Towards Automated Semantic Analysis on Biomedical Research Articles

Donghui Feng **Gully Burns** **Jingbo Zhu** **Eduard Hovy**
Information Sciences Institute
University of Southern California
Marina del Rey, CA, 90292
{donghui, burns, jingboz, hovy}@isi.edu

Abstract

In this paper, we present an empirical study on adapting Conditional Random Fields (CRF) models to conduct semantic analysis on biomedical articles using active learning. We explore uncertainty-based active learning with the CRF model to dynamically select the most informative training examples. This abridges the power of the supervised methods and expensive human annotation cost.

1 Introduction

Researchers have experienced an increasing need for automated/semi-automated knowledge acquisition from the research literature. This situation is especially serious in the biomedical domain where the number of individual facts that need to be memorized is very high.

Many successful information extraction (IE) systems, work in a supervised fashion, requiring human annotations for training. However, human annotations are either too expensive or not always available and this has become a bottleneck to developing supervised IE methods to new domains.

Fortunately, active learning systems design strategies to select the most informative training examples. This process can achieve certain levels of performance faster and reduce human annotation (e.g., Thompson et al., 1999; Shen et al., 2004).

In this paper, we present an empirical study on adapting CRF model to conduct semantic analysis on biomedical research literature. We integrate an uncertainty-based active learning framework with the CRF model to dynamically select the most informative training examples and reduce human annotation cost. A systematic study with exhaustive experimental evaluations shows that it can achieve satisfactory performance on biomedical data while requiring less human annotation.

Unlike direct estimation on target individuals in traditional active learning, we use two heuristic certainty scores, *peer comparison certainty* and *set comparison certainty*, to indirectly estimate sequences labeling quality in CRF models.

We partition biomedical research literature by experimental types. In this paper, our goal is to analyze various aspects of useful knowledge about tract-tracing experiments (TTE). This type of experiments has prompted the development of several curated databases but they have only partial coverage of the available literature (e.g., Stephan et al., 2001).

2 Related Work

Knowledge Base Management Systems allow individual users to construct personalized repositories of knowledge statements based on their own interaction with the research literature (Stephan et al., 2001; Burns and Cheng, 2006). But this process of data entry and curation is manual. Current approaches on biomedical text mining (e.g., Srinivas et al., 2005; OKanohara et al., 2006) tend to address the tasks of named entity recognition or relation extraction, and our goal is more complex: to extract computational representations of the minimum information in a given experiment type.

Pattern-based IE approaches employ seed data to learn useful patterns to pinpoint required fields values (e.g. Ravichandran and Hovy, 2002; Mann and Yarowsky, 2005; Feng et al., 2006). However, this only works if the data corpus is rich enough to learn variant surface patterns and does not necessarily generalize to more complex situations, such as our domain problem. Within biomedical articles, sentences tend to be long and the prose structure tends to be more complex than newsprint.

The CRF model (Lafferty et al., 2001) provides a compact way to integrate different types of features for sequential labeling problems. Reported work includes improved model variants (e.g., Jiao et al., 2006) and applications such as web data extraction (Pinto et al., 2003), scientific citation extraction (Peng and McCallum, 2004), word alignment (Blunsom and Cohn, 2006), and discourse-level chunking (Feng et al., 2007).

Pool-based active learning was first successfully applied to language processing on text classification (Lewis and Gale, 1994; McCallum and Nigam, 1998; Tong and Koller, 2000). It was also gradually applied to NLP tasks, such as information extraction (Thompson et al., 1999); semantic parsing (Thompson et al., 1999); statistical parsing (Tang et al., 2002); NER (Shen et al., 2004); and Word Sense Disambiguation (Chen et al., 2006). In this paper, we use CRF models to perform a more complex task on the primary TTE experimental results and adapt it to process new biomedical data.

3 Semantic Analysis with CRF Model

3.1 What knowledge is of interest?

The goal of TTE is to chart the interconnectivity of the brain by injecting tracer chemicals into a region of the brain and then identifying corresponding labeled regions where the tracer is transported to. A typical TTE paper may report experiments about one or many labeled regions.

