Ontology Extraction for Coreference Chaining

Till Christopher Lech and Koenraad de Smedt

CognIT and University of Bergen

Abstract

The KunDoc project investigates coreference chaining with ontology-based methods. In this paper, we discuss knowledge-based methods for coreference chaining and in particular the use of ontologies and their acquisition from a corpus. We present the KunDoc methodology and its implementation. We use concepts and their interrelations extracted from a corpus of Norwegian newspaper articles to build up domain-specific ontologies which contribute with selectional restrictions on possible co-referents. We expect to see an improvement over methods that do not employ any semantic knowledge.

1   Introduction

The problem is sketched by Examples (1), where semantic and world knowledge are indispensible to resolve the pronoun *He* in either (1-b) or (1-c) after (1-a).

(1) a. The police officer was searching for the suspect.
     b. He had been investigating the murder since Tuesday.
     c. He had committed the second murder on Tuesday.

Methods using heuristic rules such as salience factors (Mitkov 1998, Lappin and Leass 1994) are limited in that they cannot resolve the differences illustrated in (1). Of course, not every coreference can be resolved locally, since predicate-argument combinations are not always exclusive even in a single domain, as illustrated by the possibilities in Examples (2). Clearly, our method is not meant to be used by itself, but to enhance other methods where they fall short.

(2) a. The police officer left the scene of the crime.
     b. The murderer left the scene of the crime.

CognIT has developed the CORPORUM system, a toolkit for semantic analysis of natural language text (Engels and Lech 2003). The CORPORUM Onto-Extract tool extracts the most relevant concepts and proper nouns as well as associations between these concepts from text. In the course of the KunDoc project, these tools are being extended with powerful ontology extraction mechanisms that serve an important purpose in improving the capacity to grasp discourse threads.

The current stage of our research is aimed at exploration. To that extent we have experimented using a limited corpus of Norwegian newspaper texts. We report on our development of methods and tools and the results of our experiments so far. In later research we hope to perform evaluations based on larger corpus studies.

2   Related Work

2.1   Coreference Chaining and Anaphora Resolution

The idea of using world knowledge in order to disambiguate natural language text has its roots in the early days of Natural Language Processing, when it was strongly tied to Artificial Intelligence techniques. *Scripts* have been used to map stereotypical situations and identify their typical participants (Schank and Abelson 1977). Also in a Schankian framework, semantic cues were exploited for anaphora resolution in *Preference Semantics* (Wilks 1975).

An exhaustive and historical overview of techniques for coreference chaining and anaphora resolution is provided elsewhere (Mitkov 2002). Here, a distinction is made between *traditional* and *alternative* approaches to anaphora resolution, where traditional methods make use of heuristics such as centering or focus, whereas alternative approaches compute the likely candidates based on statistics or AI models. Without dismissing the relevance of these heuristics, we would like to focus on *knowledge-based* models, where knowledge may be acquired through corpus analysis (Dagan and Itai 1990) or by using external sources such as Word-Net, sometimes in combination with machine learning techniques such as rule induction (Ng and Cardie 2002). Lexico-semantic knowledge proves useful here; however, WordNet does not exist for all languages, as is the case for Norwegian. An efficient option may consist of generating the necessary lexico-semantic resources from a corpus.

Approaches that rely on explicit domain or world knowledge have been criticised by several authors for being somewhat impractical, as this knowledge usually is hard to come by (Mitkov 2002). Meanwhile, the field of Knowledge Representation and Reasoning (KRR) has made great advances, providing both tools and methodologies for efficient storage and manipulation of knowledge bases as well as sound logical frameworks for reasoning and inference. In today's KRR landscape, ontologies have become widely used for a variety of knowledge-intensive purposes.

In the present context, an ontology can be defined as a specification of a conceptualisation in a given domain (Gruber 1993). The rising popularity of ontologies have brought knowledge-based methods back into the discourse. Markert provides a detailed discussion of knowledge sources for (nominal) anaphora resolution and concludes that ontologies may be useful, however as they are often designed in a rather task-specific manner, they do not necessarily support coreference chaining (Markert and Nissim 2005). On the other hand, ontologies can often be extended through corpus-based methods in order to provide the knowledge for anaphora resolution. Moreover, the initial results of the KunDoc project suggest that ontologies derived from domain specific text corpora need not be extremely explicit in order to support coreference chaining.

