Optimizing Algorithms for Pronoun Resolution

Michael Schiehlen*
Institute for Computational Linguistics
University of Stuttgart
Azenbergstraße 12, D-70174 Stuttgart
mike@ims.uni-stuttgart.de

Abstract

The paper aims at a deeper understanding of several well-known algorithms and proposes ways to optimize them. It describes and discusses factors and strategies of factor interaction used in the algorithms. The factors used in the algorithms and the algorithms themselves are evaluated on a German corpus annotated with syntactic and coreference information (Negra) (Skut et al., 1997). A common format for pronoun resolution algorithms with several open parameters is proposed, and the parameter settings optimal on the evaluation data are given.

1 Introduction

In recent years, a variety of approaches to pronoun resolution have been proposed. Some of them are based on centering theory (Strube, 1998; Strube and Hahn, 1999; Tetreault, 2001), others on Machine Learning (Aone and Bennett, 1995; Ge et al., 1998; Soon et al., 2001; Ng and Cardie, 2002; Yang et al., 2003). They supplement older heuristic approaches (Hobbs, 1978; Lappin and Leass, 1994). Unfortunately, most of these approaches were evaluated on different corpora making different assumptions so that direct comparison is not possible. Appreciation of the new insights is quite hard. Evaluation differs not only with regard to size and genre of corpora but also along the following lines.

Scope of application: Some approaches only deal with personal and possessive pronouns (centering and heuristic), while others consider coreference links in general (Soon et al., 2001; Ng and Cardie, 2002; Yang et al., 2003). A drawback of this latter view is that it mixes problems on different levels of difficulty. It remains unclear how much of the success is due to the virtues of the approach and how much is due to the distribution of hard and easy problems in the corpus. In this paper, we will only deal with coreferential pronouns (i.e. possessive, demonstrative, and third person pronouns).

* My thanks go to Melvin Wurster for help in annotation and to Ciprian Gerstenberger for discussion.

Quality of linguistic input: Some proposals were evaluated on hand annotated (Strube and Hahn, 1999) or tree bank input (Ge et al., 1998; Tetreault, 2001). Other proposals provide a more realistic picture in that they work as a backend to a parser (Lappin and Leass, 1994) or noun chunker (Mitkov, 1998; Soon et al., 2001; Ng and Cardie, 2002)). In evaluation of applications presupposing parsing, it is helpful to separate errors due to parsing from intrinsic errors. On the other hand, one would also like to gauge the end-to-end performance of a system. Thus we will provide performance figures for both ideal (hand-annotated) input and realistic (automatically generated) input.

Language: Most approaches were evaluated on English where large resources are available, both in terms of pre-annotated data (MUC-6 and MUC-7 data) and lexical information (WordNet). This paper deals with German. Arguably, the free word-order of German arguably leads to a clearer distinction between grammatical function, surface order, and information status (Strube and Hahn, 1999).

The paper is organized as follows. Section 2 describes the evaluation corpus. Section 3 describes several factors relevant to pronoun resolution. It assesses these factors against the corpus, measuring their precision and restrictiveness. Section 4 describes and evaluates six algorithms on the basis of these factors. It also captures the algorithms as parametric systems and proposes parameter settings optimal on the evaluation data. Section 5 concludes.

2 Evaluation Corpus

We chose as an evaluation base the NEGRA tree bank, which contains about 350,000 tokens of German newspaper text. The same corpus was also processed with a finite-state parser, performing at 80% dependency f-score (Schiehlen, 2003).

All personal pronouns (PPER), possessive pronouns (PPOSAT), and demonstrative pronouns (PDS) in Negra were annotated in a format geared to the MUC-7 guidelines (MUC-7, 1997). Proper

names were annotated automatically by a named entity recognizer. In a small portion of the corpus (6.7%), all coreference links were annotated. Thus the size of the annotated data (3,115 personal pronouns[1], 2,198 possessive pronouns, 928 demonstrative pronouns) compares favourably with the size of evaluation data in other proposals (619 German pronouns in (Strube and Hahn, 1999), 2,477 English pronouns in (Ge et al., 1998), about 5,400 English coreferential expressions in (Ng and Cardie, 2002)).

