Active Learning for Anaphora Resolution

Caroline Gasperin
Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue
Cambridge CB3 0FD, UK
cvg20@cl.cam.ac.uk

Abstract

In this paper we present our experiments with active learning to improve the performance of our probabilistic anaphora resolution system. We have adopted entropy-based uncertainty measures to select new instances to be added to our training data. The actively selected instances, however, were not more successful in improving the performance of the system than the same amount of randomly selected instances. The uncertainty measures we used behave differently from each other when selecting new instances, but none of them achieved remarkable performance. Further studies on active sample selection for anaphora resolution are necessary.

1 Introduction

Anaphora is the relation between two linguistic expressions in the discourse where the reader is referred back to the first of them when reading the second later in the text. Anaphora resolution can be understood as the process of identifying an anaphoric relation between two expressions in the text and consequently linking the two of them, one being the anaphor and the other being the antecedent. Manually annotating corpora with anaphoric links in order to use it as training or test data for a corpus-based anaphora resolution system is a particulary difficult and time consuming task, given the complex nature of the phenomenon.

We have developed a probabilistic model for resolution of non-pronominal anaphora and aim to improve its performance by acquiring incrementally and selectively more training data using active learning. We have adopted an uncertainty-based active learning approach in order to do that, and it uses our probabilistic model as the base classifier.

The uncertainty-based approach has been applied to, for instance, named-entity recognition by Shen et al. (2004) who report at least 80% reduction in annotation costs, parsing by Tang et al. (2002) who reports 67% savings, and parse selection by Baldridge and Osborne (2003) who report 60% savings. We are not aware of any work that has applied active learning to anaphora resolution.

For calculating the uncertainty of an anaphora resolution model, we feel the need to combine the information about the confindence of the model for the classification of each antecedent candidate associated to a given anaphor. We have tested three entropy-based uncertainty measures in order to select the instances to be added to the training data.

Our training corpus is composed of five full-length scientific articles from the biomedical domain. We have used this corpus to simulate active learning: we have divided our training data into two parts, one for the initial training and the other for active learning (simulating unlabelled data), and have compared the classifier performance when trained on a sample selected by active learning to its performance when trained on the same amount of randomly selected instances.

In the next section we describe our probabilistic model for anaphora resolution. In Section 3 we detail our training corpus. In Section **??** we describe the strategy we have adopted to select the samples to take part in the active learning, and in Section 5

Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing, pages 1–8,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

we describe our experiments.

2 Anaphora resolution model

We have inplemented a probabilistic model for anaphora resolution in the biomedical domain (Gasperin and Briscoe, 2008). This model aims to resolve both coreferent and associative (also called bridging (Poesio and Vieira, 1998)) cases of non-pronominal anaphora. Table 1 shows examples of these types of anaphoric relations. Coreferent are the cases in which the anaphor and the antecedent refer to the same entity in the world, while associative cases are the ones in which the anaphor and antecedent refer to different but somehow related entities. We only take into account noun phrases referring to biomedical entities, since this was the focus of our resolution model. We consider two types of associative relations: biotype relations, which are anaphoric associative relations between noun phrases that share specific ontological relations in the biomedical domain; and set-member relations, in which the noun phrases share a set-membership relation. It is frequent however that some noun phrases do not have an antecedent, these are considered discourse-new cases, which we also aim to identify.

The probabilistic model results from a simple decomposition process applied to a conditional probability equation that involves several parameters (features). It is inspired by Ge et al.'s (1998) probabilistic model for pronoun resolution. The decomposition makes use of Bayes' theorem and independence assumptions, and aims to decrease the impact of data sparseness on the model, so that even small training corpora can be viable. The decomposed model can be thought of as a more sophisticated version of the naive-Bayes algorithm, since we consider the dependence among some of the features instead of full independence as in naive Bayes. Probabilistic models can return a confidence measure (probability) for each decision they make, which allow us to adopt techniques such as active learning for further processing.

Our model seeks to classify the relation between an anaphoric expression and an antecedent candidate as coreferent, biotype, set-member or neither. It computes the probability of each pair of anaphor and candidate for each class. The candidate with the highest overall probability for each class is selected as the antecedent for that class, or no antecedent is selected if the probability of no relation overcomes the positive probabilities; in this case, the expression is considered to be new to the discourse.

