Instance Sampling Methods for Pronoun Resolution

Holger Wunsch
University of Tübingen
wunsch@sfs.uni-tuebingen.de

Sandra Kübler
Indiana University
skuebler@indiana.edu

Rachael Cantrell
Indiana University
rcantrel@indiana.edu

Abstract

Instance sampling is a method to balance extremely skewed training sets as they occur, for example, in machine learning settings for anaphora resolution. Here, the number of negative samples (i.e. *non-anaphoric* pairs) is usually substantially larger than the number of positive samples. This causes classifiers to be biased towards negative classification, leading to suboptimal performance.

In this paper, we explore how different techniques of instance sampling influence the performance of an anaphora resolution system for German given different classifiers. All sampling methods prove to increase the F-score for all classifiers, but the most successful method is random sampling. In the best setting, the F-score improves from 0.541 to 0.608 for memory-based learning, from 0.561 to 0.611 for decision tree learning and from 0.511 to 0.584 for maximum entropy learning.

1 Introduction

Machine learning approaches to anaphora resolution are generally defined as deciding, for a pair consisting of a pronoun and a possible antecedent (*markable*), whether or not they share an anaphoric relation. In the sentence "When the car hit the tree in the dark, it lost a tire.", for example, the task is to decide whether the pronoun *it* refers to one of the noun phrases *the car*, *the tree*, *the dark*, or to any of the noun phrases in the preceding sentences in the text. This means that it is paired with each of these noun phrases individually, and the classifier decides for each case whether there is an anaphoric relationship between the two. Training data is produced in the same way. This results in a highly skewed class distribution with many more negative examples than positive ones. For example, Zhao and Ng [17] report a ratio of positive and negative examples of 1:29 for their Chinese data set. Ng and Cardie [7] report that in the MUC-6 data set, only 2% of the pairs are positive examples, all the others are negative (approximate ratio: 1:48); for MUC-7, there are 3% positive examples (approximate ratio: 1:48.5). Such an extreme skewedness of the data set tends to cause suboptimal performance in machine learning approaches. For this reason, instance sampling is often used in order to create a more balanced training set. In our case, this means restricting the number of negative examples while the positive ones remain unchanged. In the case of coreference resolution for definite noun phrases, in which case all preceding markables of a coreference chain are used for positive examples, sampling those positive examples can have a positive effect, too, as Ng and Cardie [7] show.

One often-used linguistically motivated approach to restrict the negative examples is to use only the markables between the pronoun and the actual antecedent. (cf. for example [7, 10]) Another possibility is to sample the negative examples randomly until a certain, predefined ratio is reached. To our knowledge, no systematic comparison of sampling methods has been performed. This is the aim of the work presented here. For these experiments, we used a system for pronoun resolution for German [3, 15] as the basis. The system combines a rule-based morphological pre-filter with a pronoun resolution module based on a classifier. In the original system (cf. section 5), memory-based learning was used. For this reason, we first investigated the full range of sampling methods considered here using the memory-based system. In a second round of experiments, we used the two most successful sampling methods, online sampling and random sampling, on a range of classifiers. This shows whether the success of different sampling methods is dependent on the classifier or whether there are general trends.

In the remainder of this paper, we first present the full range of sampling methods (section 3), then the data set (section 4) and the original pronoun resolution system used for the comparison (section 5). In section 6, we present and discuss the results for the following classifiers: a memory-based classifier, a decision tree classifier, and a maximum entropy classifier.

2 Pronoun resolution: Task description

The first step in pronoun resolution is a syntactic analysis, which provides the pronouns and their possible antecedents, so-called *markables*. Since we are only interested in the influence of instance sampling, we used gold standard data to identify the pronouns and the markables. The syntactic information as well as the referential information is taken from the *Tübingen Treebank of Written German, TüBa-D/Z* [11] (for more details cf. section 4).

In machine learning approaches, the task of pronoun resolution is normally defined as a classification task. Normally, this leads to the approach of pairing each anaphoric pronoun with all markables preceding it in a certain window in turn. Then for each pronoun-markable pair, the classifier decides whether there is an anaphoric relation between the two (cf. e.g. [9, 10]). We follow this approach.

