Optimising Selective Sampling for Bootstrapping Named Entity Recognition

Markus Becker M.BECKER@ED.AC.UK
Ben Hachey BHACHEY@INF.ED.AC.UK
Beatrice Alex V1BALEX@INF.ED.AC.UK
Claire Grover GROVER@INF.ED.AC.UK
School of Informatics, University of Edinburgh, 2 Buccleuch Place, Edinburgh, EH8 9LW, UK

Abstract

Training a statistical named entity recognition system in a new domain requires costly manual annotation of large quantities of in-domain data. Active learning promises to reduce the annotation cost by selecting only highly informative data points. This paper is concerned with a real active learning experiment to bootstrap a named entity recognition system for a new domain of radio astronomical abstracts. We evaluate several committee-based metrics for quantifying the disagreement between classifiers built using multiple views, and demonstrate that the choice of metric can be optimised in simulation experiments with existing annotated data from different domains. A final evaluation shows that we gained substantial savings compared to a randomly sampled baseline.

1. Introduction

The training of statistical named entity recognition (NER) systems requires large quantities of manually annotated data. Manual annotation however is typically costly and time-consuming. Furthermore, successful application of NER is dependent on training data from the same domain. Thus, bootstrapping NER in a new domain typically requires acquisition of new annotated data. Active learning promises to reduce the total amount of labelled data by selectively sampling the most informative data points.

We introduce the newly created Astronomical Bootstrapping Corpus (ABC), which contains abstracts of radio astronomical papers, and report on our assessment of active learning methods for bootstrapping a statistical named entity recognition (NER) system for this new domain.

Appearing in *Proceedings of the Workshop on Learning with Multiple Views*, 22^{nd} ICML, Bonn, Germany, 2005.

As part of our methodology, we experimented with a NER system in a known domain with existing corpus resources, namely the Genia corpus of biomedical abstracts (Kim et al., 2003). We tested relevant active learning parameters in simulation experiments with a view to arrive at an optimal setting for a real active learning experiment in the new astronomical domain. This was of particular importance since we were budgeted only 1000 sentences for active learning annotation.

We employ a committee-based method where trained classifiers are caused to be different by employing multiple views of the feature space. The degree of deviation of the classifiers with respect to their analysis can tell us if an example is potentially useful. We evaluate various metrics to quantify disagreement and demonstrate that the choice of metric can be optimised in simulation experiments with existing annotated data from distinct domains.

In the following section, we present the new corpus of astronomy abstracts developed for the bootstrapping task. In section 3, we introduce our active learning set-up for bootstrapping named entity recognition. Next, section 4 contains experimental results for a series of simulated active learning experiments used for parameter optimisation and section 5 contains the bootstrapping results. Finally, section 6 contains conclusions and future work.

2. The Corpus

2.1. Astronomical Named Entities

The main purpose of the corpus development work was to provide materials for assessing methods of porting a statistical NER system to a new domain. To do this we needed to create a small annotated corpus in a new domain which would serve as a basis for experiments with bootstrapping. Our chosen new domain was abstracts of radio astronomical papers and our corpus consists of abstracts taken from the NASA Astrophysics Data System archive, a digital library for physics, astrophysics, and instrumentation (http://adsabs.harvard.edu/preprint_service.html).

On the Column Density of <**Source-type**>AGN</**Source-type**> Outflows: The Case of <**Source-name**>NGC 5548</**Source-name**>

We reanalyze the <**Instrument-name**>Hubble Space Telescope</**Instrument-name**> high-resolution spectroscopic data of the intrinsic absorber in <**Source-name**>NGC 5548</**Source-name**> and find that the <**Spectral-feature**>C IV absorption</**Spectral-feature**> column density is at least 4 times larger than previously determined. This increase arises from accounting for the kinematical nature of the absorber and from our conclusion that the outflow does not cover the narrow emission line region in this object. The improved column density determination begins to bridge the gap between the high column densities measured in the X-ray and the low ones previously inferred from the <**Spectral-feature**>UV lines</**Spectral-feature**>. Combined with our findings for outflows in high-luminosity <**Source-type**>quasars</**Source-type**>, these results suggest that traditional techniques for measuring column densities – equivalent width, curve of growth, and Gaussian modeling – are of limited value when applied to UV absorption associated with <**Source-type**>active galactic nucleus</**Source-type**> outflows.