In the experiments, systems only looked for single NP antecedents. Hence, propositional or predicative antecedents (8.4% of the pronouns annotated) and split antecedents (0.2%) were inaccessible, which reduced optimal success rate to 91.4%.

3 Factors in Pronoun Resolution

Pronoun resolution is conditioned by a wide range of factors. Two questions arise: Which factors are the most effective? How is interaction of the factors modelled? The present section deals with the first question, while the second question is postponed to section 4.

Many approaches distinguish two classes of resolution factors: *filters* and *preferences*. Filters express linguistic rules, while preferences are merely tendencies in interpretation. Logically, filters are monotonic inferences that select a certain subset of possible antecedents, while preferences are non-monotonic inferences that partition the set of antecedents and impose an order on the cells.

In the sequel, factors proposed in the literature are discussed and their value is appraised on evaluation data. Every factor narrows the set of antecedents and potentially discards correct antecedents. Table 1 lists both the success rate maximally achievable (broken down according to different types of pronouns) and the average number of antecedents remaining after applying each factor. Figures are also given for parsed input. Preferences are evaluated on filtered sets of antecedents.

3.1 Filters

Agreement. An important filter comes from morphology: Agreement in gender and number is generally regarded as a prerequisite for coreference. Exceptions are existant but few (2.5%): abstract pronouns (such as *that* in English) referring to non-neuter or plural NPs, plural pronouns co-referring with singular collective NPs (Ge et al., 1998), antecedent and anaphor matching in natural gender rather than grammatical gender. All in all, a maximal performance of 88.9% is maintained. The filter is very restrictive, and cuts the set of possible antecedents in half. See Table 1 for details.

Binding. Binding constraints have been in the focus of linguistic research for more than thirty years. They provide restrictions on co-indexation of pronouns with clause siblings, and therefore can only be applied with systems that determine clause boundaries, i.e. parsers (Mitkov, 1998). Empirically, binding constraints are rules without exceptions, hence they do not lead to any loss in achievable performance. The downside is that their restrictive power is quite bad as well (0.3% in our corpus, cf. Table 1).

Sortal Constraints. More controversial are sortal constraints. Intuitively, they also provide a hard filter: The correct antecedent must fit into the environment of the pronoun (Carbonell and Brown, 1988). In general, however, the required knowledge sources are lacking, so they must be hand-coded and can only be applied in restricted domains (Strube and Hahn, 1999). Selectional restrictions can also be modelled by collocational data extracted by a parser, which have, however, only a very small impact on overall performance (Kehler et al., 2004). We will neglect sortal constraints in this paper.

3.2 Preferences

Preferences can be classified according to their requirements on linguistic processing. Sentence Recency and Surface Order can be read directly off the surface. NP Form presupposes at least tagging. A range of preferences (Grammatical Roles, Role Parallelism, Depth of Embedding, Common Path), as well as all filters, presuppose full syntactic analysis. Mention Count and Information Status are based on previous decisions of the anaphora resolution module.

Sentence Recency (SR). The most important criterion in pronoun resolution (Lappin and Leass, 1994) is the textual distance between anaphor and antecedent measured in sentences. Lappin and Leass (1994) motivate this preference as a dynamic expression of the attentional state of the human hearer: Memory capability for storage of discourse referents degrades rapidly.

Several implementations are possible. Perhaps most obvious is the strategy implicit in Lappin and Leass (1994)'s algorithm: The antecedent is searched in a sentence that is as recent as possible, beginning with the already uttered part of the current sentence, continuing in the last sentence, in the one but last sentence, and so forth. In case no

[1]Here, we only count anaphoric pronouns, i.e. third person pronouns not used expletively.