International Conference RANLP 2009 - Borovets, Bulgaria, pages 478–483

German is morphologically richer than English, for which most of the work has been done, and it possesses grammatical gender. Pronouns must agree with their antecedent in gender and number. For this reason, it is effective to apply a morphological filter before using the classifier. This can lead to a considerable reduction of suitable markables, as can be seen in example (1). In this example, the relative pronoun *das* can only refer to *dem Auto (the car)* (since both are neuter), and the personal pronoun *sie* only to *die Frau (the woman)* (since both are feminine).

(1) Die Frau nimmt den Ball aus dem
The woman(f) takes the ball(m) from the
Auto, das sie gekauft hatte.
car(n), which(n) she(f) bought had.
"The woman takes the ball out of the car, which she had bought."

Morphological filtering can reduce the number of pronoun-markable pairs to about half the size of the original set. However, there are exceptions to this agreement restriction. One example can be found in the sentences *Finnland schlägt Schweden. Die Finnen haben gezeigt, daß sie spielen können.* (Finland beats Sweden. The Finns have shown that they know how to play.) where the plural personal pronoun *sie* is coreferent to *Finnland (Finland)* and to *die Finnen (the Finns)*. However, *Finnland* is singular, thus it is not compatible in number with the plural pronoun. This means that the morphological filter will exclude some of the correct pairs, thus lowering the upper limit for the machine learning approach to 95.22%.

3 Instance sampling methods

The goal of instance sampling is to reduce the number of negative examples in the training set in order to reach a more balanced ratio of positive and negative examples. Since in a highly skewed training set, the number of positive examples is low, the classifier will choose a positive answer only in a few clear cases. Thus, precision is high, but recall is low: The anaphoric relations suggested are fairly reliable, but the system finds only a subset of all relations. When we use sampling techniques, the set of positive examples remains unchanged, but we reduce the number of negative examples. We expect sampling to increase recall but also to decrease precision. We used four different sampling methods:

Local sampling is based on the intuition that anaphoric relations are closely tied to proximity: On the one hand, two entities are more likely to share an anaphoric relation if they are closer, but on the other hand, negative samples close to the pronoun are thought to be especially informative on which configurations lead to no relation, in spite of proximity. We follow Soon et al. [10] in restricting the negative examples to a linguistically defined context: Given a pair of a pronoun and a correct antecedent, we include as negative samples in the training data only those non-coreferent pairs that are located *between* the pronoun and the correct antecedent. This sampling method resulted in a sampling ratio of 1:2.1 for our data set (as compared to 1:4.29 for the whole data set).

Distance sampling tests whether a negative example is especially useful if it is very close or very far from positive examples in the search space. For this sampling method, we trained the memory-based classifier on the positive examples only, using the optimal feature settings and the feature weights from the baseline experiment without sampling. Then we classified all the negative examples from the original training set against the positive examples and looked at their distances to the closest positive example. We then selected only those negative examples that had a distance greater than 0.002[1]. The distance was chosen to reach a sampling ratio close to 1:2 (close to the ratio of the other techniques). The actual ratio with distance sampling for our data set is 1:1.82.

Incremental Learning (IB2) is a modification of the standard memory-based learning algorithm by Aha et al. [1], in which the examples are presented incrementally, and only those examples are kept for the training set that are misclassified by the current training set. While this method was originally devised for memory-based learning, it can be used for any supervised machine learning paradigm. In our case, we use a slight modification of the algorithm, in which we keep all the positive examples and add the negative ones incrementally. This is performed by a training regime in which each new example is tested against the current training set and added only when it is misclassified. Since the sampling is dependent on the individual classifier's decisions, this sampling method was carried out individually for each of the classifiers used in section 6.3, thus resulting in different sampling ratios for the individual classifiers. This method results in a comparatively low sampling ratio of 1:0.96 for memory-based learning, 1:0.83 for decision-tree learning, and 1:0.70 for maximum entropy learning.