We have chosen 11 features to describe the anaphoric relations between an antecedent candidate a and an anaphor A. The features are presented in Table 2. Most features are domain-independent, while one, $gp_{a,A}$, is specific for the biomedical domain. Our feature set covers the basic aspects that influence anaphoric relations: the form of the anaphor's NP, string matching, semantic class matching, number agreement, and distance.

Given these features, we compute the probability P of an specific class of anaphoric relation C between a (antecedent candidate) and A (anaphor). For each pair of a given anaphor and an antecedent candidate we compute P for C=‘coreferent’, C=‘biotype’, and C=‘set-member’. We also compute C=‘none’, that represents the probability of no relation between the NPs. P can be defined as follows:

$$P(C = \text{`class'} | f_A, f_a, hm_{a,A}, hmm_{a,A}, mm_{a,A}, \\ num_{a,A}, sr_a, bm_{a,A}, gp_{a,A}, d_{a,A}, dm_{a,A})$$

If we were to use P as above we would suffer considerably data sparseness. In order to reduce that, we decompose the probability P and assume independence among some of the features in order to handle the sparseness of the training data. For more detail on the decomposition process refer to (Gasperin and Briscoe, 2008).

Applying Bayes' rule and selectively applying the chain rule to the above equation, as well as assuming independece among some features, we reach the following equation:

$$P(C | f_A, f_a, hm, hmm, mm, num, sr, bm, gp, d, dm) =$$

$$\frac{\begin{array}{c} P(C)\ P(f_A|C)\ P(f_a|C, f_A)\ P(d, dm|C, f_A, f_a) \\ P(sr|C, d, dm)\ P(bm, gp|C)\ P(num|C, f_A, f_a) \\ P(hm, hmm, mm|C, f_A, f_a, bm) \end{array}}{\begin{array}{c} P(f_A)\ P(f_a|f_A)\ P(d, dm|f_A, f_a) \\ P(sr|d, dm)\ P(bm, gp)\ P(num|f_A, f_a) \\ P(hm, hmm, mm|f_A, f_a, bm) \end{array}} \quad (1)$$

C	"The expression of **reaper** has been shown ... **the gene** encodes ...
B	"**Drosophila gene Bok** interacts with ... expression of **Bok protein** promotes apoptosis ..."
S	"... **ced-4** and **ced-9** ... **the genes** ..."
	"... **the mammalian anti-apoptotic protein Bcl-2** ... **Bcl-2 family** ..."

Table 1: Examples of coreferent (C), associative biotype (B) and associative set-member (S) anaphoric relations

Feature	Possible values
f_A	Form of noun phrase of the anaphor A: 'pn', 'defnp', 'demnp', 'indefnp', 'quantnp', or 'np'.
f_a	Form of noun phrase of the antecedent candidate a: same values as for f_A.
$hm_{a,A}$	Head-noun matching: 'yes' if the anaphor's and the candidate's head nouns match, 'no' otherwise.
$hmm_{a,A}$	Head-modifier matching: 'yes' if the anaphor's head noun matches any of the candidate's pre-modifiers, or vice-versa, 'no' otherwise.
$mm_{a,A}$	Modifier matching: 'yes' if anaphor and candidate have at least one head modifier in common, 'no' otherwise.
$num_{a,A}$	Number agreement: 'yes' if anaphor and candidate agree in number, 'no' otherwise.
$sr_{a,A}$	Syntactic relation between anaphor and candidate: 'none', 'apposition', 'subj-obj', 'pp', and few others.
$bm_{a,A}$	Biotype matching: 'yes' if anaphor's and candidate's biotype (semantic class) match, 'no' otherwise.
$gp_{a,A}$	is biotype *gene* or *product*? 'yes' if the anaphor biotype or candidate biotype is *gene* or *product*, 'no' otherwise. This feature is mainly to distinguish which pairs can hold biotype relations.
$d_{a,A}$	Distance in sentences between the anaphor and the candidate.
$dm_{a,A}$	Distance in number of entities (markables) between the anaphor and the candidate.

Table 2: Feature set

This equation is the basis of our resolution model. We collect the statistics to train this model from a corpus annotated with anaphoric relations that we have created. The corpus is described in the next section.

3 Our corpus

There are very few biomedical corpora annotated with anaphora information, and all of them are built from paper abstracts (Cohen et al., 2005), instead of full papers. As anaphora is a phenomenon that develops through a text, we believe that short abstracts are not the best source to work with and decided to concentrate on full papers.