Figure 1. An example abstract.

Our choice of new domain was driven partly by longer-term plans to build an information extraction system for the astronomy domain and partly by the similarities and differences between this domain and the biomedical domain that the initial NER tagger is trained on. The main point of similarity between the two data sets is that they are both comprised of scientific language taken from abstracts of academic papers. The main difference lies in the technical terms and in the named entities that are recognised.

Following consultation with our astronomy collaborators, we created a cohesive dataset in the radio astronomy domain, and established an inventory of four domain-specific named entity types. The dataset was created by extracting abstracts from the years 1997-2003 that matched the query "quasar AND line". 50 abstracts from the year 2002 were annotated as seed material and 159 abstracts from 2003 were annotated as testing material. 778 abstracts from the years 1997-2001 were provided as an unannotated pool for bootstrapping. On average, these abstracts contain 10 sentences with an average length of 30 tokens. The corpus was annotated for the four entity types below (frequencies in the seed set in brackets). Fig. 1 shows an example text from this corpus.

Instrument-name Names of telescopes and other measurement instruments, e.g. *Superconducting Tunnel Junction (STJ) camera, Plateau de Bure Interferometer, Chandra, XMM-Newton Reflection Grating Spectrometer (RGS), Hubble Space Telescope.* [136 entities, 12.7%]

Source-name Names of celestial objects, e.g. *NGC 7603, 3C 273, BRI 1335-0417, SDSSp J104433.04-012502.2, PC0953+ 4749.* [111 entities, 10.4%]

Source-type Types of objects, e.g. *Type II Supernovae (SNe II), radio-loud quasar, type 2 QSO, starburst galaxies, low-luminosity AGNs.* [499 entities, 46.8%]

Spectral-feature Features that can be pointed to on a spectrum, e.g. *Mg II emission, broad emission lines, radio continuum emission at 1.47 GHz, CO ladder from (2-1) up to (7-6), non-LTE line.* [321 entities, 30.1%]

2.2. Corpus Preparation and Annotation

The files were converted from their original HTML to XHTML using Tidy (`http://www.w3.org/People/Raggett/tidy/`), and were piped through a sequence of processing stages using the XML-based tools from the LT TTT and LT XML toolsets (Grover et al., 2000; Thompson et al., 1997) in order to create tokenised XML files. It turned out to be relatively complex to achieve a sensible and consistent tokenisation of this data. The main source of complexity is the high density of technical and formulaic language (e.g. $(N(H_2) \simeq 10_{24} cm^{-2})$, 17.8 h_{70}^{-1} kpc, for $\Omega_m = 0.3$, $\Lambda = 0.7$, 1.4 GHz of 30 μ Jy) and an accompanying lack of consistency in the way publishers convert from the original LaTex encoding of formulae to the HTML which is published on the ADS website. We aimed to tokenise in such a way as to minimise noise in the data, though inevitably not all inconsistencies were removed.

The seed and test data sets were annotated by two astrophysics PhD students using the NITE XML toolkit annotation tool (Carletta et al., 2003). In addition, they annotated 1000 randomly sampled sentences from the pool to provide a baseline for active learning. Inter-annotator agreement was obtained by directly comparing the two annotator's data. Phrase-level f-score is 86.4%. Token-level accuracy is 97.3% which corresponds to a Kappa agreement of K=.925 (N=44775, k=2; where K is the kappa coefficient, N is the number of tokens and k is the number of annotators).

3. Active Learning with Multiple Views

Supervised training of named entity recognition (NER) systems requires large amounts of manually annotated data. However, human annotation is typically costly and time-consuming. Active learning promises to reduce this cost by requesting only those data points for human annotation which are highly informative. Example informativity can be estimated by the degree of uncertainty of a single learner as to the correct label of a data point (Cohn et al., 1995) or in terms of the disagreement of a committee of learners (Seung et al., 1992). Active learning has been successfully applied to a variety of similar tasks such as document classification (McCallum & Nigam, 1998), part-of-speech tagging (Argamon-Engelson & Dagan, 1999), and parsing (Thompson et al., 1999).