Constraint		Upper Bound				∅ number	Parser	
		total	PPER	PPOSAT	PDS	of antec.	UpperB	antec.
no VP		91.6	98.4	100.0	48.5	123.2	85.5	128.4
no split		91.4	98.3	100.0	47.8	123.2		
agreement		88.9	96.8	99.5	37.6	53.0	79.1	61.8
binding		88.9				52.7	78.7	61.4
sentence recency	SR	78.8	84.6	90.2	32.3	2.4	66.2	2.7
grammatical role	GR	74.0	82.32	87.9	13.0	14.5	51.2	9.0
role parallelism	RP	64.3	77.4	–	20.0	12.5	47.0	10.3
surface order →	LR	53.5	62.8	56.6	15.3	1	42.6	1
surface order ←	RL	45.9	45.9	55.7	22.7	1	35.2	1
depth of embedding	DE	51.6	51.3	67.7	14.1	2.4	41.7	4.0
common path	CP	51.7	52.3	64.2	19.9	5.3	46.8	11.3
equivalence classes	EQ	63.6	67.5	78.4	15.7	1.3	51.3	1.5
mention count	MC	32.9	40.3	34.0	4.6	5.5	35.7	7.1
information status	IS	65.3	71.1	77.4	16.7	16.6	49.7	16.3
NP form	NF	42.4	49.9	44.4	12.8	7.4	20.6	8.3
NP form (pronoun)	NP	73.7	82.4	79.8	30.2	29.7	59.7	36.6

Table 1: Effect of Factors

antecedent is found in the previous context, subsequent sentences are inspected (cataphora), also ordered by proximity to the pronoun.

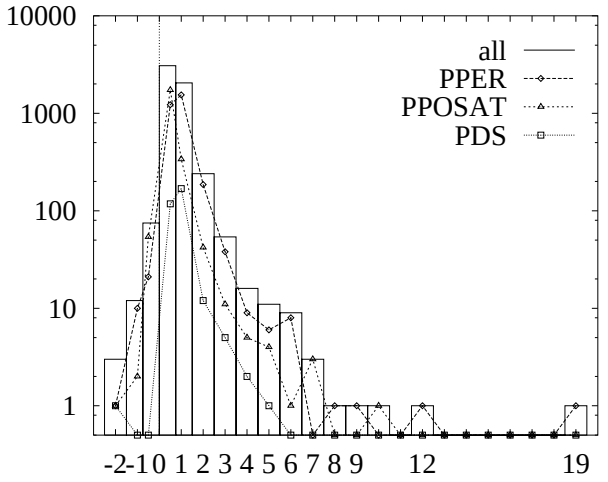

Figure 1: Sentence Recency

Figure 1 shows the absolute frequencies of sentence recency values when only the most recent antecedent (in the order just stated) is considered. In Negra, 55.3% of all pronominal anaphora can be resolved intrasententially, and 97.6% within the last three sentences. Since only 1.6% of all pronouns are cataphoric, it seems reasonable to neglect cataphora, as is mostly done (Strube and Hahn, 1999; Hobbs, 1978). Table 1 underscores the virtues of Sentence Recency: In the most recent sentence with antecedents satisfying the filters, there are on average only 2.4 such antecedents. However, the benefit also comes at a cost: The upper ceiling of performance is lowered to 82.0% in our corpus: In many cases an incorrect antecedent is found in a more recent sentence.

Similarly, we can assess other strategies of sentence ordering that have been proposed in the literature. Hard-core centering approaches only deal with the last sentence (Brennan et al., 1987). In Negra, these approaches can consequently have at most a success rate of 44.2%. Performance is particularly low with possessive pronouns which often only have antecedents in the current sentence. Strube (1998)'s centering approach (whose sentence ordering is designated as SR2 in Table 2) also deals with and even prefers intrasentential anaphora, which raises the upper limit to a more acceptable 80.2%. Strube and Hahn (1999) extend the context to more than the last sentence, but switch preference order between the last and the current sentence so that an antecedent is determined in the last sentence, whenever possible. In Negra, this ordering imposes an upper limit of 51.2%.

