Active Learning Selection Strategies for Information Extraction

Aidan Finn Nicholas Kushmerick
Smart Media Institute, Computer Science Department, University College Dublin, Ireland
{aidan.finn, nick}@ucd.ie

Abstract

The need for labeled documents is a key bottleneck in adaptive information extraction. One way to solve this problem is through active learning algorithms that require users to label only the most informative documents. We investigate several document selection strategies that are particularly relevant to information extraction. We show that some strategies are biased toward recall, while others are biased toward precision, but it is difficult to ensure both high recall and precision. We also show that there is plenty of scope for improved selection strategies, and investigate the relationship between the documents selected and the relative performance between two strategies.

1 Introduction

Information extraction (IE) is the process of identifying a set of pre-defined relevant items in text documents. For example, an IE system might convert free text resumes into a structured form for insertion in a relational database. Numerous machine learning (ML) algorithms have been developed that promise to eliminate the need for hand-crafted extraction rules. Instead, users are asked to annotate a set of training documents selected from a large collection of unlabeled documents. From these annotated documents, an IE learning algorithm generalizes a set of rules that can be used to extract items from unseen documents.

It is not feasible for users to annotate large numbers of documents. IE researchers have therefore investigated active learning (AL) techniques to automatically identify documents for the user to annotate [13, 12, 3].

The essence of AL is a strategy for selecting the next document to be presented to the user for annotation. The selected documents should be those that will maximize the future performance of the learned extraction rules. Document selection algorithms attempt to find regions of the instance space that have not yet been sampled in order to select the most informative example for human annotation. The nature of IE means that, compared to text classification, it becomes even more important to employ AL techniques. Documents are more expensive to mark-up for IE as rather than being a member of a single class, a document may contain several examples of fields to be extracted.

Several selection strategies have been studied in the more general context of machine learning. For example, confidence-based approaches select for annotation the unlabeled instance of which the learner is least confident. While such techniques are clearly applicable to IE, we focus on novel selection algorithms that exploit the fact that the training data in question is text.

AL in the context of IE is problematic, but also offers new opportunities. It is problematic in that generic approaches require feature encoding of all instances. But for LP^2 [2] and other IE systems, we need to know the details of how the learning algorithm represents a document to compute those features. This does not facilitate completely learner-independent selection strategies.

IE also offers new opportunities for AL. Because the objects in question are text, this can give rise to the possibility of using selection strategies that don't necessarily make sense in a 'generic' ML setting. For example, one of our strategies selects documents according to the frequency of common personal names.

In this paper, we investigate several selection strategies and their application to IE (Sec. 3). We show that different strategies offer a trade-off between precision or recall (Sec. 4). Some strategies improve recall at the expense of precision, while others improve precision at the expense of recall. We also estimate the optimal performance of an IE algorithm and show that there is plenty of scope for improving existing selection strate-

gies.

Furthermore, we show that the difference in performance between two selection strategies can be (weakly) predicted from the correlation between the documents they select (Sec. 5).

2 Related work

There has been a large amount of work on adaptive information extraction, e.g. [2, 1, 9] and many others. These algorithms generally perform well, but all have the potential for further improvement through active learning techniques.

Active learning refers to a variety of ways that a learning algorithm can control the training data over which it generalizes. For example, a learner might construct synthetic instances and ask the user to label them. We focus on so-called selective-sampling strategies [5], in which the learner picks an instance for the user to label from a large pool of unlabeled instances.

Selective sampling techniques are generally regarded as being of two types: confidence- or certainty-based [10], or committee-based [6]. In each case, the learner has built a model using a certain number of labeled training documents, and must select the next document to be labeled with the goal of choosing the document that will give the maximum information.

In confidence-based approaches, the learner examines unlabeled examples and attaches a confidence (usually based on the certainty with which a prediction can be made about the document) to them. Documents with low confidence are chosen to be labeled. Typically, methods for estimating certainty are based on the probability that the learner will classify a new example correctly.

In committee-based approaches, a committee of learners is constructed and each member attempts to label unlabeled documents. Documents that maximize disagreement between committee members are chosen for labeling. In fact, committee-based approaches can be regarded as confidence-based, where the confidence in a prediction is based on the agreement among committee members about that prediction.