We employ a committee-based method where the degree of deviation of different classifiers with respect to their analysis can tell us if an example is potentially useful. Trained classifiers can be caused to be different by bagging (Abe & Mamitsuka, 1998), by randomly perturbing event counts (Argamon-Engelson & Dagan, 1999), or by producing different views using different feature sets for the same classifiers (Jones et al., 2003; Osborne & Baldridge, 2004). In this paper, we present active learning experiments for NER in astronomy texts following the last approach.

3.1. Feature split

We use a conditional Markov model tagger (Finkel et al., 2004) to train two different models on the same seed data by applying a feature split. The feature split as shown in Table 1 was hand-crafted such that it provides different views while empirically ensuring that performance is sufficiently similar. While the first feature set comprises of character sub-strings, BNC frequencies, Web counts, gazetteers and abbreviations, the second set contains features capturing information about words, POS tags, word shapes, NE tags, parentheses and multiple references to NEs. These features are describe in more detail in (Finkel et al., 2004).

Feature Set 1	
Prefix/Suffix	Up to a length of 6
Frequency	Frequency in BNC
Web Feature	Based on Google hits of pattern instantiations
Gazetteers	Compiled from the Web
Abbreviations	abbr_i
	abbr_i + abbr_{i+1}
	abbr_{i-1} + abbr_i + abbr_{i+1}
Feature Set 2	
Word Features	w_i, w_{i-1}, w_{i+1}
	Disjunction of 5 prev words
	Disjunction of 5 next words
TnT POS tags	POS_i, POS_{i-1}, POS_{i+1}
Word Shape	shape_i, shape_{i-1}, shape_{i+1}
	shape_i + shape_{i+1}
	shape_{i-1} + shape_i + shape_{i+1}
Prev NE	NE_{i-1}, NE_{i-2} + NE_{i-1}
	NE_{i-3} + NE_{i-2} + NE_{i-1}
Prev NE + Word	NE_{i-1} + w_i
Prev NE + POS	NE_{i-1} + POS_{i-1} + POS_i
	NE_{i-2} + NE_{i-1} + POS_{i-2} + POS_{i-1} + POS_i
Prev NE + Shape	NE_{i-1} + shape_i
	NE_{i-1} + shape_{i+1}
	NE_{i-1} + shape_{i-1} + shape_i
	NE_{i-2} + NE_{i-1} + shape_{i-2} + shape_{i-1} + shape_i
Paren-Matching	Signals when one parenthesis in a pair has been assigned a different tag in a window of 4 words
Occurrence Patterns	Capture multiple references to NEs

Table 1. Feature split for parameter optimisation experiments

3.2. Level of annotation

For the manual annotation of named entity examples, we needed to decide on the level of granularity. The question arises what constitutes an example that will be submitted to the annotators. Reasonable levels of annotation include the document level, the sentence level and the token level. The most fine-grained annotation would certainly be on the token level. This requires semi-supervised training to allow for partially annotated sentences, as in (Scheffer et al., 2001). However, there are no directly applicable semi-supervised training regimes for discriminative classifiers. On the other extreme, one may submit an entire document for annotation. A possible disadvantage is that a document with some interesting parts may well contain large portions with redundant, already known structures for which knowing the manual annotation may not be very useful. In the given setting, we decided that the best granularity is on the sentence level.

3.3. Sample Selection Metric

There are various metrics that could be used to quantify the degree of deviation between classifiers in a committee (e.g. KL-divergence, information radius, f-measure). The work reported here uses two sentence-level metrics based on KL-divergence and one based on f-score. In the following, we describe these metrics.

