Competitive Self-Trained Pronoun Interpretation

Andrew Kehler[*]
UC San Diego
akehler@ucsd.edu

Douglas Appelt
SRI International
appelt@ai.sri.com

Lara Taylor[*]
UC San Diego
lmtaylor@ucsd.edu

Aleksandr Simma[†]
UC San Diego
asimma@ucsd.edu

Abstract

We describe a system for pronoun interpretation that is self-trained from raw data, that is, using no annotated training data. The result outperforms a Hobbsian baseline algorithm and is only marginally inferior to an essentially identical, state-of-the-art supervised model trained from a substantial manually-annotated coreference corpus.

1 Introduction

The last several years have seen a number of feature-based systems for pronoun interpretation in which the feature weights are determined via manual experimentation or supervised learning (see Mitkov (2002) for a useful survey). Reliable estimation of the weights in both paradigms requires a substantial manually-annotated corpus of examples. In this short paper we describe a system for (third-person) pronoun interpretation that is self-trained from raw data, that is, using no annotated training data whatsoever. The result outperforms a Hobbsian baseline algorithm and is only marginally inferior (2.3%) to an essentially identical, state-of-the-art supervised model trained from a manually-annotated coreference corpus. This result leaves open the possibility that systems self-trained on very large datasets with more finely-grained features could eventually outperform supervised models that rely on manually-annotated datasets.

The remainder of the paper is organized as follows. We first briefly describe the supervised system (described in more detail in Kehler et al. (2004)) to which we will compare the self-trained system. Both systems use the same learning algorithm and feature set; they differ with respect to whether the data they are trained on is annotated by a human or the algorithm itself. We then describe our Hobbsian baseline algorithm, and present the results of all three systems.

2 The Supervised Algorithm

The supervised model was trained using the improved iterative scaling algorithm for Maximum Entropy (MaxEnt) models described by Berger et al. (1996) with binary-valued features. As is standard, the model was trained as a binary coreference classifier: for each possible antecedent of each pronoun, a training instance was created that consisted of the pronoun, the possible antecedent phrase, and a binary coreference outcome. (Such a model can be seen as providing a probabilistic measure of antecedent salience.) Because we are ultimately interested in identifying the correct antecedent among a set of possible ones, during testing the antecedent assigned the highest probability is chosen.

The algorithm receives as input the results of SRI's TEXTPRO system, a shallow parser that recognizes low-level constituents (noun groups, verb groups, etc.). No difficult syntactic attachments are attempted, and the results are errorful. There was no human-annotated linguistic information in the input.

The training corpus consists of 2773 annotated third-person pronouns from the newspaper and newswire segments of the Automatic Content Extraction (ACE) program training corpus. The annotated blind corpus used for evaluation consists of 762 annotated third-person pronouns from the ACE February 2002 evaluation set. The annotated pronouns in both sets include only those that are ACE "markables", i.e., ones that refer to entities of the following types: PERSONS, ORGANIZATIONS, GEOPOLITICALENTITIES (politically defined geographical regions, their governments, or their people), LOCATIONS, and FACILITIES.

The system employs a set of HARD CONSTRAINTS and SOFT FEATURES. The hard constraints filter out

[*] Department of Linguistics.

[†] Department of Computer Science and Engineering.

those noun groups that fail conservative number and gender agreement checks before training, whereas the soft features are used by the MaxEnt algorithm. A set of forty soft features were developed and optimized manually; they fall into five categories that have become fairly standard in the literature:

Gender Agreement: Includes features to test a strict match of gender (e.g., a masculine pronoun and a masculine antecedent), as well as mere compatibility (e.g., a masculine pronoun with an antecedent of unknown gender). These features are more liberal than the gender-based hard constraint mentioned above.

Number Agreement: Includes features to test a strict match of number (e.g., a singular pronoun and a singular antecedent), as well as mere compatibility (e.g., a singular pronoun with an antecedent of unknown number). These features are likewise more liberal than the number-based hard constraint mentioned above.

Distance: Includes features pertaining to the distance between the pronoun and the potential antecedent. Examples include the number of sentences between them and the "Hobbs distance", that is, the number of noun groups that have to be skipped before the potential antecedent is found per the search order used by the Hobbs algorithm (Hobbs, 1978; Ge et al., 1998).

Grammatical Role: Includes features pertaining to the syntactic position of the potential antecedent. Examples include whether the potential antecedent appears to be the subject or object of a verb, and whether the potential antecedent is embedded in a prepositional phrase.

