Active learning for ontological event extraction

Xu Han
School of Computer Engineering
Nanyang Technological University
Singapore 639798
HANX0017@e.ntu.edu.sg

Jung-jae Kim
School of Computer Engineering
Nanyang Technological University
Singapore 639798
jungjae.kim@ntu.edu.sg

Abstract

Text mining community in the biomedical domain are targeting more event types, but the cost of manual annotation for each new event type, which is required for supervised learning systems, hinders the progress. To reduce the amount of required annotations, we propose a novel active learning method for ontological event extraction. Our method can significantly reduce the amount of annotated corpora to saturate event extraction performance, compared to random selection of corpora for annotation, which is the common practice, and previous active learning methods for corpus selection. We tested the methods using the TEES event extraction system against the BioNLP Shared Tasks datasets, showing that our method can help the system achieve its previously reported performance only with 60%-70% of the original training data.

1 Introduction

The most common framework of information extraction systems is supervised learning, which requires training data that are annotated with information to be extracted. In the biomedical information extraction, such training data are usually manually annotated, where the annotation process is time-consuming and expensive. On the other hand, recent research efforts are extending from protein-protein interactions (PPI) (Hirschman et al., 2005) to more complicated biological events, which are defined in ontologies (Kim et al., 2011a). However, the manual annotation cost hinders the progress. There is thus the need of reducing the amount of training data that is required for event extraction systems to reach performance saturation point. To address this need, we propose a novel active learning method that selects 'informative' data to maximise system performance.

Active learning (Settles, 2012) is to choose 'most informative' documents for manual annotation. It has been studied in many research areas in natural language processing, such as word sense disambiguation (Chen et al., 2013), named entity recognition (Tomanek and Hahn, 2009a; Tomanek and Hahn, 2009b; Tomanek and Hahn, 2010), speech summarization (Zhang and Yuan, 2014) and sentiment classification . In the biomedical information extraction, it has been applied to the task of extracting PPIs (Cui et al., 2009; Zhang et al., 2012), where the 'informativity' of a document is measured based on whether the document contains any expression of PPI or not.

In our work, the goal is to find documents that 'most informatively' express event concepts of a given ontology, which involves the following issues:

1. An ontology may have multiple event concepts, and a sentence may express multiple concepts in different parts. We thus consider the 'informativity' of a document as collective likelihood of containing individual concepts, ignoring the locations of the concept expressions in the document.

2. The manual annotation process can be progressive, and if an event extraction system trained on earlier annotations can extract an event concept from a new document, the document is no longer 'informative' about that concept. The

'informativity' is thus limited to those events unrecognizable by the event extraction system of our interest, in order to avoid overfitting.

3. An ontology has a hierarchical structure. We also propose a method that considers the hierarchical structure.

This paper is organized as follows: Section 2 describes the related work. In Section 3 we discuss the method and algorithm, followed by the experiment results and discussions in Section 4. Finally, Section 5 concludes this paper.

2 Related Work

2.1 Active Learning

In biomedical information extraction, some researchers have applied active learning to the extraction of PPIs. For instance, (Cui et al., 2009) proposed an uncertainty sampling-based approach of active learning, and (Zhang et al., 2012) proposed maximum uncertainty based and density based sample selection strategies. However, the extraction of PPI is a simple task, while recent biomedical event and relation extraction tasks (Kim et al., 2009) are much more complicated. In this paper, we propose an active learning method for extracting complex events.

As for the sample selection in active learning, its existing works can be roughly classified into two approaches: committee-based approach (Seung et al., 1992) and certainty-based approach (Lewis and Catlett, 1994). In committee-based approach, a committee of classifiers are maintained and documents whose classifications have the greatest disagreements among the classifiers are selected out and passed to external human experts for annotation. The certainty-based approach is to label the most uncertain samples by using uncertainty schemes such as entropy (Fu et al., 2013).

Due to the lack of many classifiers for the event extraction tasks of our interests, we follow the certainty-based approach. Our approach is based on an event extraction system and a language model for predicting ontology concepts, and considers a document as uncertain about an event concept if the system and the model disagree on the presence of the concept in the document. Particularly, when the language model predicts that a document expresses an event concept, but the system cannot extract any instance of the concept from the document possibly due to the lack of training data about the concept, we pass this document to human experts to check if the document expresses the concept or not.

