Active Learning-Based Approach for Named Entity Recognition on Short Text Streams

Cuong Van Tran, Tuong Tri Nguyen, Dinh Tuyen Hoang, Dosam Hwang and Ngoc Thanh Nguyen

Abstract The named entity recognition (NER) problem has an important role in many natural language processing (NLP) applications and is one of the fundamental tasks for building NLP systems. Supervised learning methods can achieve high performance but they require a large amount of training data that is time-consuming and expensive to obtain. Active learning (AL) is well-suited to many problems in NLP, where unlabeled data may be abundant but labeled data is limited. The AL method aims to minimize annotation costs while maximizing the desired performance from the model. This study proposes a method to classify named entities from Tweet streams on Twitter by using an AL method with different query strategies. The samples were queried for labeling by human annotators based on query by committee and diversity-based querying. The experiments evaluated the proposed method on Tweet data and achieved promising results that proved better than the baseline.

Keywords Named entity recognition · Active learning · Query strategy · Text streams

C. Van Tran · T.T. Nguyen · D.T. Hoang · D. Hwang (✉)
Department of Computer Engineering, Yeungnam University, Gyeongsan, South Korea
e-mail: dosamhwang@gmail.com

C. Van Tran
e-mail: vancuongqbuni@gmail.com

T.T. Nguyen
e-mail: tuongtringuyen@gmail.com

D.T. Hoang
e-mail: hoangdinhtuyen@gmail.com

N.T. Nguyen
Faculty of Computer Science and Management, Wrocław University of Science and Technology, Wrocław, Poland
e-mail: Ngoc-Thanh.Nguyen@pwr.edu.pl

A. Zgrzywa et al. (eds.), *Multimedia and Network Information Systems*,
Advances in Intelligent Systems and Computing 506,
DOI 10.1007/978-3-319-43982-2_28

1 Introduction

Named entity recognition also known as entity identification and entity extraction is a subtask of information extraction. It identifies entities in documents and classifies them into predefined categories such as person names, locations, organizations, etc. [1, 16]. The NER problem plays an important role in many NLP applications and is a fundamental task of NLP systems. The extracted named entities can be utilized for various purposes such as entity relation extraction, document summarization [10, 15], speech recognition [9], and term indexing in information retrieval systems [3].

Many different machine learning approaches such as maximum entropy, hidden Markov models, support vector machines and conditional random fields (CRF) have been adopted for NER and achieved high accuracy based on a large annotated corpus. The supervised learning methods achieve high performance if they are applied to well-formatted text. However, achievement results are not as expected when applied to short and noisy messages. Twitter is a social network service that provides access to large volumes of data in real-time, but it is notoriously noisy and hard to tackle in NLP problems. For example, performance by the Stanford NER that uses the CRF model to train a classifier for CoNLL03 data dropped from 90.8 % to 45.8 % when it was applied to Tweets [8]. The length of a Tweet is 140 characters at most and Tweets contain different kinds of information, such as text, hyperlinks, user mentions, and hashtags. In addition, users often write Tweets with a freestyle and acronyms, and do not include extra information to explain the author's opinion. Another challenge for NLP systems is the large volume and the dynamic content in terms of time [5, 14]. The data from Twitter could be fed into processing systems as a data stream.

The unlabeled data are often easily obtained; however, annotating these texts can be rather tedious and time-consuming. The shortage of labeled data is an obstacle to supervised learning methods in developing application systems. Active learning is an attractive technique that addresses the shortage of labeled data for the training phase. Instead of training that relies on randomly labeled samples from a large corpus, the AL method chooses samples to label via optimal algorithms. Using different strategies, AL may determine a much smaller and the most informative subset from a large amount of unlabeled data. The motivation of this work is to query the most informative samples from Tweet streams using the AL method. This paper presents an AL method for classifying named entities from Tweet streams with different query strategies: query by committee and diversity-based querying. First, two classifiers trained on the CRF model and the maximum entropy model are used to classify unlabeled data in an arrival stream, and then they select dissimilar results in order to ask a human annotator to correct them. Second, the Tweets contain proper nouns in which the context is the least similar when compared to the existing training data that are queried for labeling. As a case study, experiments were conducted on Tweet data to assess proposed strategies. The method greatly reduced the training data and achieved results better than the random sampling.

