A Constraint-based Approach to Noun Phrase Coreference Resolution in German Newspaper Text

Yannick Versley
versley@sfs.uni-tuebingen.de
SFB 441 / Seminar für Sprachwissenschaft
Universität Tübingen

Abstract

In this paper, we investigate the usefulness of a wide range of features for their usefulness in the resolution of nominal coreference, both as hard constraints (i.e. completely removing elements from the list of possible candidates) as well as soft constraints (where a cumulation of violations of soft constraints will make it less likely that a candidate is chosen as the antecedent). We present a state of the art system based on such constraints and weights estimated with a maximum entropy model, using lexical information to resolve cases of coreferent bridging.

1 Introduction

From the wider range of tasks that are subsumed under the term coreference resolution, coreference among full (i.e. non-pronominal) noun phrases has received quite a bit of attention, not only with respect to the MUC and ACE data sets and evaluation, but also for German (Hartrumpf, 2001; Strube et al., 2002). However, there is still quite a large difference between the resolution of pronominal and non-pronominal anaphora in terms of systems' accuracy.

The question arises whether this difference is simply due to the fact that the processing involved in the resolution of nominal references is more complex than with pronominal anaphora, and therefore also prone to more errors, or maybe that it involves several different kinds of (non-)anaphoricity. This question has been partially answered with the insight (Vieira and Poesio, 2000; Bean and Riloff, 1999; Ng and Cardie, 2002a) that some definite descriptions are definite because of their uniqueness in the context and that these are not anaphoric in the stricter sense, but may still be coreferent if they are mentioned multiple times. Most current systems capitalize on this insight by including a module to determine whether a given definite description is unique (and thus non-anaphoric) when no candidate with the same head is found.

Poesio et al. (2005) show that using Vieira and Poesio's syntactic heuristics together with Bean and Riloff's corpus-based approach to identify unique noun phrases and web-based definiteness counts (Uryupina, 2003), resolution results for the GNOME corpus can be improved by not considering some mentions for resolution.

One goal of the work reported here is to see what impact commonly posited resolution constraints have on the region of possible behaviours that a system may exhibit, but also to explore some constraints that may be useful for the resolution of non-same-head anaphoric definite descriptions, as lexical information alone does not seem to be enough to reliably resolve them (Poesio et al., 1997; Gasperin and Vieira, 2004).

2 Constraint-based coreference resolution

We formulate our system in terms both of hard constraints, which cannot be violated (i.e. candidates that violate them are filtered out), and soft constraints, which influence the choice that is made among the remaining candidates. To allow the soft constraints to (possibly) indicate that a given markable should not be resolved at all, we use a pseudo-candidate that indicates non-resolution (and uses different features than normal candidates), but is handled identically to other candidates otherwise.

For our experiments, we used a referentially and syntactically annotated corpus of texts from the German newspaper 'die tageszeitung', the TüBa-D/Z treebank of written German (Telljohann et al., 2003; Hinrichs et al., 2005), where mentions are marked up and grouped into sets according to (co-)reference. Appositional constructions (as in 'Peter, the Englishman') are treated as a single mention, and predicative noun phrases in copular constructions were not considered for resolution (as their

resolution is only influenced by syntactic structure), but considered as possible antecedents.

2.1 Evaluation Method

To see the impact of hard constraints, we give upper and lower bounds for precision and recall for each variant we consider, based on the candidate sets in the training corpus, after filtering with the hard constraints (Rmax, Pmax as upper bounds on recall and precision, as well as Rmin and Pmin as lower bounds).

Assessing the influence of the soft constraints is more difficult since there is a significant interaction both with the hard constraints (since the soft constraint can only be used to choose among those candidates that have not been filtered out by them), and among the soft constraints themselves.

In addition to the actual precision and recall on an evaluation corpus (distinct from the training corpus, which was used to determine the weights of the soft constraints, and the bounds on precision and recall), we also provide a figure for the perplexity of the classifier decisions[1]. This is exponential in the number of bits that an actual resolution system would need in average for each decision, *in addition to the information from the soft constraints* to achieve the given maximal precision and recall on the training data. In the case where candidates are not weighted using soft constraints, the perplexity is the geometric mean of the number of candidates one has to choose from, in the case of weighting it can be thought of as the geometric mean of the probability of choosing a 'good' candidate when randomly choosing. Because the perplexity is an average over *all* decisions that the classifier has to make, filtering out "easy" cases (i.e. those from the majority class) can also raise the perplexity when the remaining cases are more ambiguous.