Random Sampling is a method in which first the ratio is determined, then negative examples are randomly chosen (without replacement) until the ratio is reached. This method has been used successfully, for example, by Zhao and Ng [17] for Chinese zero pronoun resolution. Since our data was prefiltered by a morphological filter, our original ratio was considerable lower than theirs: We used sampling ratios between 1:1 and 1:4.29 (the full data set).

4 The data

Since we are only interested in the influence of instance sampling, we used gold standard data for the syntactic identification of the pronouns and the possible antecedents (*markables*). As our gold data source, we used the newspaper corpus *Tübingen Treebank of Written German (TüBa-D/Z)* [12], which is based on German newspaper articles from the newspaper *die tageszeitung (taz)*. The treebank contains 27 125 sentences in version 3. TüBa-D/Z is manually annotated syntactically as well as for referential relations. On the latter level, the treebank encodes the relations of coreference and anaphora between nominal antecedents and

[1] We also experimented with the complement, which resulted in inferior F-scores.

```
refl,OD,ON,proper,def,top,ana,diff,loc,-2,yes
refl,OD,HD,common,na,top,cata,diff,loc,-2,no
```

Fig. 1: *Feature vectors for the positive pair "sich – die AWO" and the negative pair "sich – Seniorenreisen"*

definite noun phrases and pronouns, respectively. The treebank provides full coreference chains. However, we follow standard practice in anaphora resolution and consider the closest coreferent markable as the sole antecedent of a pronoun. For pronoun resolution, the only two of the six categories in TüBa-D/Z that are relevant are *anaphoric* and *cataphoric*. A complete description of the annotation scheme for the referential annotation can be found in the annotation manual [5].

In our experiments, we only consider third person reflexive, possessive, and personal pronouns. These three types together make up 53% of all the 44 424 pronouns in the treebank; other frequent pronouns are demonstrative and relative pronouns. Out of the set of pronouns treated here, personal pronouns constitute the largest subset with 54.3%, followed by possessive pronouns with 23.4% and reflexive pronouns with 22.3%. We consider all noun phrases annotated in the treebank, including pronouns, to be markables. Each pronoun is paired with all markables in a context of three sentences previous to the pronoun. This results in 661 205 pronoun-markable pairs.

5 System description

We use a hybrid approach to pronoun resolution, combining a rule-based morphological filter with a machine learning resolution module. The system, which serves as the baseline system, was fully optimized, including an attempt to switch to a pairwise competition model. More details can be found in [16]. The modules operate sequentially on the core data set, the set of pairs of pronouns and candidate antecedents.

The morphological filter removes pronoun-markable pairs that do not agree in gender and number from further processing. In example (1), the relative pronoun *das* can only refer to the car (since both are neuter), and the personal pronoun *sie* only to the woman (since both are feminine). All other pairs are removed from the set.

The morphological filter removes 358 843 morphologically incompatible pairs, reducing the set to 46% of its original size.

The pronoun resolution module uses a trainable classifier to decide on the pairs that remain after morphological filtering. The task of pronoun resolution is reformulated as a binary classification task: a pair of a candidate antecedent and a pronoun is assigned one of the two possible classes: *anaphoric* or *not anaphoric*. For each pair in the pre-filtered list of candidates, a set of features is extracted and then bundled in a feature vector. The most informative features (determined in a non-exhaustive search) are listed in detail in Table 1. Figure 1 shows two feature vectors that correspond to two candidate pairs in the following sentence:

Feature	Description
PronType	pronoun type
PronGF	gramm. function of pronoun
NPGF	gramm. function of NP
NPType	type of NP
Definite	type of article
Embedding	embedding of NP
Direction	direction of relation
ParaGF	parallelism of gramm. function
SentDist	sentence distance
WordDist	word distance

Table 1: *Features used for the classifiers*

(2) Die AWO hat sich für Seniorenreisen
The AWO has itself for senior citizen travels
nach Mallorca von Hapaq Lloyd Provisionen
to Mallorca by Hapaq Lloyd commissions
zahlen lassen.
pay let.
"The AWO accepted commissions from Hapaq Lloyd for trips by senior citizens to Mallorca."

