Active Learning for Coreference Resolution

Timothy A. Miller and **Dmitriy Dligach** and **Guergana K. Savova**
Children's Hospital Boston
and Harvard Medical School
300 Longwood Ave.
Enders 141
Boston, MA 02115, USA
{Timothy.Miller,Dmitriy.Dligach,Guergana.Savova}@childrens.harvard.edu

Abstract

Active learning can lower the cost of annotation for some natural language processing tasks by using a classifier to select informative instances to send to human annotators. It has worked well in cases where the training instances are selected one at a time and require minimal context for annotation. However, coreference annotations often require some context and the traditional active learning approach may not be feasible. In this work we explore various active learning methods for coreference resolution that fit more realistically into coreference annotation workflows.

1 Introduction

Coreference resolution is the task of deciding which entity mentions in a text refer to the same entity. Solving this problem is an important part of the larger task of natural language understanding in general. The clinical domain offers specific tasks where it is easy to see that correctly resolving coreference is important. For example, one important task in the clinical domain is template filling for the Clinical Elements Model (CEM).[1] This task involves extracting various pieces of information about an entity and fitting the information into a standard data structure that can be reasoned about. An example CEM template is that for *Disease* with attributes for *Body Location*, *Associated Sign or Symptom*, *Subject*, *Negation*, *Uncertainty*, and *Severity*. Since a given entity may have many different attributes and relations, it may be mentioned multiple times in a text. Coreference resolution is important for this task because it must be known that all the attributes and relations apply to the same entity so that a single CEM template is filled in for an entity, rather than creating a new template for each mention of the entity.

2 Background

2.1 Coreference Resolution

Space does not permit a thorough review of coreference resolution, but recent publications covered the history and current state of the art for both the general domain and the clinical domain (Ng, 2010; Pradhan et al., 2011; Zheng et al., 2011).

The system used here (Zheng et al., 2012) is an end-to-end coreference resolution system, meaning that the algorithm receives no gold standard information about mentions, named entity types, or any linguistic information. The coreference resolution system is a module of the clinical Textual Analysis and Knowledge Extraction System (cTAKES) (Savova et al., 2010) that is trained on clinical data. It takes advantage of named entity recognition (NER) and categorization to detect entity mentions, and uses several cTAKES modules as feature generators, including the NER module, a constituency parser module, and a part of speech tagging module.

The system architecture is based on the pairwise discriminative classification approach to the coreference resolution problem. In that paradigm, pairs of mentions are classified as coreferent or not, and then some reconciliation must be done on all of the

[1] http://intermountainhealthcare.org/cem

Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pages 73–81,
Montréal, Canada, June 8, 2012. ©2012 Association for Computational Linguistics

links so that there are no conflicts in the clusters. The system uses support vector machines (SVMs) as the pairwise classifiers, and conflicts are avoided by only allowing an anaphor to link with one antecedent, specifically that antecedent the classifier links with the highest probability.

There are separate pairwise classifiers for named entity and pronominal anaphor types. In the domain of clinical narratives, person mentions and personal pronouns in particular are not especially challenging – the vast majority of person mentions are the patient. In addition, pronoun mentions, while important, are relatively rare. Thus we are primarily interested in named entity coreference classification, and we use that classifier as the basis of the work described here.

The feature set of this system is similar to that used by Ng and Cardie (2002). That system includes features based on surface form of the mentions, shallow syntactic information, and lexical semantics from WordNet. The system used here has a similar feature set but uses Unified Medical Language System (UMLS)[2] semantic features as it is intended for clinical text, and also incorporates several syntactic features extracted from constituency parses extracted from cTAKES.

To generate training data for active learning simulations, mention detection is run first (cTAKES contains a rule-based NER system) to find named entities and a constituency parser situates entities in a syntax tree). For each entity found, the system works backwards through all other mentions within a ten sentence window. For each candidate anaphor-antecedent pair, a feature vector is extracted using the features briefly described above.