There has been some work in the application of active learning to IE (e.g. [13, 11, 12]). [12] use learning-algorithm-specific heuristics to choose the next document for annotation. Specifically, their AL algorithm for learning Hidden Markov Models (HMM) identifies "difficult" unlabeled tokens and asks the user to label them. Difficulty is estimated by the difference between the most likely and second most likely state of the HMM.

Other applications of AL and IE do not rely on a specific learning algorithm. [13] use certainty-based sampling, where the certainty of an extracted field is the minimum of the training-set accuracies of the rules that extracted the fragment. [11] describe a multi-view approach to IE. Multi-view AL is a committee-based approach in which the committee members are formed by training on different sets of features. Muslea et al. learn two different models for extraction based on two different views of the data, and select the document where both models disagree, but are most confident in their predictions.

3 Selection strategies

3.1 Notation and terminology

The aim of an active learning selection strategy is to select documents in a way that improves performance over random selection. A selection strategy should select the document for labeling that is most informative. The difficulty is estimating how informative a document will be without knowing the labels associated with that document or the features that will represent the document. We have identified two main approaches to estimating the informativeness of a document: confidence-based and distance-based.

Confidence-based. The first approach is to try to directly estimate the informativeness of a document x using some measure of uncertainty $f(x)$. From information theory, the amount of information gained from labeling a document is equal to the uncertainty about that document before labeling it [10]. Most learning learning algorithms support some method of estimating confidence on unseen documents. For example, one can invoke a set of learned rules on a document, and then compute a confidence for the document based on the training-set accuracies of the rules that apply to that document. Other types of approaches such as multi-view and committee-based can also be regarded as confidence-based. Multi-view approaches estimate uncertainty using using some measure of disagreement between models built using different views, while committee-based approaches estimate the confidence using agreement between committee members.

Given some confidence measure f and a pool of unlabeled documents U, a confidence-based selection strategy will pick the unlabeled document x that minimizes this measure:

$$x \equiv \arg\min_{x' \in U} f(x')$$

Distance-based. The second approach is based on the idea that for any set of instances, there is (by definition) some set of documents O that optimizes performance over the unselected documents. Furthermore, one can assume that O can be generated from some distance metric $d_O(x, x')$ over documents, by selecting the $|O|$ documents that maximize the pair-wise distance between the members of O. For example, if the learning algorithm is a covering algorithm, then performance should be maximized with a sample that covers the instance space uniformly. So the second approach is to define some distance metric $d(x, x')$ that closely approximates $d_O(x, x')$, and then sampling uniformly from that space. Rather than trying to find documents that we have low confidence in, we are trying to find documents that are different to those already seen. Specifically, given some distance metric $d(x, x')$, a set of previously selected documents S, and a pool of unlabeled data U, a distance-based selection strategy will pick the unlabeled document x that maximizes the distance from x to the members of S:

$$x \equiv \arg\max_{x' \in U} \sum_{x'' \in S} d(x', x'')$$

Of course, distance-based approaches can also be thought of as confidence-based where confidence is estimated as distance from previously seen instances. This is a less direct measure of confidence than other approaches so we feel that it warrants separate categorization.

3.2 The strategies

We introduce several novel AL document selection strategies for IE. Some of the strategies are applicable only in an IE or text classification context. While they are tailored for IE, they are generic in that they do not assume any specific IE algorithm. The learning algorithm that we use is LP^2 [2] but the active learning strategies that we investigate are not particular to our choice of learning algorithm and so we could easily substitute another IE algorithm such as BWI [9] or Rapier [1].

COMPARE. This strategy selects for annotation the document that is textually least similar to the documents that have already been annotated. We select the document that is textually most dissimilar to the documents already in the corpus. The idea is to sample uniformly from the document space, using the notion of textual similarity to approximate a uniform distribution. This is a distance-based selection strategy. Similarity can be measured in various ways, such as raw term overlap, or using TFIDF weighting but our distance metric $d(x, x')$ is the inverse of the number of words that occur in both x and x' divided by the number of words that occur in x or x'. Note that COMPARE is very fast, because the learning algorithm does not need to be invoked on the previously-selected documents in order to select the next document.

EXTRACTCOMPARE. This strategy selects for annotation the document where what is extracted from the document is textually most dissimilar to the documents in the training corpus. This is similar to Compare, except that the distance metric is $d(x, extract(x'))$, where $extract(x')$ applies the learned extraction rules to the document x'. The idea here is to select documents that don't contain text that we are already able to extract. EXTRACTCOMPARE is quite slow, because the learning algorithm must be invoked on the previously-selected documents in order to select the next document.