3 Maximum likelihood estimation of constraint weights

In order to choose among the candidates that remain after filtering using the hard constraints, the latter are ranked using weighted constraints. Each candidate is represented as a vector of numerical features, and these feature values are multiplied with the feature weights to get the score of a candidate, so that we can choose the candidate with the largest score:

$$\hat{y} = \arg\max_{y \in Y} \langle w, f(y) \rangle$$

(where w is the vector of the constraint weights, f is a function that maps a candidate to a feature vector, $\langle \cdot, \cdot \rangle$ is the dot product in euclidean space, and Y is the set of possible antecedents).

To choose the constraint weights, we interpret our score in a probabilistic fashion. Given the measure

$$\mu(y) := \mathrm{e}^{\langle w, f(y) \rangle}$$

we can define a probability distribution

$$\hat{P}(y) := \frac{\mu(y)}{\sum_{y' \in Y} \mu(y')} = \frac{\mathrm{e}^{\langle w, f(y) \rangle}}{\sum_{y' \in Y} \mathrm{e}^{\langle w, f(y') \rangle}}$$

by normalizing the measure so that the probabilities of all $y \in Y$ sum up to 1. This kind of model is called in the literature a loglinear model; in case of binary features, the constraint weights can be interpreted as (the logarithm of) an odds ratio, whereas in the case of continuous features, the constraint weights can be seen as the parameter of an exponential distribution.

The probability that randomly choosing with the distribution $\hat{P}$ will yield the choices from the data (equivalently, the likelihood of the data given the model) can be calculated as the product of the model probabilities for each single decision. Together with a prior (an independently motivated probability distribution for w), the logarithm of the combined probability of model and data can be stated as a function $Loss(w)$, and, happily, this loss function is concave, which means that it only has a single (global) maximum and can be efficiently estimated using numerical techniques (see Malouf, 2002).

In coreference resolution, it is possible that we have multiple candidates that are all coreferent to the description we are looking at, and we do no longer have a single 'good' candidate y, but we can readily extend our model by simply considering the probability that the model chooses *any* of the correct candidates:

$$P_{\mathrm{good}} = \sum_{y \in Y_{\mathrm{good}}} \hat{P}(y) = \frac{\sum_{y \in Y_{\mathrm{good}}} \mu(y)}{\sum_{y \in Y} \mu(y)}$$

However, the resulting loss function is no longer guaranteed to be concave, which means that it can

[1] We are slightly twisting the meaning of *perplexity* here, since the term is normally used for the exponential of the entropy of a distribution and we use it for the exponential of the cross-entropy wrt. our corpus. As our model is discriminative rather than generative, the former would hardly be possible and we still like the idea of the perplexity corresponding to the average branching factor in the case of a uniform distribution

have multiple local maxima and the algorithms used for the simpler case could possibly get stuck in a local maximum that is not the global one. Since most decisions only involve a single positive candidate, we just naïvely assume that concavity of the loss function holds nevertheless, at least in a local environment that comprises both our starting point, the global maximum and a region around it that is large enough that we do not run into problems with the optimization algorithm we use[2].

If we compare the results of choosing the weights in the way described above with other methods, for example having the model make a binary decision about anaphora-candidate pairs as done by Morton (2000), or a binary ranking decision (requiring that the real antecedent is ranked before any non-antecedent candidate), we find that our model performs slightly better. Optimizing using a unary loss function (as in Morton's system) with exactly the same features would result in similar precision (61.0%), but visibly lower recall (69.1% vs. 70.0% in our system).

4 Results

In knowledge-poor approaches like those of Strube et al. (2002); Ng and Cardie (2002b), nominal coreference is usually determined by considering pairs of mentions which share the same lexical head, or part(s) of a name.[3]:

(1) a. [$_1$ Der koreanische Autokonzern Daewoo] wollte auf keinen Fall mit seinem Autoumschlag in Bremerhaven bleiben (...).

b. Was sollte [$_1$ Daewoo] gegen den betriebswirtschaftlich günstigeren Standort Bremerhaven haben?

(2) a. [$_2$ Ein Bremer Nazibunker] dient Johann Kresnik als Spielstätte für "Die letzten Tage der Menschheit".

b. Für viele in Bremen war [$_2$ der Bunker] lange Zeit "Der Valentin".

