Daniela Goecke, Maik Stührenberg, Tonio Wandmacher

A hybrid approach to resolve nominal anaphora

In order to resolve nominal anaphora, especially definite description anaphora, various sources of information have to be taken into account. These range from morphosyntactic information to domain knowledge encoded in ontologies. As the acquisition of ontological knowledge is a time-consuming task, existing resources often model only a small set of information. This leads to a knowledge gap that has to be closed: We present a hybrid approach that combines several knowledge sources in order to resolve definite descriptions.[1]

1 Resolving nominal anaphora

1.1 Nominal anaphora and semantic knowledge

The term nominal anaphora comprises both pronominal anaphora as well as NP or definite description anaphora (DDA henceforth). In order to resolve DDA not only morphological or syntactic knowledge is needed but also information on (lexical) semantics and domain knowledge. A large amount of work in the domain of anaphora resolution has been done in the area of pronominal anaphora achieving good results (see Mitkov, 2002, for an overview); extensive work is still done in the area of resolving definite description anaphora (Vieira and Poesio, 2000; Ng and Cardie, 2002; Versley, 2007). Many of these approaches make use of information from pre-established lexical resources (WordNet or GermaNet), or try to acquire lexical knowledge by applying automated extraction methods on large corpora (or the web, cf. Markert et al., 2003; Versley, 2007). Other approaches rely on methods that determine semantic relatedness from cooccurrence information in corpora (cf. Poesio et al., 1998).

DDA relations hold between nominal discourse entities (or referents, cf. Karttunen, 1976)[2] and can be of various types: In example (1), the antecedent is explicitly mentioned and can be resolved via the same head noun (*direct anaphora* in terms of Vieira and Poesio, 2000). In the second example the antecedent is not explicit mentioned, however

[1]The work presented in this article is a joint effort of the projects A2 and C2 of the Research Group *Text-technological modelling of information* funded by the German Research Foundation. The corpus under investigation was developed by the projects A2 and C1.

[2]Discourse entities are constants within a discourse model evoked by NPs and which can be referred to in the subsequent discourse. NPs can either evoke new discourse entities in the discourse model or can "refer to ones that are already there" (Webber, 1988).

the anaphoric relation can be resolved on the basis of the hyperonymy relation between *questionnaire* and *form*. In order to resolve examples (3) and (4) additional semantic knowledge is needed. As opposed to examples (1) and (2) the anaphoric element and its antecedent do not refer to the same referent in the latter examples. Following the terminology of Clark (1977) we will refer to examples (3) to (4) using the term *bridging relations*.

(1) a questionnaire - the questionnaire

(2) a questionnaire - the form

(3) an interview - the questionnaire

(4) an interview - the respondent

Bridging relations occur when the antecedent is not explicitly mentioned in the text but has to be inferred from the context. Cues to solve bridging relations are domain knowledge (frames, scripts or schemata, e.g. interviews are often done using questionnaires) or (lexical-)semantic knowledge encoded in lexical nets like GermaNet or WordNet. The classification of these lexical nets ranges from "terminological ontology" (Sowa, 2000) to "full ontology" (Oard, 1997). We follow the terminology introduced by Erdmann (2001, p. 72) who uses the term *light weight ontology* to define ontologies that consist primarily of a representation schema to define taxonomies as well as attributes or relations. In contrast, *heavy weight ontologies* include complex logical descriptions that are specified in more expressive logical formalisms. However, using GermaNet alone as resource for detecting semantic relations is not sufficient considering the coverage in regard to the corpus under investigation. In order to close this gap we present a hybrid approach for automatically determining semantic relatedness in order to identify the most likely antecedent from a set of antecedent candidates.

1.2 Acquiring semantic knowledge

In the past few years a variety of approaches has been presented to automatise the extraction of ontological knowledge from structured and unstructured data. The output of these systems is usually rather rudimentary and noisy. Nevertheless, this kind of information coming from automated approaches can be considered as a valuable resource for our task. Regarding the current approaches to derive ontological knowledge from unstructured data two main classes can be made up:

The first one is based on distributional or structural similarity, starting from the assumption that words being semantically similar tend to occur in similar contexts and structural settings. In this family we find completely knowledge-free approaches relying on cooccurrence only (e.g. Paaß et al., 2004); Poesio et al. (1998) showed how this kind of

information can be used to resolve nominal anaphora (using the *HAL*-model; cf. Lund and Burgess, 1996).