The organization of this paper is as follows. Section 2 briefly presents related works. An introduction to the AL method is given in Sect. 3. Section 4 presents the proposed system, and the results are analyzed in the subsequent section. The conclusion and future work are presented in Sect. 6.

2 Related Work

Named entity recognition has attracted more interest from researchers in recent years, especially the problem recognizes named entities in microtexts, such as Tweets on Twitter. The first work to mention here was contributed by Ritter et al. [12]. They rebuilt an NLP tool beginning with parts of speech tagging. The NER system leverages the redundancy inherent in Tweets to achieve high performance by using labeled latent Dirichlet allocation to exploit freebase dictionaries in a semi-supervised learning method. Another approach was described by Liu et al. [8], who proposed combining the K-nearest neighbors algorithm with a linear CRF model in a semi-supervised learning method. Li et al. [7] also proposed a novel two-step unsupervised NER approach to recognizing named entities in Twitter data based on gregarious properties of named entities in a targeted Tweet streams. The method deals with streams, however, it does not determine the class of the identified entity, determining only if a phrase is an entity or not.

Yao et al. [17] presented an alternative AL strategy and combined this method with semi-supervised learning to reduce the labeling effort for a Chinese NER task. They utilized a strategy based on information density for the sample selection in a sequential labeling problem, which is suitable for both AL and self-training. They achieved an F1 score of 77.4 % with the proposed hybrid method on a Sighan bake-off 2006 Microsoft research of Asia NER corpus. Chen et al. developed and evaluated AL methods for a clinical NER task to identify concepts of medical problems and treatments from clinical notes [4]. They simulated AL experiments using a number of existing and novel algorithms in three different categories. Based on the learning curves of F1 score and the number of sentences, uncertainty sampling algorithms outperformed all other methods, and most diversity-based methods also performed better than random sampling. In another study [6], Hassanzadeh and Keyvanpour presented a variance-based AL method for the NER task that chooses informative entities to minimize the variance of the classifier currently built from labeled data. By finding entities where labeling by the current model was certain, they used self-training to resolve unlabeled samples. The experiments, when applied to the CoNLL03 English corpus showed that the method used considerably fewer numbers of manually labeled samples to produce the same results as when samples were selected in a random manner.

This paper presents an AL method to extract named entities from Tweet streams on Twitter. The proposed method is an effective way to solve the NER task on Twitter.

3 Active Learning-Based Approach

3.1 Overview

Active learning is a supervised learning method in which the learner controls the selection of necessary data for the learning phase. The key issue is how to recognize necessary data in order to ask the human annotator to annotate them. The learner will ask an expert in the related domain about labels of samples for which the learned model has made unreliable predictions so far. The main purpose of AL is to create as good a classifier as possible without supplying more labeled samples and more human effort in annotating data. Active learning has been successfully applied to a number of NLP tasks such as information extraction, named entity recognition, text categorization, and so on [11].

3.2 Scenarios

Different scenarios that have considered in the literature for the learner to make queries are (i) membership query synthesis, (ii) stream-based selective sampling, and (iii) pool-based sampling [13].

(i) Membership query synthesis: The learning system asks whether a particular domain sample belongs to the unknown concept or not. The learning system may request the labels for any unlabeled sample in the input space, including queries that the learning system generates de novo, rather than those sampled from some underlying natural distribution.

(ii) Stream-based selective sampling: The stream-based AL is important when data is continuously available and cannot be easily stored. The data can first be sampled from the actual distribution one at a time from the data source, and then the learning system can decide whether to request its label or not. If the input distribution is uniform, selective sampling may well behave like membership query learning. However, if the distribution is non-uniform and unknown, it is guaranteed that queries will still be sensible since they come from a real underlying distribution.

(iii) Pool-based sampling: Assume there is a huge pool of unlabeled data, which is usually assumed to be closed and queries select samples from this pool. The samples are queried according to an informativeness-measure technique, evaluating all samples in the pool. There is a main difference between this scenario and stream-based selective sampling. Stream-based selective sampling queries samples in a sequential way at a time the data arrives, and makes the query decisions individually, whereas pool-based sampling queries on the pool of available samples; therefore, it can evaluate the entire data set before selecting the most informative samples.