KL-divergence has been suggested for active learning to quantify the disagreement of classifiers over the probability distribution of output labels (McCallum & Nigam, 1998) and has been applied to information extraction (Jones et al., 2003). KL-divergence measures the divergence between two probability distributions p and q over the same event

space χ:

$$D(p||q) = \sum_{x \in \chi} p(x) \log \frac{p(x)}{q(x)} \quad (1)$$

KL-divergence is a non-negative metric. It is zero for identical distributions; the more different the two distributions, the higher the KL-divergence. KL-divergence is maximal for cases where distributions are peaked and prefer different labels. Taking a peaked distribution as an indicator for certainty, using KL-divergence thus bears a strong resemblance to the co-testing setting (Muslea, 2002). Intuitively, a high KL-divergence score indicates an informative data point. However, in the current formulation, KL-divergence only relates to individual tokens. In order to turn this into a sentence score, we need to combine the individual KL-divergences for the tokens within a sentence into one single score. We employed mean and max.

The *f-complement* has been suggested for active learning in the context of NP chunking as a structural comparison between the different analyses of a committee (Ngai & Yarowsky, 2000). It is the pairwise f-score comparison between the multiple analyses for a given sentence:

$$f^{\mathcal{M}}_{comp} = \frac{1}{2} \sum_{M,M' \in \mathcal{M}} (1 - F_1(M(t), M'(t))) \quad (2)$$

where F_1 is the balanced f-score of $M(t)$ and $M'(t)$, the preferred analyses of data point t according to different members M, M' of ensemble $\mathcal{M}$. The definition assumes that in the comparison between two analyses, one may arbitrarily assign one analysis as the gold standard and the other one as a test case. Intuitively, examples with a high f-complement score are likely to be informative.

4. Parameter Optimisation Experiments

In the previous section, we described a number of parameters for our approach to active learning. Bootstrapping presents a difficult problem as we cannot optimise these parameters on the target data. The obvious solution is to use a different data set but there is no guarantee that experimental results will generalise across domains. The work reported here addresses this question. We simulated active learning experiments on a data set which consists of biomedical abstracts marked up for the entities DNA, RNA, cell line, cell type, and protein (Kim et al., 2003).[1] Seed, pool, and test sets contained 500, 10,000, and 2,000 sentences respectively, roughly the same size as for the astronomical data. As smaller batch sizes require more retraining iterations and larger batch sizes increase the amount of annotation necessary at each round and could lead to unnecessary strain for the annotators, we settled on a batch size of 50 sentences for the real AL experiment as a compromise between computational cost and work load for the annotator.

We then ran simulated AL experiments for each of the three selection metrics discussed in section 3. The performance was compared to a baseline where examples were randomly sampled from the pool data. Experiments were run until there were 2000 sentences of annotated training material including the sentences from the seed data and the sentences selected from the pool data.

4.1. Costing Active Learning

For quality evaluation, we used the established f-score metric as given by the evaluation scripts developed for the CoNLL NER tasks (Tjong Kim Sang & De Meulder, 2003). In order to assess the relative merits of various active learning scenarios, we will plot learning curves, i.e. the increase in f-score over the invested effort. Ideally, a cost metric should reflect the effort that went into the annotation of examples in terms of time spent. However, a precise time measurement is difficult, or may be not available in the case of simulation experiments. We will therefore consider a number of possible approximations.

A sentence-based cost metric may seem like an obvious cost function, but this may pose problems when different sample selection metrics have a tendency to choose longer or shorter sentences. Thus, we will also consider more fine-grained metrics, namely the number of tokens in a sentence and the number of entities in a sentence.

4.2. Comparison of Selection Metrics

The plots in figure 2 show the learning curves for random sampling and the three AL selection metrics we examined for the parameter optimisation experiments. The first takes the number of sentences as the cost metric and the second and third take the number of tokens and the number of entities respectively.

Random sampling is clearly outperformed by all other selection metrics. The random curve for the sentence cost metric, for example, reaches an f-score of 69% after approximately 1500 sentences have been annotated while the maximum KL-divergence curve reaches this level of performance after only $\approx$ 1100 sentences. This represents a substantial reduction in sentences annotated of 26.7%. In addition, at 1500 sentences, maximum KL-divergence offers an error reduction of 4.9% over random sampling with a 1.5 point improvement in f-score. Averaged KL-divergence offers the same error reduction when using the sentence cost metric, but at 19.3%, a lower reduction of sentences annotated. F-complement performs worst giving 10% cost reduction and 1.6% error reduction.