Name	Description
injectionLocation	the named brain region where the injection was made.
tracerChemical	the tracer chemical used.
labelingLocation	the region/location where the labeling was found.
labelingDescription	a description of labeling, density or label type.

Table 1. Minimum knowledge schema for a TTE.

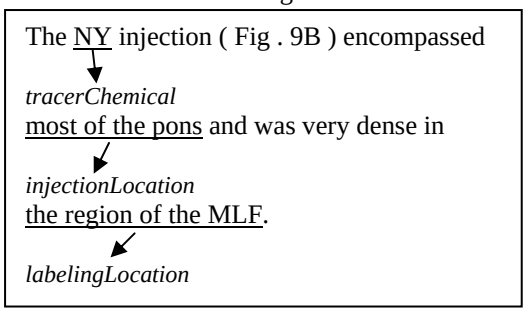

Figure 1. An extraction example of TTE description.

In order to construct the minimum information required to interpret a TTE, we consider a set of specific components as shown in Table 1.

Figure 1 gives an example of description of a complete TTE in a single sentence. In the research articles, this information is usually spread over many such sentences.

3.2 CRF Labeling

We use a plain text sentence for input and attempt to label each token with a field label. In addition to the four pre-defined fields, a default label, "O", is used to denote tokens beyond our concern.

In this task, we consider five types of features based on language analysis as shown in Table 2.

Name	Feature	Description
Lexical Knowledge	TOPOGRAPHY	Is word topographic?
	BRAIN_REGION	Is word a region name?
	TRACER	Is word a tracer chemical?
	DENSITY	Is word a density term?
	LABELING_TYPE	Does word denote a labeling type?
Surface Word	Word	Current word
Context Window	CONT_INJ	If current word if within a window of injection context
Window Words	Prev-word	Previous word
	Next-word	Next word
Dependency Features	Root-form	Root form of the word if different
	Gov-verb	The governing verb
	Subj	The sentence subject
	Obj	The sentence object

Table 2. The features for system labeling.

Lexical Knowledge. We define lexical items representing different aspects of prior knowledge. To this end we use names of brain structures taken from brain atlases, standard terms to denote neuroanatomical topographical spatial relationships, and common sense words for labeling descriptions. We collect five separate lexicons as shown in Table 3.

Lexicons	# of terms	# of words
BRAIN_REGION	1123	5536
DENSITY	8	10
LABELING_TYPE	9	13
TRACER	30	30
TOPOGRAPHY	9	36
Total	1179	5625

Table 3. The five lexicons.

Surface word. The word token is an important indicator of the probable label for itself.

Context Window. The TTE is a description of the inject-label-findings context. Whenever we find a word with a root form of "injection" or "deposit", we generate a context window around this word and all the words falling into this window are assigned a feature of "CON_INJ". This means when labeling these words the system should consider the very current context.

Window Words. We also use all the words occurring in the window around the current word. We set the window size to only include the previous and following words (window size = 1).

Dependency Features. To untangle word relationships within each sentence, we apply the dependency parser MiniPar (Lin, 1998) to parse each sentence, and then derive four types of features. These features are (a) root form of word, (b) the subject in the sentence, (c) the object in the sentence, and (d) the governing verb for each word.

4 Uncertainty-based Active Learning

Active learning was initially introduced for classification tasks. The intuition is to always add the most informative examples to the training set to improve the system as much as possible.

We apply an uncertainty/certainty score-based approach. Unlike traditional classification tasks, where disagreement or uncertainty is easy to obtain on target individuals, information extraction tasks in our problem take a whole sequence of tokens that might include several slots as processing units. We therefore need to make decisions on whether a full sequence should be returned for labeling.

Estimations on confidence for single segments in the CRF model have been proposed by (Culotta and McCallum, 2004; Kristjannson et al., 2004). However as every processing unit in the data set is at the sentence level and we make decisions at the sentence level to train better sequential labeling models, we define heuristic scores at the sentence level.