Grammatical Roles (GR). Another important factor in pronoun resolution is the grammatical role of the antecedent. The role hierarchy used in centering (Brennan et al., 1987; Grosz et al., 1995) ranks subjects over direct objects over indirect objects over others. Lappin and Leass (1994) provide a more elaborate model which ranks NP complements and NP adjuncts lowest. Two other distinctions in

their model express a preference of rhematic[2] over thematic arguments: Existential subjects, which follow the verb, rank very high, between subjects and direct objects. Topic adjuncts in pre-subject position separated by a comma rank very low, between adjuncts and NP complements. Both positions are not clearly demarcated in German. When the Lappin&Leass hierarchy is adopted to German without changes, a small drop in performance results as compared with the obliqueness hierarchy used in centering. So we will use the centering hierarchy. Table 1 shows the effect of the role-based preference on our data. The factor is both less restrictive and less precise than sentence recency.

The definition of a grammatical role hierarchy is more involved in case of automatically derived input, as the parser cannot always decide on the grammatical role (determining grammatical roles in German may require world knowledge). It proposes a syntactically preferred role, however, which we will adopt.

Role Parallelism (RP). Carbonell and Brown (1988) argue that pronouns prefer antecedents in the same grammatical roles. Lappin and Leass (1994) also adopt such a principle. The factor is, however, not applicable to possessive pronouns.

Again, role ambiguities make this factor slightly problematic. Several approaches are conceivable: Antecedent and pronoun are required to have a common role in one reading (weak match). Antecedent and pronoun are required to have the same role in the reading preferred by surface order (strong match). Antecedent and pronoun must display the same role ambiguity (strongest match). Weak match restricted performance to 49.9% with 12.1 antecedents on average. Strong match gave an upper limit of 47.0% but with only 10.3 antecedents on average. Strongest match lowered the upper limit to 43.1% but yielded only 9.3 antecedents. In interaction, strong match performed best, so we adopt it.

Surface Order (LR, RL). Surface Order is usually used to bring down the number of available antecedents to one, since it is the only factor that produces a unique discourse referent. There is less consensus on the preference order: (sentence-wise) left-to-right (Hobbs, 1978; Strube, 1998; Strube and Hahn, 1999; Tetreault, 1999) or right-to-left (recency) (Lappin and Leass, 1994). Furthermore, something has to be said about antecedents which embed other antecedents (e.g. conjoined NPs and their conjuncts). We registered performance gains (of up to 3%) by ranking embedding antecedents higher than embedded ones (Tetreault, 2001).

Left-to-right order is often used as a surrogate for grammatical role hierarchy in English. The most notable exception to this equivalence are fronting constructions, where grammatical roles outperform surface order (Tetreault, 2001). A comparison of the lines for grammatical roles and for surface order in Table 1 shows that the same is true in German.

Left-to-right order performs better (upper limit 56.8%) than right-to-left order (upper limit 49.2%). The gain is largely due to personal pronouns; demonstrative pronouns are better modelled by right-to-left order. It is well-known that German demonstrative pronouns contrast with personal pronouns in that they function as topic-shifting devices. Another effect of this phenomenon is the poor performance of the role preferences in connection with demonstrative pronouns.

Depth of Embedding (DE). A prominent factor in Hobbs (1978)'s algorithm is the level of phrasal embedding: Hobbs's algorithm performs a breadth-first search, so antecedents at higher levels of embedding are preferred.

Common Path (CP). The syntactic version of Hobbs (1978)'s algorithm also assumes maximization of the common path between antecedents and anaphors as measured in NP and S nodes. Accordingly, intra-sentential antecedents that are syntactically nearer to the pronoun are preferred. The factor only applies to intrasentential anaphora.

The anaphora resolution module itself generates potentially useful information when processing a text. Arguably, discourse entities that have been often referred to in the previous context are topical and more likely to serve as antecedents again. This principle can be captured in different ways.

Equivalence Classes (EQ). Lappin and Leass (1994) make use of a mechanism based on equivalence classes of discourse referents which manages the attentional properties of the individual entities referred to. The mechanism stores and provides information on how recently and in which grammatical role the entities were realized in the discourse. The net effect of the storage mechanism is that discourse entities are preferred as antecedents if they recently came up in the discourse. But the mechanism also integrates the preferences Role Hierarchy and Role Parallelism. Hence, it is one of the best-performing factors on our data. Since the equivalence class scheme is tightly integrated in the parser, the problem of ideal anaphora resolution data does not arise.