The reflexive pronoun *sich* is anaphoric to *die AWO*, which yields a positive pair. The pronoun is not in an anaphoric (or rather cataphoric here) relation to *Seniorenreisen*, so this gives a negative pair. The training set for the classifier will then be reduced by the different instance sampling techniques.

The original system uses the *Tilburg Memory Based Learner* (TiMBL) [2]. For the experiments reported in section 6.1, we also use TiMBL, with the IB1 algorithm, with $k = 20$ and modified value distance metric (MVDM) as the similarity metric. For the decision tree experiments, we used Weka's [14] J48 algorithm with $c = 0.25$, $m = 2$ and no subtree raising. For the maximum entropy classifier, we used Weka's Logistic algorithm with $R = 1.0E - 8$ and $M = -1$.

6 The sampling experiments

The experiments were carried out in a 10-fold cross-validation setting. As classification takes place pairwise, it is not necessary to split folds along article boundaries. Each training set contains 90% of the total number of pairs (146 153 training instances per fold). The remaining 10% are assigned to the test sets (16 239 pairs each). We evaluate the performance of the system by computing pairwise precision and recall of the classifier output against the manually annotated gold standard.[2] All experiments reported below use

[2] Since our system does not generate full coreferential chains, a pairwise evaluation metric is preferable over strategies such as the MUC-6 model-theoretic coreference scoring scheme by Vilain et al. [13].

	ratio	prec.	recall	F-score
baseline	1:4.29	**0.664**	0.457	0.541
local s.	1:2.1	0.511	0.707	0.593
distance s.	1: 2.47	0.458	**0.801**	0.583
IB2	1:0.96	0.592	0.511	0.547
random s.	1:1	0.479	0.783	0.593
	1:1.5	0.502	0.751	0.602
	1:1.75	0.521	0.720	**0.604**
	1:2	0.542	0.683	**0.604**
	1:2.25	0.552	0.662	0.602
	1:2.5	0.567	0.632	0.598
	1:3	0.598	0.570	0.584
	1:4	0.653	0.477	0.552

Table 2: *Results for training the memory-based classifier with instance sampling*

the same data split, i.e. the data for the 10-fold cross-validation was produced once and then reused for all experiments.

6.1 A comparison of all sampling methods using memory-based learning

The results of the experiments with TiMBL are shown in Table 2. For the **baseline**, we used the system as described in section 5, without any sampling. This means that the training set has a ratio of 1:4.29 of positive to negative examples. This setting results in a precision of 0.664 and a considerably lower recall of 0.457. A comparison of the baseline to the sampling experiments shows that the baseline reaches the highest precision and the lowest recall of all experiments. Thus, the experiments corroborate our assumption that instance sampling increases recall while decreasing precision.

Local sampling, i.e. reducing the negative examples to the ones found between the pronoun and its correct antecedent, increases recall by 25 percent points, but it also decreases precision by approximately 15 percent points, resulting in an increase of the F-score of 5 percent points.

Surprisingly, **distance sampling** fares considerably better, which is due to the high recall of 0.801, the highest recall of all the experiments. These results show that in order to increase recall, we need examples that are clearly different from the positive examples. However, this selection of such negative examples is also detrimental to precision: With 0.458, we get the lowest precision of all experiments.

The **incremental learning** approach IB2 presents the next surprise: The sampling ratio is the lowest of all experiments (1:0.96), which should result in high recall and low precision, but the opposite is the case: At 0.592, precision is higher than for all other sampling approaches except for random sampling with almost the complete set of negative instances (1:3) or higher. Correspondingly, recall is lower (0.511) than for most other sampling approaches, with the same exception. And while the F-score is fairly stable across the 10 folds, both precision and recall vary considerably more across the 10 folds than in all other experiments: Precision varied between 0.642 and 0.544, recall between 0.549 and 0.498.