The second class of methods basically relies on lexico-syntactic patterns, the so-called *Hearst* patterns (Hearst, 1992). Here, a text corpus is scanned for characteristic word combinations, typically containing a semantic relation between two terms (e.g. X *such as* Y, X being a hyperonym of Y). Recently, hybrid approaches can be found using both techniques to enhance the quality of the extractions. In Cimiano and Staab (2005), nouns are first clustered by cooccurrence methods and Hearst patterns are applied afterwards to extract the most useful relations. Cederberg and Widdows (2003) go the other way around: Based on patterns they extract word pairs from text and filter them by a cooccurrence based threshold being able to raise precision by 30%, compared to a standard pattern-based approach.

1.3 Objectives and organization of the article

The objective of our approach is to increase information on semantic relatedness of terms by a combination of – amongst others – extracted relations and cooccurrence information, and to use it in an anaphora resolution system. In general, the anaphora resolution process can be subdivided into three steps: (1) For each anaphoric element, determine an antecedent candidate list (ACL) and (2) apply constraints to exclude incompatible candidates from the ACL; (3) identify the most likely antecedent.

This paper concentrates on step 3, i.e. on the identification of the most likely antecedent candidate. We use a fixed search window to collect the candidate list and we do not apply constraints to downsize the list thus leading to forced test conditions for step 3. Ongoing work focuses on the implementation of a variable search window size as well as on the implementation of constraints for step 2. The remainder of the paper is structured as follows: Section 2 introduces the corpus under investigation as well as the annotation scheme and procedure, Section 3 describes the methods applied in our approach: GermaNet lookup, Hearst patterns, recency information and a semantic similarity measure, based on Latent Semantic Analysis (LSA). Finally, section 4 discusses the results of our approach, and Section 5 presents a conclusion and ongoing work.

2 The corpus under investigation

2.1 Annotation Scheme and Procedure

The evaluation of the approach described above is based on a corpus of German scientific and newspaper articles annotated for training and evaluation of an anaphora resolution system. The subset used for the evaluation presented in this paper includes three scientific articles and one newspaper article. For the purpose of anaphora resolution

the corpus has been annotated for discourse entities (DEs) and anaphoric relations between DEs. Several annotation schemes for annotating anaphoric relations have been developed in the last years, e.g. the UCREL anaphora annotation scheme (Fligelstone, 1992; Garside et al., 1997), the SGML-based MUC annotation scheme (Hirschmann, 1997), and the XML-based MATE/GNOME scheme for anaphoric annotation (Poesio, 2004), amongst others. The annotation scheme used for our approach is based on the one presented by Holler et al. (2004) and has been adapted for the annotation of bridging relations (Goecke et al., 2007). The versions of the annotation scheme are used within our research group both for the task of hypertextualization (project B1) as well as for the task of anaphora resolution (project A2). Therefore, the annotation of anaphora and coreference is distinguished explicitly: "Although anaphoric and coreferential relations can coincide, it is not generally the case that all coreferential relations are anaphoric, nor are all anaphoric relations coreferential" (Holler et al., 2004).[3] This distinction especially holds for bridging relations that can be inferred due to semantic role assignment (*a wedding - the bride*) or the meronymy relation (*a car - the wheel*): In these examples the anaphor and the antecedent do not refer to the same discourse entity even if an anaphoric relation holds between them. For cospecification and bridging relations two types of primary relations have been defined:

- *cospecLink*: Cospecification; anaphor and antecedent refer to the same referent;
- *bridgingLink*: Bridging; associative or indirect anaphora (Clark, 1977); anaphor and antecedent do not refer to the same referent.

For each of these relations a set of secondary relation types has been defined (see Table 1).

The corpus has been preprocessed using the dependency parser *Machinese Syntax*[4] which provides lemmatization, POS information, dependency structure, morphological information and grammatical function. Based on this information, discourse entities have been detected automatically by identifying nominal heads (i.e. nouns or pronouns) and their premodifiers. The anaphoric relations are annotated using the annotation tool *Serengeti* described in Stührenberg et al. (2007)[5]. The annotation procedure is subdivided into four steps: First, it is checked for each discourse entity (DE) whether it is used anaphorically. For each anaphoric DE the correct antecedent is identified, and for each anaphor/antecedent pair (AC pair henceforth) the primary relation is chosen. As the last step, the secondary relation is chosen. Listing 1 shows a sample annotation from a german linguistic article. In this example a bridging relation holds between the

[3] The MATE scheme states the distinction between anaphoric relations and reference proper, however the distinction is not made explicit in the annotation scheme; the term *coreference* is used to denote anaphoric annotation (Poesio, 2004).