Symons et al. (2006) presents multi-criterion for active learning with CRF models, but our motivation is from a different perspective. The labeling result for every sentence corresponds to a decoding path in the state transition network. Inspired by the decoding and re-ranking approaches in statistical machine translation, we use two heuristic scores to measure the degree of correctness of the top labeling path, namely, *peer comparison certainty* and *set comparison certainty*.

Suppose a sentence S includes n words/tokens and a labeling path at position m in the ranked N-best list is represented by $L^m = (l_0, l_1, ..., l_{n-1})$. Then the probability of this labeling path is represented by $P(L^m)$, and we have the following two equations to define the peer comparison certainty score, $Score_{peer}(S)$ and set comparison certainty score, $Score_{set}(S)$:

$$Score_{peer}(S) = \frac{P(L^1)}{P(L^2)} \quad (1)$$

$$Score_{set}(S) = \frac{P(L^1)}{\sum_{k=1}^{N} P(L^k)} \quad (2)$$

For peer comparison certainty (Eq. 1), we calculate the ratio of the top-scoring labeling path probability to the second labeling path probability. A high ratio means there is a big jump from the top labeling path to the second one. The higher the ratio score, the higher the relative degree of correctness for the top labeling path, giving system higher confidence for those with higher peer comparison certainty scores. Sentences with lowest certainty score will be sent to the oracle for manual labeling.

In the labeling path space, if a labeling path is strong enough, its probability score should dominate all the other path scores. In Equation 2, we compute the set comparison certainty score by considering the portion of the probability of the path in the overall N-best labeling path space. A large value means the top path dominates all the other labeling paths together giving the system a higher confidence on the current path over others.

We start with a seed training set including k labeled sentences. We then train a CRF model with the training data and use it to label unlabeled data. The results are compared based on the certainty scores and those sentences with the lowest certainty scores are sent to an oracle for human labeling. The new labeled sentences are then added to the training set for next iteration.

5 Experimental Results

We first investigated how the active learning steps could help for the task. Second, we evaluated how the CRF labeling system worked with different sets of features. We finally applied the model to new

biomedical articles and examined its performance on one of its subsets.

5.1 Experimental Setup

We have obtained 9474 *Journal of Comparative Neurology (JCN)*[1] articles from 1982 to 2005. For sentence labeling, we collected 21 TTE articles from the JCN corpus. They were converted from PDF files to XML files, and all of the article sections were identified using a simple rule-based approach. As most of the meaningful descriptions of TTEs appear in the Results section, we only processed the Results section. The 21 files in total include 2009 sentences, in which 1029 sentences are meaningful descriptions for TTEs and 980 sentences are not related to TTEs.

We randomly split the sentences into a training pool and a testing pool, under a ratio 2:1. The training pool includes 1338 sentences, with 685 of them related to TTEs, while 653 not. Testing was based on meaningful sentences in the testing pool. Table 4 gives the configurations in the data pools.

	# of Related Sentences	# of Unrelated Sentences	Sum
Training Pool	685	653	1338
Testing Pool	344	327	671
Sum	1029	980	2009

Table 4. Training and testing pool configurations.

5.2 Evaluation Metrics

As the label "O" dominates the data set (70% out of all tokens), a simple accuracy score would provide an inappropriate high score for a baseline system that always chooses "O". We used Precision, Recall, and F_Score to evaluate only meaningful labels.

5.3 How well does active learning work?

For the active learning procedure, we initially selected a set of seed sentences related to TTEs from the training pool. At every step we trained a CRF model and labeled sentences in the rest of the training pool. As described in section 4, those with the lowest rank on certainty scores were selected. If they are related to a TTE, human annotation will be added to the training set. Otherwise, the system will keep on selecting sentences until it finds enough related sentences.

People have found active learning in batch mode is more efficient, as in some cases a single additional training example will not improve a classifier/system that much. In our task, we chose the bottom k related sentences with the lowest certainty scores. We conducted various experiments for k = 2, 5, and 10. We also compared experiments with passive learning, where at every step the new k related sentences were randomly selected from the corpus. Figures 2, 3, and 4 give the learning curves for precision, recall, and F_Scores when k = 10.