2.2 BioNLP-ST datasets and event extraction system

The BioNLP shared tasks (BioNLP-ST) were organized to track the progress of information extraction in the biomedical text mining. In this paper, we used the datasets of two tasks, namely GRO'13 (Gene Regulation Ontology) and CG'13 (Cancer Genetics). Each corpus was manually annotated with an underlying ontology, whose number of concepts and hierarchy are different from the others. Those differences in the underlying ontologies bring about difference in the results of our experiments as shown in Section 4. A comparison between the two is given in Table 1.

Task	Number of ontology concepts	Ontology depth	Corpus size (abstracts)
GRO'13	507	7	300
CG'13	18	2	400

Table 1: Summary of task datasets

In this study, we propose a novel active learning method and test it against the two datasets above, using the state of the art biomedical event extraction system, namely the Turku Event Extraction System (TEES) (Björne et al., 2011). The TEES is based on SVM and was the only system that participated in all the tasks of BioNLP-ST'13, showing the best performance in many tasks (Björne et al., 2012).

3 Method

The tasks of our interests are to annotate all relevant concepts and relations from given ontologies on specific spans of text (Kim et al., 2011b). To simplify the problem, our proposed method of active learning does not locate the exact text spans that express ontology concepts/relations, but checks if a document

as a whole expresses any concept/relation.

As explained above, our approach is based on a language model, which measures the probability of the presence of an ontology concept in a document, and an event extraction system (i.e. TEES), which automatically extracts events of ontology concepts from a document. We assume that if the TEES can annotate a concept on a document, the document is not informative for the system to be further trained about the concept. Our method thus ignores the concepts that the TEES can annotate on a document, even if the document is likely to express the concepts according to the language model. Figure 1 depicts the workflow of the proposed method.

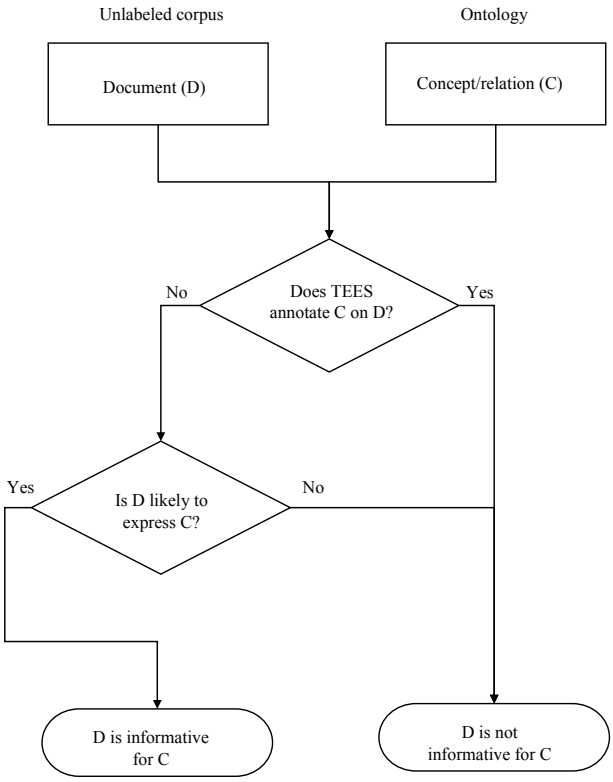

Figure 1: Overview of proposed active learning method

Our method works iteratively as follows: We train the TEES and build a language model based on an initial training dataset. We measure the informativity of each unlabelled document for an ontology and choose the top documents as feed for manual annotation and for retraining the TEES and rebuilding the language model. We then update the informativity of unlabelled documents using the retrained systems and continue to increase the training data size until the system performance is saturated.