2.2 Active Learning

Active Learning (AL) is a popular approach to selecting unlabeled data for annotation (Settles, 2010) that can potentially lead to drastic reductions in the amount of annotation that is necessary for training an accurate statistical classifier. Unlike passive learning, where the data is sampled for annotation randomly, AL delegates data selection to the classifier. AL is an iterative process that operates by first training a classifier on a small sample of the data known as the seed examples. The classifier is subsequently applied to a pool of unlabeled data with the purpose of selecting additional examples the classifier views as informative. The selected data is annotated and the cycle is repeated, allowing the learner to quickly refine the decision boundary between classes. One common approach to assessing the informativeness is *uncertainty sampling* (Lewis and Gale, 1994; Schein and Ungar, 2007), in which the learner requests a label for the instance it is most uncertain how to label. In this work, we base our instance selection on the distance to the SVM decision boundary (Tong and Koller, 2002), assuming that informative instances tend to concentrate near the boundary.

Most AL work focuses on *instance selection* where the unit of selection is one instance represented as a feature vector. In this paper we also attempt *document selection*, where the unit of selection is a document, typically containing multiple coreference pairs each represented as a feature vector. The most obvious way to extend a single instance informativeness metric to the document scenario is to aggregate the informativeness scores. Several uncertainty metrics have been proposed that follow that route to adapt single instance selection to multiple instance scenarios (Settles et al., 2008; Tomanek et al., 2009). We borrow some of these metrics and propose several new ones.

To the best of our knowledge only one work exists that explores AL for coreference resolution. Gasperin (2009) experiments with an instance based approach in which batches of anaphoric pairs are selected on each iteration of AL. In these experiments, AL did not outperform the passive learning baseline, probably due to selecting batches of large size.

3 Active Learning Configurations

3.1 Instance Selection

The first active learning model we considered selects individual training instances – putatively coreferent mention pairs. This method is quite easy to simulate, and follows naturally from most of the theoretical active learning literature, but it has the drawback of being seemingly unrealistic as an annotation paradigm. That is, since coreference can span across an entire document, it is probably not practical to

[2] http://www.nlm.nih.gov/research/umls/

have a human expert annotate only a single instance at a time when a given instance may require many sentences of reading in order to contextualize the instance and properly label it. Moreover, even if such an annotation scheme proved viable, it may result in an annotated corpus that is only valuable for one type of coreference system architecture.

Nonetheless, active learning for coreference at the instance level is still useful. First, since this method most closely follows the successful active learning literature by using the smallest discrete problems, it can serve as a proof of concept for active learning in the coreference task – if it does not work well at this level, it probably will not work at the document level. Previous results (Gasperin, 2009) have shown that certain multiple instance methods do not work for coreference resolution, so testing on smaller selection sizes first can ensure that active learning is even viable at that scale. In addition, though instance selection may not be feasible for real world annotations, individual instances and metrics for selecting them are usually used as building blocks for more complex methods. In order for this to be possible it must be shown that the instances themselves have some value.

3.2 Document Selection

Active learning with document selection is a much more realistic representation of conventional annotation methods. Conventionally, a set of documents is selected, and each document is annotated exhaustively for coreference (Pradhan et al., 2011; Savova et al., 2011). Document selection fits into this workflow very naturally, by selecting the next document to annotate exhaustively based on some metric of which document has the best instances. In theory, this method can save annotation time by only annotating the most valuable documents.

Document selection is somewhat similar to the concept of *batch-mode active learning*, wherein multiple instances are selected at once, though batch-mode learning is usually intended to solve a different problem, that of an asymmetry between classifier training speed and annotation speed (Settles, 2010). A more important difference is that document selection requires that all of the instances in the batch must come from the same document. Thus, one might expect a priori that document selection for active learning will not perform as well as instance selection. However, it is possible that even smaller gains will be valuable for improving annotation time, and the more robust nature of a corpus annotated in such a way will make the long term benefits worthwhile.

In this work, we propose several metrics for selecting documents to annotate, all of which are based on instance level uncertainty. In the following descriptions, D is the set of documents, d is a single document, $\hat{d}$ is the selected document, $Instances(d)$ is a function which returns the set of pair instances in document d, i is an instance, $dist(i)$ is a function which returns the distance of instance i from the classification boundary, and I is the indicator function, which takes the value 1 if its argument is true and 0 otherwise. Note that high uncertainty occurs when $Abs(dist(i))$ approaches 0.