In order to collect the statistics to train our model, we have manually annotated anaphoric relations between biomedical entities in 5 full-text articles (approx. 33,300 words)[1], which are part of the Drosophila molecular biology literature. The corpus and annotation process are described in (Gasperin et al., 2007). To the best of our knowledge, this corpus is the first corpus of biomedical full-text articles to be annotated with anaphora information.

Before annotating anaphora, we have preprocessed the articles in order to (1) tag gene names, (2) identify all NPs, and (3) classify the NPs according to their domain type, which we call biotype. To tag all gene names in the corpus, we have applied the gene name recogniser developed by Vlachos et al. (2006). To identify all NPs, their subconstituents (head, modifiers, determiner) and broader pre- and post-modification patterns, we have used the RASP parser (Briscoe et al., 2006). To classify the NPs according to their type in biomedical terms, we have adopted the Sequence Ontology (SO)[2] (Eilbeck and Lewis, 2004). SO is a fine-grained ontology, which contains the names of practically all entities that participate in genomic sequences, besides the relations among these entities (e.g. is-a, part-of, derived-from relations). We derived from SO seven biotypes to be used to classify the entities in the text, namely: "gene", "gene product", "part of gene", "part of product", "gene variant", "gene subtype", and "gene

[1] Corpus available via the FlySlip project website http://www.wiki.cl.cam.ac.uk/rowiki/NaturalLanguage/FlySlip

[2] http://www.sequenceontology.org/

Class	Relations
coreferent	1678
biotype	274
set-member	543
discourse new	436
Total	3048
none	873,731

Table 3: Training instances, according to anaphoric class

supertype”. We also created the biotype “other-bio” to be associated with noun phrases that contain a gene name (identified by the gene name recogniser) but whose head noun does not fit any of the other biotypes. All NPs were tagged with their biotypes, and NPs for which no biotypes were found were excluded.

The gene-name tags, NP boundaries and biotypes resulting from the preprocessing phase were revised and corrected by hand before the anaphoric relations were annotated.

For each biotyped NP we annotated its closest coreferent antecedent (if found) and its closest associative antecedent (if found), from one of the associative classes. From our annotation, we can infer coreference chains by merging the coreferent links between mentions of a same entity.

The annotated relations, and the features derived from them, are used as training data for the probabilistic model above. A special characteristic of data annotated with anaphora information is the overwhelming amount of negative instances, which result from the absence of an anaphoric relation between a NP that precedes an anaphoric expression and was not marked as its antecedent (nor marked as part of the same coreference chain of its antecedent). The negative instances outnumber considerably the number of positive instances (annotated cases). Table 3 presents the distribution of the cases among the classes of anaphoric relations.

To balance the ratio between positive and negative training samples, we have clustered the negative samples and kept only a portion of each cluster, proportional to its size. All negative samples that have the same values for all features are grouped together (consequently, a cluster is formed by a set of identical samples) and only one-tenth of each cluster members is kept, resulting in 85,314 negative samples. This way, small clusters (with less than 10 members), which are likely to represent noisy samples (similar to positive ones), are eliminated, and bigger clusters are shrunk; however the shape of the distribution of the negative samples is preserved. For example, our biggest cluster (feature values are: f_A=‘pn’, f_a=‘pn’, hm=‘no’, hmm=‘no‘, mm=‘no’, bm=‘yes’, gp=‘yes’, num=‘yes’, sr=‘none’, d=‘16<’, dm=‘50<’) with 33,998 instances is reduced to 3,399 – still considerably more numerous than any positive sample. Other works have used a different strategy to reduce the imbalance between positive and negative samples (Soon et al., 2001; Ng and Cardie, 2002; Strube et al., 2002), where only samples composed by a negative antecedent that is closer than the annotated one are considered. Our strategy is more flexible and is able to the reduce further the number of negative samples. The higher the number of negative samples, the higher the precision of the resolution, but the lower the recall.

4 Active learning

When trained using all our annotated corpus on a 10-fold cross-validation setting our anaphora resolution model, presented above, reached the results shown in Table 4[3].

We would like to improve this results without having to annotate too much more data, therefore we decided to experiment with active learning. We defined three entropy-based measures to calculate the uncertanty of our model for each decidion is makes.

