Using the Web to Resolve Coreferent Bridging in German Newspaper Text

Yannick Versley

Sonderforschungsbereich 441
Universität Tübingen
versley@sfs.uni-tuebingen.de

Abstract. We adopt Markert and Nissim (2005)'s approach of using the World Wide Web to resolve cases of coreferent bridging for German and discuss the strength and weaknesses of this approach. As the general approach of using surface patterns to get information on ontological relations between lexical items has only been tried on English, it is also interesting to see whether the approach works for German as well as it does for English and what differences between these languages need to be accounted for.
We also present a novel approach for combining several patterns that yields an ensemble that outperforms the best-performing single patterns in terms of both precision and recall.

1 Introduction

While coreference resolution in general is thought of as well understood, certain aspects are difficult to model only with cheap features: in the case of a definite noun phrase that has an antecedent with a different lexical head, one needs to use some conceptual knowledge. Vieira and Poesio (2000) call the cases of anaphoric definite noun phrases where the antecedent's lexical head is different *Coreferent bridging*. These cases actually account for half of all anaphoric definite noun phrases, which means that a system using only surface-based features will always have poor recall for common nouns (as opposed to named entities, where surface-based features work quite well).

The obvious approach to this problem is to use an ontology – either a full-blown ontology for smaller domains, or a wordnet if we are considering open-domain text, as Poesio et al. (1997) also do. While the coverage of wordnets (the English WordNet as well as the German GermaNet) is quite impressive, it is not perfect, and often, a given relation between concepts (in our case, hypernymy or synonymy) is not encoded in the way that the text we work on seems to suggest it should.

Markert and Nissim (2005) argue that in part, this limitation stem from the fact that the conceptual relations expressed in the text are often context-dependent, and that instead of trying to automatically improve an ontology from text and getting all these context-dependent relations (which are ultimately unwanted since they would add noise to an ontology that tries to abstract from

the context at hand), we would fare better if we just use the mechanism to extract potential relations from text directly.

The approach they propose is to use the web as a noisy but very large corpus, together with simple lexicosyntactic patterns like those used by Hearst (1992), to achieve useful results with low-recall patterns that yield sensible results due to the large amount of text available on the web.

To our knowledge, the work on German that is presented in this paper is the first that puts this approach to test for a language other than English[1]. Several differences between German and English make this comparison interesting in its own right: The use of shallow surface patterns may be less useful, possibly because of German's richer morphology and flexible word order, but also because the amount of text available on the World Wide Web is by far greater for English than for any other language.

2 Antecedent Selection in German

For our experiments, we use the referential layer of the TüBa-D/Z treebank (Hinrichs et al., 2005) together with GermaNet, a German wordnet(Kunze and Lemnitzer, 2002). In a similiar setting to Markert and Nissim (2005)'s experiment II, we want to decide for a given definite NP (excluding named entities), which antecedent from the preceding sentences is the correct one. This setting is different from the full task of coreference resolution, not only because we exclude pronouns and named entities (which are both easier to detect and to resolve), but also because (following Markert and Nissim) we assume perfect knowledge on whether a noun phrase is discourse-new. Since it is not always evident whether a given definite NP is discourse-old or not, and since there are many more discourse-new NPs than anaphoric ones (about 27% of the definite NPs in the TüBa-D/Z corpus are coreferent to an earlier mention), the precision of a full system would be much lower than the precision for antecedent selection. Perfect knowledge on the discourse status of a definite noun phrase is not a realistic assumption, and usually there is ample room for interaction between the classification of discourse-newness and the resolution of coreferent bridging (either definite noun phrases wrongly recognized as discourse-old and subsequently resolved by the resolution component, or that an anaphoric NP that the resolution component would be able to resolve reliably were wrongly recognized as discourse-new). Because of this interaction, we think that the integration of information concerning discourse status, be it in a modular approach like the one presented by Uryupina (2003), Poesio et al. (2005), or an integrated approach like the one presented in Versley (2006), is best considered as a separate research topic.

