MMR-based Active Machine Learning for Bio Named Entity Recognition

Seokhwan Kim[1] Yu Song[2] Kyungduk Kim[1] Jeong-Won Cha[3] Gary Geunbae Lee[1]
[1] Dept. of Computer Science and Engineering, POSTECH, Pohang, Korea
[2] AIA Information Technology Co., Ltd. Beijing, China
[3] Dept. of Computer Science, Changwon National University, Changwon, Korea
megaup@postech.ac.kr, Song-Y.Song@AIG.com, getta@postech.ac.kr
jcha@changwon.ac.kr, gblee@postech.ac.kr

Abstract

This paper presents a new active learning paradigm which considers not only the uncertainty of the classifier but also the diversity of the corpus. The two measures for uncertainty and diversity were combined using the MMR (Maximal Marginal Relevance) method to give the sampling scores in our active learning strategy. We incorporated MMR-based active machine-learning idea into the biomedical named-entity recognition system. Our experimental results indicated that our strategies for active-learning based sample selection could significantly reduce the human effort.

1 Introduction

Named-entity recognition is one of the most elementary and core problems in biomedical text mining. To achieve good recognition performance, we use a supervised machine-learning based approach which is a standard in the named-entity recognition task. The obstacle of supervised machine-learning methods is the lack of the annotated training data which is essential for achieving good performance. Building a training corpus manually is time consuming, labor intensive, and expensive. Creating training corpora for the biomedical domain is particularly expensive as it requires domain specific expert knowledge.

One way to solve this problem is through active learning method to select the most informative samples for training. Active selection of the training examples can significantly reduce the necessary number of labeled training examples without degrading the performance.

Existing work for active learning explores two approaches: certainty or uncertainty-based methods (Lewis and Gale 1994; Scheffer and Wrobel 2001; Thompson *et al.* 1999) and committee-based methods (Cohn *et al.* 1994; Dagan and Engelson 1995; Freund *et al.* 1997; Liere and Tadepalli 1997). Uncertainty-based systems begin with an initial classifier and the systems assign some uncertainty scores to the un-annotated examples. The *k* examples with the highest scores will be annotated by human experts and the classifier will be retrained. In the committee-based systems, diverse committees of classifiers were generated. Each committee member will examine the un-annotated examples. The degree of disagreement among the committee members will be evaluated and the examples with the highest disagreement will be selected for manual annotation.

Our efforts are different from the previous active learning approaches and are devoted to two aspects: we propose an entropy-based measure to quantify the uncertainty that the current classifier holds. The most uncertain samples are selected for human annotation. However, we also assume that the selected training samples should give the different aspects of learning features to the classification system. So, we try to catch the most representative sentences in each sampling. The divergence measures of the two sentences are for the novelty of the features and their representative levels, and are described by the minimum similarity among the examples. The two measures for uncertainty and diversity will be combined using the MMR (Maximal Marginal Relevance) method (Carbonell and Goldstein 1998) to give the sampling scores in our active learning strategy.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 69–72,
New York, June 2006.

We incorporate MMR-based active machine-learning idea into the POSBIOTM/NER (Song *et al.* 2005) system which is a trainable biomedical named-entity recognition system using the Conditional Random Fields (Lafferty *et al.* 2001) machine learning technique to automatically identify different sets of biological entities in the text.

2 MMR-based Active Learning for Biomedical Named-entity Recognition

2.1 Active Learning

We integrate active learning methods into the POSBIOTM/NER (Song *et al.* 2005) system by the following procedure: Given an active learning scoring strategy *S* and a threshold value *th*, at each iteration *t*, the learner uses training corpus T_{M_t} to train the NER module M_t. Each time a user wants to annotate a set of un-labeled sentences *U*, the system first tags the sentences using the current NER module M_t. At the same time, each tagged sentence is assigned with a score according to our scoring strategy *S*. Sentences will be marked if its score is larger than the threshold value *th*. The tag result is presented to the user, and those marked ones are rectified by the user and added to the training corpus. Once the training data accumulates to a certain amount, the NER module M_t will be retrained.

2.2 Uncertainty-based Sample Selection

We evaluate the uncertainty degree that the current NER module holds for a given sentence in terms of the entropy of the sentence. Given an input sequence **o**, the state sequence set *S* is a finite set. And $p_\wedge(\mathbf{s}\,|\,\mathbf{o})$, $\mathbf{s} \in \mathbf{S}$ is the probability distribution over *S*. By using the equation for CRF (Lafferty *et al.* 2001) module, we can calculate the probability of any possible state sequence s given an input sequence **o**. Then the entropy of $p_\wedge(\mathbf{s}\,|\,\mathbf{o})$ is defined to be:

$$H = -\sum_{\mathbf{s}} P_\wedge(\mathbf{s}\,|\,\mathbf{o}) \log_2[P_\wedge(\mathbf{s}\,|\,\mathbf{o})]$$