The learning curves also allow us to easily visualise the

[1] Simulated AL experiments use 5-fold cross-validation.

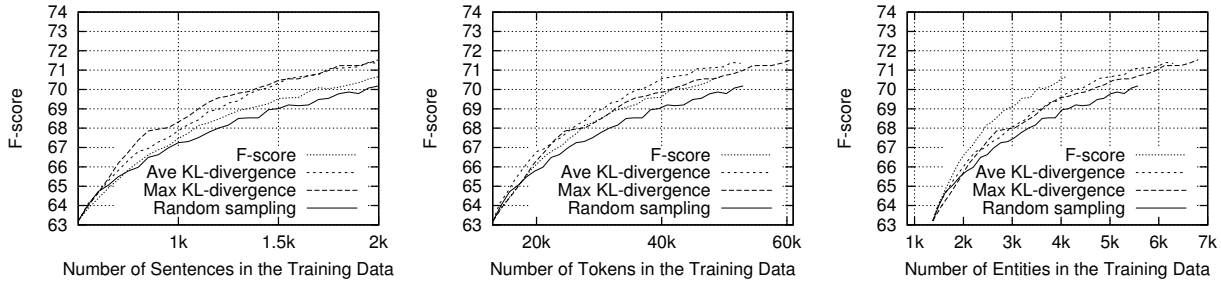

Figure 2. Parameter optimisation learning curves for sentence, token, and entity cost metrics

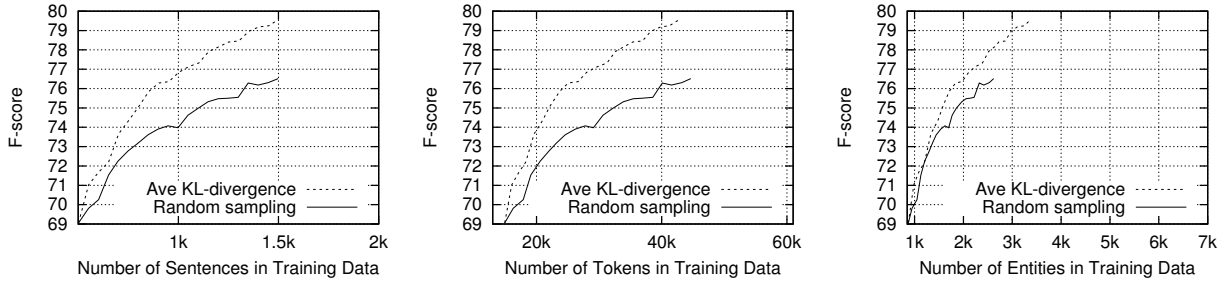

Figure 3. Active annotation learning curves for sentence, token, and entity cost metrics

performance difference of the three selection metrics with respect to each other. The f-complement metric clearly underperforms with respect to KL-divergence based metrics.

According to the learning curves with number of sentences as the cost metric, maximum KL-divergence performs the best. However, when choosing a different cost metric, as for example the number of tokens or entities that occur in each selected sentence, the learning curves behave completely differently as can be seen in the second and third plots in figure 2. This illustrates the fact that the selection metrics operate in different ways preferring shorter or longer sentences with more or less entities. With number of tokens as the cost metric, averaged KL-divergence performs the best with a 23.5% reduction in annotation cost to reach an f-score of 69% and an error reduction of 4.9% at $\approx$ 40,000 tokens. And with entities as the cost metric, the f-complement selection metric seems to perform best. So, the question arises: how do we combine this information to prepare for a real annotation task where we only have a single opportunity to get the best performing and most cost effective system possible.