The experimental setting is also different from other experiments concentrating on the resolution of coreferent bridging (Poesio et al., 1997, Gasperin

[1] as it appears, research groups that pursue similar directions, even those in Germany such as Cimiano and Staab (2004) exclusively work on English texts, which we think is a pity

Table 1. Results for the task of NP antecedent selection

	Prec.	Recall	$F_{\beta=1}$	Prec*
SameHead	0.865	0.498	0.632	—
SameHead+GWN(1)	0.832	0.575	**0.680**	0.67
SameHead+GWN(2)	0.845	0.545	0.663	0.68
GWN only(1)	0.798	0.526	0.634	—
SameHead+"*X* u.a. *Y*s"(3)	0.843	0.510	0.635	0.42
SameHead+"*Y*s wie *X*"(3)	0.825	0.543	**0.655**	0.55
SameHead+both patterns(3)	0.821	0.542	0.653	0.52
SameHead+"*X* u.a. *Y*s"(4)	0.809	0.543	0.650	0.48
combined(5)	0.814	0.601	**0.692**	0.63

Prec*: Precision on coreferent bridging cases
(1): no sentence limit (2): cur+4prev sent.
(3): cur+4prev sent., distance-based
(4): cur+8prev sent., distance-based
(5): first same-head resolution, then web-based ("*Y*s wie *X*", cur+4prev sent.)

and Vieira, 2004), since these only look at coreferent bridging cases, excluding the cases of same-head resolution, which are easier and account for a very large portion of the correct resolutions handled by current resolvers for definite noun phrases.

Similar to Markert and Nissim, we include a lemma-based string comparison as a baseline, where we use the same-head resolution component from a state-of-the-art coreference resolver for German (Versley, 2006), which also checks modifier compatibility using the heuristics from (Vieira and Poesio, 2000). In order to cope with German synthetic compounds, the antecedent lemma may also have the lemma of the anaphor as a suffix.

As the second baseline, we see if a possible antecedent is a hypernym of the anaphor's head, using GermaNet. In order to cope with compounds, GermaNet is searched not only for an exact match to the word lemma in question, but also for the longest possible suffix, increasing the coverage to 88% of the noun tokens occuring in anaphoric definite NPs. Number agreement is always checked, and in the case of multiple potential antecedents being hyponyms, the nearest is chosen.

Using this algorithm for cases that the same-head resolver could not handle, we get an increase in recall of about 7.5%, and the hypernymy test has about 65% precision in our experimental setting. Limiting the search window for antecedents to the 8 preceding sentences does not improve the precision, but seems to limit the recall that can be gained. Using only the hypernymy lookup in GermaNet, and no specialized same-head resolver, both precision and recall are worse than for the combination of same-head resolver and hypernymy lookup.

2.1 The web-based approach

For the web-based algorithm, we need to normalise the surface form to minimise the influence of case inflection and/or the strong/weak inflection in the case of deadjectival nouns. For this purpose, we use SMOR (Schmid et al., 2004), a finite-state morphology model for German, to normalize the nouns to nominative form and form plurals. Using SMOR for generation results in spurious ambiguities, for example, generating the nominative singular form of *Radrennen* - bycicle race - results in 3 additional surface forms, *Räderrennen*, *Radsrennen*, *Radesrennen*. We then choose a form by preferring surface forms that are nearer to the original string in terms of edit distance, more frequent globally (to allow generating the dissimilar "Länder" from the genitive "Landes", rather than the similar, and equally valid, but marked, "Lande"), and more similar frequencies (to prevent spurious matches by confusion with high-frequency words).

Using the pattern "X and other Ys" (X und andere Ys), we can correctly resolve a few more instances (improving recall by a comparatively meager 1.2%), but the precision is much lower than with GermaNet, making it worse both in terms of recall and precision than the wordnet-based approach. Another pattern that Markert and Nissim mention, but did not include in their study, "Ys wie X", yields a higher gain in recall and better precision, as it is found more frequently, but with a recall gain more than three time as high, and better precision (54% for the part that is handled by the web-based resolver),is still inferior to hyponymy-based lookup in a wordnet. The main reason for this seems to be that data sparsity is still a problem, more so for "X und andere Ys" than for the higher-recall pattern "Ys wie X".

The reason why low recall for a pattern presents not only a recall problem, but also a precision problem is to be seen in cases where one potential (but wrong) antecedent is found by the pattern, but the correct one isn't, for example in *die Region* (the region) wrongly being resolved to *China* instead of *das Kosovo* because China occurs so much more frequently than the Kosovo that the pattern is seen with the former, but not with the latter.