[2]Carbonell and Brown (1988) also argue that clefted or fronted arguments should be preferred.

Mention Count (MC). Ge et al. (1998) try to factorize the same principle by counting the number of times a discourse entities has been mentioned in the discourse already. However, they do not only train but also test on the manually annotated counts, and hence presuppose an optimal anaphora resolution system. In our implementation, we did not bother with intrasentential mention count, which depends on the exact traversal. Rather, mention count was computed only from previous sentences.

Information Status (IS). Strube (1998) and Strube and Hahn (1999) argue that the information status of an antecedent is more important than the grammatical role in which it occurs. They distinguish three levels of information status: entities known to the hearer (as expressed by coreferential NPs, unmodified proper names, appositions, relative pronouns, and NPs in titles), entities related to such hearer-old entities (either overtly via modifiers or by bridging), and entities new to the hearer. Like (Ge et al., 1998), Strube (1998) evaluates on ideal hand annotated data.

NP Form (NF, NP). A cheap way to model information status is to consider the form of an antecedent (Tetreault, 2001; Soon et al., 2001; Strube and Müller, 2003). Personal and demonstrative pronouns are necessarily context-dependent, and proper nouns are nearly always known to the hearer. Definite NPs may be coreferential or interpreted by bridging, while indefinite NPs are in their vast majority new to the hearer. We considered two proposals for orderings of form: preferring pronouns and proper names over other NPs over indefinite NPs (Tetreault, 2001) (NF) or preferring pronouns over all other NPs (Tetreault, 2001) (NP).

4 Algorithms and Evaluation

In this section, we consider the individual approaches in more detail, in particular we will look at their choice of factors and their strategy to model factor interaction. According to interaction potential, we distinguish three classes of approaches: Serialization, Weighting, and Machine Learning.

We re-implemented some of the algorithms described in the literature and evaluated them on syntactically ideal and realistic German[3] input. Evaluation results are listed in Table 2.

With the ideal treebank input, we also assumed ideal input for the factors dependent on previous anaphora resolution results. With realistic parsed input, we fed the results of the actual system back into the computation of such factors.

4.1 Serialization Approaches

Algorithmical approaches first apply filters unconditionally; possible exceptions are deemed nonexistant or negligible. With regard to interaction of preferences, many algorithms (Hobbs, 1978; Strube, 1998; Tetreault, 2001) subscribe to a scheme, which, though completely rigid, performs surprisingly well: The chosen preferences are applied one after the other in a certain pre-defined order. Application of a preference consists in selecting those of the antecedents still available that are ranked highest in the preference order.

Hobbs (1978)'s algorithm essentially is a concatenation of the preferences Sentence Recency (without cataphora), Common Path, Depth of Embedding, and left-to-right Surface Order. It also implements the binding constraints by disallowing sibling to the anaphor in a clause or NP as antecedents. Like Lappin and Leass (1994), we replaced this implementation by our own mechanism to check binding constraints, which raised the success rate.

The Left-Right Centering algorithm of Tetreault (1999) is similar to Hobbs's algorithm, and is composed of the preferences Sentence Recency (without cataphora), Depth of Embedding, and left-to-right Surface Order. Since it is a centering approach, it only inspects the current and last sentence.

Strube (1998)'s S-list algorithm is also restricted to the current and last sentence. Predicative complements and NPs in direct speech are excluded as antecedents. The primary ordering criterion is Information Status, followed by Sentence Recency (without cataphora) and left-to-right Surface Order.

Since serialization provides a quite rigid frame, we conducted an experiment to find the best performing combination of pronoun resolution factors on the treebank and the best combination on the parsed input. For this purpose, we checked all permutations of preferences and subtracted preferences from the best-performing combinations until performance degraded (greedy descent). Greedy descent outperformed hill-climbing. The completely annotated 6.7% of the corpus were used as development set, the rest as test set.