2.2 Ontology Extraction

The idea of deriving semantic classes from noun phrase/verb co-occurrences is based on the distributional hypothesis, i.e. that nouns are similar to the extent that they occur in similar contexts. We assume that certain actions or processes — denoted by verbs — typically involve a semantically restricted set of entities. One of the first significant attempts to exploit the distributional hypothesis describes a methodology for generating semantic classes based on predicate-argument structures (Hindle 1990). The starting point for Hindle's approach is the pointwise Mutual Information (MI) of verb–object and verb–subject co-occurences. MI, as shown in Equation 1, is a symmetric, non-negative measure of the common information in two variables (Manning and Schütze 1999).

$$I(x|y) = log_2 \frac{P(x|y)}{P(x)P(y)} \tag{1}$$

where $P(x|y)$ is the joint probability of events x and y, and $P(x)$ and $P(y)$ are the independent probabilities. In order to calculate a weighting for each verb-object co-occurence, Hindle derives a co-occurence score (2) from the observed frequencies,

$$C_{obj}(n|v) = log_2 \frac{\frac{f(n|v)}{N}}{\frac{f(n)}{N} \frac{f(v)}{N}} \tag{2}$$

where $f(n|v)$ is the frequency of a noun n occuring as object of verb v. A similar co-occurence weighting can be derived for the verbs and their subjects. Based on the verb-object co-occurence weighting, Hindle computes the similarity of objects for a certain verb (3):

$$SIM_{obj}(v_i n_j n_k) = \begin{cases} min(C_{obj}(v_i n_j)), (C_{obj}(v_i n_k)), \\ \quad \text{if } Cobj(v_i n_j) > 0 \wedge C_{obj}(v_i n_k) > 0 \\ |max(C_{obj}(v_i n_j)), (C_{obj}(v_i n_k)), \\ \quad \text{if } Cobj(v_i n_j) < 0 \wedge C_{obj}(v_i n_k) < 0 \\ 0, \text{otherwise} \end{cases} \tag{3}$$

Analogously, the similarity for subjects are computed. Subsequently, Hindle derives a measure for noun similarity that computes the sums of the respective subject and object similarity for a pair of nouns (4):

$$SIM(n_1, n_2) = \sum_{i=0}^{N} SIM_{subj}(v_i n_1 n_2) + SIM_{obj}(v_i n_1 n_2) \tag{4}$$

Although aimed at the generation of classes rather than taxonomies, Hindle's method provides a framework for the initial experiments described in the following paragraphs.

3   Acquisition of Predicate-Argument Structures

The extraction of Predicate-Argument Structures (PAS) requires somewhat accurate parses of the sentences in the corpus, for which a deep parser would be ideal. This approach has been tried for Norwegian with the LFG-based XLE parsing environment, together with the large NORGRAM grammar (Eiken 2005). However, the coverage of the rules and especially the lexicon in NORGRAM are in practice insufficient for parsing real texts. As an alternative—and more robust—approach, a more shallow parsing of the text was chosen by using the Oslo-Bergen Tagger (OBT) (Johannessen, Hagen, Haaland, Nøklestad, Jónsdottir, Kokkinakis, Meurer, Bick and Haltrup 2005).

The OBT is a PoS-Tagger developed within a cooperation between the Universities of Oslo and Bergen, Norway. The OBT consists of a pre-processor for tokenisation, sentence boundary detection as well as a morphologic tagger and a CG-based module for disambiguation of tags. The CG module delivers an annotation of sentence constituents such as subjects, objects or modifiers. The annotation of syntactic functions is by far not exact, as shown in Example (3) and its annotation.

(3)   Medelevene tente lys for Anne Slåtten under dagens minnestund.
         'Classmates lit candles for Anne Slåtten during today's obsequies.'