[2]The optimization code used is the L-BFGS routine by Liu and Nocedal (1989), which is available at http://www.ece.northwestern.edu/~nocedal/lbfgs.html

[3]from the TüBa-D/Z treebank; translations: The Korean automobile company Daewoo did not want to keep its car shipment centre in Bremerhaven (...). / What should Daewoo have against Bremerhaven, which is economically more advantageous as a location?
A Nazi bunker in Bremen serves as a stage for Johann Kresnik's "The Last Days of Humanity". / For a long time, many in Bremen thought of the bunker as the "Valentin".

Typically, but not always, names and other descriptions are shortened in subsequent mentions. In comparison to English, distinguishing between named entities and nominal mentions is more difficult since all nouns are capitalised in German, and compounding, together with morphology, makes it less obvious in some cases that two mentions share the same head.

The first two parts of table 1 contain an overview of the system's results when considering same head resolution only. In the first part, the choice of introducing a new referent for the mention was always allowed, with a soft constraint weighting that choice against resolving to an earlier mention, and one soft constraint weighting resolution candidates according to their distance in sentences. In the second part, resolution is always attempted when possible, and the ranking is purely by distance.

Of all names and definite descriptions, 27% corefer with an earlier mention, which normally is a name or full noun phrase (except for 1.3% in the cases, which are due cataphoric pronouns). As a baseline for same head resolution, we simply included every candidate markable that shared at least one letter-4gram with the anaphor, yielding an upper bound of 76.5% recall that is achievable using knowledge-poor same-head matching techniques. Because of German compounding and morphology, checking for exact identity of the head also results in a large loss in recall. Using morphological analysis and suffix matching (this is the same_head version), it is possible to achieve optimal recall in conjunction with much improved precision bounds.

Besides generic mentions (where no reference to a specific entity is made, and coreference should not be annotated following the annotation guidelines), there are some spurious matches due to cases where two mentions share the head noun, but do not corefer. Checking number agreement solves some of these cases, improving the upper precision bound to that for the 'head identity' variant, without any noticeable impact on recall. Pairs of mentions where two instances of a single concept are mentioned (the red car/the blue car) are another source of spurious matches, which we filter out using some heuristics due to Vieira and Poesio (2000), notably requiring that all modifiers present in the anaphor are also present in the antecedent. An exception is made for adjectives, where it is allowed that the anaphor has some attributive adjectives when the antecedent doesn't have any at all.

	P_{max}	R_{max}	P_{min}	R_{min}	Perp	Prec	Recl
always allow non-resolution							
head identity	100.0	54.4	0.0	0.0	1.89	62.5	38.5
same head	100.0	76.9	0.0	0.0	1.98	58.3	40.5
uniq_name	100.0	74.3	0.0	0.0	1.88	66.8	58.4
force resolution							
all	27.0	98.7	0.0	0.0	23.68	1.2	4.9
4gram	31.1	76.6	13.3	37.5	2.28	26.3	54.7
head identity	52.1	54.4	32.1	47.1	1.68	58.2	50.5
same_head	49.0	76.9	33.6	59.0	1.65	51.6	69.4
+agr_num	52.1	76.5	36.3	60.4	1.62	56.0	69.7
+comp_mod	56.4	71.4	38.2	57.7	1.57	62.1	64.8
uniq_name	57.1	74.3	40.5	61.6	1.57	62.0	68.6
+hard_seg(8)	64.9	68.7	43.8	59.0	1.61	67.8	63.2
+loose_seg(8)	62.8	71.1	43.0	59.8	1.58	66.6	65.8
include coreferent bridging							
no filter	62.3	92.5	14.3	61.6	1.42	62.0	68.6
+gwn only	62.3	92.5	14.3	61.6	1.28	62.0	68.6
filter_ne	61.7	90.1	17.1	61.6	1.68	62.0	68.6
+gwn only	61.7	90.1	17.1	61.6	1.31	62.0	68.6
unique_mod	60.7	86.3	21.2	61.6	1.51	62.0	68.6
+segment	60.6	85.6	21.4	61.6	1.49	62.0	68.6
+num	60.6	85.6	21.4	61.6	1.49	62.0	68.6
+gwn	59.8	83.0	21.7	61.6	1.28	61.7	69.2
+syn_role	59.8	83.0	21.7	61.6	1.27	61.9	69.5
NE_semdist	59.8	83.0	21.7	61.6	1.27	61.9	69.7
+pred_arg	59.8	83.0	21.7	61.6	1.26	61.9	70.0

Table 1: Upper and lower bounds / evaluation results

Up to here, we treated coreference for definite noun phrases and for named entities in the same way. But named entities are special in that names are usually unique to an entity (e.g. Miller, the CEO is different from Smith, the CEO, although they both share the common noun CEO). Therefore, two named entities are only allowed to match if they share the name, not if they share any other common noun. Named entities also occur more frequently in conjunction with modifiers that are indicative of uniqueness, even when they are discourse-old, which is why we do not check for modifier compatibility in the case of named entities.