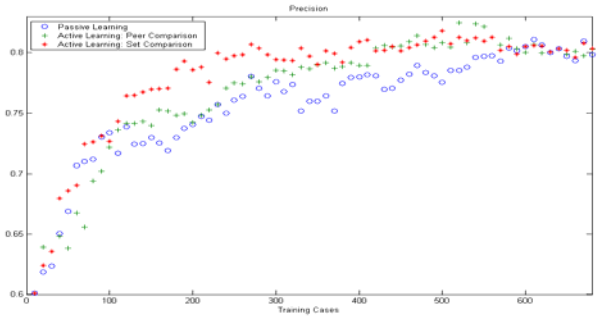

Figure 2. Learning curve for Precision.

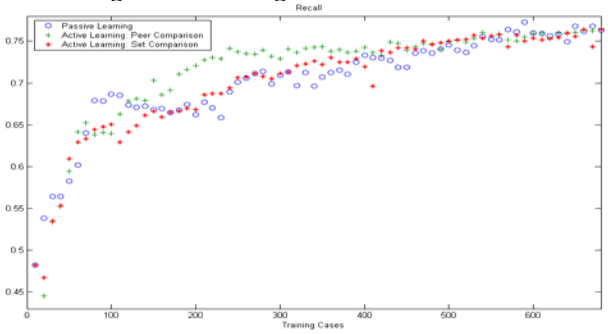

Figure 3. Learning curve for Recall.

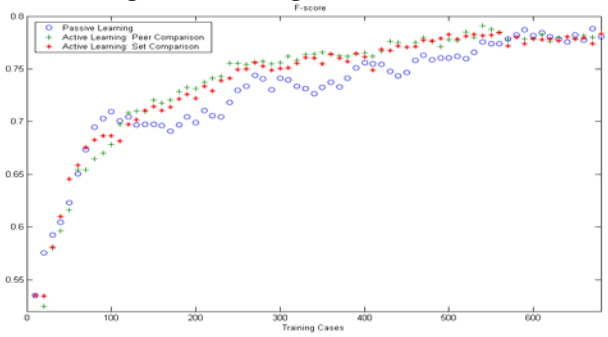

Figure 4. Learning curve for F_Score.

From these figures, we can see active learning approaches required fewer training examples to achieve the same level of performance. As we iteratively added new labeled sentences into the training set, the precision scores of active learning were steadily better than that of passive learning as the uncertain examples were added to strengthen

[1] http://www3.interscience.wiley.com/cgi-bin/jhome/31248

existing labels. However, the recall curve is slightly different. Before some point, the recall score of passive learning was a little better than active learning. The reason is that examples selected by active learning are mainly used to foster existing labels but have relatively weaker improvements for new labels, while passive learning has the freedom to add new knowledge for new labels and improve recall scores faster. As we keep on using more examples, the active learning catches up with and overtakes passive learning on recall score.

These experiments demonstrate that under the framework of active learning, examples needed to train a CRF model can be greatly reduced and therefore make it feasible to adapt to other domains.

5.4 How well does CRF labeling work?

As we added selected annotated sentences, the system performance kept improving. We investigated system performance at the final step when all the related sentences in the training pool are selected into the training set. The testing set also only includes the related sentences. This results in 685 training sentences and 344 testing sentences.

To establish a baseline for our labeling task, we simply scanned every sentence for words or phrases from each lexicon. If the term was present, then we labeled the word based on the lexicon in which it appeared. If words appeared in multiple lexicons, we assigned labels randomly.

System Features	Prec.	Recall	F_Score
Baseline	0.4067	0.1761	0.2458
Lexicon	0.5998	0.3734	0.4602
Lexicon + Surface Words	0.7663	0.7302	0.7478
Lexicon + Surface Words + Context Window	0.7717	0.7279	0.7491
Lexicon + Surface Words + Context Window + Window Words	0.8076	0.7451	0.7751
Lexicon + Surface Words + Context Window + Window Words + Dependency Features	**0.7991**	**0.7828**	**0.7909**

Table 5. Precision, Recall, and F_Score for labeling.