"Medelevene"	"medelev" subst appell mask be fl @obj @subj
"tente"	"tenne" verb pret tr1 tr11 pa5 tr15 @fv
"lys"	"lys" subst appell nøyt ub fl @obj @subj
	"lys" subst appell nøyt ub ent @obj @subj
"for"	"for" prep @adv
"Anne"	"Anne" subst prop fem @p-utfyll person
"Slåtten"	"Slåtten" subst prop @obj @subj
"under"	"under" prep @adv
"dagens"	"dag" subst appell mask be ent gen @det
"minnestund"	"minnestund" subst appell mask ub ent @p-utfyll
	"minnestund" subst appell fem ub ent @p-utfyll
"."	"$." clb punkt

In the annotations, we find both subject and object (@subj, @obj) tags for the sentence's subject *medelevene* (classmates). The situation is the same for the object of the sentence, *lys* (candles). As the analysis of he data set will show, some of these mistakes will be filtered out as noise, whereas others will obscure the results.

In order to extract verb-subject-structures (VSS) from the corpus, the texts were tagged by the OBT. Hereafter, the Spartan script (Velldal 2003) was used for the extraction of subject-verb-object structures from the annotated texts. The Spartan script takes the annotated texts as input and does a heuristic search in order to find verbs and their respective subjects and objects as well as modifiers

and prepositions for nouns. As the annotation of the syntactic functions is not always distinct, this will necessarily lead to mistakes if the object of a sentence is topicalised.

Two data sets were used in the same domain, namely newspaper articles on murder cases in Norway. The first 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 (http://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 twelve conceptual classes:

1. politi (police)

2. offer (victim)

3. etterforskning (investigation)

4. spor (trace)

5. pårørende (relatives)

6. gjerningsmann (perpetrator)

7. forbrytelse, sak (crime, case)

8. media (media)

9. personer (persons)

10. tid (time)

11. sted (places)

12. annet (others)

We assume that subjects in most cases denote the agents of the predicates described by the respective verbs. In our first experiments, a co-occurrence score is calculated for subject–verb pairs only, according to Hindle's method. These scores provide probable semantic contexts (VSS) for each of the concepts in the ontology as depicted in Table 1 for the three top concepts in the *politi* (police) cluster.

Furthermore, these co-occurence weightings provide the basis for a similarity matching between the extracted subjects. As an indicator, the most frequent subject, *politi* (police) with its 15 most similar subjects is presented in Table 2.

Whereas concepts like *etterforsker, lensmann, Fonn* (the latter being the name of the investigating sergeant) are conceivably similar to the *police* class, the concepts *drapsmann or gjerningsmann* are certainly not. Other classes show a comparable error rate. Some possible reasons for this poor performance can be the following:

Table 1: Probable semantic contexts for concepts in the *politi* (police) cluster

etterforsker	fatte	6.49
	overse	6.49
	etterforske	4.90
lensmann	utdype	7.18
	avtale	6.18
	erfare	6.18
	antyde	6.18
Broberg	fastholde	6.67
	oppfordre	5.67
	bekrefte	5.25

1. Size of the corpus

2. Lack of precision with regards to the extracted VSS

3. Omission of object similarity measure

4. Quality of similarity measure

In an attempt to verify if corpus size is relevant for class generation, we added a second corpus, consisting of 69 newspaper texts about another murder case in the village of Sogndal, Norway. The Sogndal case and thus its media coverage was similar in a number of ways: In both cases, the victim was a young female student, both cases happened in small villages in Western Norway, thus providing similar investigation backgrounds. To ensure utter domain uniformity with the Førde texts, the Sogndal articles were taken from the same newspaper (VG Nett). The Sogndal-corpus yielded another 1430 VSS, thus almost doubling the original data set. The new result for the concept politi is shown in Table 3.

Obviously, the results improve, with more police-related names and concepts in the list of subjects. However, there is still noise represented by e.g. *somalier* or *Hashin* (both referring to the suspect in the case).