Many cases where a wrong antecedent is found are due to this kind of error, for example the correct antecedent *Pioneer* being less known as a company dealing in seeds (*Saatguthersteller*) than *Monsanto*, which is further away and the incorrect antecedent, *Bonn* being chosen instead of the lesser-known *Gummersbach* as antecedent for *die Stadt*, or *Hamburg* instead of the small city *Elmshorn* for *der Ort*.

Combining both patterns (i.e. choosing a potential antecedent if either pattern can be found) does not lead to increased recall, and precision is between the values for each pattern alone. Also, enlarging the sentence window (from 4 to 8 sentences) only yields more false positives, which suggests that the web counts capture different information than the hyponym search in GermaNet.

When we combine the higher-recall pattern "Ys wie X with the approach using GermaNet by first looking for hypernyms in GermaNet and only considering the web-based method for cases that could not be resolved in this way, we can maintain the good precision of the GermaNet-based resolver (0.64) while

Table 2. Counts for selected patterns

	X u. a. *Y*s	*Y*s wie *X*	*Y*s einschl. *X*	*Y*s insb. *X*	*Y*s außer *X*
Monsanto – Firma	30	243	—	—	—
Bremen – Stadt	22	231	5	2	3
Korsika – Insel	1	51	1	—	—
Magath – Trainer	1	31	—	—	—

	das *Y* *X*	die *Y*s *X*	*X* ein *Y*	*X* ist ein *Y*	*X* ist das *Y*
Monsanto – Firma	571	76	6	1	2
Bremen – Stadt	39.800	846	231	465	67
Korsika – Insel	674	130	66	142	17
Magath – Trainer	112	1	11	9	5

achieving better recall – this is partly due to the fact that the web-based method only works well for rather close antecedents (as seen in Table 1, using a larger window size does not improve the recall), whereas the GermaNet-based method also works for antecedents that are further away.

2.2 Can we do better?

The method we presented here resolves some of the cases where lemma-based matching of heads does not help, with a precision that seems usable, even keeping in mind that determining the anaphoricity of definite noun phrases cannot be done perfectly and we will get more false positives because of antecedents proposed for discourse-new definites.

In some cases, several antecedents are equally plausible on lexical semantic grounds (*die Partei* with possible antecedents *die CDU*, *die SPD*, both political parties in Germany) and the closest one is wrong. These cases are also expected to be resolved erronously by other sources of lexical information, and to handle these cases, it is probably necessary to take into account discourse structure or approximations thereof. In the example above, *die CDU* and *die Partei* stand in a sentence coordination which corresponds to a *parallel* discourse relation, and the subject of the second conjunct (*meine Partei*) would normally be replaced with a pronoun if it were coreferent with the subject of the first conjunct.

Other matches are simply spurious, as in the case of *der Professor* (the professor) being matched to *Cottbus* (a city in eastern Germany) due to a spurious match (in the relevant document, of which 3 copies are found on the web, *Cottbus* is in a PP modifying the real conjunct, something which only a (more expensive) full syntactic analysis could have revealed. In the case of *Mannschaften wie Bremen*, on the other hand, the relation is pertinent but not warranted by the context, where *Bremen* was used in a non-metonymic way.

But even then, we only propose a candidate for 73% of the definite NPs that we want to resolve. It is certainly not a good solution to always propose the closest NP that matches in number, as Markert and Nissim propose, but it

Table 3. Comparing different patterns

	Prec.	Recall	$F_{\beta=1}$	Prec*	Prec**
(SH+pat)					
*Y*s wie *X*	0.825	0.543	**0.655**	0.55	—
X u. a. *Y*s	0.843	0.510	0.636	0.42	—
das *Y X*	0.813	0.523	0.637	0.37	—
X ist ein *Y*	0.829	0.521	0.640	0.44	—
(SH+GWN+pat)					
*Y*s wie *X*	0.814	0.602	**0.692**	0.63	0.55
das *Y X*	0.808	0.592	0.683	0.60	0.41
X ist ein *Y*	0.815	0.588	0.683	0.62	0.42
(SH+GWN+combination)					
best only (*Y*s wie *X*)	0.814	0.602	0.692	0.63	0.55
*Y*s wie *X*, all others	0.802	0.615	0.696	0.61	0.52
PMI (all patterns)	0.812	0.615	**0.700**	0.64	0.60

Prec*: Precision on coreferent bridging cases
Prec**: Precision for web pattern alone

would be desirable to have some means to resolve the other 27% of the anaphoric definites with at least somewhat adequate accuracy.