Class	Perfect			Relaxed		
	P	R	F	P	R	F
coreferent	56.3	54.7	55.5	69.4	67.4	68.3
biotype	28.5	35.0	31.4	31.2	37.9	34.2
set-member	35.4	38.2	36.7	38.5	41.5	40.0
discourse new	44.3	53.4	48.4	44.3	53.4	48.4

Table 4: Performance of the probabilistic model

[3] ‘Perfect’ scores shows the result of a strict evaluation, where we consider as correct all pairs that match exactly an antecedent-anaphor pair in the annotated data. On the other hand, column ‘Relaxed’ treats as correct also the pairs where the assigned antecedent is not the exact match in the annotated data but is coreferent with it.

4.1 Uncertainty measures

In order to measure how confident our model is about the class it assigns to each candidate, and consequently the one it chooses as the antecedent of an anaphor, we experiment with the following entropy-based measures.

We first compute what we call the "local entropy" among the probabilities for each class—*P(C="coreferent")*, *P(C="biotype")*, *P(C="set-member")* and *P(C="none")*—for a given pair anaphor(A)-candidate(a), which is defined as

$$LE(A,a) = -\sum_{C} P(C) log_2 P(C) \tag{2}$$

where $P(C)$ represents Equation 1 above, that is, the probability assigned to the anaphor-candidate relation by our probabilistic model for a particular class. The more similar the probabilities are, the more uncertain the model is about the relation, so the higher the local entropy. This measure is similar to others used in previous work for different problems.

We also compute the "global entropy" of the distribution of candidates across classes for each anaphor. The global entropy aims to combine the uncertainty information from all antecedent candidates for a given anaphor (instead of considering only a single candidate-anaphor pair as for LE). The higher the global entropy, the higher the uncertainty of the model about the antecedent for an anaphor. The global entropy combines the local entropies for all antecedent candidates of a given anaphor. We propose two versions of the global entropy measure. The first is simply a sum of the local entropies of all candidates available for a given anaphor, it is defined as

$$GE1(A) = \sum_{a} LE(A,a) \tag{3}$$

The second version averages the local entropies across all candidates, it is defined as

$$GE2(A) = \frac{\sum_a LE(A,a)}{|a|} \tag{4}$$

where $|a|$ corresponds to the number of candidates available for a given anaphor.

We consider that in general the further away a candidate is from the anaphor, the lower the local entropy of the pair is (given that when distance increases, the probability of the candidate not being the antecedent, *P(C="none")*, also increases), and consequently the less it contributes to the global entropy. This is the intuition behind $GE1(A)$.

However, in some cases, mainly when the anaphor is a proper name, there can be several candidates at a long distance from the anaphor that still get a reasonable probability assigned to them due to positive string matching. Therefore we decided to experiment with averaging the sum of the local probabilities by the number of candidates, so $GE2(A)$.

5 Experiments

Initially, our training data was divided in 10-folds for cross-validation evaluation of our probabilistic model for anaphora resolution. For the active learning experiments we kept the same folds, using one for the initial training, eight for the active learning phase, and the remaining one for testing. We have experimented with 10 different initial-training/active-learning/testing splits, selected randomly from all combinations of the 10 folds, and the results in this section correspond to the average of the results from the different data splits. A fold contains the positive and negative samples derived from about 270 anaphors, it contains about 7000 candidate-anaphor pairs (an average of about 26 antecedent candidates per anaphor). The anaphors that are part of each fold were randomly selected.

The purpose of our experiments is to check whether the samples selected by using the entropy-based measures described above, when added to our training data, can improve the performance of the model more than in the case of adding the same amount of randomly selected samples. For that, we computed (1) the performance of our model using one fold of training data, (2) the performance of the model over 10 iterations of active learning using each of the uncertainty measures above, and (3) the performance of the model over 10 iterations adding the same amount of randomly selected instances as for active learning. At each active learning iteration, when using $LE(A,a)$ we selected the 1500 candidate-anaphor pairs for which uncertainty was the highest, and when using $GE1(A)$ and $GE2(A)$ we selected the 50 anaphors for which the model was most uncertain and generated the positive and negative instances that were associated to the anaphors.

We expected (2), entropy-based sample selection, to achieve better performance than (3), random sample selection, however this has not happened. The graphs in Figure 1 compare the precision, recall and F-measure scores for (2) and (3) along the 10 iterations for each class of anaphoric relation. The lines corresponding to random sampling plot the results of the experiments done in the same way as for $GE1(A)$ and $GE2(A)$, that is, where 50 anaphors are selected at each iteration, although we also tested random sampling in the $LE(A, a)$ fashion, selecting 1500 candidate-anaphor pairs.