[4] `http://www.connexor.eu/technology/machinese/machinesesyntax/`

[5] `http://coli.lili.uni-bielefeld.de/serengeti/annotator.pl`

cospecLink		
ident	pronouns same head noun of anaphor and antecedent	a man – he a man – the man
namedEntity	anaphor is an NP referring to a proper noun antecedent	Peter Jones – the man
propName	anaphor is a proper noun antecedent may be either an NP or a proper noun	the CTO – Peter Jones Peter Jones – Jones
synonym	synonymy holds between head nouns	a car – the automobile
paraphrase	anaphor is a paraphrase	the HTML-editor – the web site creation tool
hyperonym	anaphor is an hyperonym of the antecedent	a horse – the animal
hyponym	anaphor is an hyponym of the antecedent	an animal – the horse
addInfo	anaphor adds further information	Peter Jones – the 67 year old CTO
bridgingLink		
possession	possessive relation	Peter – his car
meronym	anaphor is part of the antecedent	a room – the window
holonym	anaphor has the antecedent as one of its parts	the window – the room
setMember	anaphor is an element of a set	two cars – the red car
hasMember	anaphor is a set consisting of its antecedents	Paul [...] Susan – the two children
bridging	associative link (e.g. role assignment, schema)	a wedding – the bride

Table 1: Secondary relation types for cospecLink and bridgingLink

discourse entities denoted by *die Befragung* ('interview'; lines 4-6) and *der Fragebogen* ('questionnaire'; lines 27-29).

2.2 Corpus Design

The evaluation set comprises a total amount of 4196 DEs. Based on these DEs, a total amount of 1433 *cospecLinks* and 541 *bridgingLinks* could be found. In our study we focus on those relation types between anaphor and antecedent that can be found in GermaNet: synonymy, hyperonymy, hyponymy, meronymy, holonymy, bridging. The subset that contains the semantic relations under investigation comprises a total amount of 224 anaphoric links. As distance between anaphor and antecedent is a crucial point, we defined a fixed distance for our evaluation. Especially for bridging relations in scientific articles, distances between anaphor and antecedent can be extremely large. For our corpus, distances up to hundred DEs could be found, therefore, not all of the relations have been taken into account. Corpus investigation shows that limiting the distance to 15 DEs results in a reasonable subset: 50% of the cospecLinks and 55.78% of the bridgingLinks find their antecedent within this window. Thus, for each anaphoric DE a candidate list of (at most) 15 possible antecedents has been created (including

the correct one that has been marked during the annotation process).[6] This leads to an evaluation set of 115 anaphoric DEs and 1428 antecedent candidates (app. 12,5 candidates per anaphor). For the corpus study presented here we have chosen this fixed window; however one has to include more sophisticated methods in order to find suitable sets of antecedent candidates in a complete anaphora resolution system due to varying distances between anaphor and antecedent. Modelling the search space for candidate sets that cover both anaphors with small distances as well as anaphors with long distances should not be grounded solely on the linear structure of text but should be flexible in size according to structural elements, e.g. on the basis of discourse structure (cf. Cristea et al., 2000; Chiarcos and Krasavina, 2005) or logical document structure (Goecke and Witt, 2006).

3 Method

Our approach makes use of four information sources and combines them into one measure. It is a forced choice algorithm, i.e. to any input pair of anaphor and antecedent candidate a score will be assigned. In the following we describe the four single methods separately, and then we show how we combine their information.

3.1 GermaNet relations

As we have already shown, many bridging phenomena are based on synonymy, hyponymy or meronymy. These relations are encoded in a lexical resource like GermaNet, making it our first source of information, since the information being found here are very reliable and noise-free, despite of their low coverage. For each AC pair the underlying lemmas are looked up in GermaNet and – if both are included – the distance between the corresponding nodes is computed (cf. Poesio et al., 2004, for node-node distance measures using WordNet). Nevertheless, distance information does not include information on the relation holding between two lemmas, this information has to be computed from the path information separately. In our study node-node distances have been computed using the implementation provided by the project A4 of our research group (cf. Mehler et al., 2007)[7]. The resulting distance values (in terms of path length) have been normalised for each set of AC pairs belonging to a given anaphor. A value of 1 indicates the shortest path within a given set and a value of 0 indicates either maximum length or the fact that one token of the AC pair is not found in GermaNet.