During each iteration of active learning, we measure the informativity score of a document at the sentence level, that is, the sum of the informativity scores of all the sentences in the document. For each sentence (S_k), we measure its informativity score $I(S_k)$ in two dimensions: the ontology event concepts (E_i) and relations (R_j), as expressed in (1).

$$I(S_k) = \sum^{E_i \in S_k} P(E_i|S_k) + \sum^{R_j \in S_k} P(R_j|S_k) \quad (1)$$

The conditional probability in (1) is estimated using the Bayes' theorem as shown in formula (2) and (3).

$$P(E_i|S_k) = \frac{P(E_i)P(S_k|E_i)}{P(S_k)} \quad (2)$$

$$P(R_j|S_k) = \frac{P(R_j)P(S_k|R_j)}{P(S_k)} \quad (3)$$

We decompose a sentence in two ways: n-grams (NG) and predicate-argument relations (PAS) produced by the Enju parser (Sagae et al., 2007). Equations (4, 5) show the conditional probabilities of event concept based on the two decomposition methods, respectively. The conditional probabilities of relation are calculated likewise.

$$P(S_k|E_i) = \sum_{l;NG_l \in S_k} W(NG_l, E_i) \quad (4)$$

$$P(S_k|E_i) = \sum_{l;PAS_l \in S_k} W(PAS_l, E_i) \quad (5)$$

The weight score $W(NG_l, E_i)$ (similarly for PAS_l and R_j) is dependent on co-occurrences between the n-gram (NG_I) and the ontology event concept (E_i). The score is thus calculated in three ways: 1) Yates' chi-square test, 2) relative risk, and 3) odds ratio (Corder and Foreman, 2009).

In addition, we incorporate the hierarchy structure of ontology into the statistical active learning framework as follows: Given an event concept E_i and a sentence S_k, we replace $P(E_i|S_k)$ with the sum of $P(E_m|S_k)$ for all the ancestor concepts of E_i, as shown in Equation (6).

$$P(E_i|S_k) = \sum_{m;E_i \subseteq E_m} P(E_m|S_k) \qquad (6)$$

Table 2 summarises the calculation of informativity scores in pseudo codes.

4 Results and discussion

The experiments are carried out in two phases: 1) Parameter optimization and 2) evaluation against the BioNLP-ST datasets.

4.1 Experiment 1: Parameter optimization

We first take a separate parameter optimization step, in order to determine the most appropriate measure for the calculation of the aforementioned weight score and the most effective n-gram size. A simulation of ontology concept prediction is carried out at the sentence level. In this task, for a sentence S_i that contains N manually annotated ontology concepts, only the original string of S_i and the number N are given to the informativity calculator. The calculator predicts top N candidate ontology concepts for this given sentence.

Using 10-fold cross validation, the average prediction rate is calculated and reported in Table 3. Each column corresponds to a n-gram size, and each row to one of the three co-occurrence analysis methods used for the prediction. Note that when N=2 (i.e. bi-grams), it does not include unigrams for the calculation. This experiment is carried out using the GRO'13 dataset.

Calculation Method	N-gram				
	N=1	N=2	N=3	N=4	N=5
chi-square	**0.507**	0.413	0.159	0.036	0.009
relative ratio	0.341	0.395	0.307	0.128	0.038
odds	0.420	0.395	0.274	0.117	0.035

Table 3: Paremeter optimization results

As shown in Table 3, for all weight score calculation methods, the average accuracy mostly drops as the length of N-grams increases. This may happen due to the data sparseness problem for large N-grams. We choose to use chi-square test and unigrams for the following experiments based on the results.

4.2 Experiment 2: Evaluation of application of active learning

In this section, we compare the proposed active learning method with other sample selection strategies, including random selection (i.e. baseline) and entropy-based active learning (Zhang et al., 2012). Each experiment has ten rounds, where in each round, 10% of the original training data are added for training the TEES system. The followings are considered for the selection of additional 10% training data in each round:

- Random selection: We randomly split the labeled documents into 10 bins in advance, and in each round during the training phase, one bin is randomly chosen. By using 10-fold cross validation, we report the averaged performance of random selection (hereafter referred as RS_Average).
- Entropy-based active learning: We calculate the maximum entropy of each document regardless of the given ontology, sort documents by their entropy values and feed from documents with top values to those with bottom values as training data. Note that the sorting will be done only once. (designated as Entropy)
- Proposed active learning: We vary the method in two orthogonal dimensions: 1) Using either unigrams (Unigram) or predicate-argument relations (PAS), and 2) the weight scores for events only (Event), relations only (Relation), and both events and relations (Event + Relation). In each round, we re-calculate the informativity of all remaining documents in the training data and feed those with top scores as training data.