MELITA [4]. MELITA selects for annotation the document that matches the fewest patterns that the learning algorithm has learned from the training corpus. This is a confidence based metric. $f(x) = |extract(x)|$. This approach is similar to EXTRACTCOMPARE. It selects documents that do not match patterns that we have already learned. Like EXTRACTCOMPARE, MELITA is quite slow. Note that MELITA is essentially a special case of the approach described in [13] in that the confidences of the extracted items are ignored.

NAMEFREQ. Often the items to be extracted are people's names, but these can be difficult to extract, because they are likely to be words that the learner has not seen before. NAMEFREQ selects for annotation the document with the most unusual personal names. Specifically, NAMEFREQ assigns a part of speech tag to each term in document x, and then uses $f(x) = 1/\sum_{p \in x} n(p)$, where $p \in x$ iterates over the proper nouns in document x, and $n(p)$ is the frequency of

proper noun p as a personal name according to recent US Census data. We assume that the learner is less confident about names that are unusual as it is less likely to have seen these names before. Like COMPARE, NAMEFREQ is very fast.

BAG. Bagging is a standard approach in machine learning. We apply it to IE by invoking the learning algorithm on different partitions of the available training data and selecting the document that maximizes disagreement between the models built on different partitions of the training data. The training set is divided into two partitions and a model built using each as it's training set. The document is selected where the two learners extract the most dissimilar text. This is a committee-based strategy (and thus confidence-based), where the members of the committee comprise learners built on different partitions of the training data. The confidence of prediction is estimated based on agreement between the two learned models. BAG is very slow.

ENSEMBLE. It is common in machine learning to use the combined predictions of different learning algorithms to improve performance. We can similarly with IE seek to combine selections of different selection strategies to improve learning rate. This approach is an ensemble learner based on the MELITA and NAMEFREQ strategies. It selects half of those documents that NAMEFREQ would pick and half of those that MELITA would pick. This strategy was designed after examination of the performance of the other selection strategies. The aim to to try to simultaneously maximize both precision and recall. ENSEMBLE is quite slow.

4 Experiments

We have evaluated our selection algorithms on two information extraction tasks, and report our results in the form of the learning curve for each selection strategy.

Each learning curve was averaged over ten runs. Documents are added to the training-set in batches of size 10. For each selection strategy, the first 10 documents are picked at random, while subsequent batches are chosen according to the selection strategy. Each point on the learning curve shows the accuracy of the learning algorithm when trained on the selected documents and tested on the rest.

We compare our results to two baselines: a trivial strategy that selects documents randomly, and an "omniscient" optimal strategy. Because finding the true optimal is combinatorially prohibitive, we use a greedy estimate of the optimal (at each step, the greedy algorithms selects the one document that will result in the largest increase in performance). That is, the optimal selection x given a set of previously selected documents S and a pool U of unlabelled documents with respect to some measure M (eg, precision, recall or F1) is

$$x \equiv \arg\max_{x' \in U} M(S \cup \{x'\}).$$

We include this data as an estimate of the upper bound on the performance of any selection strategy. Finally, because even the greedy implementation requires a large amount of CPU time, we report the optimal results for just a small number of documents.

4.1 Seminar announcements

The SA dataset consists of 473 seminar announcements [7]. For each document we wish to extract the speaker, location, start-time and end-time.

Fig. 1 shows the learning curves for F1, precision and recall generated on this dataset. Looking at F1 shows that random selection is one of the better strategies. In fact only MELITA and COMPARE perform better than the random selection strategy on this extraction task, but the difference is small. However, recall that COMPARE is much faster than MELITA, so COMPARE is more suitable for the interactive scenarios that motivate MELITA [4]. NAMEFREQ performs considerably worse that the other selection strategies.

If we look at precision and recall separately, we get a clearer picture of the performance of each strategy. MELITA performs best when recall is considered followed by COMPARE and EXTRACTCOMPARE. All of these are significantly better than random. NAMEFREQ is the worst performer.

If we look at the precision learning curve, this trend is reversed. NAMEFREQ gives the highest precision, while MELITA and EXTRACTCOMPARE give the worst precision. COMPARE gives slightly better precision than random and better recall than random.

