Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 14 (2): 177-182, 2014
ISSN 1812-5654 / DOI: 10.3923/jas.2014.177.182

Active Learning Based Semi-automatic Annotation of Event Corpus

Jianfeng Fu, Nianzu Liu and Shuangcheng Wang
School of Mathematics and Information, Shanghai Lixin University of Commerce, Shanghai, China

Abstract: In the area of Natural Language Processing, building corpus by hand was a hard and time-consuming task. Active learning promised to reduce the cost of annotating dataset for it was allowed to choose the data from which it learned. This study presented a semi-automatic annotation method based on active learning for labeling events in Chinese text. Particularly, it focused on uncertainty-based sampling and query-by-committee based sampling algorithm to evaluate which instance was informative and could be labeled by hand in the unlabeled dataset. The selected informative instances were labeled manually for obtaining a more effective classifier. Experimental results not only demonstrated that active learning improved the accuracy of Chinese event annotation, but also showed that it reduced the number of labeling actions dramatically.

Key words: Semi-automatic annotation, event corpus, active learning, uncertainty-based sampling, query-by-committee sampling

INTRODUCTION

With the development of Information Extraction (IE), event extraction, as a subtask of IE, is become more and more important. It has attracted much attention in recent years (Ahn, 2006; Ji and Grishman, 2008; Ritter *et al*., 2012; Arendarenko and Kakkonen, 2012). Most of the relevant developmental work has focused supervised machine learning method in which the event corpus is required. However preparing the corpus (for example, the Automatic Content Extraction (ACE) (Doddington *et al*., 2004) and TimeBank (Pustejovsky *et al*., 2003) event corpus) is an expensive undertaking since manually annotating event corpus is a hard and time-consuming task. It is a pity that very little of the work has been done on the semi-automated annotation of event corpus to accelerate the corpus annotation progress and reduce the burden of the annotator, while the event extraction systems have been extensively studied.

Active learning Settles (2010), Balcan *et al*. (2010) is a subfield of machine learning. The learning algorithm is allowed to choose the data from which it learns and gives suggestion of which data are valuable and should be tagged. The instances are selected since they have made unreliable predictions of the learned model. Without supplying the learner with more labeled data, it can produce a classifier as good as possible. Nowadays, active learning has been successfully applied to the tasks of speech recognition, information extraction, classification, filtering and so on.

The objective of the present study is to investigate the use of active learning based semi-automatic annotation technique that could be used to optimize the progress of event annotation.

EVENT CORPUS

Although, event corpus is widely distributed, there are mainly two annotation models in the available event corpus. One is Time bank model, in which an event is a node in a network of temporal relations. Every event which denotes by a specific term, temporal expressions and temporal relations is annotated in the Time Bank model. It focuses on the temporal relation of events. The other is ACE model, in which an event is more complex. In addition to temporal relation, the event arguments (such as participants and place) and the event properties (such as polarity, tense and modality) are all involved in this model.

In the previous work, Fu *et al*. (2010) studied the Chinese event taggability. The study proposed a serial of tagging rules, developed an annotation tool, collected 200 news articles of Chinese and annotated the articles manually. Difference with ACE (which is limited to specific event types), the study has annotated all the events which are involved in the articles. An example of the annotated Chinese event article is shown in Fig. 1.

As is shown in Fig. 1, the corpus is tagged by XML (Extensible Markup Language). The detail specification of Chinese event annotation scheme in BNF (Backus-Naur Form) is given below:

Corresponding Author: Jianfeng Fu, School of Mathematics and Information, Shanghai Lixin University of Commerce, Shanghai, China

```
<?xml version="1.0" encoding="GB2312"?>

<Body><Title>昆明震感强烈 许多市民冒雨走上街头</Title>

<ReportTime type="absTime">2008年08月30日18:00</ReportTime>

<Content>

    <Paragraph          甘娜 刘子倩)<Sentence><Event eid="e1"><Time
type="relTime">八月三十日十六时三十分许</Time>，<Location 四川省攀枝花市仁和区、四川省凉山彝族自治州会
理县交界处</Location>发生六点一级<Denoter type="emergency">地震</Denoter></Event>，<Event
eid="e2"><Location>昆明</Location>有强烈<Denoter
type="emergency">震感</Denoter></Event>。</Sentence></Paragraph>

    <Paragraph><Sentence><Event eid="e3"><Participant type="agent">记者</Participant><Denoter
type="action">致电</Denoter><Participant type="recipient">云南省地震局相关负责人</Participant></Event>，<Event
eid="e4"><Participant>该负责人</Participant><Denoter type="statement">表示</Denoter></Event>，<Event
eid="e5">关于地震的<Object>相关情况</Object>及<Object>数据</Object>收集正在<Denoter
type="operation">统计</Denoter>中</Event>。</Sentence><Sentence><Event eid="e6">此次<Denoter
type="emergency">地震</Denoter></Event>，<Event eid="e7"><Location>昆明</Location><Denoter
type="emergency">震感</Denoter>强烈</Event>，<Event eid="e8"><Participant>许多市民</Participant>冒雨<Denoter
type="movement">走</Denoter>上<Location type="destination">街头</Location></Event>，<Event
eid="e9"><Object>通讯</Object>短暂<Denoter type="stateChange">中断</Denoter>，</Event><Event eid="e10"><Time
type="relTime">现</Time>已<Denoter type="stateChange">恢复</Denoter>正常。</Event></Sentence></Paragraph>

</Content>

</Body>
```

Fig. 1: An example of annotated Chinese event article

Event
attributes ::= eid Polarity
eid ::= EventID
EventID ::= e<integer>
Polarity ::= ["Positive"] | "Negative"
Child_Nodes ::= <Denoter> [Time] [Location] [Participant] [Object]
Denoter
attribute ::= Type
Type ::= "Emergency" | "Movement" | "Statement" | "Act" | "Operation" | "StateChange" | "Perception"
Time
attribute ::= Type
Type ::= "absTime" | "relTime" | "timeInterval"
Location
attribute ::= Type
Type ::= ["Origin" | "Destination"]
Participant
attribute ::= Type
Type ::= [" Agent" | " Recipient"]
Object
attribute ::= Type
Type ::= ["Agent" | "Recipient"]

EVENT ANNOTATION

As well as part-of-speech tagging, this study considers the Chinese event annotation as a sequence labeling task. Not only the event arguments (for example, “time”, “location” and “participant”), but also the event “denoter” (a word or phrase in the text expresses an event happening) and “event extend” are required to be tagged. In particular, the identification of event extend is crucial in the annotation task for it will determine whether or not entities in the text can be used as arguments in nearby events. For simplicity, the identification of the “attribute” tag is not involved in this study. Figure 2 illustrates how the Chinese event annotation can be treated as a sequence labeling task in this study.

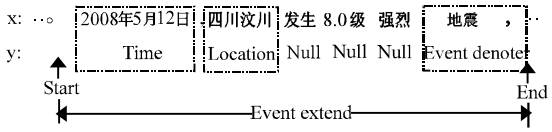

Fig. 2: A example of Chinese event annotation which is treated as sequence labeling task

SEMI-AUTOMATIC ANNOTATION WITH ACTIVE LEARNING

As is mentioned before, the Chinese event corpus was labeled by hand. The annotation work has consumed 12 man-months on manually labeling the 200 news articles. However, the corpus of 200 news articles is too insufficient to support a robust machine learner. For accelerate the annotation progress and enrich the corpus, this study presents a method of semi-automatic annotating of Chinese event in free text based on active learning. In particular, the method focuses active learning for sequence labeling on Conditional Random Fields (CRFs) (Lafferty *et al*., 2001). It allows the annotator label