Algorithm 1 Active learning algorithm

Input: T - Time interval
Output: C - Classifiers

1: Label initial samples - L
2: Train initial classifiers C on L
3: **repeat**
4: Get arrival data - U
5: Preprocessing U
6: Apply query strategies to unlabeled data U to obtain D
7: Ask the annotator for labeling samples in D to obtain D'
8: Add D' to L
9: Retrain classifiers C on L
10: **until** Out of T or Annotator stops
11: **return** C

4 System Descriptions

4.1 System Procedure

This section presents a brief description of the proposed method. This work experiments with a practical stream-based AL scenario. The workflow of the idea is illustrated in Fig. 1, and the algorithm for the training phase is described in Algorithm 1. Some main steps in the algorithm are as follows.

Initial phase. The model parameters and data are initialized for the system. Initially, a set of random samples is selected to annotate as initial training data for the classification models. Two models are utilized to train classifiers: the CRF model and the maximum entropy model, respectively. This work uses the CRF model provided by Stanford[1] and the maximum entropy model provided by OpenNLP.[2]

Preprocessing. Elements such as mentions, hyperlinks, hashtags, and symbols are removed from Tweets.

Querying. To increase the training dataset, new samples are selected from unlabeled arrival data based on two query strategies: query by committee and diversity-based querying. The selected samples satisfy some criterion such as the classification results of models are dissimilar or the context of the proper noun is the least similar to the training data. The queried samples are labeled by the human annotator and then added to the training data to retrain models.

Training. The classification models are applied to the updated training data to retrain the classifiers.

[1] http://nlp.stanford.edu/software/CRF-NER.shtml.

[2] https://opennlp.apache.org/documentation/1.5.3/manual/opennlp.html.

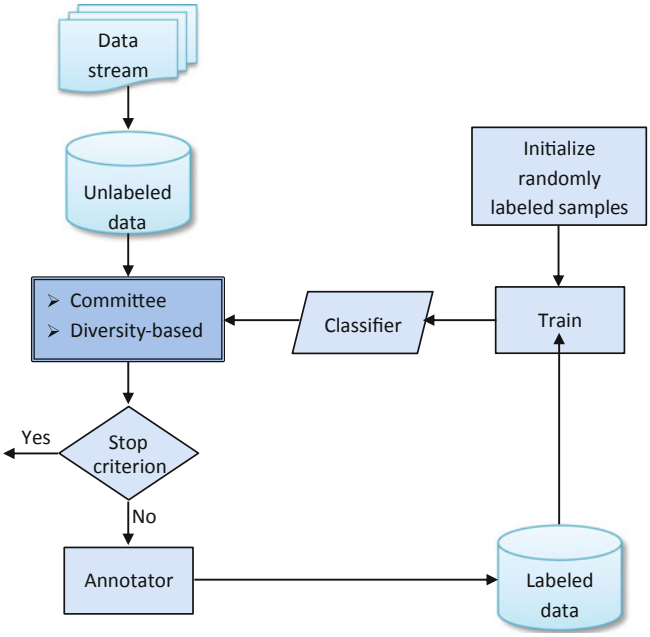

Fig. 1 The workflow of the system for the training phase

Iteration. The sampling and labeling processes are iterated until the stopping criterion is met. In this experiment, the system's processes are finished when the annotator stops working or they are beyond the considered time interval.

4.2 Context Features

A part of speech (POS) tagger assigns the parts of speech to each word or token, such as noun, verb, adjective, etc. The proper nouns are extracted based on the results of POS tagger (i.e., the experiments use a publicly available POS tagger developed by Stanford[3]).

With each proper noun in the unlabeled data and the named entity in the labeled data, a contextual vector is constructed to represent its contextual information. The size of the contextual vector is the size of the considered window. In the experiments, the window size is set to six (i.e., three words on the left and three words on the right of the proper noun or the named entity are examined). The elements of the vector are the POS tag of words in the considered window. Proper noun *PN* is represented by the contextual vector of the POS tag as follows:

[3] http://nlp.stanford.edu/software/.

$$PN = (..., pos_{-3}, pos_{-2}, pos_{-1}, pos_{+1}, pos_{+2}, pos_{+3}, ...) \quad (1)$$

where pos_i is the POS tag of the word at location i from the proper noun. The negative sign and the positive sign, respectively, mean that the words are to the left-hand side and to the right-hand side of the proper noun.