- Best instance – This method uses the uncertainty sampling criteria on instances, and selects the document containing the instance the classifier is least certain about. $\hat{d} = \underset{d \in D}{\operatorname{argmin}}[\min_{i \in Instances(d)} Abs(dist(i))]$
- Highest average uncertainty – This method computes the average uncertainty of all instances in a document, and selects the document with the highest average uncertainty. $\hat{d} = \underset{d \in D}{\operatorname{argmin}} \frac{1}{|Instances(d)|} \sum_{i \in Instances(d)} Abs(dist(i))$
- Least bad example – This method uses uncertainty sampling criteria to find the document whose most certain example is least certain, in other words the document whose most useless example is least useless. $\hat{d} = \underset{d \in D}{\operatorname{argmin}} \max_{i \in Instances(d)} Abs(dist(i))$
- Narrow band – This method creates an uncertainty band around the discriminating boundary and selects the document with the most examples inside that narrow band. $\hat{d} = \underset{d \in D}{\operatorname{argmax}} \sum_{i \in Instances(d)} I(Abs(dist(i) < 0.2))$
- Smallest spread – This method computes the distance between the least certain and most certain instances and selects the document minimizing that distance.

$\hat{d} = \underset{d \in D}{\operatorname{argmin}}[\max_{i \in Instances(d)}(Abs(dist(i))) - \min_{i \in Instances(d)}(Abs(dist(i)))]$

- Most positives – This method totals the number of positive predicted instances in each document and selects the document with the most positive instances. $\hat{d} = \underset{d \in D}{\operatorname{argmax}} \sum_{i \in Instances(d)} I(dist(i) > 0)$
- Positive ratio – This method calculates the percentage of positive predicted instances in each document and selects the document with the highest percentage. $\hat{d} = \underset{d \in D}{\operatorname{argmax}} \frac{\sum_{i \in Instances(d)} I(dist(i) > 0)}{|Instances(d)|}$

Many of these are straightforward adaptations of the instance uncertainty criteria, but others deserve a bit more explanation. The *most positives* and *positive ratio* metrics are based on the observation that the corpus is somewhat imbalanced – for every positive instance there are roughly 20 negative instances. These metrics try to account for the possibility that instance selection focuses on positive instances. The *average uncertainty* is an obvious attempt to turn instance metrics into document metrics, but *narrow band* and *smallest spread* metrics attempt to do the same thing while accounting for skew in the distribution of “good” and “bad” instances.

3.3 Document-Inertial Instance Selection

One of the biggest impracticalities of instance selection is that labeling any given instance may require reading a fair amount of the document, since the antecedent and anaphor can be quite far apart. Thus, any time savings accumulated by only annotating an instance is reduced since the reading time per instance is probably increased.

It is also possible that document selection goes too far in the other direction, and requires too many useless instances to be annotated to achieve gains. Therefore, we propose a hybrid method of document-inertial instance selection which attempts to combine aspects of instance selection and document selection.

This method uses instance selection criteria to select new instances, but will look inside the current document for a new instance within an uncertainty threshold rather than selecting the most uncertain instance in the entire training set. Sticking with the same document for several instances in a row can potentially solve the real world annotation problem that marking up each instance requires some knowledge of the document context. Instead, the context learned by selecting one instance can be retained if useful for annotating the next selected instance from the same document.

This also preserves one of the biggest advantages of instance selection, that of re-training the model after every selected instance. In batch-mode selection and document selection, many instances are selected according to criteria based on the same model starting point. As a result, the selected instances may be redundant and document scores based on accumulated instance scores may not reflect reality. Re-training the model between selected instances prevents redundant instances from being selected.

4 Evaluation

Evaluations of the active learning models described above took place in a simulation context. In active learning simulations, a labeled data set is used, and the unlabeled pool is simulated by ignoring or “covering” the labels for part of the data until the selection algorithm selects a new instance for annotation. After selection the next data point is simply put into the training data and its label is uncovered.

The data set used was the Ontology Development and Information Extraction (ODIE) corpus (Savova et al., 2011) used in the 2011 i2b2/VA Challenge on coreference resolution.[3] We used a set of 64 documents from the training set of the Mayo Clinic notes for our simulations.