4.2 Weighting Approaches

Compared with the serialization approaches, the algorithm of Lappin and Leass (1994) is more sophisticated: It uses a system of hand-selected weights to control interaction among preferences, so that in principle the order of preference application can

[3] A reviewer points out that most of the algorithms were proposed for English, where they most likely perform better. However, the algorithms also incorporate a theory of saliency, which should be language-independent.

Algorithm	Definition	F-Scores – treebank				F-Score
		total	PPER	PPOSAT	PDS	Parser
(Hobbs, 1978)	SR∘CP∘DE∘LR	59.9	65.1	70.5	17.4	45.4
(Tetreault, 1999)	SR2∘DE∘LR	57.0	64.1	61.9	17.2	43.3
(Strube, 1998)	IS∘SR2∘LR	57.9	65.9	63.7	12.0	39.1
optimal algor. (treebank)	SR∘CP∘IS∘DE∘MC∘RP∘GR∘RL	70.4	75.6	82.0	22.7	43.7
optimal algor. (parsed)	SR∘CP∘GR∘IS∘DE∘LR	67.7	74.3	82.0	10.6	50.6
(Lappin and Leass, 1994)	EQ∘SR∘RL	65.4	71.0	78.0	16.6	50.8
(Ge et al., 1998)	Hobbs+MC	43.4	45.7	53.6	12.1	36.3
(Soon et al., 2001)	(SR+NP)∘RL	24.8	30.8	23.6	0.0	26.8
optimal algor. (C4.5)	(SR/RL+GR+NF/IS)∘RL	71.1	78.2	79.0	9.8	51.7

Table 2: Performance of Algorithms

switch under different input data. In the actual realization, however, the weights of factors lie so much apart that in the majority of cases interaction boils down to serialization. The weighting scheme includes Sentence Recency, Grammatical Roles, Role Parallelism, on the basis of the equivalence class approach described in section 3.2. Final choice of antecedents is relegated to right-to-left Surface Order.

Interestingly, the Lappin&Leass algorithm outperforms even the best serialization algorithm on parsed input.

4.3 Machine Learning Approaches

Machine Learning approaches (Ge et al., 1998; Soon et al., 2001; Ng and Cardie, 2002) do not distinguish between filters and preferences. They submit all factors as features to the learner. For every combination of feature values the learner has the freedom to choose different factors and to assign different strength to them.

Thus the main problem is not choice and interaction of factors, but rather the formulation of anaphora resolution as a classification problem. Two proposals emerge from the literature. (1) Given an anaphor and an antecedent, decide if the antecedent is the correct one (Ge et al., 1998; Soon et al., 2001; Ng and Cardie, 2002). (2) Given an anaphor and two antecedents, decide which antecedent is more likely to be the correct one (Yang et al., 2003). In case (1), the lopsidedness of the distribution is problematic: There are much more negative than positive training examples. Machine Learning tools have to surpass a very high baseline: The strategy of never proposing an antecedent typically already yields an f-score of over 90%. In case (2), many more correct decisions have to be made before a correct antecedent is found. Thus it is important in this scenario, that the set of antecedents is subjected to a strict filtering process in advance so that the system only has to choose among the best candidates and errors are less dangerous.

Ge et al. (1998)'s probabilistic approach combines three factors (aside from the agreement filter): the result of the Hobbs algorithm, Mention Count dependent on the position of the sentence in the article, and the probability of the antecedent occurring in the local context of the pronoun. In our re-implementation, we neglected the last factor (see section 3.1). Evaluation was performed using 10-fold cross validation.

Other Machine Learning approaches (Soon et al., 2001; Ng and Cardie, 2002; Yang et al., 2003) make use of decision tree learning[4]; we used C4.5 (Quinlan, 1993). To construct the training set, Soon et al. (2001) take the nearest correct antecedent in the previous context as a positive example, while all possible antecedents between this antecedent and the pronoun serve as negative examples. For testing, potential antecedents are presented to the classifier in Right-to-Left order; the first one classified positive is chosen. Apart from agreement, only two of Soon et al. (2001)'s features apply to pronominal anaphora: Sentence Recency, and NP Form (with personal pronouns only). We used every 10th sentence in Negra for testing, all other sentences for training. On parsed input, a very simple decision tree is generated: For every personal and possessive pronoun, the nearest agreeing pronoun is chosen as antecedent; demonstrative pronouns never get an antecedent. This tree performs better than the more complicated tree generated from treebank input, where also non-pronouns in previous sentences can serve as antecedents to a personal pronoun.