In addition to the increased corpus, we tried another measure of similarity in order to verify the fourth of the identified reasons for the lack of performance. Inspired by Cimiano, we try the cosine similarity (Cimiano, Tane and Staab 2003). However, in contrast to this work, we do not compute the cosine similarity of the conditional probality of the VSS, but the weighting of the VSS, computed by (5), as mentioned above with respect to the MI.

$$SIM = \frac{\sum_{v \in A(n_1) \cap A(n_2)} C_{subj}(n_1|v) \cdot C_{subj}(n_2|v)}{\sqrt{\sum_{v \in A(n_1)} C_{subj}(n_1|v)^2 \cdot \sum_{v \in A(n_1)} C_{subj}(n_2|v)^2}} \qquad (5)$$

Table 2: Subjects similar to *politi* (police) from Førde corpus

politi (police)	247.47
etterforsker (detective)	37.67
vitne (witness)	31.09
lensmann (sergeant)	22.02
drapsmann (murderer)	21.43
kvinne (woman)	20.32
Fonn (person name)	19.34
mann (man)	19.00
person (person)	18.06
gjerningsmann (perpetrator)	13.28
etterforskning (investigation)	11.81
lensmannskontor (sergeant's office)	9.92
lapp (note)	9.90
drapsetterforsker (homicide investigator)	9.49
VG (name of newspaper)	9.48

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 4.

This approach seems more promising as there is only little noise in the twenty most similar subjects, such as *VG*, or the more generic concept *person*.

Finally, in Figure 1 we present an architectural overview of the ontology extraction mechanisms used in KunDoc.

4   Coreference Chaining

The domain-specific ontologies extracted by the mechanisms sketched in the previous sections can be used in coreference chaining as follows.

During the text analysis with CORPORUM, the text is tokenised and PoS-tagged. In addition, proper nouns such as person or place names are identified and put into a list of possible antecedents. In addition to the morpho-syntactic features, each candidate is then looked up in the ontology in order to get information on the following items:

- Class/subclass membership

- Properties (extracted from analysis of adjectives)

- Predicates

The choice between possible antecedents can be positively influenced by exploiting the similarity between the the semantic context of a pronoun and its antecedent in terms of predicate-argument relations derived from a deep syntactic

Table 3: Subjects similar to *politi* (police) from combined Førde/Sogndal corpus

politi (police)	355.35
etterforsker (detective)	54.30
person (person)	37.97
lensmann (sergeant)	32.98
vitne (witness)	31.55
mann (man)	30.59
Hashin (person name)	26.88
Politiet (police, definite form)	26.49
drapsmann (murderer)	25.62
Fonn (person name)	25.10
kvinne (woman)	24.34
VG (name of newspaper)	19.73
Broberg (person name)	18.98
somalier (Somalian)	18.91
gjerningsmann (perpetrator)	18.64
Kripos (crime division)	18.05

and semantic analysis of sentences (Eiken 2005). It is thus correctly predicted in Example (4) that the most likely antecedent for the pronoun *hun* (she) is *vitne* (witness), based on the semantic co-ocurrences in the corpus. More specifically, *vitne* is the most frequent first argument of the predicate *høre* (hear) when *rop* (cries) occurs as its second argument.

(4)   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. The pronoun *hun* (she) in Example (5) is first linked to the concept *kvinne* (woman), based on co-occurence with the predicate *funnet* (found). Although this concept is not among the candidates, the correct concept *Slåtten* (a woman's name), which is among the candidates, is clustered together with *kvinne* and can therefore be selected.

(5)   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 matching coreferents in a set of candidates.

    In the KunDoc project, Eiken's clustering extension is not used, but a similar extension is achieved by using the relations in the extracted ontologies.

    On the one hand, we use the extracted co-occurrence pairs in order to extend

Table 4: Subjects similar to *politi* (police), based on cosine similarity for the combined corpus

politi (police)	1.000
etterforsker (detective)	0.280
lensmann (seargeant)	0.185
Politiet (police, definite form)	0.181
Broberg (person name)	0.175
tekniker (technician)	0.159
VG (name of newspaper)	0.157
mannskap (squad)	0.153
Fonn (person name)	0.151
vitne (witness)	0.150
Borlaug (person name)	0.149
Naustdal (person name)	0.148
person (person)	0.146
etterforskning (investigation)	0.146
kriminaltekniker (crime technician)	0.143
tiltale (accusation)	0.142
dag (day)	0.141
etterforskningsledelse (investigation leaders)	0.140
drapsetterforsker (homicide detective)	0.139
Kripos (crime division)	0.139
tjenestemann (officer)	0.133

a manually constructed concept hierarchy of the Førde domain with predicates. This hierarchy was constructed and visualized with the Protégé ontology editor (http://protege.stanford.edu).