Besides ranking candidates by their sentence distance, it is also possible to cut off candidates that are more than a certain number of sentences away. Imposing a hard 8-sentence window improves the precision by more than 5%, but has a detrimental effect of the same size on recall. Vieira and Poesio (2000) introduce a loose segmentation heuristic where they consider antecedents that are either not further away than a certain number of sentences or have been mentioned multiple times. Such a loose segmentation heuristic has the potential to improve precision by the same amount as in the case of hard segmentation, with a much smaller loss of recall (using the grouping from the gold data, the loose segmentation heuristic gives 67.8% precision and 67.0% recall). Because of the propagation of resolution errors, however, these improvements are only partly realized.

4.1 Resolving coreferent bridging descriptions

Ultimately, a resolver for NP coreference should be able to also handle cases that involve non-same-head coreference (which Vieira and Poesio call coreferent bridging).
Consider the following example[4]:

[4]contiguous sentences from the TüBa-D/Z treebank; translation: An 88-year-old [female] pedestrian has been gravely injured in a collision with a car. When crossing the Waller Heerstraße, the woman had obviously overlooked the automobile.

(3) a. Lebensgefährliche Körperverletzungen hat sich [$_1$ eine 88jährige Fußgängerin] bei einem Zusammenstoß mit [$_2$ einem Pkw] zugezogen.
b. [$_1$ Die Frau] hatte [$_2$ das Auto] beim Überqueren der Waller Heerstraße offensichtlich übersehen.

The referents 1 (the woman) and 2 (the car) are mentioned again in the second sentence, not pronominalized, nor repeated identically, but in a semantically poorer form (the [female] pedestrian - the woman), or as a synonym (the car - the automobile). In contrast to pronominal reference or same-head coreference, it is possible that anaphor and antecedents have a differing grammatical gender ('Pkw' has male gender, while 'Auto' has neuter gender), and there are also (rare) cases of number disagreements when an organization is metonymously referred to by a plural person reference ('the GOP' - 'the Republicans').

Another problem is the fact that, in the absence of head similarity, nearly all earlier markables are possible antecedents, and all definite descriptions in a text could possibly be anaphoric. Vieira and Poesio (2000) use syntactic heuristics to see whether a definite description is unique (which means that it would always get a definite article and not necessarily be anaphoric). Unique descriptions usually only corefer with earlier mentions if they are repeated verbatim, but sometimes the inferred discourse structure and world knowledge allow or even force an interpretation where two different unique descriptions corefer[5]:

(4) a. [$_3$ Nikolaus W. Schües] bleibt Präsident der Hamburger Handelskammer.
b. [$_3$ Der Geschäftsführer der Reederei "F. Laeisz"] wurde gestern für drei Jahre wiedergewählt.

Without including detailed world knowledge, we have no chance of actually resolving example (4), whereas it should eventually be possible to resolve cases like example (3).

A useful starting point for resolving such bridging cases would be to find an antecedent that is synonymous or strictly more specific (the woman – the [female] pedestrian), or possibly semantically similar. A minimal approach to this would be to just distinguish between a fixed number of semantic classes, or even just between inanimate, animate and abstract entities as do Strube et al. (2002).

While knowledge-poor approaches usually include semantic class labels (minimally a distinction between animate and inanimate objects, which is crucial for pronoun resolution in English), it seems that these features, although very useful for the resolution of pronominal anaphora even in German, do not allow the resolution of coreferent bridging cases – Ng and Cardie (2002b) include the decision tree that their system uses in their paper and it is clear that their animacy feature is only used in the case of pronominal anaphora.

In our case, we automatically classify the markables into five semantic classes (persons, organizations, events, temporal entities and others), which we use as features (the semantic classes of anaphor and antecedent in the case of resolution, and the semantic class of the discourse-new description in the case of non-resolution). With this feature alone, no additional NPs are resolved since marking the definite NP as discourse-new is always preferred.