Note that all the datasets were already annotated by the shared task organisers, and so there is no need to annotate them again. In the cases of real application to unlabelled data, after selecting a subset of the data as training data in a round, the documents in the subset should be annotated by domain experts. In the testing phase, the F-score will be measured as the performance of the methods.

Input: labeled document pool L, unlabeled document pool U, batch size b

// Initialization

ER_0 = the set of events/relations annotated on L

Learn a TEES model M_0 from ER_0

// Active Learning Loop

while U is **not** empty:

  for each document D_{ij} in U:

    Document informativity score $I(D_{ij}) = 0$

    for each sentence S_k in D_{ij}:

      Apply M_{i-1} to S_k and collect the resultant events/relations set ER_{S_k}

      for each event/relation er s.t. $er \in ER_{i-1}$ and $er \notin ER_{sk}$:

        $I(D_{ij})$ += informativity score $I(S_k, er)$

    $I(D_{ij})$ = $I(D_{ij})$ / sizeOf(D_{ij})

  Rank D_{ij} in U based on $I(D_{ij})$ and select the top b documents, designated as B

  Remove B from U, add B to L, and add the annotations on B to ER_{i-1}, designated as ER_i

  Learn a new model M_i from ER_i

Table 2: Proposed algorithm of active learning with TEES

We first applied those methods to the dataset of GRO'13 (Kim et al., 2013) and measured the performance change with incremental feed of the training data by using the TEES system. The evaluation results are plotted in Figures 2 and 3. As shown in the figures, the PAS models generally perform better than the unigram models. In Figure 2, the method using AL(Event_PAS + Relation_PAS) reaches the saturated performance only with 60% of the original training data and does not drop as more training data are added for the system training. Saturated performance indicates the performance of a system when it is trained with 100% of the training data. Note that the proposed active learning models perform better than the random selection and the entropy-based active learning in most cases.

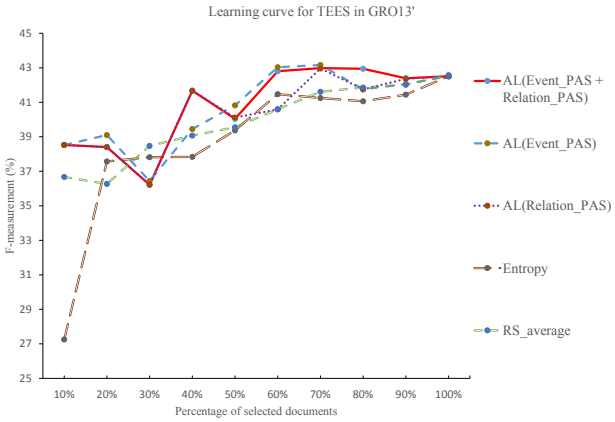

Figure 2: Comparison of active learning with PAS, entropy-based method and random selection in GRO'13

At the point of using 30% of training data in Figure 2, the performance of the active learning methods all fall significantly, though quickly restored afterwards. We examined the results at the point and found that the performance of relation identification was relatively low. It may mean that the relation identification is unstable with small amount of training data.

We then carried out a similar experiment using the CG'13 dataset. Figures 4 and 5 plot the results of the experiments. In these experiments, we can again confirm that the PAS models outperform the unigram models and that the proposed active learning methods outperform the random selection and the entropy-based active learning. In Figure 4, the method using AL(Event_PAS + Relation_PAS) reaches the saturated performance only with 70% of the original training data.