To explore the behaviour of the three selection metrics further, we also look at the number of tokens and the number of entities in the sentences chosen by each metric. Table 2 contains the number of tokens and entities contained within the selected sentences averaged across the 5 cross-validation results. Comparing these numbers, one can observe the types of sentences preferred by each selection

Metric	Tokens	Entities
Random	26.7 (0.8)	2.8 (0.1)
F-comp	25.8 (2.4)	2.2 (0.7)
KL-max	30.9 (1.5)	3.5 (0.2)
KL-ave	27.1 (1.8)	3.3 (0.2)

Table 2. Average tokens and entities per sentence for different selection metrics (standard deviation in brackets)

metric. While the maximum KL-divergence metric selects the longest sentences containing the most number of entities, the f-complement selection metric chooses the shortest sentences with the least number of entities in them. The averaged KL-divergence metric, on the other hand, generally selects average length sentences which still contain relatively many entities.

As averaged KL-divergence does not affect sentence length, we expect the sentences selected to take less time to annotate than the sentences selected by maximum KL-divergence. And, since these sentences have relatively many entity phrases, we expect to have more positive examples than with the f-complement metric and thus have higher informativity and therefore performance increase per token. Furthermore, sentence length is not the best single unit cost metric. The number of sentences is too coarse as this gives the same cost to very long and very short sentences and does not allow us to consider the types of sentences selected by the various metrics. Likewise, the number of entities does not reflect the fact that every selected

sentence needs to be read regardless of the number of entities it contains, which again covers up effects of specific selection metrics.

5. Active Annotation Results

We developed NEAL, an interactive Named Entity Active Learning tool for bootstrapping NER in a new domain. The tool manages the data and presents batches of selectively sampled sentences for annotation in the same annotation tool used for the seed and test data. The entire abstract is presented for context with the target sentence(s) highlighted. On the basis of the findings of the simulated experiments we set up the real AL experiment using averaged KL-divergence as the selection metric. The tool was initialised with the 50 document seed set described in section 2 and given to the same annotators that prepared the seed and test sets.

As we do not have a model of temporal or monetary cost in terms of our three cost metrics, we evaluate with respect to all three cost metrics. Figure 3 contains learning curves for random sampling and for selective sampling with the averaged KL-divergence selection metric plotted against number of sentences, number of tokens, and number of entities. The initial performance (given only the seed data for training) amounts to an f-score of 69.1%. 50 sentences (with an average of 28 tokens and 2.5 entities per sentence as compared to 29.8 and 2.0 for the randomly sampled data) are added to the training data at each round. After 20 iterations, the training data therefore comprises of 1,502 sentences (containing approx. 43,000 tokens) which leads to an f-score of 79.6%.

Comparing the selective sampling performance to the baseline, we confirm that active learning provides a significant reduction in the number of examples that need annotating. Looking first at the token cost metric, the random curve reaches an f-score of 76% after approximately 39,000 tokens of data has been annotated while the selective sampling curve reaches this level of performance after only $\approx$ 24,000 tokens. As for the optimisation data, this represents a dramatic reduction in tokens annotated of 38.5%. In addition, at 39,000 tokens, selectively sampling offers an error reduction of 13.0% with a 3 point improvement in f-score. Selective sampling with the averaged KL-divergence selection metric also achieves dramatic cost and error rate reductions for the sentence (35.6% & 12.5%) and entity cost metrics (23.9% & 5.0%).

These improvements are comparable to the cost and error reduction achieved in the optimisation data. While it should be taken into account that these domains are relatively similar, this suggests that a different domain can be used to optimise parameters when using active learning to bootstrap NER. This is confirmed not only by an improvement over baseline for the token cost metric but also by an improvement for the sentence and entity cost metrics.

In a companion paper, we report in some more detail about the effects of selective sampling on annotator's performance (Hachey et al., 2005). Even though we find that active learning may result in a slightly higher error rate in the annotation, we demonstrate that active learning still incurs substantial reductions in annotation effort as compared to random sampling.

6. Conclusions and Future Work

We have presented an active learning approach to bootstrapping named entity recognition for which a new corpus of radio astronomical texts has been collected and annotated. We employ a committee-based method that uses two different feature sets for a conditional Markov model tagger and we experiment with several metrics for quantifying the degree of deviation: averaged KL-divergence, maximum KL-divergence, and f-complement.