Random sampling shows the trade-off between precision and recall dependent on the sampling rate. The more we restrict the number of negative samples, the lower precision, but the higher recall. The highest F-score is reached with a ratio between 1:2 and 1:1.75, i.e. by reducing the number of negative examples to less than half of the original set. This random combination of negative examples from all areas of the search space provides the most balanced results.

6.2 Discussion

One hypothesis to pursue would be the assumption that the only relevant factor is the sampling ratio, and there is no other significant difference between the different sampling methods. This is clearly not the case for memory-based learning. Distance sampling, for example, results in a ratio of 1:2.47 but shows results that correspond to a ratio below 1:1 in random sampling. This shows clearly that random sampling is considerably more informative than restricting the negative samples to the ones that have the longest distance from the positive examples. A comparison of local sampling and random sampling shows that the former (with a ratio of 1:2.1) results in precision and recall figures that are in the area of random sampling ratio between 1:1.5 and 1:1.75. Thus, while it is considerably more competitive than distance sampling, it reaches results that random sampling reaches with considerably fewer negative examples. The most compelling argument, however, is provided by IB2, whose results most closely resemble a random sampling ratio in the range of 1:3 and 1:4. This finding is important if efficiency is a concern. In memory-based learning, the size of the instance base is directly correlated to classification times. Thus, if efficiency is important, a smaller sampling ratio can be chosen without losing too much performance.

We can conclude that while it is linguistically reasonable to assume that the negative examples between pronoun and actual antecedent are the most informative, the best results can be reached by choosing the negative examples randomly, i.e. by including examples from all areas of the search space. One reason for this may be that both distance sampling and local sampling lead to a restricted training set, but the test set cannot be restricted in the same way since we do not know which markable is the correct antecedent or what are the closest examples. This may force the classifier to make decisions on types of pairs that it did not encounter in the training set. The results also show that random sampling with a specific ratio reaches comparable results to the other sampling methods, but with a considerably lower number of examples, such as in the comparison of the random sampling results for 1:1 with distance sampling (ratio 1:2.47).

6.3 Random sampling and IB2 sampling with different classifiers

In the previous sections, we have shown that memory-based learning profits considerably from instance sampling. The next question that arises from these results concerns the general applicability of the sampling

		memory-based			decision tree			maximum entropy		
	ratio	prec.	recall	F-score	prec.	recall	F-score	prec.	recall	F-score
baseline	1:4.29	**0.664**	0.457	0.541	**0.658**	0.489	0.561	**0.637**	0.414	0.502
IB2	1:0.96/0.65/0.7	0.592	0.511	0.547	0.476	0.789	0.594	0.380	0.737	0.501
random s.	1:1	0.479	**0.783**	0.593	0.478	**0.803**	0.600	0.443	**0.801**	0.570
	1:1.5	0.502	0.751	0.602	0.530	0.722	**0.611**	0.485	0.730	0.583
	1:1.75	0.521	0.720	**0.604**	0.551	0.680	0.608	0.514	0.667	0.581
	1:2	0.542	0.683	**0.604**	0.569	0.650	0.607	0.526	0.656	**0.584**
	1:2.25	0.552	0.662	0.602	0.584	0.627	0.604	0.548	0.614	0.579
	1:2.5	0.567	0.632	0.598	0.595	0.610	0.602	0.561	0.580	0.570
	1:3	0.598	0.570	0.584	0.618	0.559	0.587	0.594	0.513	0.550
	1:4	0.653	0.477	0.552	**0.658**	0.491	0.562	0.630	0.430	0.511