However, the performance improvement is less significant compared to that of the GRO'13. Note that the ontology of CG'13 is much smaller than that of GRO'13, and thus that the CG'13 task would require less training data for performance saturation and so there is less room for contribution from active learning. This claim can be supported by com-

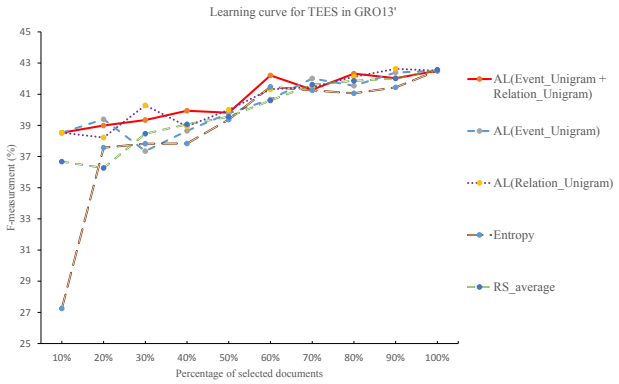

Figure 3: Comparison of active learning with n-grams, entropy-based method and random selection in GRO'13

paring the performance changes for the random selection models against the two datasets. As shown in Figures 3 and 5, the performance of the random selection model for the CG'13 is getting saturated, while that for the GRO'13 is not.

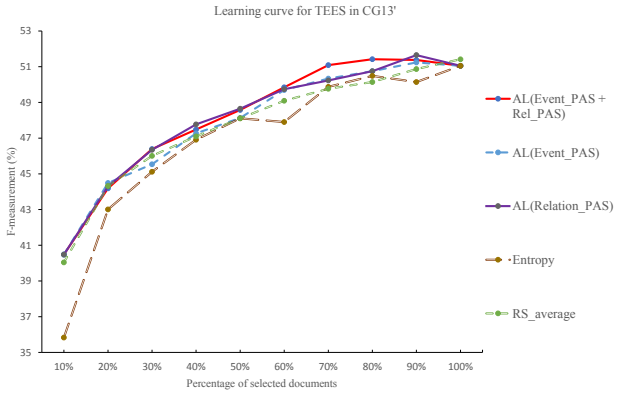

Figure 4: Comparison of active learning with PAS, entropy-based method and random selection in CG'13

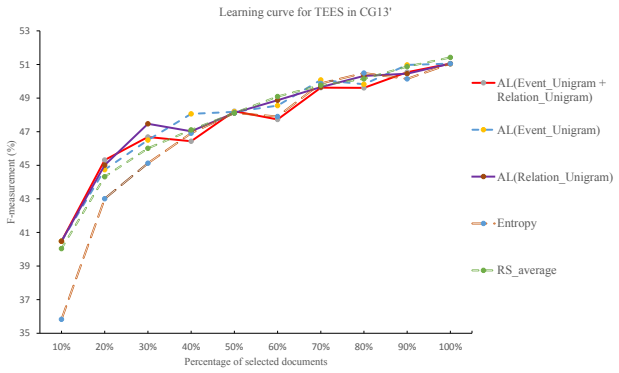

Figure 5: Comparison of active learning with n-grams, entropy-based method and random selection in CG'13

4.3 Incorporation of ontology hierarchy

Apart from the previous comparisons, we incorporate ontology hierarchies into the active learning method. We carried out experiments using the GRO'13 dataset, as its ontology depth is bigger than that of CG'13. Note that in this experiment, only the extraction of events but not relations is evaluated, since only event concepts have a hierarchy structure in the GRO. We use PAS models for this experiment as they show better performance than unigram models in the previous comparisons.

The results of the ontology hierarchy incorporation is plotted in Figure 6. The incorporation helps system performance reach the reported performance of the TEES system from the point of using 60% of training data and not fall below the reported one, while the active learning method of AL(Event_PAS) without the ontology hierarchy information shows unstable performance. Additionally, we have also tried to use descendants, but the performance is worse than using ancestors.