Hearst (1992) mentions additional patterns: "*Y*s, especially *X*" and "*Y*s, including *X*". A quick check shows that the recall for these (and for *X*s except *Y*) is vastly inferior to the other, more frequent ones, like "*X* and other *Y*s". Cimiano and Staab (2004) use Hearst-like patterns for the original purpose of ontology population, and although the way they combine the evidence – summing up the query results to a total for each candidate – is not suited for our task (the two first patterns led to a loss in precision even for cases that could be handled correctly with only the higher-recall pattern "*X*s wie *Y*"), Cimiano and Staab also use patterns other than those proposed by Hearst. They aim at appositional and copular constructs, using patterns like "the *Y X*" (i.e. the city Berlin) and "the *X Y*" (the Ramada hotel – this pattern is not easily translatable into German, and mostly concerns cases that we are less interested in). The definite apposition "the *Y X*" occurs much more often than the Hearst patterns, and even the plural variant (which Cimiano and Staab did not consider) or the copular construction "*X* ist ein *Y*" seem to give good results for some semantic classes, if not for all.

The results from incorporating the patterns "das *Y X*" and "*X* ist ein *Y*" into the system and evaluating on coreference data, however, seem to indicate that even these patterns, including the more frequent "das *Y X*" cannot bring a larger improvement. This could indicate that, firstly, "*Y*s wie *X*" is better at capturing context-dependent relations than the appositional and copular constructions, and secondly, that the recall problem we are facing for the cases of coreferent bridging is probably not due only to data sparseness, but also to the fact that

hyponymy and instance relations do only account for part of the coreferent bridging cases.

When we use almost all usable patterns from table 2 (X u.a. Ys, Ys wie X, das Y X, die Ys X, X ist ein Y), trying the higher-precision pattern "Ys wie X" first, it is possible to correctly resolve 61.4% of all anaphoric definite noun phrases, which is more than using GermaNet and the "Ys wie X" pattern alone (60.2%), and we propose an antecedent for 76.6% of all cases (compared to 73.9% for the approach using GermaNet and the best-performing pattern, 65.9% for using GermaNet and 57.5% for same-head resolution alone). This seems to suggest that a significant part of the anaphoric definite noun phrases are idiosyncratic and not resolvable using encyclopedic or lexical knowledge if we assume a strict hyponymy or instance relation.

Garera and Yarowsky (2006), who use a pattern-based approach on the British National Corpus as a baseline for their work of combining wordnet information with an association statistic, suggest the use of pointwise mutual information (PMI) to rank candidate antecedents. We did not directly implement this idea here as we always rank candidates using sentence distance. But it is certainly possible to eliminate some of the false positives (at least in the case of the pattern combination) by filtering out cases with low PMI values, which are probably due to spurious matches. To do this, we first estimate a smoothed PMI value for each pattern, regardless of whether it has been seen or not, and then combine these PMI values into a combined estimate to allow for variation in the relative frequencies of the patterns for different anaphor-antecedent pairs.

For each pattern, we calculate an expected number of pattern matches based on the frequencies of both parts (for example, *Branchen* and *die Druckindustrie* in the case of "*Branchen wie die Druckindustrie*") and a correction factor that is meant to account for the size of the German WWW and the relative frequency of the given pattern, which we determined so that the geometric mean of the seen vs. expected ratio over a set of selected concept-instance pairs (cf. table 2) is exactly 1. In order to account for unseen patterns that are due to the involved parts being rare, we smooth the ratio by adding 0.5 to each pattern or part count – this improves the chance that our PMI estimate for unseen patterns, which would be 0 in the unsmoothed case, contributes meaningful information to the ensemble: An unseen pattern for high frequency terms gets a very small ratio (i.e. indicating that the match for the pattern that has been seen could be a false positive) and an unseen pattern for lower frequency terms, which could just be due to insufficient sample size, gets small but comparatively higher ratio.

The logarithms of the smoothed expected-to-seen ratios (the latter correspond to our estimation of the pointwise mutual information values) are then added up and compared to a threshold (currently -1). If there are pattern matches, but the combined value is lower than the threshold, the candidate is discarded. This approach yields exactly the same amount of true positives that we get with the simple combination of all patterns, but with a reduced rate of false positives.