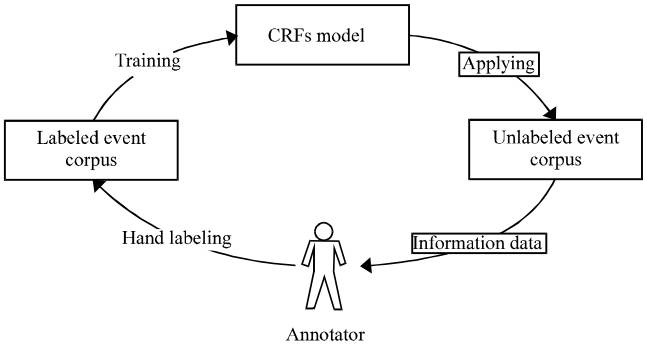

Fig. 3: Framework of semi-automatic annotation of event corpus

the instances rapidly and easily and enhances the system's prediction, while reducing the amount of effort spent on labeling events. The framework of semi-automatic annotation of event corpus is showed in Fig. 3.

The framework contains the following steps:

- Train a CRFs model with initial hand-labeled events corpus
- Apply the CRFs model to unlabeled data
- Evaluate potential informative instances ("time", "location", "participant", "denoter", "event extend" and so on) to be labeled
- Remove top n instances from unlabeled data and give to annotator
- Add the n instances which are labeled by the annotator into training dataset
- Retrain CRFs model with the training dataset
- Repeat step 2 to step 6 until stopping criterion is satisfied

In the method, the CRFs is employed to annotate event and its arguments. The CRFs are statistical graphical models which have demonstrated state-of-the-art accuracy on the sequence labeling tasks. This study uses linear-chain CRFs, which correspond to conditionally trained probabilistic finite state machine.

Difference from "ordinary" machine learning, the active learning which is employed in the method gives suggestion of which data are valuable and should be tagged for obtaining a model as well as possible, without supplying the learner with more data than necessary. This study employs two strategies of uncertainty-based and query-by-committee based sampling to evaluate unlabeled data.

Uncertainty-based sampling: An active learning must have a strategy of measuring informative instances. One of the most common general frameworks for measuring informative instances is uncertainty-based sampling. In this framework, an active learner queries the instances about which is least certain how to label. A simple uncertainty-based sampling for sequence models is Least Confident (LC):

$$Q_{LC}(X) = 1\text{-}P(Y^*|X;\theta) \quad (1)$$

Where Y* is the most likely label sequence. It can be efficiently computed by using dynamic programming (for example, Viterbi algorithm). The approach queries the instance for which the current model (LC) has the least confidence in its most likely labeling. For CRFs, the confidence can be calculated by using the posterior probability of the following equation (Lafferty *et al.*, 2001).

$$P(Y|X;\theta) = \frac{1}{Z_X}\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\theta_k f_k(y_{t-1}, y_t, X_t)\right) \quad (2)$$

where, $X = x_1...x_n$ is an input sequence, $Y = y_1...y_n$ is a label sequence. Z_X is the per-input normalization that makes the probability of all label sequences sum to one; f_x (y_{n-1}, y_n, X_t) is a feature function which is often binary-valued and θ_k is a learned weight associated with feature f_k.

Query-by-committee based sampling: Another general active learning framework is query-by-committee based sampling. In this framework, the most informative instance is the most disagreement about how to label in the committee. The committee models are trained on the labeled data, but represent competing hypotheses. Each committee member is allowed to vote on the labelings of query candidates. In this study, the committee is consisted of CRFs models. This study employs Sequence Vote Entropy (SVE), (Settles and Craven, 2008) which have demonstrated the best strategy

in query-by-committee based sampling, for measuring the level of disagreement. Let committee of models $\theta = \{\eta_1, \ldots \eta_i\}$ represent different hypotheses that are consistent with the labeled data, the SVE query instance as follows:

$$Q_{SVE}(X) = -\sum_{Y \in N^*} P(Y|X;\theta)\log P(Y|X;\theta) \quad (3)$$

where, N_θ is the union of the N-best parses from all models in the committee θ and $P(Y|X;\theta) = 1/i\sum_i P(Y|X;\eta^{(i)})$, or the "consensus" posterior probability for some label sequence.

EXPERIMENTS AND ANALYSIS

The chinese event corpus (which is collected from sina, yahoo, sohu and so on) is used for the experiments. It contains 200 hand-labeled articles (set L) and 600 unlabeled articles (set U). There are 3133 events and 4878 event arguments (total 8011 instances) involved in the L.