We observe that none of the uncertainty measures that we tested have performed consistently better than random sampling. $LE(A, a)$ presents the most dramatic results, it worsens the general performance of the model for all classes, although it causes a considerable increase in precision for coreferent and set-member cases. $GE1(A)$ and $GE2(A)$ have a less clear pattern, but it is possible to notice that $GE1(A)$ tends to bring improvements in precision while $GE2(A)$ causes the opposite, improvements in recall and drops in precision.

6 Discussion

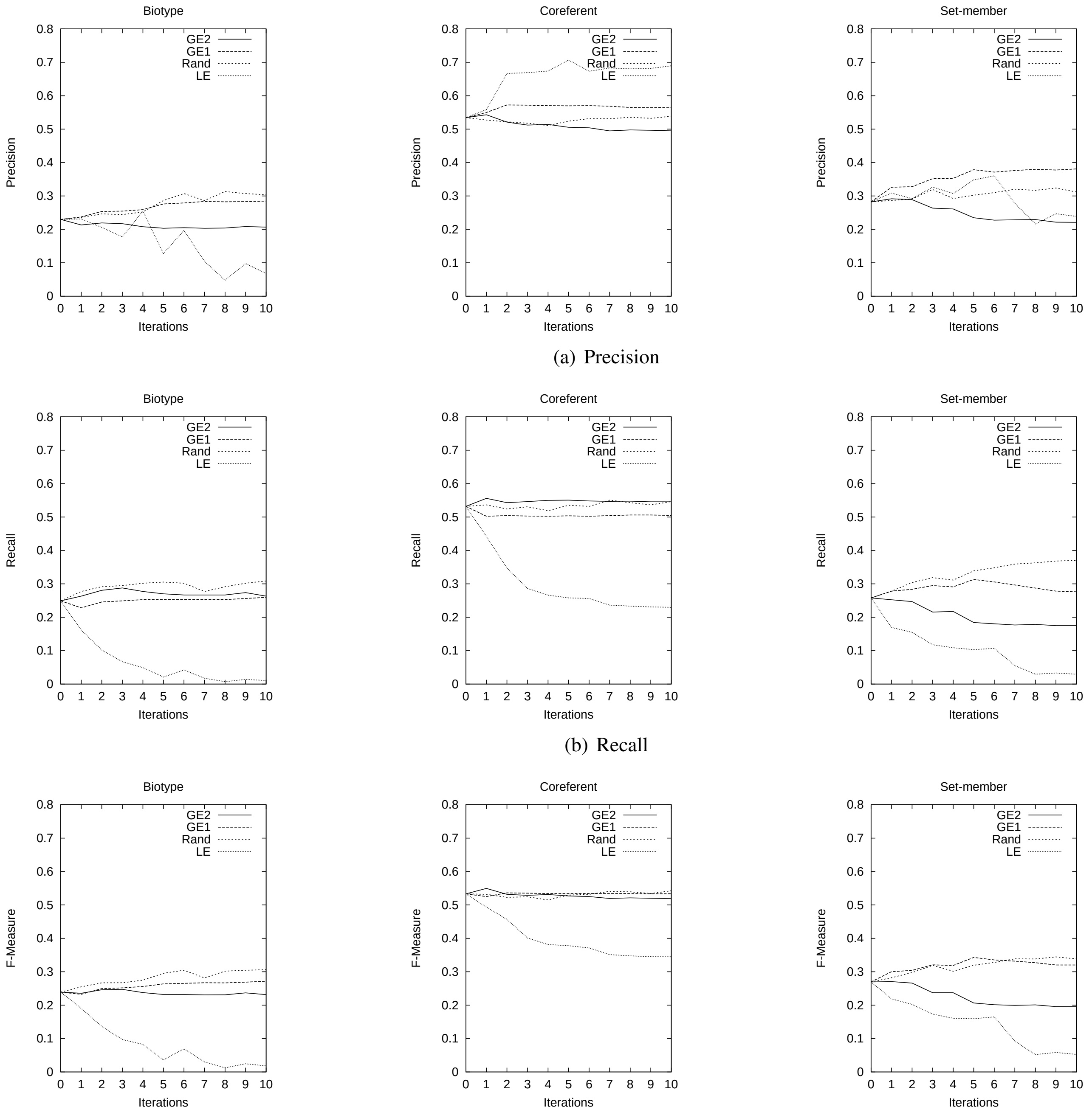
When looking at the instances selected by each active learning strategy, we observe the following. $LE(A, a)$, which considers anaphor-candidate pairs, selects mostly negative instances, given the fact that these are highly frequent. This can explain the increase in precision and drop in recall for the positive cases (observed for coreferent and set-member, the most frequent positive classes), since that is expected with the increase of negative instances.