[6] Only non-pronominal DEs can serve as antecedents, thus the candidate list may be shorter than 15 elements.
[7] http://www.scientific-workplace.org/

```
<chs:chs>
<chs:text>
 <cnx:token ref="w2732">In</cnx:token>
 <chs:de deID="de764" deType="nom" headRef="w2734">
  <cnx:token ref="w2733">die</cnx:token><cnx:token ref="w2734">Befragung</cnx:token>
 </cnx:de>
 <cnx:token ref="w2735">wurden</cnx:token><cnx:token ref="w2736">nur</cnx:token>
 <chs:de deID="de765" deType="nom" headRef="w2738">
  <cnx:token ref="w2737">solche</cnx:token><cnx:token ref="w2738">Kurse</cnx:token>
 </cnx:de>
 <cnx:token ref="w2739">einbezogen</cnx:token><cnx:token ref="w2740">,</cnx:token>
 <chs:de deID="de766" deType="nom" headRef="w2741">
  <cnx:token ref="w2741">die</cnx:token>
 </cnx:de>
 <cnx:token ref="w2742">bereits</cnx:token><cnx:token ref="w2743">über</cnx:token>
 <chs:de deID="de767" deType="nom" headRef="w2745">
  <cnx:token ref="w2744">gute</cnx:token><cnx:token ref="w2745">Grundkenntnisse
  </cnx:token>
 </cnx:de>
 <cnx:token ref="w2746">in</cnx:token>
 <chs:de deID="de768" deType="nom" headRef="w2749">
  <cnx:token ref="w2747">der</cnx:token><cnx:token ref="w2748">deutschen</cnx:token>
  <cnx:token ref="w2749">Sprache</cnx:token>
 </cnx:de>
 <cnx:token ref="w2750">verfügten</cnx:token><cnx:token ref="w2754">,</cnx:token>
 <cnx:token ref="w2755">da</cnx:token>
 <chs:de deID="de770" deType="nom" headRef="w2757">
  <cnx:token ref="w2756">der</cnx:token><cnx:token ref="w2757">Fragebogen</cnx:token>
 </cnx:de>
 <cnx:token ref="w2758">nur</cnx:token><cnx:token ref="w2759">auf</cnx:token>
 <chs:de deID="de771" deType="nom" headRef="w2760">
  <cnx:token ref="w2760">Deutsch</cnx:token>
 </cnx:de>
 <cnx:token ref="w2761">vorlag</cnx:token><cnx:token ref="w2762">.</cnx:token>
</chs:text>
<chs:standoff>
 <chs:semRel>
  <chs:bridgingLink relType="bridging" phorIDRef="de770" antecedentIDRefs="de764"/>
 </chs:semRel>
 <cnx:token_ref id="w2757" head="w2761" pos="N" syn="@NH" lemma="frage#bogen"
  depV="subj" morph="MSC␣SG␣NOM"/>
 <cnx:token_ref id="w2734" head="w2735" pos="N" syn="@NH" lemma="befragung"
  depV="advl" morpho="FEM␣SG␣ACC"/>
</chs:standoff>
</chs:chs>
```

Listing 1: The annotation format for anaphoric relations. Shortened and manually revised output

3.2 Relation extraction by patterns

Our second information source relies on pattern-based information. We follow the approaches of Markert et al. (2003) and Versley (2007), who look up patterns on the web. We first generate patterns of the types "*X und andere Y*", "*X wie Y*", "*X insbesondere Y*", "*X einschließlich Y*" for all AC pairs of our text corpus and submit them as queries via the *Google* API. We then compute a normalized score from the added hit counts of each pattern.

3.3 Recency information

Since linear distance between an anaphor and a potential candidate also provides valuable information, we took a closer look at the distance distribution in our corpus. We determined the distance (in DEs) between each AC pair; the (standardized) distribution is shown in Figure 1 (columns). It can be seen that the most frequent distance between anaphor and antecedent is 5 DEs. We can assume that the distances are (roughly) normally distributed after this peak. However, assuming normal distribution with the same standard deviation σ beforehand would result in an overestimation of very short distances (1-4). For this reason we apply two different σs (σ_-, σ_+) in order to best adapt to this distribution. Equation 1 displays our recency function, the curve in Figure 1 shows the developing of the function for $x = 0 - 20$.

$$Rec(x) = e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2},\ with\ \mu = 4, \sigma_- = 1\ and\ \sigma_+ = 5. \tag{1}$$