The number of possible state sequences grows exponentially as the sentence length increases. In order to measure the uncertainty by entropy, it is inconvenient and unnecessary to compute the probability of all the possible state sequences. Instead we implement N-best Viterbi search to find the *N* state sequences with the highest probabilities. The entropy *H*(*N*) is defined as the entropy of the distribution of the N-best state sequences:

$$H(N) = -\sum_{i=1}^{N} \frac{P_\wedge(\mathbf{s}_i\,|\,\mathbf{o})}{\sum_{i=1}^{N} P_\wedge(\mathbf{s}_i\,|\,\mathbf{o})} \log_2\left[\frac{P_\wedge(\mathbf{s}_i\,|\,\mathbf{o})}{\sum_{i=1}^{N} P_\wedge(\mathbf{s}_i\,|\,\mathbf{o})}\right]. \quad (1)$$

The range of the entropy *H*(*N*) is [0, $-\log_2 \frac{1}{N}$] which varies according to different *N*. We could use the equation (2) to normalize the *H*(*N*) to [0, 1].

$$H(N)' = \frac{H(N)}{-\log_2 \frac{1}{N}}. \quad (2)$$

2.3 Diversity-based Sample Selection

We measure the sentence structure similarity to represent the diversity and catch the most representative ones in order to give more diverse features to the machine learning-based classification systems.

We propose a three-level hierarchy to represent the structure of a sentence. The first level is NP chunk, the second level is Part-Of-Speech tag, and the third level is the word itself. Each word is represented using this hierarchy structure. For example in the sentence "I am a boy", the word "boy" is represented as $\vec{w}$=[NP, NN, boy]. The similarity score of two words is defined as:

$$sim(\vec{w}_1 \cdot \vec{w}_2) = \frac{2 * Depth(\vec{w}_1, \vec{w}_2)}{Depth(\vec{w}_1) + Depth(\vec{w}_2)}$$

Where $Depth(\vec{w}_1, \vec{w}_2)$ is defined from the top level as the number of levels that the two words are in common. Under our three-level hierarchy scheme above, each word representation has depth of 3.

The structure of a sentence S is represented as the word representation vectors $[\vec{w}_1, \vec{w}_2, \ldots, \vec{w}_N]$. We measure the similarity of two sentences by the standard cosine-similarity measure. The similarity score of two sentences is defined as:

$$similarity(\vec{S}_1, \vec{S}_2) = \frac{\vec{S}_1 \cdot \vec{S}_2}{\sqrt{\vec{S}_1 \cdot \vec{S}_1}\sqrt{\vec{S}_2 \cdot \vec{S}_2}},$$

$$\vec{S}_1 \cdot \vec{S}_2 = \sum_i \sum_j sim(\vec{w}_{1i} \cdot \vec{w}_{2j}).$$

2.4 MMR Combination for Sample Selection

We would like to score the sample sentences with respect to both the uncertainty and the diversity. The following MMR (Maximal Marginal Relevance) (Carbonell and Goldstein 1998) formula is used to calculate the active learning score:

$$score(s_i) \overset{def}{=} \lambda * \text{Uncertainty}(s_i, M) - (1-\lambda) * \max_{s_j \in T_M} \text{Similarity}(s_i, s_j) \quad (3)$$

where s_i is the sentence to be selected, Uncertainty is the entropy of s_i given current NER module M, and Similarity indicates the divergence degree between the s_i and the sentence s_j in the training corpus T_M of M. The combination rule could be interpreted as assigning a higher score to a sentence of which the NER module is uncertain and whose configuration differs from the sentences in the existing training corpus. The value of parameter λ coordinates those two different aspects of the desirable sample sentences.

After initializing a NER module M and an appropriate value of the parameter λ, we can assign each candidate sentence a score under the control of the uncertainty and the diversity.

3 Experiment and Discussion

3.1 Experiment Setup

We conducted our active learning experiments using pool-based sample selection (Lewis and Gale 1994). The pool-based sample selection, in which the learner chooses the best instances for labeling from a given pool of unlabelled examples, is the most practical approach for problems in which unlabelled data is relatively easily available.

For our empirical evaluation of the active learning methods, we used the training and test data released by JNLPBA (Kim *et al.* 2004). The training corpus contains 2000 MEDLINE abstracts, and the test data contains 404 abstracts from the GENIA corpus. 100 abstracts were used to train our initial NER module. The remaining training data were taken as the pool. Each time, we chose *k* examples from the given pool to train the new NER module and the number *k* varied from 1000 to 17000 with a step size 1000.

We test 4 different active learning methods: Random selection, Entropy-based uncertainty selection, Entropy combined with Diversity, and Normalized Entropy (equation (2)) combined with Diversity. When we compute the active learning score using the entropy based method and the combining methods we set the values of parameter N (from equation (1)) to 3 and λ (from equation (3)) to 0.8 empirically.