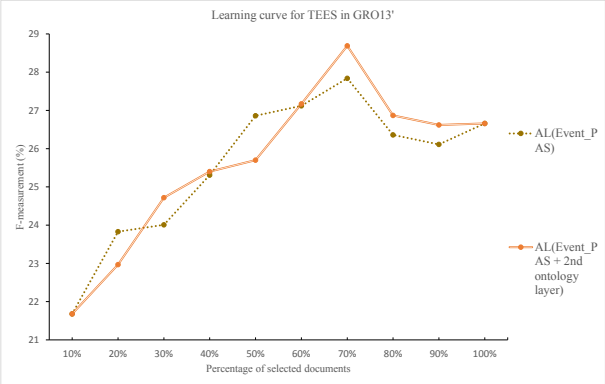

Figure 6: Incorporation of ontology hierarchy into active learning

5 Conclusion

In this study, we propose a novel active learning method for ontological event extraction, which is more complicated than the simple PPI extraction. Our method measures the collective 'informativity' for unrecognizable biological events expressed in documents. We found that the proposed method using the predicate-argument structure and the ontology hierarchy is able to make the underlying event extraction system reach saturated performance with

significantly less documents and hence reduce the amount of needed documents for manual annotation.

References

Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. 2011. Extracting Complex Biological Events with Rich Graph-Based Feature Sets. *Computational Intelligence*, 27(4):541–557.

Jari Björne, Filip Ginter, and Tapio Salakoski. 2012. University of Turku in the BioNLP'11 Shared Task. *BMC Bioinformatics*, 13(Suppl 11):S4.

Yukun Chen, Hongxin Cao, Qiaozhu Mei, Kai Zheng, and Hua Xu. 2013. Applying active learning to supervised word sense disambiguation in medline. *Journal of the American Medical Informatics Association*, 20(5):1001–1006.

Gregory W Corder and Dale I Foreman. 2009. *Nonparametric statistics for non-statisticians: a step-by-step approach*. John Wiley & Sons.

Baojin Cui, Hongfei Lin, and Zhihao Yang. 2009. Uncertainty sampling-based active learning for protein-protein interaction extraction from biomedical literature. *Expert Systems with Applications*, 36(7):10344–10350, September.

Yifan Fu, Xingquan Zhu, and Bin Li. 2013. A survey on instance selection for active learning. *Knowledge and information systems*, 35(2):249–283.

Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. 2005. Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1):S1.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of bionlp'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, Boulder, Colorado, June. Association for Computational Linguistics.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011a. Overview of genia event task in bionlp shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 7–15, Portland, Oregon, USA, June. Association for Computational Linguistics.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011b. Overview of genia event task in bionlp shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 7–15. Association for Computational Linguistics.

Jung-Jae Kim, Xu Han, Vivian Lee, and Dietrich Rebholz-Schuhmann. 2013. GRO Task: Populating the Gene Regulation Ontology with events and relations. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 50–57, Sofia, Bulgaria, August. Association for Computational Linguistics.

David D Lewis and Jason Catlett. 1994. Heterogenous uncertainty sampling for supervised learning. In *ICML*, volume 94, pages 148–156.

Kenji Sagae, Yusuke Miyao, and Junichi Tsujii. 2007. HPSG parsing with shallow dependency constraints. In *In Proceedings of ACL 2007*, pages 624–631.

Burr Settles. 2012. Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, June.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. 1992. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM.

Katrin Tomanek and Udo Hahn. 2009a. Reducing Class Imbalance During Active Learning for Named Entity Annotation. In *Proceedings of the Fifth International Conference on Knowledge Capture*, K-CAP '09, pages 105–112, New York, NY, USA. ACM.

Katrin Tomanek and Udo Hahn. 2009b. Semi-supervised Active Learning for Sequence Labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1039–1047, Stroudsburg, PA, USA. Association for Computational Linguistics.

Katrin Tomanek and Udo Hahn. 2010. A Comparison of Models for Cost-Sensitive Active Learning. In *Coling 2010: Posters*, pages 1247–1255, Beijing, China, August. Coling 2010 Organizing Committee.

Jian Zhang and Huaqiang Yuan. 2014. A Certainty-Based Active Learning Framework of Meeting Speech Summarization. In W.Eric Wong and Tingshao Zhu, editors, *Computer Engineering and Networking SE - 28*, volume 277 of *Lecture Notes in Electrical Engineering*, pages 235–242. Springer International Publishing.

Hong-Tao Zhang, Min-Lie Huang, and Xiao-Yan Zhu. 2012. A unified active learning framework for biomedical relation extraction. *Journal of Computer Science and Technology*, 27(6):1302–1313.