Instances were created by using the training pipeline from the coreference system described in Section 2.1. As previously mentioned, this work uses the named entity anaphor classifier as it contains the most data points. This training set resulted in 6820 instances, with 311 positive instances and 6509 negative instances. Baseline ten-fold cross validation performance on this data set using an SVM with RBF kernel is an F-score of 0.48.

Simulations are performed using ten fold cross-validation. First, each data point is assigned to one

[3] https://www.i2b2.org/NLP/Coreference/

of ten folds (this is done randomly to avoid any autocorrelation issues). Then, for each iteration, one fold is made the seed data, another fold is the validation data, and the remainder are the unlabeled pool. Initially the labeled training data contains only the seed data set. The model is trained on the labeled training data, tested on the validation set, then used to select the next data point from the pool data set. The selected data point is then removed from the pool and added to the training data with its gold standard label(s), and the process repeats until the pool of unlabeled data is empty. Performance is averaged across folds to minimize the effects of randomness in seed and validation set selection. Typically, active learning is compared to a baseline of passive learning where the next data point to be labeled is selected from the unlabeled pool data set randomly.

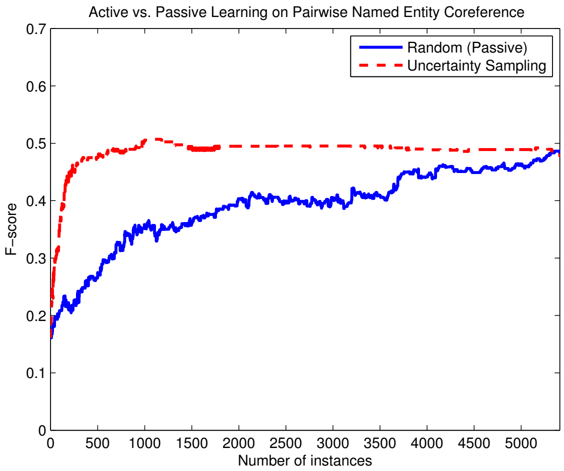

Figure 1: Instance selection simulation results. The x-axis is number of instances and the y-axis is ten-fold averaged f-score of the pairwise named entity classifier.

4.1 Instance Selection Experiments

Instance selection simulations follow the general template above, with each instance (representing a putative antecedent-anaphor pair) randomly assigned to a fold. After scoring on the validation set, uncertainty sampling is used to select a single instance from the unlabeled pool, and that instance is added to the training set.

Figure 1 shows the results of active learning using uncertainty selection on instances versus using passive learning (random selection). This makes it clear that if the classifier is allowed to choose the data, top performance can be achieved much faster than if the data is presented in random order. Specifically, the performance for uncertainty selection levels off at around 500 instances into the active learning, out of a pool set of around 5500 instances. In contrast, the passive learning baseline takes basically the entire dataset to reach the same performance.

This is essentially a proof of concept that there is such a thing as a “better” or “worse” instance when it comes to training a classifier for coreference. We take this as a validation for attempting a document selection experiment, with many metrics using instance uncertainty as a building block.

4.2 Document Selection Experiments

Document selection follows similarly to the instance selection above. The main difference is that instead of assigning pair vectors to folds, we assign documents to folds. To make a selection, each instance is labeled according to the model, document level metrics described in Section 3.2 are computed per document, and the document is selected which optimizes the metric being evaluated. All of that document’s instances and labels are added to the training data, and the process repeats as before.

The results of these experiments are divided into two plots for visual clarity. Figure 2 shows the results of these experiments, roughly divided into those that work as well as a random baseline (left) and those that seem to work worse than a random baseline (right). The best performing metrics (on the left side of the figure) are *Positive Ratio*, *Least Worst*, *Highest Average*, and *Narrow Band*, although none of these performs noticeably better than random. The remaining metrics (on the right) seem to do worse than random, taking more instances to reach the peak performance near the end.

The performance of document selection suggests that it may not be a viable means of active learning. This may be due to a model of data distribution in which useful instances are distributed very uniformly throughout the corpus. In this case, an average document will only have 8–10 useful instances and many times as many that are not useful.