On this task, NAMEFREQ gives the best improvement in precision, while it is the worst when recall is considered. Conversely MELITA offers the best improvement in recall, but performs worst when precision is considered.

Each strategy seems to bias toward either improving precision or improving recall. Some strategies can be

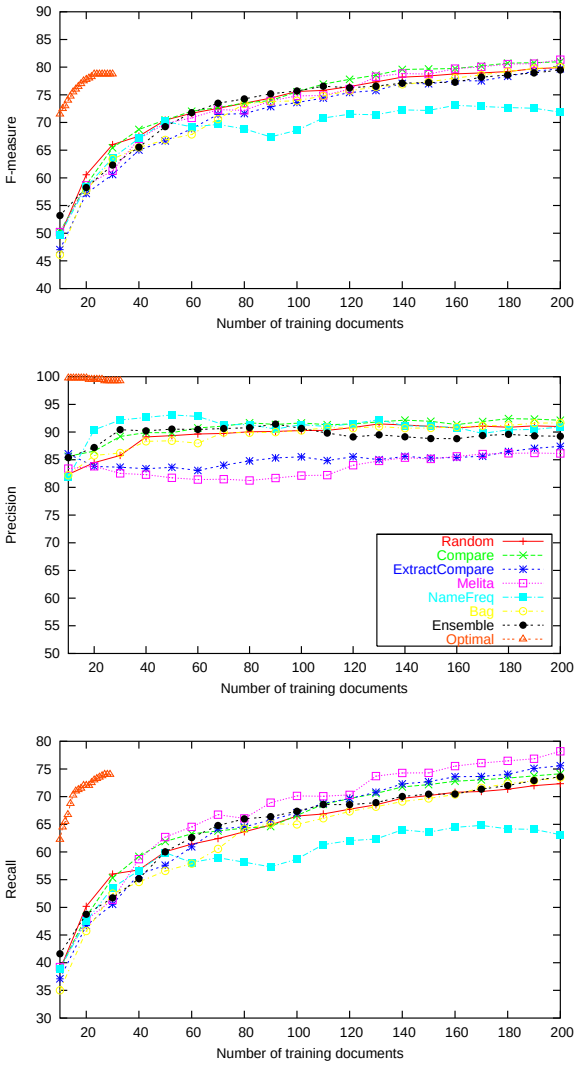

Figure 1: Learning curves for the SA dataset.

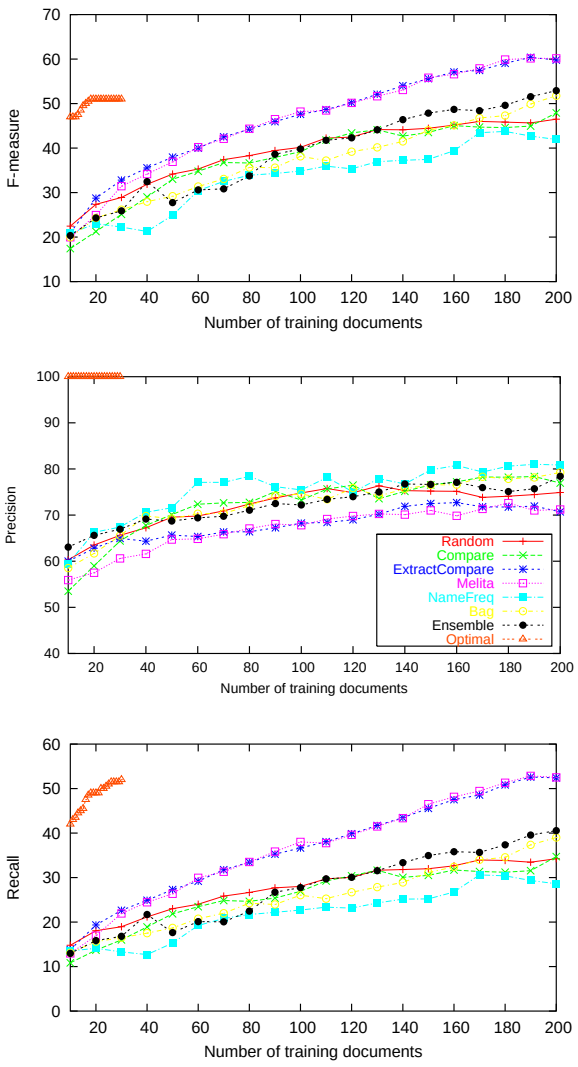

Figure 2: Learning curve for the ACQ dataset.

used to improve recall performance, while others can be used to improve precision performance. Other strategies that perform closer to random don't offer significant improvements in either precision or recall.