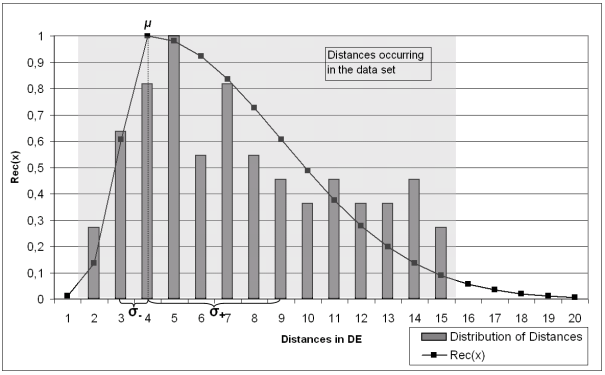

Figure 1: Graph of our recency function and distribution of distances (0-20 DEs)

3.4 LSA-based similarity

Since the early 1990s, Latent Semantic Analysis (LSA) has become a well-known technique in NLP. When it was first presented by Deerwester et al. (1990), it aimed mainly at improving the vector space model in information retrieval. Its abilities to enhance retrieval performance are remarkable; results could be improved by up to 30%, compared to a standard vector space technique (Dumais, 1995). Moreover, meaningful documents could be retrieved that did not share a single word with the query.

LSA is based on the vector space model from information retrieval (Salton and McGill, 1983). Here, a given corpus of text is first transformed into a term×context matrix A, displaying the occurrences of each word in each context. Usually, this matrix is then weighted by one of the standard weighting methods used in IR (cf. Salton and McGill, 1983). The decisive step in the LSA process is then a *singular value decomposition* (SVD) of the weighted matrix. Thereby the original matrix A is decomposed as follows:

$$SVD(A) = U\Sigma V^T \tag{2}$$

The matrices U and V consist of the eigenvectors of the columns and rows of A. Σ is a diagonal matrix, containing in descending order the singular values of A. By only keeping the k strongest (k usually being 100 to 300) eigenvectors of either U or V, a so-called semantic space can be constructed for the terms or the contexts, respectively. Each term or each context then corresponds to a vector of k dimensions, whose distance to others can be compared by a standard vector distance measure. In most LSA approaches the *cosine* measure is used.

We use a slightly different setting, close to the one described by Schütze (1998) and Cederberg and Widdows (2003), where the original matrix is not based on occurrences of terms in documents but on other cooccuring terms (term×term-matrix). We thus count the frequency with which a given term occurs with others in a predefined context window ($\pm 10 - 100$ words). After applying *singular value decomposition*, each word is represented as a vector of k dimensions, and for every word pair w_i, w_j of our vocabulary we can calculate a similarity value $Sim(w_i, w_j)$, based on the *cosine* between their respective vectors.

Treatment of compounds: As German compounds are lexicalized as a single graphical unit, they are often a tricky problem for NLP applications. Many algorithms rely at some point on string matching in order to identify lexical units in a given text; many compounds are not part of any predefined vocabulary, therefore they are neglected in further processing stages. Our LSA component, however, is able to deal with compound words, since we make the (somewhat simplifying) assumption that the meaning of a compound word is the sum of its parts. This idea is straightforward in a vectorial setting: Every time we encounter a compound which is not contained in the vocabulary, we split

it up into its parts (by partial matching) and take the vector sum of the corresponding vectors. This simple measure works surprisingly well, as can be seen in the Section 4.

3.5 Combining information sources

So far we have four information sources at hand, which can describe possible anaphoric relations: GermaNet, lexico-syntactic patterns, linear distance or recency information, and LSA similarity. We now have to combine this information into one measure in order to be able to calculate the most likely antecedent out of our candidate list. A well-known way to combine information from several sources is interpolation. We describe in the following how this can be done in our setting:

So, for a given anaphoric expression b and a set of candidates of antecedents $A = (a_1, ..., a_n)$,

1. we consult for each candidate $a_1, ..., a_n$ if a path to b can be found in GermaNet. We define a function $GN(a_i, b)$ whose values range from 0 to 1, according to the normalised path length;

2. we define a function $Pat(a_i, b)$ returning the normalized frequency score of matching candidate strings including a_i and b;

3. we determine the LSA-similarity $Sim(a_i, b)$ between a_i and b with respect to a previously calculated reference semantic space;

4. finally a recency function $Rec(a_i, b)$ determines the recency factor for the distance between a_i, b, as described in Formula 1.