4.3 Query Strategies

Query by committee is a multi-classifier approach, where the classifiers are trained on the current training data by individual models (i.e., the CRF model and the maximum entropy model) and then they are used to examine the unlabeled arrival data. The disagreement between classifiers with respect to the value and the category of a named entity is utilized to decide whether the sample is to be labeled by the human annotator.

Diversity-based querying selects samples where the training data are the least similar. A vector model is used to measure the similarity between a Tweet and all the training data. The Tweets that contain a proper noun are examined for the similar context between the proper noun and named entities that exist in the current training data. Each proper noun and named entity is represented by a contextual vector, as presented above. The contextual vectors of proper nouns are then compared with all contextual vectors of the named entities in the training data. The Tweets where similar scores are less than a certain threshold will be subjected to labeling by the human annotator.

5 Experimental Results

5.1 Dataset and Baselines

The performance of the proposal method was evaluated by applying the system to Tweet data. The data consisted of Tweets of 20 users that were collected on Twitter from January 1st, 2014 to September 30th, 2015, by using the public Java library for the Twitter API[4] and then dropping Tweets with only hashtags, mentions, hyperlinks, or emoticons. Finally, 10,813 unlabeled Tweets were selected. In addition, 4,716 labeled Tweets were used as initial labeled data. The test set (*TS*) included 1,153 Tweets that were also collected on Twitter from October 1st, 2015 to December 31th, 2015, and annotated as the gold standard (*GS*) to assess the performance. The dataset was annotated with three named entity categories: Person, Location, and Organization.

[4] http://twitter4j.org.

The two systems implemented are the query by committee (QBC) algorithm and the diversity-based querying (DBQ) algorithm. A random sampling (RS) algorithm that presents a passive learning method was also implemented as a baseline compared to our proposed method.

5.2 Evaluation and Results

The performance of this task was calculated following #MSM2013's measures [2]. Precision, Recall, and F-measure are calculated for each entity category, and the final results for all entity categories are the average performance of the defined categories.

The entity is represented in a tuple (entity value, entity category), and the strict matching is performed between named entities in the *TS* and answers in the *GS* for both detection of the correct entity value and the correct entity category.

Precision ($\overline{P}$) and Recall ($\overline{R}$) for all entity categories are the average value of the precision and the recall of all entity categories, respectively. The F-measure score (also called F_1 score) is the harmonic mean of $\overline{P}$ and $\overline{R}$, defined as follows:

$$F_1 = 2 \times \frac{\overline{P} \times \overline{R}}{\overline{P} + \overline{R}} \tag{2}$$

The results for each entity category and overall results from systems are shown in Table 1. All systems were tested in the same AL framework (i.e., the same initial training data, model, default parameters for models, time interval, and test set). The experiments used the CRF model for testing data and set the time interval at 20 (i.e., the Tweets were queried during 20 days for each iteration).

The performance of all query methods (i.e., QBC and DBQ) outperformed the RS method, in which the selected Tweets of the AL method were less than or equal to the RS method. The F_1 score of the QBC was the best, where the F_1 score was 64 %.

Table 1 The performance of the systems

	System	Person	Location	Organization	All
Precision	*QBC*	**83.8**	83.3	78.6	81.9
	DBQ	79.7	**85.6**	**80.8**	**82**
	RS	79.5	83.3	66.7	76.5
Recall	*QBC*	59.6	**68.9**	**28.9**	**52.5**
	DBQ	**60.3**	67.7	27.6	51.9
	RS	57.1	65.9	23.7	48.9
F_1	*QBC*	**69.7**	75.4	**42.3**	**64**
	DBQ	68.6	**75.6**	41.2	63.6
	RS	66.4	73.6	35	59.7

Table 2 The number of selected tweets of each system

System	#Selected Tweets
QBC	2,165 (20.02 %)
DBQ	3,252 (30.1 %)
RS	3,252 (30.1 %)

That was better than the baseline 4.3 %. The QBC outperformed the DBQ since it required fewer labeled Tweets than the DBQ. The F_1 score of the two query methods are very similar (i.e., 64 % for QBC and 63.6 % for DBQ).

5.3 Discussions

The number of selected Tweets at the end of the training process from the AL method for each query method is shown in Table 2. Although the QBC method only selects 20.02 % of the unlabeled Tweets, its performance is better than the DBQ and the RS, which select 30.1 % of unlabeled Tweets.