$GE1(A)$ and $GE2(A)$ select a proportional number of positive and negative instances, since these measures consider an anaphor and all possible antecedent candidates, generating all instances that derive from each selected anaphor (usually one or two positive intances and several negative ones). However, we can observe some differences between the impact of using $GE1(A)$ and $GE2(A)$ to select instances. We observe that about 70% of the samples selected by $GE1(A)$ were proper names, while the distribution of NP types among the samples selected by $GE2(A)$ is similar to the original distribution in the data. This confirms the problem we expected to have with $GE1(A)$, since exact matches of proper names that occur at a considerable distance from the anaphor still get a higher probability assigned to them, which does not happen so often with other types of NPs. On the other hand, the correct antecedent of about 30% of $GE2(A)$-selected samples were in the same sentence as the anaphor, while the same occurs with only 8% of $GE1(A)$-selected samples. $GE2(A)$ behaviour in this case is counter intuitive, since antecedents in the same sentence should be found by the model with lower uncertainty than antecedents further away from the anaphor. Another counter intuitive behaviour of $GE2(A)$ is that only 3% of the selected anaphors have no string matching with their antecedents (33% have no head-noun matching), while these cases correspond to about 30% of samples selected by $GE1(A)$ (62% of samples have no head-noun matching). We expected samples involving no string matching to be selected because they are usually the ones the model is mostly uncertain about.

Despite the different behaviour among the measures none was successful in improving the performance of the model in relation to the performance of random sampling.

While entropy-based measures for sample selection seem the obvious option given that we use a probabilistic model, they did not give positive results in our case. A future study of different ways to combine the local entropies is necessary, as well as the study of other non-entropy-based measures for sample selection.

The main difference between our application of active learning to anaphora resolution and previous successful applications of active learning to other tasks is the amount of probabilities involved in the calculation of the uncertainty of the model. We believe this is the reason why our active learning experiments were not succesfull. While, for example, name entity recognition involves a binary decision, and parse selection involves a few parsing options, in our case there are several antecedent candidates to be considered. For anaphora resolution, when using a pairwise resolution model, it is necessary to combine the predictions for one candidate-anaphor



(a) Precision

(b) Recall

(c) F-measure

Figure 1: Graphs of the performance of active learning using $LE(A, a)$, $GE1(A)$, $GE2(A)$ and random sampling.

pair to the others in order to predict the global uncertainty of the model.

Acknowledgments

This work was part of the BBSRC-funded FlySlip project. Caroline Gasperin was funded by a CAPES award from the Brazilian government. Thanks to Ted Briscoe for comments on this manuscript.

References

Jason Baldridge and Miles Osborne. 2003. Active learning for hpsg parse selection. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 17–24. Edmonton, Canada.

Edward J. Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of ACL-COLING 06*, Sydney, Australia.

K. Bretonnel Cohen, Lynne Fox, Philip Ogren, and Lawrence Hunter. 2005. Corpus design for biomedical natural language processsing. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases*, Detroit.

Karen Eilbeck and Suzanna E. Lewis. 2004. Sequence ontology annotation guide. *Comparative and Functional Genomics*, 5:642–647.

Caroline Gasperin and Ted Briscoe. 2008. Statistical anaphora resolution in biomedical texts. In *Proceedings of COLING 2008*, Manchester, UK.

Caroline Gasperin, Nikiforos Karamanis, and Ruth Seal. 2007. Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. In *Proceedings of DAARC 2007*, pages 19–24, Lagos, Portugal.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora - COLING-ACL'98*, Montreal, Canada.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of ACL 2002*, Philadelphia.

Massimo Poesio and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of ACL 2004*, Barcelona.

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, Stefan Rapp, and Christoph Mller. 2002. The influence of minimum edit distance on reference resolution. In *Proceedings of the EMNLP 2002*, Philadelphia.

Min Tang, Xiaoqiang Luo, and Salim Roukos. 2002. Active learning for statistical natural language parsing. In *Proceedings of ACL 2002*, pages 120–127, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Andreas Vlachos and Caroline Gasperin. 2006. Bootstrapping and evaluating named entity recognition in the biomedical domain. In *Proceedings of BioNLP at HLT-NAACL 2006*, pages 138–145, New York.