4.2 Reuters acquisitions articles

The ACQ dataset consists of 300 Reuters articles describing corporate acquisitions [8]. The task is to extract the name of the purchasing and acquired companies.

Fig. 2 shows the learning curves for the various selection strategies on this dataset. In this case, the results are somewhat more clear cut. When looking at F1, MELITA and EXTRACTCOMPARE are significantly better than the other strategies. NAMEFREQ is again the worst. On this task, the difference in recall performance is large enough to be reflected as a large difference in the F1 performance. The boost in recall using these strategies is greater than the resulting drop in precision. As on the SA dataset, when precision is considered, NAMEFREQ performs best, with MELITA and EXTRACTCOMPARE performing worst. The relative performance of the selection strategies is reversed

```
--------------------------------------------------------------------------
KEY CENTURION {KEYC} COMPLETES ACQUISITIONS
CHARLESTON, W.Va., April 2 - Key Centurion Bancshares Inc said it has
completed the previously-announced acquisitions of Union Bancorp of West
Virginia Inc and Wayne Bancorp Inc.
Reuter
--------------------------------------------------------------------------
                          HCI & G SEMINAR
                     Wednesday, January 13, 1993
                           3:30 - 5:00pm
                          Wean Hall 5409

               Aiding Performance in Complex Dynamic Worlds:
                         Some HCI Challenges

                           Emilie Roth
                     Information Technology Dept.
                Westinghouse Science and Technology Center

We have been studying crew problem-solving and decision-making in
simulated power plant emergencies with the objective of developing the
next generation computerized control room.  Power plant control rooms
offer some unique challenges to HCI. Because they are complex ...
--------------------------------------------------------------------------
```

Figure 3: The most-informative ACQ (top) and SA (bottom) documents.

when we consider precision instead of recall. The two strategies that perform best when recall is considered are those that perform worst when precision is considered.

Again this indicates that the various strategies are suited to optimizing either precision or recall. Given this trend, we investigate whether selecting documents according to both kinds of strategy will improve both precision and recall. The ensemble selection strategy selects documents according to both MELITA (improves recall) and NAMEFREQ (improves precision). This approach performs slightly better than random for both precision and recall, but not as well as NAMEFREQ for precision or MELITA for recall.

4.3 Discussion

For each task, we have shown the first few points of the optimal learning curve. On each task, the optimal curve is several times better than the best selection strategy in the early stages of learning. This indicates that there is plenty of scope for improved selection strategies. Indeed the optimal curve shows that the choice of initial training documents can lead to very good performance. For example, on the SA dataset there is a single document (see Fig. 3) that when the learner is trained on, it performs with F1 of 24.25% on the rest of the training corpus. On the ACQ dataset, there is a single document that gives an F-score of 21.5%. On the SA dataset, best performing strategy (MELITA) requires 130 documents to achieve the same performance as the optimal after 20 documents. On the ACQ dataset, MELITA requires 130 documents to achieve the same F1 performance as the optimal strategy after 30 documents. For recall, it requires 190 documents to achieve the same performance as the optimal recall strategy. Even after 200 documents it does not reach the level of performance of the optimal precision curve. This indicates that there are a small number of highly informative examples in the dataset, while all the other documents contribute only very small incremental increases in performance.

There is clear trade-off between optimizing precision, recall or F1. Fig. 4 shows the learning curves when optimizing for F1, precision and recall respectively for the ACQ dataset. The optimal precision curve results in low recall, and vice-versa. This trend is to be expected, but Fig. 4 shows that the trade-off is not complete. While we can maximize precision at 100% if we are prepared to accept very low recall, the optimal recall curve is much lower. We cannot achieve very high recall, even if we are prepared to accept very low precision. We conjecture that this is because, as a covering algorithm, LP^2 is inherently biased to favor precision over recall.

The choice of strategy depends on whether we wish to optimize for precision or recall. We have shown that some strategies perform better than random at improving precision, while others perform better at improving recall.