Each candidate a_i then receives a score $Sc(a_i)$ by interpolating the results from the single functions defined above. The parameters λ_{GN}, λ_{Pat}, λ_{LSA}, λ_{Rec} will be set empirically. It is clear that advanced optimization techniques such as the EM algorithm could be employed here. However, since our test set is rather small, we could not assure to reach converged values, therefore we adjust the values manually.

$$Sc(a_i) = \lambda_{GN} \cdot GN(a_i, b) + \lambda_{Pat} \cdot Pat(a_i, b) + \lambda_{LSA} \cdot LSA(a_i, b) + \lambda_{Rec} \cdot Rec(a_i, b)$$

It is important to note that this function assigns a score to any pair of anaphor and antecedent. Apart from the maximum distance of 15 DEs we apply no further exclusion criteria, our algorithm is forced to make a choice among the candidates, according to their respective score, even though none of the semantic components might be able to assign a value (due to an unknown word in the pair). The choice is based on recency information only, which is necessarily rather unreliable.

4 Results

GermaNet Relations For 71% of the DE in our corpus the underlying lemma of the head noun is stored in GermaNet. For 759 out of 1428 AC pairs (53,15%) a path length could be computed.

Relations generated by patterns As described before we generated candidate strings comprising one out of 4 patterns and an AC pair each. We submitted each of the strings as a query to the *Google* API, and we summed up the total hit counts for each AC pair.[8] The summed up hit counts were logarithmized and normalized in order to have a meaningful score that can be used in the interpolation formula. As expected, most of the hit counts were 0, only for 119 out of 1428 AC pairs (8,3%) we could find at least one matching pattern.

LSA-based similarity factor Using the Infomap[9] toolkit, we calculated a term×term-cooccurrence matrix of 80.000×3.000 words over a corpus of 101 million token (from *Wikipedia* and *Tageszeitung*). This matrix was then reduced by *singular value decomposition* to 150 dimensions, giving us a vector for each of the 80.000 words. We now calculated for each of the 1428 AC pairs their LSA-similarity using the cosine distance of their respective vectors.

For compound words we calculated the normalized sum of the vectors of each component and used it instead of the word vector. This tremendously reduced blind spots in the calculation process: Only 94 out of the 1428 word pairs (6,5%) could not be assigned a similarity value, whereas this would have been the case for 910 pairs (63%) without compound treatment.

Recency function For each of the 1428 AC pairs, its recency factor was calculated, using the recency function in Formula 1 (see p. 50), with $\mu = 4$, $\sigma_- = 1$ and $\sigma_+ = 5$. We admit that, due to limited data resources, we could not estimate the parameters on a held out test set, however we would expect these parameters to be quite stable over different corpora.

Overall results To get a first impression of the effectiveness of each component, we set successively each of the four coefficients to 1, the others to 0. For the GermaNet component we get 20 right candidates, for the pattern approach we get 10. 51 of the correct candidates could be found by the LSA component only. The recency component by itself finds 17 correct candidates, however it seemed to interfere with the LSA

[8] Thanks to Henrik Dittmann, Universität Osnabrück for his help.
[9] `http://infomap-nlp.sourceforge.net/`

component: When we gave equal strength to both the LSA and the recency component ($\lambda_{LSA} = 0,5; \lambda_{Rec} = 0,5$), only 34 correct candidates could be found. The maximum number of correct candidates (57) could be found using the parameters given in the last line of Table 2.

Coefficients				# correct	# wrong
λ_{GN}	λ_{Pat}	λ_{LSA}	λ_{Rec}		
1,0	0	0	0	20	95
0	1,0	0	0	10	105
0	0	1,0	0	51	64
0	0	0	1,0	17	98
0,25	0,05	0,65	0,05	57	58

Table 2: Overall results for our test set of 115 anaphors

When we split up the results for the different relation types (cf. Table 3), we see immediately that there is an important difference between the semantic and the bridging relations: Whereas 34 out of the 56 anaphors based on a straightforward semantic

Relation type	# correct	# wrong	# total
Hypo-/Hyperonyms	1 (33,3%)	2	(3)
Mero-/Holonyms	4 (36,4%)	7	(11)
Synonyms	29 (69,0%)	13	(42)
All sem. relations	34 (60,8%)	22	(56)
Bridging	23 (38,9%)	36	(59)
Overall	57 (49,6%)	58	(115)

Table 3: Results for each of the relation types considered

relation could be resolved (61%), this was the case for only 23 out of 59 bridging anaphors (39%).

Another remarkable fact is that among the semantic relations the synonyms scored far better than the meronymic or hyponymic relations. This shows the effectiveness of the LSA to measure semantic similarity between terms, since the meaning of two synonyms will be more similar than that of mero- or hyponyms.

