only little noise in the ten most similar subjects, such as *VG*. In addition to the semantic classification of the concepts in the corpus, based on the co-occurrence measure between verbs and subjects, we are able to establish probable semantic contexts that can be added to concepts and instances in the ontology as depicted in Table 2 for the three top terms in the "police" cluster:

etterforsker	fatte
	overse
	etterforske
lensmann	utdype
	avtale
	erfare
	antyde
Broberg	fastholde
	oppfordre
	bekrefte

Table 2. Semantic contexts for concepts in the "police" cluster.

Coreference Chaining

As (Eiken 2005) has shown, the choice between possible antecedents can be positively influenced by exploiting the similarity between the semantic context of a pronoun and its antecedent in terms of predicate-argument relations. For example, in (3) it is correctly predicted that the most likely antecedent for the pronoun *hun* is *vitne* (witness). Even without any further information, this is derived entirely from co-occurrence relationships in the corpus.

(3) Hun skal ha hørt rop.
(She is supposed to have heard cries.)

This analysis was extended by Eiken by a clustering of concepts, which implies that concepts no longer need to be matched perfectly, but the coreferent must be part of a concept group. In (4), for example the pronoun *hun* (she) is first linked to the concept *kvinne* (woman), which is not among the candidates. However, the correct concept *Slåtten*, which is among the candidates is clustered together with *kvinne* and can therefore be selected.

(4) Hun ble funnet omkommet.
(She was found dead.)

In this way, a certain fuzziness of the matching is achieved, which enhances the possibility of finding coreferents in a set of candidates. In the KunDoc project this analysis is again extended by using the relations in the extracted ontologies.

Conclusions and Further Work

We have presented a method for the enhancement of information extraction and semantic annotation through Coreference chaining. We have shown how a shallow and robust analysis of a domain specific text corpus yields Verb-Subject-Structures that can be exploited in order to extend domain-specific ontologies by adding semantic contexts to the concepts and instances in the taxonomy. The robustness of the proposed methodology is constituted by the fact that it does not require tools for extensive parsing of natural language, but only a part-of-speech tagger and rather simple statistical models. This will also ensure an easy transfer to other languages.

The next steps in the KunDoc project will be a thorough evaluation of precision and recall of the resolution of anaphoric expressions and an evaluation on how this improves the quality of information extraction. Future work in the KunDoc project will look into the feasibility of the methodology for other knowledge domains and text genres.

Acknowledgements

The KunDoc project is a co-operation between CognIT a.s and the University of Bergen, supported by the Research Council of Norway, within the KUNSTI framework.

References

Cimiano, P., S. Staab, et al. (2003). Deriving Concept Hierarchies from Text by Smooth Formal Concept Analysis. GI Workshop "Lehren - Lernen - Wissen - Adaptivität" (LLWA). Karlsruhe, Germany.

Eiken, U. (2005). Corpus-based Semantic Categorisation for Anaphora Resolution. Bergen, University of Bergen. **M.A. Thesis**.

Engels, R. H. P. and T. C. Lech (2003). Generating Ontologies for the Semantic Web: OntoBuilder. Towards the Semantic Web. J. Davies, D. Fensel and F. v. Harmelen. The Atrium, Chichester, John Wiley & Sons, Ltd**:** 91-115.

Faure, D. and C. Nédellec (1998). A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition. LREC workshop on Adapting lexical and corpus resources to sublanguages and applications. Granada, Spain.

Hindle, D. (1990). Noun classification from predicate-argument structure. 28th annual meeting of the Association for Computational Linguistics. Pittsburg, Pa.**:** 268-275.

Mitkov, R. (1998). Robust Pronoun Resolution with Limited Knowledge. 17th International Conference on Computational Linguistics (COLING'98/ACL'98). Montreal, Canada**:** 969-875.

Ng, V. and C. Cardie (2003). Bootstrapping Coreference Classifiers with Multiple Machine Learning Algorithms. 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Soon, W. M., H. T. Ng, et al. (2001). "A Machine Learning Approach to Coreference Resolution of Noun Phrases." Computational Linguistics(27): 285-291.

Stuckardt, R. (2000). Robust Anaphor Resolution: Design and Evaluation of the ROSANA System. 1st workshop on RObust Methods in Analysis of Natural language Data. Lausanne.