This was investigated by follow-up experiments on the instance selection which kept track of which

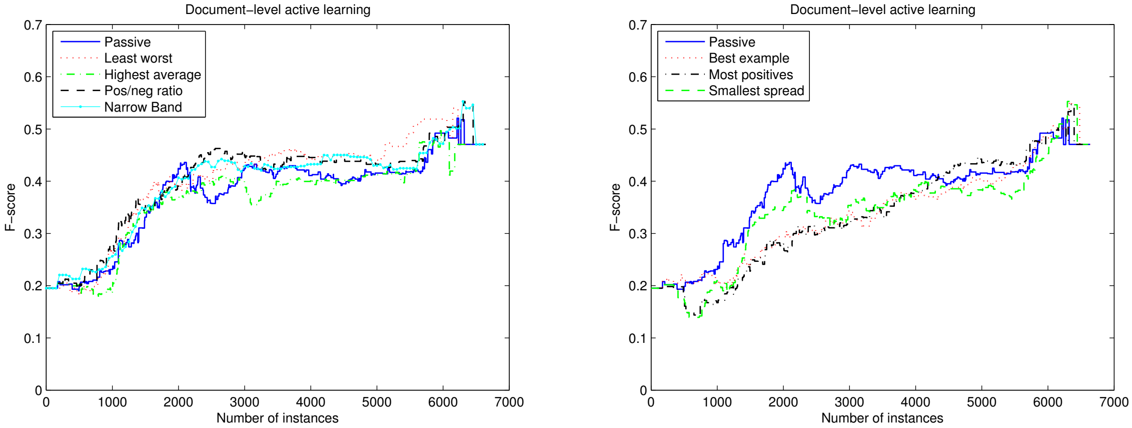

Figure 2: Two sets of document selection experiments.

document each instance came from. The experiments tracked the first 500 instances only, which is roughly the number of instances shown in Figure 1 to reach peak performance. Figure 3 (left) shows a histogram with document indices on the x-axis and normalized instance counts on the y-axis. The counts are normalized by total number of document vectors. In other words, we wanted to show whether there was a distinction between "good" documents containing lots of good instances and "bad" documents with few good instances.

The figure shows a few spikes, but most documents have approximately 10% of their instances sampled, and all but one document has at least one instance selected. Further investigation shows that the spikes in the figure are from shorter documents. Since shorter documents have few instances overall but always at least one positive instance, they will be biased to have a higher ratio of positive to negative instances. If positive instances are more uncertain (which may be the case due to the class imbalance), then shorter documents will have more selected instances per unit length.

We performed another follow-up experiment along these lines using the histogram as a measure of document value. In this experiment, we took the normalized histogram, selected documents from it in order of normalized number of items selected, and used that as a document selection technique. Obviously this would be "cheating" if used as a metric for document selection, but it can serve as a check on the viability of document selection. If the results are better than passive document selection, then there is some hope that a document level metric based on the uncertainty of its instances can be successful.

In fact, the right plot on Figure 3 shows that the "cheating" method of document selection still does not look any better than random document selection.

4.3 Document-Inertial Instance Selection Experiments

The experiments for document-inertial instance selection were patterned after the instance selection paradigm. However, each instance was bundled with metadata representing the document from which it came. In the first selection, the algorithm selects the most uncertain instance, and the document it comes from is recorded. For subsequent selections, the document which contained the previously selected instance is given priority when looking for a new instance. Specifically, each instance in that document is classified, and the confidence is compared against a threshold. If the document contains instances meeting the threshold, the most uncertain instance was selected. After each instance, the model is retrained as in normal instance selection, and the new model is used in the next iteration of the selection algorithm. For these experiments, the threshold is set at 0.75, where the distance between the classification boundary and the margin is 1.0.