Given that MELITA improves recall and NAMEFREQ improves precision, we attempted to improve both by combining both approaches. However this ENSEMBLE approach does not perform as well as either approach.

5 Predicting performance

The previous experiments concerned the relative performance of the selection strategies. From a practical perspective, it is important to be able to predict which strategy will perform best, without having to actually try the strategies and measure the results. We now turn to some preliminary results that address this issue.

In order to predict the relative performance of the different selection strategies, we need to find some informative property of the strategies that can be measured without knowing the labels of the unlabeled data. We have used the the correlation between the documents selected by each strategy. Our hypothesis is that if two strategies tend to select the same documents, then they will have similar performance, while if two strategies select very different documents, then there will be a large performance gap between the two. Our ultimate

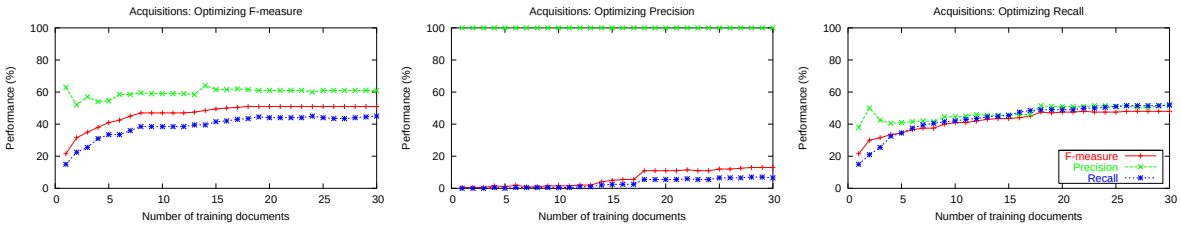

Figure 4: Optimal learning curves for F1, precision and recall on the ACQ dataset.

goal is to derive such a relationship analytically. We now consider empirical evidence that partially supports this hypothesis.

To measure the degree of agreement between two strategies, we first randomly select 50 documents. Then, in batches of 10, we selected the remaining documents using each selection strategy. This was repeated 10 times and the average Spearman rank correlation coefficient calculated for each pair of strategies. Strategies that select documents in the same order have a correlation of $+1$, while strategies that select documents in the opposite order have a correlation of -1.

On both tasks, there is a strong positive correlation between EXTRACTCOMPARE and MELITA, indicating that they both tend to pick the same documents. There is also a positive correlation between ENSEMBLE and MELITA and NAMEFREQ. This is expected as ENSEMBLE combines these two strategies.

On the SA task, there is quite a strong negative correlation between NAMEFREQ and MELITA. There is a slight negative correlation between these strategies on the ACQ dataset. This indicates that these strategies tend to select different documents.

To determine whether selection agreement is useful for predicting relative performance, we then measured the performance gap between the strategies. We define gap(x, y) as the normalized performance difference, averaged over all points on the learning curve from 50 to 200 documents.

Fig. 5 shows the selection agreement between various selection strategy pairs plotted against the gap in performance between the strategies. We display SA and ACQ in different plots, and we measure the gap in precision, recall and F1. Anecdotally, it is apparent that our ability to predict the performance gap is quite good for strategies that are highly correlated (either positively or negatively), but rather poor when the strategies are weakly correlated.

More precisely, our hypothesis that selection agreement can be used to predict performance gap is validated to the extent that these data have a correlation of -1. Fig. 6 shows the six correlations. As anticipated, all of the correlations are negative, though weakly so. Our approach is slightly better at predicting the performance gap for SA compared to ACQ, and for predicting the recall gap compared to precision and F1.

6 Conclusion

We have investigated several Active Learning selection strategies that can be applied to Information Extraction. Of these, several performed significantly better than a random selection strategy. MELITA and EXTRACTCOMPARE offer improved recall over random selection with a resulting drop in precision. NAMEFREQ offers improved precision at the expense of recall. Some strategies offer improvements in recall while others improve precision, but it is difficult to get significant improvement in both recall and precision. Most importantly, there is still however a significant difference in performance between the optimal curve and the various selection strategies. Existing selection strategies still have significant scope for improvement.

Our immediate future work involves identifying strategies that bridge the wide gap between the optimal strategy and the strategies we have investigated so far. For example, we are exploring a committee-based strategy called DUAL that has two committee members for each field: one that extracts the field itself, and one that extracts all document fragments except the particular field. We are also conducting a detailed analysis of the optimal documents to determine strategies that can bridge the gap.