We tried exhaustive feature combinations. Table 5 shows system performance with different feature combinations. All systems performed significantly higher than the baseline. The sole use of lexicon knowledge produced poor performance, and the inclusion of surface words produced significant improvement. The use of window words boosted precision and recall. The performance with all the features generated an F_score of 0.7909.

We explored how system performance reflects different labels. Figure 5 and 6 depict the detailed distribution of system labeling from the perspective of precision and recall respectively for the system with the best performance. Most errors occurred in the confusion of *injectionLocation* and *labelingLocation*, or of the meaningful labels and "O".

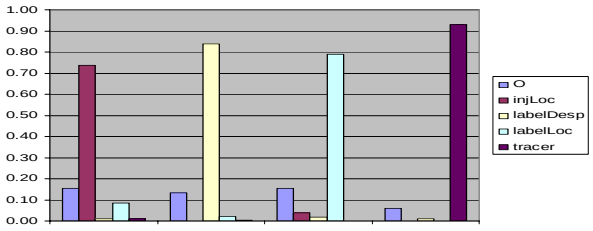

Figure 5. Precision confusion matrix distribution.

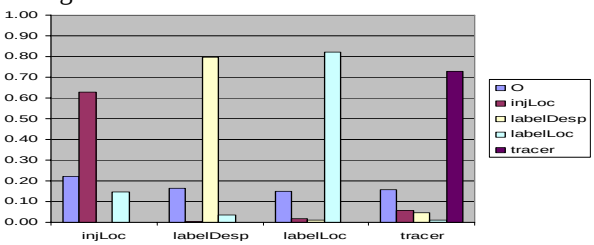

Figure 6. Recall confusion matrix distribution.

The worst performance occurred for files that distinguish themselves from others by using fairly different writing styles. We believe given more training data with different writing styles, the system could achieve a better overall performance.

5.5 On New Biomedical Data

Under this active learning framework, we have shown a CRF model can be trained with less annotation cost than using traditional passive learning. We adapted the trained CRF model to new biomedical research articles.

Out of the 9474 collected JCN articles, more than 230 research articles are on TTEs. The whole processing time for each document varies from 20 seconds to 90 seconds. We sent the new system-labeled files back to a biomedical knowledge expert for manual annotation. The time to correct one automatically labeled document is dramatically reduced, around 1/3 of that spent on raw text.

We processed 214 new research articles and examined a subset including 16 articles. We evalu-

ated it in two aspects: the overall performance and the performance averaged at the document level.

Table 6 gives the performance on the whole new subset and that averaged on 16 documents. The performance is a little bit lower than reported in the previous section as the new document set might include different styles of documents. We examined system performance at each document. Figure 7 gives the detailed evaluation for each of the 16 documents. The average F_Score of the document level is around 74%. For those documents with reasonable TTE description, the system can achieve an F_Score of 87%. The bad documents had a different description style and usually mixed the TTE descriptions with general discussion.

	Prec.	Recall	F_Score
Overall	0.7683	0.7155	0.7410
Averaged per Doc.	0.7686	0.7209	0.7418

Table 6. Performance on the whole new subset and the averaged performance per document.

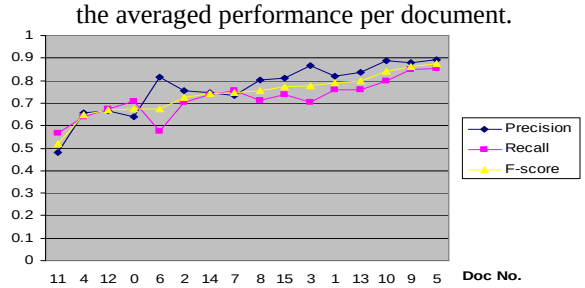

Figure 7. System performance per document.