Linguistic Form: Includes features pertaining to the referential form of the potential antecedent, e.g., whether it is a proper name, definite description, indefinite NP, or a pronoun.

The values of these features – computed from TextPro's errorful shallow constituent parses – comprised the input to the learning algorithm, along with the outcome as indicated by the annotated key.

3 The Self-Trained Algorithm

The self-trained algorithm likewise uses MaxEnt, with the same feature set and shallow parser. The two systems differ in the training data utilized. Instead of the training corpus of 2773 annotated pronouns used in the supervised experiments, the self-trained algorithm creates training data from pronouns found in a raw corpus, particularly the newswire segment of the Topic Detection and Tracking (TDT-2) corpus. The system was evaluated on the same annotated set of 762 pronouns as the supervised system; the performance statistics reported herein are from the only time an evaluation with this data was carried out.

The self-trained system embeds the MaxEnt algorithm in an iterative loop during which the training examples are acquired. The first phase of the algorithm builds an initial model as follows:

1. For each third-person pronoun:
   (a) Collect possible antecedents, that is, all of the noun groups found in the previous two sentences and to the left of the pronoun in the current sentence.
   (b) Filter them by applying the hard constraints.
   (c) If only one possible antecedent remains, create a pronoun-antecedent pair and label the coreference outcome as True.
   (d) Otherwise, with some probability (0.2 in our experiments[1]), create a pronoun-antecedent pair for each possible antecedent and label the coreference outcome as False.
2. Train a MaxEnt classifier on this training data.

The simplification assumed above – that coreference holds for all and only those pronouns for which TextPro and hard constraints find a single possible antecedent – is obviously false, but it nonetheless yields a model to seed the iterative part of the algorithm, which goes as follows:

3. For each pronoun in the training data acquired in step 1:
   (a) Apply the current MaxEnt model to each pronoun-antecedent pair.
   (b) Label the pair to which the model assigns the highest probability the coreference outcome of True. Label all other pairs (if any) for that pronoun the outcome of False.
4. Retrain the MaxEnt model with this new training data.
5. Repeat steps 3 and 4 until the training data reaches a steady state, that is, there are no pronouns for which the current model changes its preference to a different potential antecedent than it favored during the previous iteration.

[1] This choice will be explained in Section 5.

The hope is that improved predictions about which potential antecedents of ambiguous pronouns are correct will yield iteratively better models (note that the "unambiguous" pronoun-antecedent pairs collected in step 1c will be considered to be correct throughout). This hope is notwithstanding the fact that the algorithm is based on a simplifying assumption – that each pronoun is associated with exactly one correct antecedent – that is clearly false for a variety of reasons: (i) there will be cases in which there is more than one coreferential antecedent in the search window, all but one of which will get labeled as not coreferential during any given iteration, (ii) there will be cases in which the (perhaps only) correct antecedent was misparsed or incorrectly weeded out by hard constraints, and thus not seen by the learning algorithm (presumably some of the "unambiguous" cases identified in step 1c will be incorrect because of this), and (iii) some of the pronouns found will not even be referential, e.g. pleonastic pronouns. The empirical question remains, however, of how good of a system can be trained under such an assumption. After all, the model probabilities need not necessarily be accurate in an *absolute* sense, but only in a *relative* one: that is, good enough so that the antecedent assigned the highest probability tends to be correct.

4 Hobbs Baseline

For comparison purposes, we also implemented a version of Hobbs's (1978) well-known pronoun interpretation algorithm, in which no machine learning is involved. This algorithm takes the syntactic representations of the sentences up to and including the current sentence as input, and performs a search for an antecedent noun phrase on these trees. Since TextPro does not build full syntactic trees for the input, we developed a version that does a simple search through the list of noun groups recognized. In accordance with Hobbs's search procedure, noun groups are searched in the following order: (i) in the current sentence from right-to-left, starting with the first noun group to the left of the pronoun, (ii) in the previous sentence from left-to-right, (iii) in two sentences prior from left-to-right, (iv) in the current sentence from left-to-right, starting with the first noun group to the right of the pronoun (for cataphora). The first noun group encountered that agrees with the pronoun with respect to number, gender, and person is chosen as the antecedent.

5 Results

Reporting on the results of a self-trained system means only evaluating the system against annotated data once, since any system reconfiguration and reevaluation based on the feedback received would constitute a form of indirectly supervised training. Thus we had to select a configuration as representing our "reportable" system before doing any evaluation. To allow for the closest comparison with our supervised system, we opted to train the system with the same number of pronouns that we had in our supervised training set (2773), and sought to have approximately the same ratio of positive to negative training instances, which meant randomly including one-fifth of the pronouns in the raw data that had more than one possible antecedent (see step 1d). Later we report on post-hoc experiments to assess the effect of training data size on performance.