Comparing the performance and the number of selected Tweets among the query strategies of the AL method, and between the AL method and the passive learning method, the QBC method is the best (i.e., to achieve a 64 % F_1 score, QBC queried 2,165 Tweets; DBQ queried 3,252 Tweets to achieve 63.6 %, and RS also queried 3,252 Tweets to achieve 59.7 %).

One concern with applying the QBC method to real tasks is that it relies on updated models, which require more time for training. In the experiments, it takes several minutes to fully train a model; this is also suitable in reality for the stream-based scenario. The DBQ method does not depend on trained models, but in the experiments, it did not outperform the QBC method.

6 Conclusion and Future Work

The active learning method aims to minimize annotation costs while maximizing the performance of the model. This study conducted AL experiments for NER with Tweet streams from Twitter and showed that the AL method has the potential to reduce annotation costs to train the model. By using two different query strategies (query by committee and diversity-based querying), the AL method achieved the performance better than the passive learning method.

Future work includes improving the current query strategies and proposing more query methods for the NER problem on Twitter.

Acknowledgments This work was supported by the BK21+ program of the National Research Foundation (NRF) of Korea.

References

1. Abdallah, S., Shaalan, K., Shoaib, M.: Integrating rule-based system with classification for arabic named entity recognition. In: Computational Linguistics and Intelligent Text Processing, pp. 311–322. Springer (2012)
2. Cano Basave, A.E., Varga, A., Rowe, M., Stankovic, M., Dadzie, A.S.: Making sense of microposts (#msm2013) concept extraction challenge (2013)
3. Chen, H.H., Ding, Y.W., Tsai, S.C.: Named entity extraction for information retrieval. Comput. Process. Orient. Lang. **12**(1), 75–85 (1998)
4. Chen, Y., Lasko, T.A., Mei, Q., Denny, J.C., Xu, H.: A study of active learning methods for named entity recognition in clinical text. J. Biomed. Inf. **58**, 11–18 (2015)
5. Giao, B.C., Anh, D.T.: Similarity search for numerous patterns over multiple time series streams under dynamic time warping which supports data normalization. Vietnam J. Comput. Sci. pp. 1–16 (2016)
6. Hassanzadeh, H., Keyvanpour, M.: A variance based active learning approach for named entity recognition. In: Intelligent Computing and Information Science, pp. 347–352. Springer (2011)
7. Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., Lee, B.S.: Twiner: named entity recognition in targeted twitter stream. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 721–730. ACM (2012)
8. Liu, X., Zhang, S., Wei, F., Zhou, M.: Recognizing named entities in tweets. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1. pp. 359–367. Association for Computational Linguistics (2011)
9. Meyer, C., Schramm, H.: Boosting hmm acoustic models in large vocabulary speech recognition. Speech Commun. **48**(5), 532–548 (2006)
10. Nobata, C., Sekine, S., Isahara, H., Grishman, R.: Summarization system integrated with named entity tagging and ie pattern discovery. In: Proceedings of Third International Conference on Language Resources and Evaluation, pp. 1742–1745 (2002)
11. Olsson, F.: A literature survey of active machine learning in the context of natural language processing (2009)
12. Ritter, A., Clark, S., Etzioni, O., et al.: Named entity recognition in tweets: an experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1524–1534. Association for Computational Linguistics (2011)
13. Settles, B.: Active learning literature survey. Univ. Wis. Madison **52**(55–66), 11 (2010)
14. Stahl, F., Schomm, F., Vossen, G., Vomfell, L.: A classification framework for data marketplaces. Vietnam J. Comput. Sci. pp. 1–7 (2016)
15. Tran, T., Nguyen, D.T.: Algorithm of computing verbal relationships for generating vietnamese paragraph of summarization from the logical expression of discourse representation structure. Vietnam J. Comput. Sci. pp. 1–12 (2015)
16. Tran, V.C., Hwang, D., Jung, J.J.: Semi-supervised approach based on co-occurrence coefficient for named entity recognition on twitter. In: 2015 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS), pp. 141–146. IEEE (2015)
17. Yao, L., Sun, C., Wang, X., Wang, X.: Combining self learning and active learning for chinese named entity recognition. J. Softw. **5**(5), 530–537 (2010)