The articles are parsed by the LTP system (developed by IR-Lab in HIT, http://ir.hit.edu.cn/) to obtain the feature information of word POS and dependency relation. Besides, word position and context information within a window of size 3 are also contained in the features for training the CRFs learner.

Firstly, the set L is used to train an initial CRFs model, then apply the CRFs model to the U, employ the active learning algorithms (LC and SVE) to evaluate potential informative instances, remove top n (in this study, n = 10) instances from unlabeled data and label it manually, add the n instances to training dataset and retrain CRFs model M. The progress is repeated until get 500 informative instances for each algorithm.

In the set U, 500 instances are randomly selected, labeled by hand and moved into set L. The set L is then used to train a CRFs model. The model is predicted on test set T (randomly select 1000 instances from U as test set) as baseline. This study also predict the model M on T and evaluate the experimental results manually for ascertaining the usefulness of the active learning approaches (LC and SVE) explored in this study. In Fig. 4, this study first shows that active learning is beneficial in Chinese event annotation.

Annotation accuracy: Figure 4 presents the comparison results of LC, SVE and baseline on event annotation. It can be seen that, with the informative instances adding to the training set, the active learning methods (both LC and SVE) make a significant improvement (8.20% of LC and 8.14% of SVE) on system accuracy of event annotation.

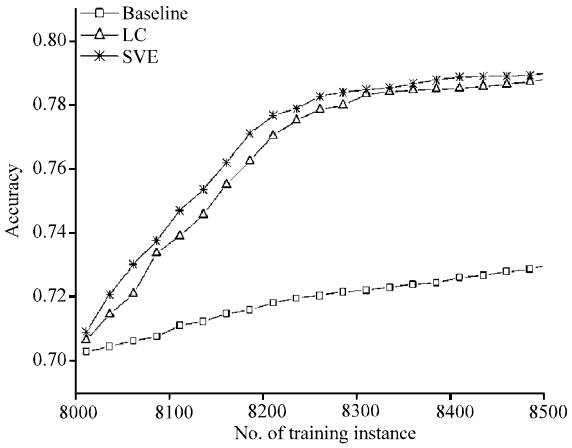

Fig. 4: Accuracy for event annotation of baseline, least confident (LC) and sequence vote entropy (SVE) (higher the better)

Although, the baseline get improvement (2.75%) with the randomly selected samples moving into training set, the active learning methods achieve better performance than baseline on the iteration progress. In addition, it also can be observed that the SVE performs better than LC on semi-automatic annotation event corpus.

Labeling action: Beside the traditional method of calculating annotation accuracy, this study also introduces the concept of atomic labeling action (Culotta and McCallum, 2005) for evaluating how the semi-automatic annotation can reduce the labeling effort. In the case of event annotation, there are three atomic labeling actions: Start, end and type, which are corresponding to labeling the start boundary, end boundary and type of an instance (event or event argument). Consider the following example.

<Event>
<Time>2008年 5月 12日</Time>,
<Location> 四川汶川</Location> 发生 8.0 级强烈
<Denoter> 地震</Denoter>
</Event>

It contains 4 start, 4 end and 4 type actions. Figure 5 shows that the presented methods can reduce the number of labeling action.

The corresponding labeling actions of baseline, LC and SVE are calculated when system accuracy reach 71, 72 and 73%, for quantifying the contribution of active learning on semi-automatic annotation of event corpus. From the Fig. 5, it can be seen that both LC and SVE need less labeling actions than baseline (for example, when system accuracy reach 73%, baseline requires 1464