The self-trained system was trained fourteen times, once using each of fourteen different segments of the TDT-2 data that we had arbitrarily apportioned at the inception of the project. The scores reported below and in Table 1 for the self-trained system are averages of the fourteen corresponding evaluations. The final results are as follows:

- Hobbs Baseline: 68.8%
- Self-Trained: 73.4%
- Supervised: 75.7%

The self-trained system beats the competitive Hobbs baseline system by 4.6% and comes within 2.3% of the supervised system trained on the same number of manually-annotated pronouns.[2]

Convergence for the self-trained system was fairly rapid, taking between 8 and 14 iterations. The number of changes in the current model's predictions started off relatively high in early iterations (averaging approximately 305 pronouns or 11% of the dataset) and then steadily declined (usually, but not always, monotonically) until convergence. Post-hoc

[2] All results are reported here in terms of accuracy, that is, the number of pronouns correctly resolved divided by the total number of pronouns read in from the key. An antecedent is considered correct if the ACE keys place the pronoun and antecedent in the same coreference class.

In the case of 64 of the 762 pronouns in the evaluation set, none of the antecedents input to the learning algorithms were coreferential. Thus, 91.6% accuracy is the best that these algorithms could have achieved.

In Kehler et al. (2004) we describe two ways in which our supervised system was augmented to use predicate-argument frequencies, one which used them in a post-processor and another which modeled them with features alongside our morphosyntactic ones. In our self-trained system, the first of these methods improved performance to 75.1% (compared to 76.8% for the supervised system) and the second to 74.1% (compared to 75.7% for the supervised system).

Number of Pronouns	Blind Test Performance
55	71.4%
138	72.3%
277	72.5%
554	72.6%
1386	73.5%
2773	73.4%
5546	73.5%
Full Segment	73.7%

Table 1: Effect of Training Data Size on Blind Test Performance

analysis showed that the iterative phase contributed a gradual (although again not completely monotonic) improvement in performance during the course of learning.

We then performed a set of post-hoc experiments to measure the effect of training data size on performance for the self-trained system. The results are given in Table 1, which show a gradual increase in performance as the number of pronouns grows. The final row includes the results when all of the "unambiguous" pronouns in each TDT segment are utilized (again, along with approximately one-fifth of the ambiguous pronouns), which amounted to between 7,212 and 11,245 total pronouns.[3] (Note that since most pronouns have more than one possible antecedent, the number of pronoun-antecedent training examples fed to MaxEnt is considerably higher than the numbers of pronouns shown in the table.) Perhaps one of the more striking facts is how well the algorithm performs with relatively few pronouns, which suggests that the generality of the features used allow for fairly reliable estimation without much data.

6 Conclusion

To conclude, a pronoun interpretation system can be trained solely on raw data using a standard set of morphosyntactic features to achieve performance that approaches that of a state-of-the-art supervised system. Although the self-acquired training data is no doubt highly noisy, the resulting model is still accurate enough to perform well at selecting correct antecedents. As a next step, we will take a closer look at the training data acquired to try to ascertain the underlying reasons for this success.

There are also a number of variants of the algorithm that could be pursued. For instance, whereas our algorithm uses the current model's probabilities in a winner-take-all strategy for positive example selection, these probabilities could instead be used to dictate the likelihood that examples are assigned a positive outcome, or they could be thresholded in various ways to create a more discerning positive outcome assignment mechanism. Such strategies would avoid the current simplification of assigning a positive outcome to exactly one potential antecedent for each pronoun.

The relative generality of our feature set was appropriate given the size of the data sets used. The availability of very large raw corpora, however, creates the prospect of using self-training with considerably more fine-grained features than is possible in a supervised scenario, due to the relative infrequency with which they would be found in any corpus of a size that could be feasibly annotated manually. It is thus at least conceivable that a self-trained approach, coupled with a large set of features and a large corpus of raw data, could eventually overtake the performance of the best supervised models.

Acknowledgments

This work was supported by the ACE program (www.nist.gov/speech/tests/ACE/).

References

Adam Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

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

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 HLT/NAACL-04*, Boston, MA.

Ruslan Mitkov. 2002. *Anaphora Resolution*. Longman, London.

[3] TDT segment 14, which is smaller than the others, provided only about 3800 pronouns in the runs corresponding to the last two rows of Table 1. The overall average performance figures are the same to the first decimal place whether or not the results from this segment are included.