Figure 4 shows the performance of this algorithm compared to passive and uncertainty sampling. Per-

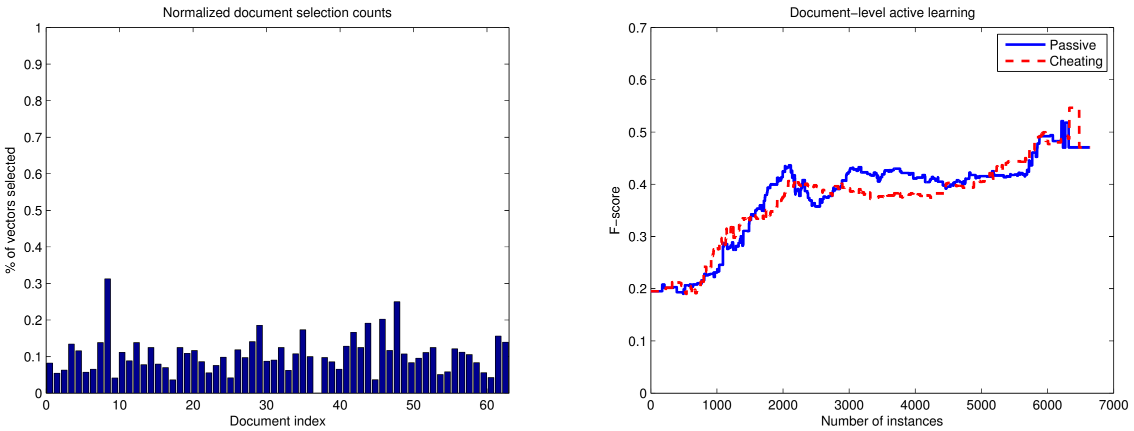

Figure 3: Left: Percentage of instances selected from each document. Right: Performance of a document selection algorithm that can 'cheat' and select the document with the highest proportion of good instances.

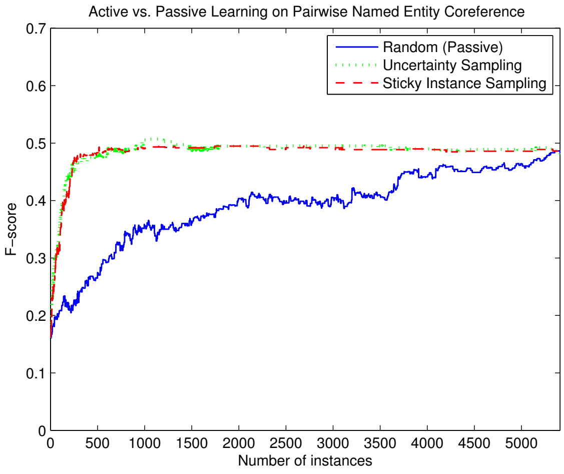

Figure 4: Document-inertial instance selection results.

formance using this algorithm is clearly better than passive learning and is similar to standard uncertainty selection ignoring document constraints.

5 Discussion and Conclusion

The results of these experiments paint a complex picture of the way active learning works for this domain and model combination. The first experiments with uncertainty selection indicate that the number of instances required to achieve classifier performance can be compressed. Selecting and training on all the good instances first leads to much faster convergence to the asymptotic performance of the classifier given the features and data set.

Attempting to extend this result to document selection met with mediocre results. Even the best performing of seven attempted algorithms seems to be about the same as random document selection. One can interpret these results in different ways.

The most pessimistic interpretation is that document selection simply requires too many useless instances to be annotated, good instances are spread too evenly, and so document selection will never be meaningfully faster than random selection. This interpretation seems to be supported by experiments showing that even if document selection uses a "cheating" algorithm to select the documents with the highest proportion of good instances it still does not beat a passive baseline.

One can also interpret these results to inspire further work, first by noting that all of the selection techniques attempt to build on the instance selection metrics. While our document selection metrics were more sophisticated than simply taking the n-best instances, Settles (2010) notes that some successful batch mode techniques explicitly account for diversity in the selections, which we do not. In addition, one could argue that our experiments were unduly constrained by the small number of documents available in the unlabeled pool, and that with a larger unlabeled pool, one would eventually encounter documents with many good instances. This may be true, but may be difficult in practice as clinical notes often need to be manually de-identified before any research use, and so it is not simply a matter of querying all records in an entire electronic medical record system.

The document-inertial instance selection showed that the increase in training speed can be maintained without switching documents for every instance. This suggests that while good training instances may be uniformly distributed, it is usually possible to find multiple *good enough* instances in the current document, and they can be found despite not selecting instances in the exact best order that plain instance selection would suggest.

Future work is mainly concerned with real world applicability. Document level active learning can probably be ruled out as being non-beneficial despite being the easiest to work into annotation work flows. Instance level selection is very efficient in achieving classifier performance but the least practical.