Quite interestingly, a version that only includes more fine-grained lexical knowledge by adding a simple graph distance measure based on the hypo-/hypernymy graph in Germanet (but no semantic classes) does not lead resolve any more anaphoric mentions (although it gets a lower perplexity), whereas a combination of the two (together with a hard recency limit of 4 sentences, number agreement, and filtering out unique descriptions with syntax-based heuristics like those put forward by Vieira and Poesio (2000)) leads to the resolution of some anaphoric mentions (see *unique_mod* and following entries in the results table). Besides just looking at the form of a (possibly anaphoric) definite NP, we also include its syntactic role to provide an approximation for its information status, as subjects are more likely carry a thematic (and thus discourse-old) referent than objects or prepositional phrases.

As a next step, we differentiate between hypernymy proper and semantical similarity, i.e. the semantical distance feature is split into one for hypernyms (where the anaphor is synonymous or more general than the antecedent), one for general semantic distance (in the case that there is no hypernymy relation between anaphor and antecedent).

[5]translation: Nikolaus W. Schües remains president of the Hamburg Chamber of Commerce. The managing director of the shipping company "F. Laeisz" was reelected for three years yesterday.

non-resolution		*resolution*	
non-resolution bias	0.26	PER→PER	1.61
new: PER,sg	–1.83	ORG→ORG	1.63
new: ORG,sg	0.16	LOC→LOC	0.92
new: LOC,sg	–0.45	EVT→EVT	0.84
new: TMP,sg	0.72	TMP→TMP	–0.18
new: SUBJ	–0.97	sentence distance	–0.75
new: PP	0.95	GWN node distance	–0.60
		GWN dist (NE)	–0.65
		Pred-Arg odds	–0.83

Table 2: Some of the constraint weights for coreferent bridging

Additionally, named entities that are not found in GermaNet are represented by a general term corresponding to the semantic class (e.g. the *person* synset for person NEs), further increasing recall by a small amount.

As a last step, we used a statistical model of selectional preferences for verbs. Using 11 millions of sentences from the German newspaper "die tageszeitung", which we automatically parsed using a PCFG parser (Versley, 2005) to get subject-verb and direct object-verb pairs, we trained models for both relations using LSC, a soft clustering software by Helmut Schmid[6] based on the Expectation Maximization algorithm.

To realize the intuition that it should be possible to exchange two coreferent descriptions against each other, while taking care of frequency differences, we used the following term:

$$q := \log \frac{p_{r'}(n_1, v_2) \cdot p_r(n_2, v_1)}{p_r(n_1, v_1) \cdot p_{r'}(n_2, v_2)}$$

If the anaphor is very likely to appear in the antecedent's context and vice versa, q should be near (or even above) zero, while a negative value of q indicates that the anaphor does not fit to the antecedent's context or vice versa.

For example, *Arbeiterwohlfahrt* (a German charity organisation) as subject of *entlassen* (to lay off) and *Mark* (currency) as subject of *fliessen* (to flow) are not exchangeable, yielding a large negative value (-5.9) since the switched version is about 370 times less likely than the original one, whereas *Siegerin* (victor) as object of *disqualifizieren* (disqualify) and a person name as subject of *landen* (to land) are well exchangeable, yielding a positive value (+1.0).

[6]http://www.ims.uni-stuttgart.de/tcl/SOFTWARE/LSC.html

The final version of our system has 70% recall, 1.4% above that of same-head resolution, and precision is lower by only 0.1%. To put this in absolute terms, 32 more definite descriptions (of 1340) have been resolved and the correct antecedent was found for 17 of them, giving a precision of 53% for coreferent bridging, which is quite near to the same-head precision for common nouns (albeit with a much lower recall).

Looking at the constraint weights from the final system (see table 2), we see that both sentence distance and semantic distance are important factors, and that there are quite large differences in the behaviours of the different semantic classes with respect to (non-)anaphoricity, with persons having a large preference to being anaphoric (both in terms of resolution, where resolving a person anaphor to a person antecedent carries a large positive weight, and in terms of introducing a new referent, where introducing a new person referent carries a large negative weight). We also see confirmed Prince (1992)'s observation that subjects tend to be discourse-old.