Regarding the N-best distribution in Figure 2, we can see that most of the correct AC pairs appear on the top of the N-best lists. When we consider the first two candidates, we find 71 correct pairs (62%), the first 4 candidates comprise already 86 (75%) and the first 6 candidates 97 correct pairs (84%). Our approach therefore seems to calculate plausible semantic relationships, however it is not precise enough in the selection process.

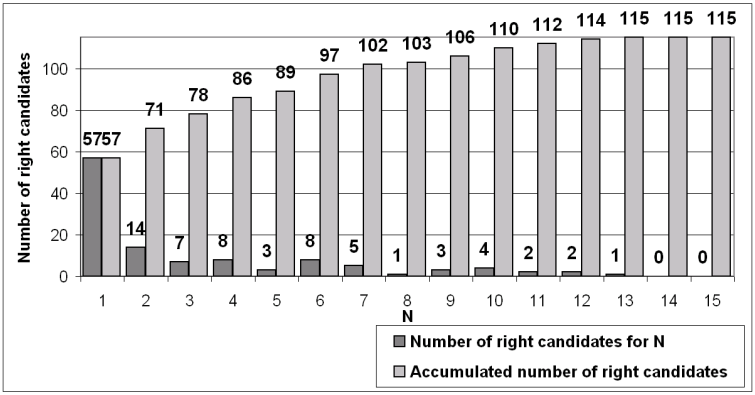

Figure 2: N-best analysis for our test set of 115 anaphors

A thorough look at the ranked lists of the candidates seems to confirm this observation: Many of the candidates are indeed ranked according to their semantic relatedness with the anaphor. Table 4 shows a typical candidate list, the candidates are ranked by their score.

correctAnte:de764 relation:bridgingLink(bridging)						Fragebogen
nbest	deID	distance	GN value	LSA	total score	text
1	**de764**	**5**	**0,4**	**0,221**	**0,294**	***Befragung***
2	de768	1	1	-0,028	0,286	*Sprache*
3	de761	8	0,6	0,027	0,203	*Unterricht*
4	de757	11	0,4	0,099	0,189	*Prüfungen*
5	de762	7	0,4	0,063	0,187	*Gruppen*
6	de767	2	0,2	0,093	0,152	*Grundkenntnisse*
7	de758	10	0,4	-0,105	0,130	*Niveaus*
8	de763	6	0,2	0,048	0,130	*Deutsch*
9	de756	12	0,4	0,015	0,128	*Vorbereitung*
10	de765	4	0,2	0,039	0,125	*Kurse*
11	de755	13	0,2	0,040	0,090	*Kurse*
12	de760	9	0	0,009	0,042	*Instituten*
13	de750	15	0	0,004	0,005	*Goethe-Institut*

Table 4: N-best list with correct antecedent found (correct antecedent in bold letters)

5 Conclusion and Outlook

This paper presents ongoing work in the domain of nominal anaphora resolution; it concentrates on the identification of the most likely antecedent from a set of antecedent candidates. Future work includes both further improvement of this component as well as work on the other two components of an anaphora resolution model: Defining the set of antecedent candidates and applying constraints to eliminate incompatible antecedent candidates from the set.

Concerning the pattern extraction component, future work focuses on the definition of more patterns and especially those extracting synonymy or meronymy relations (the results for these patterns are usually not as reliable as for the ones we used). Further experiments are needed in order to understand which patterns help and which do not.

Concerning the remaining components of the anaphora resolution system, work is done in order to define a variable search window in order to find suitable candidate sets for anaphoric items that find their antecedent at long distance. This work includes the analysis of rhetorical structure and logical document structure. Regarding the use of constraints to eliminate incompatible items from the set of candidates we assume that congruence restrictions (e.g. number agreement) might help downsizing the set of candidates and thus will help to improve the complete system; the smaller the number of elements for the semantic component the better the overall results as elements already identified as being incorrect candidates cannot interfere the LSA component.

References

Cederberg, S. and Widdows, D. (2003). Using lsa and noun coordination information to improve the precision and recall of automatic hyponymy. In *Proc. of the Conference on Natural Language Learning (CoNNL)*.

Chiarcos, C. and Krasavina, O. (2005). Rhetorical distance revisited: a parametrized approach. In *Proceedings of Workshop on Constraints in Discourse*, pages 63–70, Dortmund, Germany.

Cimiano, P. and Staab, S. (2005). Learning concept hierarchies from text with a guided agglomerative clustering algorithm. In *Proc. of the ICML Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods*, Bonn, Germany.

Clark, H. (1977). Bridging. In Johnson-Laird, P.N. & Wason, P., editor, *Thinking: Readings in Cognitive Science*, pages 411–420. Cambridge University Press, Cambridge.