We started with a NER system tested and optimised in a domain with existing corpus resources and built a system to identify four novel entity types in a new domain of astronomy texts. Experimental results from the real active learning annotation illustrate that the optimised parameters performed well on the new domain. This is confirmed for cost metrics based on the number of sentences, the number of tokens, and the number of entities.

While presenting results with respect to the three cost metrics ensures completeness, it also suggests that the real cost might be better modelled as a combination of these metrics. During annotation, we collected timing information for each sentence and we are currently using this timing information to investigate realistic models of cost based on sentence length and number of entities.

Acknowledgments

We are very grateful for the time and resources invested in corpus preparation by our collaborators in the Institute for Astronomy, University of Edinburgh: Rachel Dowsett, Olivia Johnson and Bob Mann. We would also like to thank Shipra Dingare, Jochen Leidner, Malvina Nissim and Yuval Krymolowski for helpful discussions. Many thanks to Ewan Klein, Miles Osborne, and Bonnie Webber for being instrumental in formulating and organising the task.

We are grateful to the UK National e-Science Centre for giving us access to BlueDwarf, a p690 server donated to the University of Edinburgh by IBM. This work was performed as part of the SEER project, which is supported by a Scottish Enterprise Edinburgh-Stanford Link Grant (R36759).

References

Abe, N., & Mamitsuka, H. (1998). Query learning strategies using boosting and bagging. *Proceedings of the 15th International Conference on Machine Learning*.

Argamon-Engelson, S., & Dagan, I. (1999). Committee-based sample selection for probabilistic classifiers. *Journal of Artificial Intelligence Research*, *11*, 335–360.

Carletta, J., Evert, S., Heid, U., Kilgour, J., Robertson, J., & Voormann, H. (2003). The NITE XML Toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers, special issue on Measuring Behavior*, *35*.

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1995). Active learning with statistical models. *Advances in Neural Information Processing Systems* (pp. 705–712). The MIT Press.

Finkel, J., Dingare, S., Nguyen, H., Nissim, M., Manning, C., & Sinclair, G. (2004). Exploiting context for biomedical entity recognition: From syntax to the web. *Proceedings of the COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*.

Grover, C., Matheson, C., Mikheev, A., & Moens, M. (2000). LT TTT—a flexible tokenisation tool. *Proceedings of the 2nd International Conference on Language Resources and Evaluation*.

Hachey, B., Alex, B., & Becker, M. (2005). Investigating the effects of selective sampling on the annotation task. *Proceedings of CoNLL 2005, Ann Arbor, USA*.

Jones, R., Ghani, R., Mitchell, T., & Riloff, E. (2003). Active learning for information extraction with multiple view feature sets. *ECML 2003 Workshop on Adaptive Text Extraction and Mining*.

Kim, J.-D., Ohta, T., Tateishi, Y., & Tsujii, J. (2003). Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, *19*, 180–182.

McCallum, A., & Nigam, K. (1998). Employing EM and pool-based active learning for text classification. *Proceedings of the 15th International Conference on Machine Learning*.

Muslea, I. (2002). *Active learning with multiple views*. Doctoral dissertation, University of Southern California.

Ngai, G., & Yarowsky, D. (2000). Rule writing or annotation: Cost-efficient resource usage for base noun phrase chunking. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.

Osborne, M., & Baldridge, J. (2004). Ensemble-based active learning for parse selection. *Proceedings of the 5th Conference of the North American Chapter of the Association for Computational Linguistics*.

Scheffer, T., Decomain, C., & Wrobel, S. (2001). Active hidden markov models for information extraction. *Proceedings of the International Symposium on Intelligent Data Analysis*.

Seung, H. S., Opper, M., & Sompolinsky, H. (1992). Query by committee. *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*.

Thompson, C. A., Califf, M. E., & Mooney, R. J. (1999). Active learning for natural language parsing and information extraction. *Proceedings of the 16th International Conference on Machine Learning*.

Thompson, H., Tobin, R., McKelvie, D., & Brew, C. (1997). LT XML. Software API and toolkit for XML processing. `http://www.ltg.ed.ac.uk/software/`.

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *Proceedings of the 2003 Conference on Computational Natural Language Learning*.