Table 3: *Results of different classifiers with instance sampling*

process: Does the success translate to other settings with other classifiers? For this reason, we repeated the experiments with two more classifiers. We chose classifiers that have been successfully used for coreference resolution: a decision tree learner [4, 7, 10], and a maximum entropy learner [6, 8]. For both classifiers, we used the Weka [14] implementations, J48 and logistic regression respectively. We were planning to include a SVM classifier in the set, but training times proved prohibitive. In order to be able to estimate the effects of instance sampling on the different classifiers, we kept the whole system and data sets constant and changed only the classifiers. This means that all classifiers are trained on exactly the same folds in the 10-fold CV and on the same feature sets. We are aware that the feature set that proved optimal for the memory-based classifier may not guarantee optimal performance for other classifiers. However, if we had optimized the feature sets for the different classifiers, we would have introduced another free variable into the experiment, and we would not have been able to distinguish differences based on sampling from differences based on the feature sets. However, we did optimize the classifiers' parameters. We are also aware that physically removing examples from the training set may not be optimal for all classifiers since some classifiers allow weighting examples so that positive examples could be assigned increased weights to balance the ratio. Again, we decided to use the same data sets since not all classifiers can use example weighting, and it is unclear whether the two methods are absolutely comparable.

For the comparative experiments, we concentrated on the best sampling method (random sampling), and the method that gave the most surprising results for memory-based learning (IB2 sampling). Note that we did not use the built-in IB2 option in the TiMBL classifier but rather used a script that would start with all positive instances as training examples and would test each negative instance separately. Negative examples were only added to the training set if they were misclassified in the test.

The results of the experiments with the different classifiers are shown in Table 3. A comparison of the **baseline** results shows very similar F-scores (0.541 for memory-based learning, 0.561 for decision tree learning, and 0.502 for maximum entropy learning). From these results, we can conclude that the selected features carry enough information for similarly successful anaphora resolution results. The fact that the first two classifiers outperform the maximum entropy learner is most likely due to the small feature set. In general, maximum entropy learning performs best in the presence of a high number of low level features while memory-based learning and decision tree learning both prefer small feature sets with more complex features.

The results for the **incremental learning** setting are surprising in that the decision tree learner reaches the third highest recall value in all the experiments presented in this section. It is only surpassed by the recall value of random sampling with the lowest ratio of 1:1. As a consequence, despite the very low precision (0.476), this method reaches a competitive F-score of 0.593. However, while the results for memory-based learning with IB2 are rather atypical with regard to the sampling ratio, the results for the combination of this sampling method with both decision tree learning and maximum entropy learning are very close to the ones for random sampling with a similar ratio (1:1). The F-score for the maximum entropy classifier does not show any improvement over the baseline for this sampling method. However, the precision and recall results are different from the baseline: precision is lower, and recall is higher.

A comparison of the **random sampling** results shows that the differences are small, again with the restriction that the maximum entropy learner has an overall performance that is 3-4 percent points lower than the other classifiers. Decision tree learning, in contrast outperforms the memory-based classifier by a small margin (F-score: 0.611 vs. 0.604), and it reaches the best results with a smaller ratio of negative examples (1:1.5 vs. 1:1.75).

An analysis of the whole table of results shows that while there are smaller differences between the classifiers, both sampling methods result in higher performance for all three classifiers. And while the results for the incremental learning approach are different for the different classifiers, the results for random sampling are very stable for the three classifiers. We can therefore cautiously conclude that using random sampling for anaphora resolution in general results in a higher F-score.

Since random sampling appears to be the most stable and the most successful sampling method, we decided to have a closer look at the curves for precision, recall, and F-score given different sampling ratios. The results for the memory-based classifier are

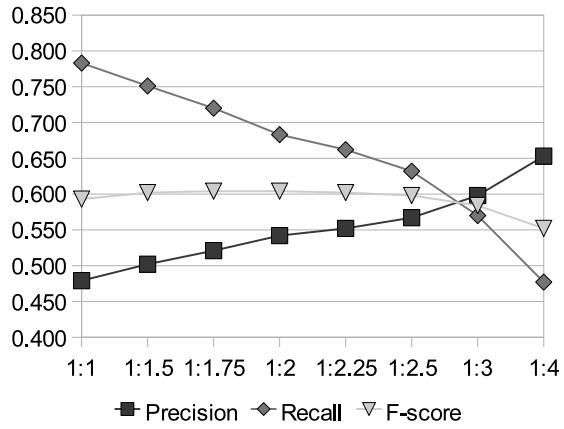

Fig. 2: *The influence of different ratios on random sampling using the memory-based classifier*

shown in Figure 2. The results for the decision tree classifier and for the maximum entropy classifier show very similar curves. The only differences are slightly lower F-scores for the maximum entropy classifier and a slight difference in ratios at which the best F-scores are reached. It is clear that across all classifiers, a low number of negative examples results in high recall, and a high number in high precision. Random sampling is therefore ideally suited for applications that may be interested in optimizing one of these measures rather than F-scores.