Cristea, D., Ide, N., Marcu, D., and Tablan, M.-V. (2000). Discourse structure and co-reference: An empirical study. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000)*, Luxembourg.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Dumais, S. T. (1995). Latent semantic indexing (lsi): Trec-3 report. In Harman, D., editor, *Proceedings of the 3rd Text REtrieval Conference (TREC-3)*, volume 500-226, pages 219–230. NIST Special Puclication.

Erdmann, M. (2001). *Ontologien zur konzeptuellen Modellierung der Semantik von XML*. Books on Demand GmbH.

Fligelstone, S. (1992). Developing a Scheme for Annotating Text to Show Anaphoric Relations. In Leitner, G., editor, *New Directions in English Language Corpora: Methodology, Results, Software Developments*, pages 153–170. Mouton de Gruyter, Berlin.

Garside, R., Fligelstone, S., and Botley, S. (1997). Discourse Annotation: Anaphoric Relations in Corpora. In Garside, R., Leech, G., and McEnery, A., editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pages 66–84. Addison-Wesley Longman, London.

Goecke, D., Stührenberg, M., and Holler, A. (2007). Koreferenz, kospezifikation und bridging: Annotationsschema. Interne Reports der DFG-Forschergruppe 437 "Texttechnologische Informationsmodellierung".

Goecke, D. and Witt, A. (2006). Exploiting logical document structure for anaphora resolution. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France.

Hirschmann, L. (1997). MUC-7 Coreference Task Definition (version 3.0). In Hirschman, L. and Chinchor, N., editors, *Proceedings of Message Understanding Conference (MUC-7)*.

Holler, A., Maas, J.-F., and Storrer, A. (2004). Exploiting coreference annotations for text-to-hypertext conversion. In *Proceeding of LREC*, volume II, pages 651–654, Lisbon, Portugal.

Karttunen, L. (1976). Discourse referents. *Syntax and Semantics: Notes from the Linguistic Underground*, 7:363–385.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behaviour Research Methods, Instruments and Computers*, 28(2):203–208.

Markert, K., Modjeska, N., and Nissim, M. (2003). Using the web for nominal anaphora resolution. In *Proc. of the EACL Workshop on the Computational Treatment of Anaphora*.

Mehler, A., Waltinger, U., and Wegner, A. (2007). A formal text representation model based on lexical chaining. In *Proceedings of the KI 2007 Workshop on Learning from Non-Vectorial Data (LNVD 2007)*, pages 17–26, Osnabrück. Universität Osnabrück.

Mitkov, R. (2002). *Anaphora resolution*. Longman, London.

Ng, V. and Cardie, C. (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111.

Oard, D. W. (1997). Alternative approaches for cross-language text retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence.

Paaß, G., Kindermann, J., and Leopold, E. (2004). Learning prototype ontologies by hierarchical latent semantic analysis. In *Knowledge Discovery and Ontologies*, Pisa, Italy.

Poesio, M. (2004). The MATE/GNOME Scheme for Anaphoric Annotation, Revisited. In *Proceedings of SIGDIAL*, Boston.

Poesio, M., Mehta, R., Maroudas, A., and Hitzeman, J. (2004). Learning to resolve bridging references. In *Proceedings of the ACL*, pages 143–150.

Poesio, M., Schulte im Walde, S., and Brew, C. (1998). Lexical clustering and definite description interpretation. In *Proc. of the AAAI Spring Symposium on Learning for Discourse*, pages 82–89, Stanford, CA.

Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124.

Sowa, J. (2000). *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Brooks Cole Publishing Co., Pacific Grove, CA.

Stührenberg, M., Goecke, D., Diewald, N., Mehler, A., and Cramer, I. (2007). Web-based annotation of anaphoric relations and lexical chains. In *Proceedings of the Linguistic Annotation Workshop (Merger of NLPXML 2007 and FLAC 2007)*, Prague, Czech Republic.

Versley, Y. (2007). Using the web to resolve coreferent bridging in german newspaper text. In Rehm, G., Witt, A., and Lemnitzer, L., editors, *Data Structures for Linguistic Resources and Applications. Proceedings of GLDV-2007*, pages 253–261, Tübingen. Gunter Narr Verlag.

Vieira, R. and Poesio, M. (2000). An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

Webber, B. L. (1988). Discourse deixis: Reference to discourse segments. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistic (ACL–88)*, pages 113–122, State University of New York at Buffalo. June 27-30 1988.