Soon et al. (2001)'s algorithm performs below its potential. We modified it somewhat to get better results. For one, we used every possible antecedent

[4]On our data, Maximum Entropy (Kehler et al., 2004) had problems with the high baseline, i.e. proposed no antecedents.

in the training set, which improved performance on the treebank set (by 1.8%) but degraded performance on the parsed data (by 2%). Furthermore, we used additional features, viz. the grammatical role of antecedent and pronoun, the NP form of the antecedent, and its information status. The latter two features were combined to a single feature with very many values, so that they were always chosen first in the decision tree. We also used fractional numbers to express intrasentential word distance in addition to Soon et al. (2001)'s sentential distance. Role Parallelism (Ng and Cardie, 2002) degraded performance (by 0.3% F-value). Introducing agreement as a feature had no effect, since the learner always determined that mismatches in agreement preclude coreference. Mention Count, Depth of Embedding, and Common Path did not affect performance either.

5 Conclusion

The paper has presented a survey of pronoun resolution factors and algorithms. Two questions were investigated: Which factors should be chosen, and how should they interact? Two types of factors, 'filters' and 'preferences', were discussed in detail. In particular, their restrictive potential and effect on success rate were assessed on the evaluation corpus. To address the second question, several well-known algorithms were grouped into three classes according to their solution to factor interaction: Serialization, Weighting, and Machine Learning. Six algorithms were evaluated against a common evaluation set so as to facilitate direct comparison. Different algorithms have different strengths, in particular as regards their robustness to parsing errors. Two of the interaction strategies (Serialization and Machine Learning) allow data-driven optimization. Optimal algorithms could be proposed for these strategies.

References

Chinatsu Aone and Scott William Bennett. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In ACL'95, pages 122–129, Cambridge, MA.

Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In ACL'87, pages 155–162, Stanford, CA.

Jaime G. Carbonell and Ralph D. Brown. 1988. Anaphora resolution: A multi-strategy approach. In COLING '88, pages 96–101.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–170.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2):203–225.

Jerry R. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44:311–338.

Andrew Kehler, Douglas Appelt, Lara Taylor, and Aleksandr Simma. 2004. The (Non)Utility of Predicate-Argument Frequencies for Pronoun Interpretation. In *Proceedings of the 2nd HLT/NAACL*, Boston, MA.

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

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In COLING '98, pages 869–875, Montreal, Canada.

MUC-7. 1997. Coreference task definition. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In ACL'02, pages 104–111, Philadelphia, PA.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Michael Schiehlen. 2003. Combining Deep and Shallow Approaches in Parsing German. In ACL'03, pages 112–119, Sapporo, Japan.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An Annotation Scheme for Free Word Order Languages. In ANLP-97, Washington, DC.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544.

Michael Strube and Udo Hahn. 1999. Functional Centering – Grounding Referential Coherence in Information Structure. *Computational Linguistics*, 25(3):309–344.

Michael Strube and Christoph Müller. 2003. A Machine Learning Approach to Pronoun Resolution in Spoken Dialogue. In ACL'03, pages 168–175, Sapporo, Japan.

Michael Strube. 1998. Never look back: An alternative to Centering. In COLING '98, pages 1251–1257, Montreal, Canada.

Joel R. Tetreault. 1999. Analysis of Syntax-Based Pronoun Resolution Methods. In ACL'99, pages 602–605, College Park, MA.

Joel R. Tetreault. 2001. A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4):507–520.

Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. 2003. Coreference Resolution Using Competition Learning Approach. In ACL'03, pages 176–183, Sapporo, Japan.