A second goal is to improve our ability to predict the performance gap between two strategies. Ultimately, we seek a theoretically-grounded model of active learning that will enable us to derive upper or lower bounds

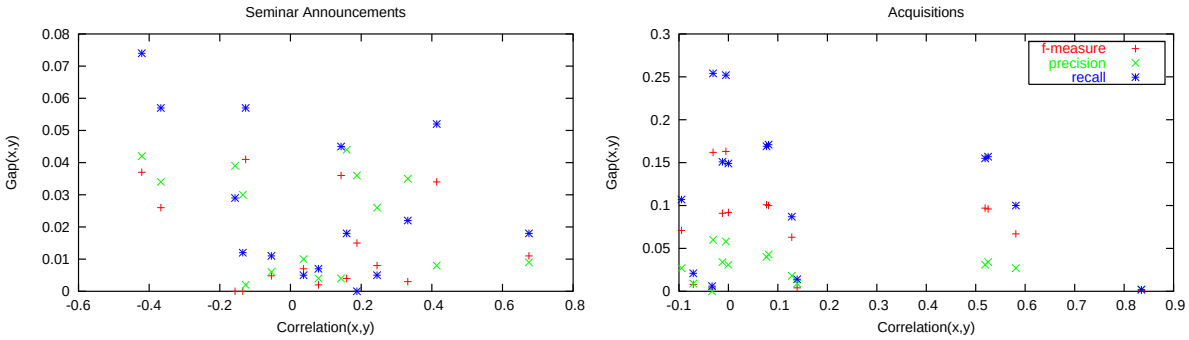

Figure 5: Performance gap vs. selection correlation.

	F1	P	R	mean
SA	-0.27	-0.32	-0.43	-0.34
ACQ	-0.22	-0.25	-0.23	-0.23
mean	-0.25	-0.29	-0.33	

Figure 6: The correlation between two strategies' performance gap and the degree to which they select the same documents.

on the performance of a given strategy.

Acknowledgements

This research was supported by grants SFI/01/F.1/C015 from Science Foundation Ireland, and N00014-03-1-0274 from the US Office of Naval Research. We thank Fabio Ciravegna for access to LP^2.

References

[1] M. Califf and R. Mooney. Relational learning of pattern-match rules for information extraction. In *Proc. 16th Nat. Conf. Artifical Intelligence*, 1999.

[2] F. Ciravegna. Adaptive information extraction from text by rule induction and generalisation. In *Proc. 17th Int. Joint Conf. Artificial Intelligence*, 2001.

[3] F. Ciravegna, A. Dingli, D. Petrelli, and Y. Wilks. Timely and non-intrusive active document annotation via adaptive information extraction. In *ECAI Workshop Semantic Authoring Annotation and Knowledge Management*, 2002.

[4] F. Ciravegna, A. Dingli, D. Petrelli, and Y. Wilks. User-system cooperation in document annotation based on information extraction. In *13th International Conference on Knowledge Engineering and Knowledge Management*, 2002.

[5] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2), 1994.

[6] I. Dagan and S. P. Engelson. Committee-based sampling for training probabilistic classifiers. In *International Conference on Machine Learning*, 1995.

[7] D. Freitag. *Machine Learning for Information Extraction in Informal Domains*. PhD thesis, Carnegie Mellon University, 1998.

[8] D. Freitag. Toward general-purpose learning for information extraction. In *35th Annual Meeting of the Association for Computational Linguistics*, 1998.

[9] D. Freitag and N. Kushmerick. Boosted wrapper induction. In *Proc. 17th Nat. Conf. Artificial Intelligence*, 2000.

[10] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *11th International Conference on Machine Learning*, 1994.

[11] I. Muslea, S. Minton, and C. Knoblock. Selective sampling with reduntant views. In *Proc. 17th Nat. Conf. Artificial Intelligence*, 2000.

[12] T. Scheffer and S. Wrobel. Active learning of partially hidden Markov models. *Active Learning, Database Sampling, Experimental Design: Views on Instance Selection*, 2001.

[13] C. Thompson, M. Califf, and R. Mooney. Active learning for natural language processing and information extraction. In *Proc. 16th Int. Conf. Machine Learning*, 1999.