3 Conclusion

For the antecedent selection task, both using a wordnet (WordNet or GermaNet), and using web counts for patterns indicating hyponymy, help improving the recall of NP coreference. Contrary to Markert and Nissim's experiment, however, where using web counts gives a greater boost than using WordNet, we find that using GermaNet actually helps more than the web counts. Besides the smaller size of the German World Wide Web, this seems to be due to the fact that we consider a larger context, which is a sensible thing to do as the precision stays on the same level when removing the 4 sentence limit[2]. Using this insight, we were able to improve upon Markert and Nissim's results by combining GermaNet and the pattern-based approach for the non-same-head cases and achieved an improvement over the single components both in terms of recall and in terms of f-measure. The results from using different patterns that capture appositional and copular constructions do not meet the expectations that the pure frequency counts may suggest, due to lower precision, but the combination of a several patterns with a statistical thresholding technique allows us to outperform the best-performing single pattern, in terms of recall and also in terms of precision of pattern-resolved instances. Looking at the combined precision of same-head resolution, GermaNet *and* the pattern-based method of course tells us a different picture since the resolution accuracy for coreferent bridging is still much worse than that for direct coreference – for applications where high precision is needed, using only the same-head component remains the best option.

Harnessing the sheer size of the Web with low-recall methods seems more difficult for languages like German, where the data is sparser by approximately one order of magnitude: Kilgrariff and Grefenstette (2003) give a factor of about 1:10 for the size ratio of German to English web indexed in AltaVista. For counts of the patterns that interest us, the numbers can vary wildly, between 4:1 for *Bremen*, and 1:250 for *Houston*. *"Monsanto and other companies"* yields 523 hits on google.com, while its German counterpart only yields 6; the higher recall pattern *"companies such as Monsanto"* yields 25,500 hits for English, while its German counterpart only gets 236 hits. The problem may be even worse for other languages like Dutch or Italian, which have an even smaller number of (native or non-native) speakers and consequently also a smaller amount of text that is available via web searches.

Given the computational cost of morphological regeneration and also the strict limit on search queries (the Google API has a limit of 1000 queries per day, while the Yahoo API has a larger limit of 5000 queries per day, but sometimes returns wildly inaccurate results for phrase queries), it is quite clear that only using web queries with shallow patterns is not a solution in itself. But even for languages with fewer speakers than English, using the Web as a last resort can help improve a system.

[2] in contrast, Markert and Nissim do never consider candidates further than 4 sentences away, and they specifically exclude anaphoric noun phrases where there is no antecedent in the last 4 sentences

Acknowledgements I would like to thank the two anonymous reviewers for critical comments on an earlier version of this paper, and Heike Zinsmeister for a comment on the final version. The research reported here was supported as part of the DFG collaborative research centre (Sonderforschungsbereich) "SFB 441: Linguistische Datenstrukturen".

Bibliography

Cimiano, P. and Staab, S. (2004). Learning by googling. *SIGKDD Explorations*, 6(2):24–33.

Garera, N. and Yarowsky, D. (2006). Resolving and generating definite anaphora by modeling hypernymy using unlabeled corpora. In *CoNLL 2006*.

Gasperin, C. and Vieira, R. (2004). Using word similarity lists for resolving indirect anaphora. In *ACL'04 workshop on reference resolution and its applications*.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proc. of the 14th International Conference on Computational Linguistics (COLING 92)*.

Hinrichs, E., Kübler, S., and Naumann, K. (2005). A unified representation for morphological, syntactic, semantic and referential annotations. In *ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, Ann Arbor.

Kilgrariff, A. and Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–348.

Kunze, C. and Lemnitzer, L. (2002). GermaNet – representation, visualization, application. In *LREC 2002*.

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

Poesio, M., Alexandrov-Kabadjov, M., Vieira, R., Goulart, R., and Uryupina, O. (2005). Does discourse-new detection help definite description resolution? In *IWCS-6*.

Poesio, M., Vieira, R., and Teufel, S. (1997). Resolving bridging descriptions in unrestricted text. In *ACL-97 Workshop on Operational Factors in Practical, Robust, Anaphora Resolution For Unrestricted Texts*.

Schmid, H., Fitschen, A., and Heid, U. (2004). SMOR: A german computational morphology covering derivation, composition and inflection. In *Proceedings of LREC 2004*.

Uryupina, O. (2003). High-precision identification of discourse new and unique noun phrases. In *Proceedings of the ACL Student Workshop*.

Versley, Y. (2006). A constraint-based approach to noun phrase coreference resolution in german newspaper text. In *Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS 2006)*.

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