On the other hand, a relatively flat ontology is extracted, where each node in the concept hierarchy is associated with its prototypical predicates, based on co-occurrence. In Figure 2, the *police* branch of the Førde ontology is visualized by means of Formal Concept Analysis (FCA) (Ganter and Wille 1999) in the following way. An FCA-lattice represents a *context* consisting of objects and attributes. Objects that share all their attributes will be placed on identical nodes in the lattice; objects whose attributes are a subset of another objects will appear higher up on the same line in the lattice than the one with the superset of attributes.

Based on the 15 highest-scoring concepts of the *police* cluster (cf. Table 4), we create a context. The set of attributes consist of the top five co-occurring predicates for any concept in the context. For each concept, a predicate is applicable as an attribute if it is among the top five for some concept and at the same time among the 20 most frequently predicates co-occuring with the concept on hand.

According to the distributional hypothesis, we make the assumption that concepts with attributes that are subsets of other concepts' attributes will be subclasses
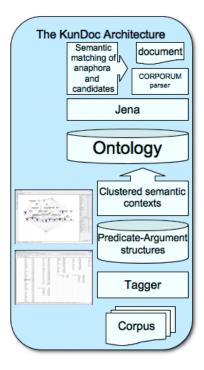
Figure 1: Architecture of Ontology Extractor

of these concepts, as their more restricted set of predicates indicate a higher degree of specificity, and thus they appear higher up in the lattice. As we see in Figure 2, this holds true for some concepts in the police cluster. Our extension of the set of possible candidates for pronoun reference is based on the inclusion of such subclasses.

Consider Example (6), where *Naustdal* is the name of a police officer and is indeed located as a subclass of the concept *politi* (police) in the FCA lattice in Figure 2. Even if there is no predicate-argument co-occurrence between *Naustdal* and the predicate *forklare* (explain), such a relation is still established since *forklare* is a predicate typical for *politi* as this subject–verb pair has a high co-occurrence value according to Hindle's method.

(6)   a.   Naustdal utelukker et selvmord.
              'Naustdal excludes the possibility of suicide.'
        b.   Han forklarte saken for pressen.
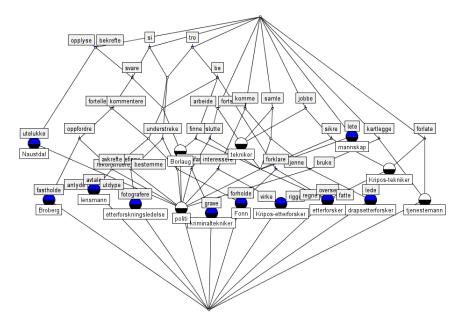   'He explained the case to the press.'

Figure 2: *Police* branch of the FCA analysis of the Førde domain

5   The KunDoc Demonstrator

The KunDoc demonstrator has been developed to illustrate the process of coreference chaining graphically. Figure 2 shows a screen shot of the demonstrator after analysing a text and marking the coreference chains in colours.

In the current version of the demonstrator, a limited version of the Mitkov algorithm is implemented in order to provide a baseline. Furthermore, the demonstrator currently includes the manually constructed Førde ontology, enhanced with the predicates based on the analysis described in the previous sections.

The user can choose to employ the background knowledge provided in the ontology or not. The demonstrator attempts to establish coreference chains for the personal pronouns *han* (he) and *hun* (she) and for parts of multi-token proper names, such as *Anne* or *Slåtten* as parts of *Anne Slåtten*. The result is displayed graphically by means of colored lines indicating the chains between words in the text.