6 Conclusions and Future Work

In this paper, we explored adapting a supervised CRF model for semantic analysis on biomedical articles using an active learning framework. It abridges the power of the supervised approach and expensive human costs. We are also investigating the use of other certainty measures, such as averaged field confidence scores over each sentence.

In the long run we wish to generalize the framework to be able to mine other types of experiments within the biomedical research literature and impact research in those domains.

References

Blunsom, P. and Cohn, T. 2006. Discriminative word alignment with conditional random fields. In *ACL-2006*.

Burns, G.A. and Cheng, W.C. 2006. Tools for knowledge acquisition within the NeuroScholar system and their application to anatomical tract-tracing data. In *Journal of Biomedical Discovery and Collaboration*.

Chen, J., Schein, A., Ungar, L., and Palmer, M. 2006. An empirical study of the behavior of active learning for word sense disambiguation. In *Proc. of HLT-NAACL 2006*.

Culotta, A. and McCallum, A. 2004. Confidence estimation for information extraction. In *HLT-NAACL-2004, short papers*.

Feng, D., Burns, G., and Hovy, E.H. 2007. Extracting data records from unstructured biomedical full text. In *Proc. of EMNLP-CONLL-2007*.

Feng, D., Ravichandran, D., and Hovy, E.H. 2006. Mining and re-ranking for answering biographical queries on the web. In *Proc. of AAAI-2006*.

Jiao, F., Wang, S., Lee, C., Greiner, R., and Schuurmans, D. 2006. Semi-supervised conditional random fields for improved sequence segmentation and labeling. In *Proc. of ACL-2006*.

Kristjannson, T., Culotta, A., Viola, P., and McCallum, A. 2004. Interactive information extraction with constrained conditional random fields. In *Proc. of AAAI-2004*.

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML-2001*.

Lewis, D.D. and Gale, W.A. 1994. A sequential algorithm for training text classifiers. In *Proc. of SIGIR-1994*.

Lin, D. 1998. Dependency-based evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems*.

Mann, G.S. and Yarowsky, D. 2005. Multi-field information extraction and cross-document fusion. In *Proc. of ACL-2005*.

McCallum, A.K. 2002. *MALLET: a machine Learning for language toolkit*. http://mallet.cs.umass.edu.

McCallum, A. and Nigam, K. 1998. Employing EM in pool-based active learning for text classification. In *Proc. of ICML-98*.

OKanohara, D., Miyao, Y., Tsuruoka, Y., and Tsujii, J. 2006. Improving the scalability of semi-markov conditional random fields for named entity recognition**.** In *ACL-2006*.

Peng, F. and McCallum, A. 2004. Accurate information extraction from research papers using conditional random fields. In *Proc. of HLT-NAACL-2004*.

Pinto, D., McCallum, A., Wei, X., and Croft, W.B. 2003. Table extraction using conditional random fields. In *SIGIR-2003*.

Ravichandran, D. and Hovy, E.H. 2002. Learning surface text patterns for a question answering system. In *ACL-2002*.

Shen, D., Zhang, J., Su, J., Zhou, G., and Tan, C.L. 2004. Multi-criteria-based active learning for named entity recognition. In *Proc. of ACL-2004*.

Srinivas, et al., 2005. Comparison of vector space model methodologies to reconcile cross-species neuroanatomical concepts. *Neuroinformatics*, 3(2).

Stephan, K.E., et al., 2001. Advanced database methodology for the Collation of Connectivity data on the Macaque brain (CoCoMac). *Philos Trans R Soc Lond B Biol Sci*.

Symons et al., 2006. Multi-Criterion Active Learning in Conditional Random Fields. In *ICTAI-2006*.

Tang, M., Luo, X., and Roukos, S. 2002. Active learning for statistical natural language parsing. In ACL-2002.

Thompson, C.A., Califf, M.E., and Mooney, R.J. 1999. Active learning for natural language parsing and information extraction. In *Proc. of ICML-99*.

Tong, S. and Koller, D. 2000. Support vector machine active learning with applications to text classification. In *Proc. of ICML-2000*.