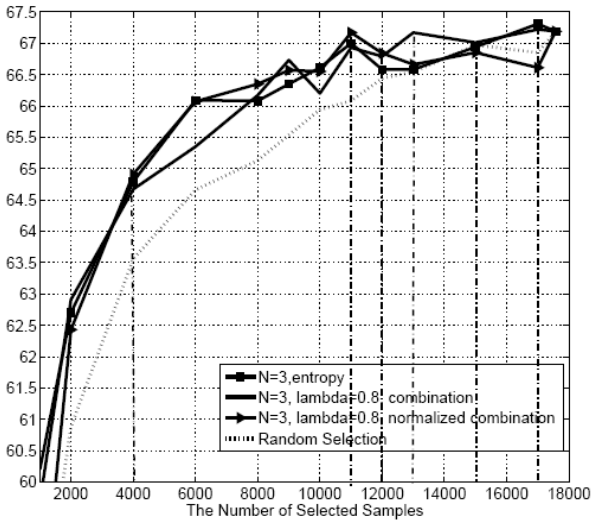

Fig1. Comparison of active learning strategies with the random selection

3.2 Results and Analyses

The initial NER module gets an F-score of 52.54, while the F-score performance of the NER module using the whole training data set is 67.19. We plotted the learning curves for the different sample selection strategies. The interval in the x-axis between the curves shows the number of examples selected and the interval in the y-axis shows the performance improved.

We compared the entropy, entropy combined with sentence diversity, normalized entropy combined with sentence diversity and random selection.

The curves in Figure 1 show the relative performance. The F-score increases along with the number of selected examples and receives the best performance when all the examples in the pool are selected. The results suggest that all three kinds of active learning strategies consistently outperform the random selection.

The entropy-based example selection has improved performance compared with the random selection. The entropy (N=3) curve approaches to the random selection around 13000 sentences selected, which is reasonable since all the methods choose the examples from the same given pool. As

the number of selected sentences approaches the pool size, the performance difference among the different methods gets small. The best performance of the entropy strategy is 67.31 when 17000 examples are selected.

Comparing with the entropy curve, the combined strategy curve shows an interesting characteristic. Up to 4000 sentences, the entropy strategy and the combined strategy perform similarly. After the 11000 sentence point, the combined strategy surpasses the entropy strategy. It accords with our belief that the diversity increases the classifier's performance when the large amount of samples is selected. The normalized combined strategy differs from the combined strategy. It exceeds the other strategies from the beginning and maintains the best performance up until 12000 sentence point.

The entropy strategy reaches 67.00 in F-score when 11000 sentences are selected. The combined strategy receives 67.17 in F-score while 13000 sentences are selected, while the end performance is 67.19 using the whole training data. The combined strategy reduces 24.64 % of training examples compared with the random selection. The normalized combined strategy achieves 67.17 in F-score when 11000 sentences are selected, so 35.43% of the training examples do not need to be labeled to achieve almost the same performance as the end performance. The normalized combined strategy's performance becomes similar to the random selection strategy at around 13000 sentences, and after 14000 sentences the normalized combined strategy behaves the worst.

4 Conclusion

We incorporate active learning into the biomedical named-entity recognition system to enhance the system's performance with only small amount of training data. We presented the entropy-based uncertainty sample selection and combined selection strategies using the corpus diversity. Experiments indicate that our strategies for active-learning based sample selection could significantly reduce the human effort.

Acknowledgement

This research was supported as a Brain Neuroinformatics Research Program sponsored by Ministry of Commerce, Industry and Energy.

References

Carbonell J., & Goldstein J. (1998). The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 335-336.

Cohn, D. A., Atlas, L., & Ladner, R. E. (1994). Improving generalization with active learning, *Machine Learning*, 15(2), 201-221.

Dagan, I., & Engelson S. (1995). Committee-based sampling for training probabilistic classifiers. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 150-157, San Francisco, CA, Morgan Kaufman.

Freund Y., Seung H.S., Shamir E., & Tishby N. (1997). Selective sampling using the query by committee algorithm, *Machine Learning*, 28, 133-168.

Kim JD., Ohta T., Tsuruoka Y., & Tateisi Y. (2004). Introduction to the Bio-Entity Recognition Task at JNLPBA, *Proceedings of the International Workshop on Natural Language Processing in Biomedicine and its Application (JNLPBA).*

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the 18th International Conf. on Machine Learning*, pages 282-289, Williamstown, MA, Morgan Kaufmann.

Lewis D., & Gale W. (1994). A Sequential Algorithm for Training Text Classifiers, In: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. pp. 3-12, Springer-Verlag.

Liere, R., & Tadepalli, P. (1997). Active learning with committees for text categorization, In *proceedings of the Fourteenth National Conference on Artificial Intelligence*, pp. 591-596 Providence, RI.

Scheffer T., & Wrobel S. (2001). Active learning of partially hidden markov models. In *Proceedings of the ECML/PKDD Workshop on Instance Selection*.

Song Y., Kim E., Lee G.G., & Yi B-k. (2005). POSBIOTM-NER: a trainable biomedical named-entity recognition system. *Bioinformatics*, 21 (11): 2794-2796.

Thompson C.A., Califf M.E., & Mooney R.J. (1999). Active Learning for Natural Language Parsing and Information Extraction, In *Proceedings of the Sixteenth International Machine Learning Conference*, pp.406-414, Bled, Slovenia.