6   Conclusion and future work

We have presented a methodology for coreference chaining that constitutes the starting point for the research in the KunDoc project. This methodology is based on background knowledge in the form of an ontology which is automatically ex-
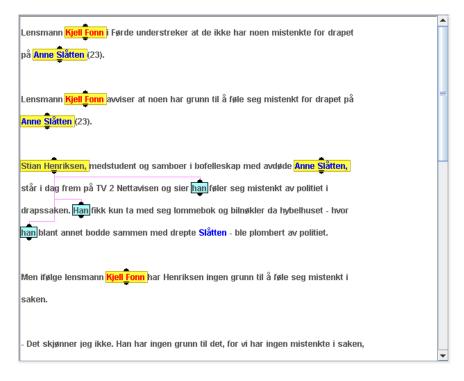
Figure 3: Screenshot of KunDoc Demonstrator

tracted from a domain-specific corpus of Norwegian texts. In the ongoing KunDoc project, progress has been made with respect to the ontology extraction and a first, incomplete version of a demonstrator. The current phase of the research is exploratory and results have not been extensively quantified. It is an open question in how far this ontology-based approach will work for other domains or independently from any domain. For the time being, we are therefore testing the methods only on limited domains for which we have manually constructed gold standards.

In the further course of the project, we plan to extend the ontology extraction to include not only subjects, but also verb–object relations as well as adjective–noun relations. The end goal is to integrate the use of the world knowledge in the demonstrator, test its performance relative to other methods and test its usability for real-life applications, such as Information Retrieval or Summarisation.

7   Acknowledgements

References

Cimiano, P., Tane, J. and Staab, S. (2003), Deriving concept hierarchies from text by smooth formal concept analysis, *Proceedings of GI Workshop Lernen-Wissen-Adaptivität (LLWA), Karlsruhe*, pp. 72–79.

Dagan, I. and Itai, A. (1990), Automatic processing of large corpora for the resolution of anaphora references, *Proceedings of the 13th conference on Computational linguistics*, Vol. 3.

Eiken, U. (2005), *Corpus-based semantic categorisation for anaphora resolution*, Master's thesis, University of Bergen.

Engels, R. and Lech, T. C. (2003), *Towards the Semantic Web*, Wiley, chapter 6 (Generating ontologies for the Semantic Web: OntoBuilder), pp. 91–115.

Ganter, B. and Wille, R. (1999), *Formal Concept Analysis: Mathematical Foundations*, Springer Verlag.

Gruber, T. (1993), A translation approach to portable ontologies, *Knowledge Acquisition* **5**(2), 199–220.

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

Johannessen, J. B., Hagen, K., Haaland, Å., Nøklestad, A., Jónsdottir, A. B., Kokkinakis, D., Meurer, P., Bick, E. and Haltrup, D. (2005), Named entity recognition for the mainland Scandinavian languages, *Literary and Linguistic Computing* **20**(1), 91–102.

Lappin, S. and Leass, H. J. (1994), An algorithm for pronominal anaphora resolution, *Computational Linguistics* **20**(4), 535–561.

Manning, C. and Schütze, H. (1999), *Foundations of statistical natural language processing*, MIT Press.

Markert, K. and Nissim, M. (2005), Comparing knowledge sources for nominal anaphora resolution, *Computational Linguistics* **31**(3), 367–402.

Mitkov, R. (1998), Robust pronoun resolution with limited knowledge, *Proceedings of the 18th International Conference on Computational Linguistics*.

Mitkov, R. (2002), *Anaphora Resolution*, Pearson Education, Edinburgh/London.

Ng, V. and Cardie, C. (2002), Combining sample selection and error-driven pruning for machine learning of coreference rules, *Conference on empirical methods in natural language processing (EMNLP)*.

Schank, R. and Abelson, R. P. (1977), *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*, Lawrence Erlbaum.

Velldal, E. (2003), *Modeling word senses with fuzzy clustering*, Master's thesis, University of Oslo.

Wilks, Y. (1975), Preference semantics, *in* E. L. Keenan (ed.), *The Formal Semantics of Natural Language*, Cambridge Univ. Press, Cambridge, pp. 329–350.