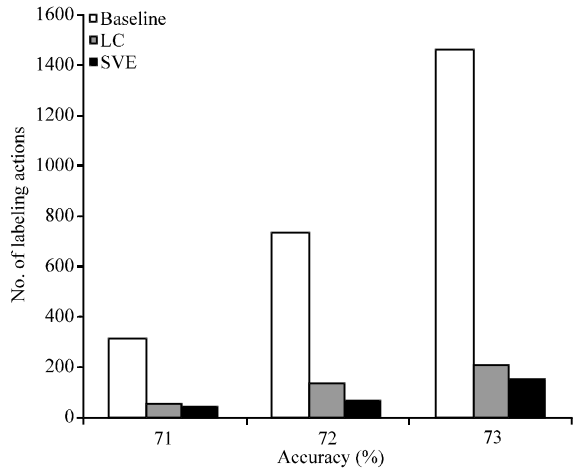

Fig. 5: Labeling actions for event annotation of baseline, least confident (LC) and sequence vote entropy (SVE) (less the better)

labeling action, LC requires 204 labeling actions and SVE requires 147 labeling actions only), since the LC and SVE select informative instances and produces better CRFs models than baseline. It demonstrates that the active learning algorithm is helpful for semi-automatic annotation of event corpus and reduces number of the labeling actions.

CONCLUSION AND FUTURE WORK

Corpus annotation manually is a hard and time-consuming task. It is meaningful to explore a semi-automatic method for event annotation in free text. This study has presented an active learning based semi-automatic annotation of event corpus. In particular, it has explored uncertainty-based sampling (LC) and query-by-committee based sampling (SVE) strategies for measuring the informative instances in unlabeled data and suggestion to annotator. These instances have been labeled and used to retrain more efficient CRFs models. Experimental results have demonstrated that it improves the accuracy of Chinese event annotation and reduces the number of labeling actions dramatically. Future work will explore more active learning algorithms on event annotation and integrate them into a union annotation platform.

ACKNOWLEDGMENTS

This study is supported by the projects of National Science Foundation of China (NSFC No. 60975033) and Shanghai Young College Teacher Training Project (SLX11010).

REFERENCES

Ahn, D., 2006. The stages of event extraction. Proceedings of the Workshop on Annotating and Reasoning about Time and Events, July 23, 2006, Sydney, Australia, pp: 1-8.

Arendarenko, E. and T. Kakkonen, 2012. Ontology-based information and event extraction for business intelligence. Artif. Intell.: Methodol. Syst. Appl., 7557: 89-102.

Balcan, M.F., S. Hanneke and J.W. Vaughan, 2010. The true sample complexity of active learning. Mach. Learn., 80: 111-139.

Culotta, A. and A. McCallum, 2005. Reducing labeling effort for structured prediction tasks. Proceedings of the 20th National Conference on Artificial Intelligence, July 9-13, 2005, Pittsburgh, PA., USA., pp: 746-751.

Doddington, G., A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel and R. Weischedel, 2004. The Automatic Content Extraction (ACE) program-tasks, data and Evaluation. Proceedings of the 4th International Conference on Language Resources and Evaluation, May 26-28, 2004, Centro Cultural de Belem, Lisbon, Portugal, pp: 837-840.

Fu, J.F., W. Liu, Z.T. Liu and S.S. Zhu, 2010. A study of Chinese event taggability. Proceedings of the 2nd International Conference on Communication Software and Networks, February 26-28, 2010, Singapore, pp: 400-404.

Ji, H. and R. Grishman, 2008. Refining event extraction through unsupervised cross-document inference. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, OH., USA., pp: 254-262.

Lafferty, J.D., A. McCallum and F.C.N. Pereira, 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning, June 28-July 1, 2001, Williamstown, MA, USA., pp: 282-289.

Pustejovsky, J., P. Hanks, R. Sauri, A. See and R. Gaizauskas *et al.*, 2003. The TIMEBANK corpus. Proceedings of the Corpus Linguistics Conference, March 28-31, 2003, Lancaster University, UK., pp: 647-656.

Ritter, A., Mausam, O. Etzioni and S. Clark, 2012. Open domain event extraction from twitter. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 12-16, 2012, Beijing, China, pp: 1104-1112.

Settles, B. and M. Craven, 2008. An analysis of active learning strategies for sequence labeling tasks. Proceedings of the Conference on Empirical Methods in Natural Language Processing, October 25-27, 2008, Waikiki, Honolulu, HI., USA., pp: 1070-1079.

Settles, B., 2010. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison, pp: 1-67.