5 Related work

The maximum entropy framework has been used by Morton (2000) for coreference resolution in a similar setting to ours, as well as by Luo et al. (2004). Coreference resolution of German texts has been investigated by Hartrumpf (2001), who uses a mixture of hard constraints and disambiguation based on a learned backoff model, as well as Strube et al. (2002), who use a decision tree with yes/no classification. While Hartrumpf only gives quantitative results for the whole system, including pronoun resolution, a comparison is only possible with Strube's system, which has been evaluated on a different text type, with lower results than those reported here (Strube et al. give figures of F=76.2%

for named entities and F=33.9% for definite noun phrases, whereas we get F=87.5% for named entities and F=46.9% for definite noun phrases) [7].

Vieira and Poesio (2000) give a relatively detailed evaluation of the heuristics they use, not in terms of precision and recall bounds, but in terms of the influence on the performance of the whole system.

Bean and Riloff (2004) use contextual information from verbs in a more elaborate fashion than it was done here, going beyond the selectional preference model presented here, but on more restricted domains (terrorism and disasters) for testing.

6 Conclusion

We presented a coreference resolution system based on a loglinear statistical model together with hard filtering of the candidates as well as the definite descriptions themselves, yielding competitive results compared to earlier approaches to coreference resolution.

The parameter estimation method we used allows for continuous features in addition to binary ones, making possible a natural combination of sentence distance, semantic distance as well as syntactic and shallow semantic class information.

In contrast to previous work using Maximum Entropy modeling for coreference resolution (Morton, 2000; Luo et al., 2004; Uryupina, 2006), our learning algorithm is able to select its positive examples from the correct antecedents, much like Harabagiu et al. (2001)'s COCKTAIL system, and is less susceptible to training data noise due to idiosyncratic antecedents being the nearest ones.

As we currently use syntactic information from the treebank, it would be interesting to see if and by how much parsing errors influence the quality of the system output. Hartrumpf (2001) as well as Luo et al. (2004) use the weights (in both cases determined by modelling local decisions) to globally rank alternatives of multiple resolution decisions, but time as well as computational complexity have kept us from considering these.

Acknowledgements I would like to thank Sandra Kübler, Piklu Gupta and Mareile Knees for critical comments on an earlier version of this paper. The research reported here was supported as part of the DFG collaborative research centre (Sonderforschungsbereich) "SFB 441: Linguistische Datenstrukturen".

[7] Because they seem to use annotated markables for their system and we do not make use of syntactic information beyond the internal structure of markables except for the resolution of coreferent bridging, where it only has a very small influence, only the difference in text type might provide an alternative explanation for the accuracy difference.

References

Bean, D. and Riloff, E. (1999). Corpus-based identification of non-anaphoric noun phrases. In *ACL-1999*.

Bean, D. and Riloff, E. (2004). Unsupervised learning of contextual role knowledge for coreference resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*, pages 297–304.

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

Harabagiu, S., Bunescu, R., and Maiorano, S. (2001). Text and knowledge mining for coreference resolution. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association of Computational Linguistics (NAACL-2001)*, Pittsburgh.

Hartrumpf, S. (2001). Coreference resolution with syntactico-semantic rules and corpus statistics. In *Proceedings of the Fifth Conference on Natural Language Learning (CoNLL-2001)*.

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.

Liu, D. C. and Nocedal, J. (1989). On the limited memory method for large scale optimization. *Mathematical Programming B*, 45(3):503–528.

Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., and Roukos, S. (2004). A mention-synchronous coreference resolution algorithm based on the bell tree. In *ACL 2004*.

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002)*.

Morton, T. S. (2000). Coreference for NLP applications. In *ACL-2000*.

Ng, V. and Cardie, C. (2002a). Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *COLING 2002*.

Ng, V. and Cardie, C. (2002b). Improving machine learning approaches to coreference resolution. In *40th Annual Meeting of the Asssociation for Computational Linguistics*.

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*.

Prince, E. F. (1992). The ZPG letter: subjects, definiteness and information-status. In Thompson, S. and Mann, W., editors, *Discourse description: diverse analyses of a fund raising text*. John Benjamins B.V.

Strube, M., Rapp, S., and Müller, C. (2002). The influence of minimum edit distance on reference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pages 312–319.

Telljohann, H., Hinrichs, E. W., and Kübler, S. (2003). Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical report, Seminar für Sprachwissenschaft, Universität Tübingen.

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

Uryupina, O. (2006). Coreference resolution with and without linguistic knowledge. In *LREC 2006*.

Versley, Y. (2005). Parser evaluation across text types. In *Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*.

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