7 Conclusion and future work

We have shown that instance sampling is important for pronoun resolution to offset the inherent bias of the machine learner. All sampling methods (with the sole exception of maximum entropy learning with online learning) improve the F-score considerably. The highest F-score is reached for all classifiers by using random instance sampling with a ratio between 1:1.5 and 1:2. The fact that random sampling outperforms the other sampling techniques, which concentrate on different areas of the search space, clearly indicates that all examples are informative for classifications. The only function that instance sampling should perform is reducing the skewedness of the data set without fundamentally changing the distribution of the examples.

For the future, we are planning to investigate the high variance in the ten folds for IB2. Here, the sampling ratio is constant but precision and recall vary in the range of 10 and 5 percent points respectively. It is unclear why only this method should result in such a variance across the folds. One factor that does influence results is the order in which the examples are presented. But if we can resolve this issue and obtain high precision in all folds, this could be an ideal setting for classifier combination.

We also want to extend the comparison of classifiers to include feature optimization. Now that we know that all classifiers used here react favorably to random sampling given the same feature set, the next question to be answered is whether they show that same behavior with feature sets that were optimized individually.

References

[1] D. Aha, D. Kibler, and M. K. Albert. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

[2] W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. TiMBL: Tilburg memory based learner– version 6.1– reference guide. Technical Report ILK 07-07, Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University, 2007.

[3] E. W. Hinrichs, K. Filippova, and H. Wunsch. A data-driven approach to pronominal anaphora resolution in German. In *Proceedings of the 5th International Conference on Recent Advances in Natural Language Processing (RANLP 2005)*, Borovets, Bulgaria, 2005.

[4] C. Müller, S. Rapp, and M. Strube. Applying co-training to reference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 352–359, Philadelphia, PA, 2002.

[5] K. Naumann. Manual for the annotation of in-document referential relations. Technical report, Seminar für Sprachwissenschaft, Universität Tübingen, 2006.

[6] V. Ng. Learning noun phrase anaphoricity to improve coreference resolution: Issues in representation and optimization. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, Barcelona, Spain, 2004.

[7] V. Ng and C. Cardie. Combining sample selection and error-driven pruning for machine learning of coreference rules. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-02)*, Philadelphia, PA, 2002.

[8] S. Ponzetto and M. Strube. Semantic role labeling for coreference resolution. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, Trento, Italy, 2006.

[9] J. Preiss. Anaphora resolution with memory based learning. In *Proceedings of the 5th UK Special Interest Group for Computational Linguistics (CLUK5)*, pages 1–8, 2002.

[10] W. M. Soon, H. T. Ng, and D. C. Y. Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.

[11] H. Telljohann, E. Hinrichs, and S. Kübler. The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2229–2235, Lisbon, Portugal, 2004.

[12] H. Telljohann, E. W. Hinrichs, S. Kübler, and H. Zinsmeister. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen, Germany, 2006.

[13] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. A model-theoretic coreference scoring scheme. In *MUC6 '95: Proceedings of the 6th Conference on Message Understanding*, pages 45–52, Columbia, MD, 1995.

[14] I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[15] H. Wunsch. Anaphora resolution – What helps in German. In *Proceedings of the International Conference on Linguistic Evidence 2006*, pages 101–105, Tübingen, Germany, 2006.

[16] H. Wunsch. *Rule-Based and Memory-Based Pronoun Resolution for German: A Comparison and Assessment of Data Sources*. PhD thesis, Universität Tübingen, 2009.

[17] S. Zhao and H. T. Ng. Identification and resolution of Chinese zero pronouns: A machine learning approach. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, 2007.