Document-inertial seems to provide some compromise. It does not completely solve the problems of instance selection, however, as annotation will still not be complete if done exactly as simulated here. In addition, the assumption of savings is based on a model that each instance takes a constant amount of time to annotate. This assumption is probably true for tasks like word sense disambiguation, where an annotator can be presented one instance at a time with little context. However, a better model of annotation for tasks like coreference is that there is a constant amount of time required for reading and understanding the context of a document, then a constant amount of time on top of that per instance.While modeling annotation time may provide some insight, it will probably be most effective to undertake empirical annotation experiments to investigate whether document-inertial instance selection actually provides a valuable time savings.

The final discussion point is that of producing complete document annotations. For coreference systems following the pairwise discriminative approach as in that described in Section 2.1, a corpus annotated instance by instance is useful. However, many recent approaches do some form of document-level clustering or explicit coreference chain building, and are not natively able to handle incompletely annotated documents.[4]

Future work will investigate this issue by quantifying the value of complete gold standard annotations versus the partial annotations that may be produced using document-inertial instance selection. One way of doing this is in simulation, by training a model on the 500 good instances that document-inertial instance selection selects, and then classifying the rest of the training instances using that model to create a "diluted" gold standard. Then, a model trained on the diluted gold standard will be used to classify the validation set and performance compared to the version trained on the full gold standard corpus. Similar experiments can be performed using other systems. The logic here is that if an instance was not in the top 10% of difficult instances it can be classified with high certainty. The fact that positive instances are rare and tend to be most uncertain is a point in favor of this approach – after all, high accuracy can be obtained by guessing in favor of negative once the positive instances are labeled. On the other hand, if document-inertial instance selection simply amounts to labeling of positive instances, it may not result in substantial time savings.

In conclusion, this work has shown that instance selection works for coreference resolution, introduced several metrics for document selection, and proposed a hybrid selection approach that preserves the benefits of instance selection while offering the potential of being applicable to real annotation. This work can benefit the natural language processing community by providing practical methods for increasing the speed of coreference annotation.

Acknowledgments

The project described was supported by award number NLM RC1LM010608, the Strategic Health IT Advanced Research Projects (SHARP) Program (90TR002) administered by the Office of the National Coordinator for Health Information Technology, and Integrating Informatics and Biology to the Bedside (i2b2) NCBO U54LM008748. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NLM/NIH/ONC.

[4] Other recent unsupervised graphical model approaches using Gibbs sampling (Haghighi and Klein, 2007) may be able to incorporate partially annotated documents in semi-supervised training.

References

Caroline Gasperin. 2009. Active learning for anaphora resolution. In *Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing*, pages 1–8.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 848–855.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task*, pages 1–27.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*, 17(5):507–513.

Guergana K. Savova, Wendy W. Chapman, Jiaping Zheng, and Rebecca S. Crowley. 2011. Anaphoric relations in the clinical narrative: corpus creation. *J Am Med Inform Assoc*, 18:459–465.

A.I. Schein and L.H. Ungar. 2007. Active learning for logistic regression: an evaluation. *Machine Learning*, 68(3):235–265.

B. Settles, M. Craven, and S. Ray. 2008. Multiple-instance active learning. *Advances in Neural Information Processing Systems (NIPS)*, 20:1289–1296.

Burr Settles. 2010. Active learning literature survey. Technical report, University of Wisconsin–Madison.

Katrin Tomanek, Florian Laws, Udo Hahn, and Hinrich Schütze. 2009. On proper unit selection in active learning: co-selection effects for named entity recognition. In *HLT '09: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 9–17, Morristown, NJ, USA. Association for Computational Linguistics.

S. Tong and D. Koller. 2002. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66.

Jiaping Zheng, Wendy Webber Chapman, Rebecca S. Crowley, and Guergana K. Savova. 2011. Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of Biomedical Informatics*, 44:1113–1122.

Jiaping Zheng, Wendy W Chapman, Timothy A Miller, Chen Lin, Rebecca S Crowley, and Guergana K Savova. 2012. A system for coreference resolution for the clinical narrative. *Journal of the American Medical Informatics Association*.