Evaluating Ensemble Based Pre-annotation on Named Entity Corpus Construction in English and Chinese

Tingming Lu[1,2], **Man Zhu**[3], **Zhiqiang Gao**[1,2], **and Yaocheng Gui**[1,2]
[1]Key Lab of Computer Network and Information Integration (Southeast University),
Ministry of Education, China
[2]School of Computer Science and Engineering, Southeast University, China
[3]School of Computer Science and Technology,
Nanjing University of Posts and Telecommunications, China
lutingming@163.com,mzhu@njupt.edu.cn,{zqgao,yaochgui}@seu.edu.cn

Abstract

Annotated corpora are crucial language resources, and pre-annotation is an usual way to reduce the cost of corpus construction. Ensemble based pre-annotation approach combines multiple existing named entity taggers and categorizes annotations into *normal annotations* with high confidence and *candidate annotations* with low confidence, to reduce the human annotation time. In this paper, we manually annotate three English datasets under various pre-annotation conditions, report the effects of ensemble based pre-annotation, and analyze the experimental results. In order to verify the effectiveness of ensemble based pre-annotation in other languages, such as Chinese, three Chinese datasets are also tested. The experimental results show that the ensemble based pre-annotation approach significantly reduces the number of annotations which human annotators have to add, and outperforms the baseline approaches in reduction of human annotation time without loss in annotation performance (in terms of F_1-measure), on both English and Chinese datasets.

1 Introduction

The current success and widespread use of machine learning techniques for processing human language make annotated corpora essential language resources. Many popular natural language processing (NLP) algorithms require large amounts of high-quality training samples, which are time-consuming and costly to build. One usual way to improve this situation is to automatically pre-annotate the corpora, so that human annotators need merely to correct errors rather than to annotate from scratch.

Named Entity Recognition (NER), one of the fundamental tasks for building NLP systems, is a task that detects Named Entity (NE) mentions in a given text and classifies these mentions to a predefined list of types. Resulted from more than two decades of research, many named entity taggers are publicly available now. Some of the taggers are integrated into NLP workflows based on Service Oriented Architecture (Ide et al., 2015; Piperidis et al., 2015). And it is well known that multiple taggers can be combined using ensemble techniques to create a system that outperforms the best individual tagger within the system (Wu et al., 2003; Speck and Ngomo, 2014). However, only a few studies have been reported on leveraging ensemble to combine multiple existing taggers to assist named entity annotation.

Lu et al. (2016) introduced ensemble based pre-annotation approach in named entity corpus construction. They conducted experiments on an English dataset, and the results showed that the ensemble based pre-annotation approach outperforms the baseline approaches in reduction of human annotation time.

In this paper, we perform a more thorough evaluation on the ensemble based pre-annotation approach. 1) We manually annotate three English datasets under various pre-annotation conditions, report the effects of ensemble based pre-annotation, and analyze the experimental results. 2) We also manually annotate three Chinese datasets, to verify the effectiveness of ensemble based pre-annotation in Chinese language.

The remaining part of this paper is organized as follows: In Section 2, we mention related work. Section 3 describes the experimental setup, followed by experimental results and analysis in Section 4. Finally, we conclude and discuss future directions in Section 5.

This work is licenced under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/

Proceedings of WLSI/OIAF4HLT,
pages 56–60, Osaka, Japan, December 12 2016.

2 Related Work

Given the importance of annotated corpora to NLP system development, many applications for different domains have been built in order to assist named entity annotation, using a single tagger (Lingren et al., 2014; Ogren et al., 2008), or multiple taggers (Ganchev et al., 2007).

The goal of an ensemble learning algorithm is to generate a classifier with a high predictive performance by combining the predictions of a set of basic classifiers. Previous work has already suggested that ensemble learning can be used to improve NER (Wu et al., 2003; Speck and Ngomo, 2014; Florian et al., 2003; Desmet and Hoste, 2010). Speck and Ngomo (2014) combined four state-of-the-art taggers by using 15 different algorithms for ensemble learning and evaluated their performance on five datasets. Their results suggested that ensemble learning can reduce the error rate of state-of-the-art NER systems by 40%.

We follow Lu et al. (2016) and perform a more thorough evaluation on the ensemble based pre-annotation. We manually annotate three English datasets and three Chinese datasets, report the performance, and analyze the results.

3 Experimental Setup

3.1 Datasets

All the datasets used in our experiments are publicly available. There are three English datasets, and three Chinese datasets. From each dataset, 60 articles are selected randomly. From each of the 60 articles, one sentence is extracted to perform the actual assisted annotation experiments. Sentences containing more NEs are preferred over ones containing less NEs. Sentences containing no NE will not be extracted. The three English datasets have been described in Lu et al. (2016). The three Chinese datasets are People's Daily (Fu and Luke, 2005), Penn Chinese Treebank 5.1 (CTB5) (Xue et al., 2005), and ITNLP[1].

3.2 Taggers

Six English NE taggers and three Chinese NE taggers are involved in our experiments. They are all public available. For outputs of these taggers, only three types are considered in our experiments, namely Person, Location, and Organization. The English NE taggers are the same as the ones described in Lu et al. (2016). The Chinese NE taggers are ICTCLAS[2] (Liu et al., 2004), FudanNLP[3] (Qiu et al., 2013) and Stanford Named Entity Recognizer[4] (Stanford(zh))(Manning et al., 2014).

3.3 Pre-annotators

For the English taggers, the pre-annotator Ensemble(en) which combines six taggers and produces *normal* and *candidate annotations* is used to evaluate the ensemble based pre-annotation approach. One baseline pre-annotator using a single tagger is denoted as Stanford(en). Another baseline pre-annotator Stanford(en)+Illinois produces annotations which are union of the outputs of two taggers, namely Stanford(en) and Illinois. No ensemble technique is applied on the pre-annotator Stanford(en) and Illinois. We choose Stanford(en) and Illinois because they are the best two taggers in terms of F_1-measure on the test datasets.

Similarly, for the Chinese taggers, the pre-annotator Ensemble(zh) combines three taggers. One baseline pre-annotator using a single tagger is ICTCLAS. Another baseline pre-annotator ICTCLAS+Stanford(zh) produces annotations which are union of the outputs of the two taggers.

For the ensemble based pre-annotators (Ensemble(en) and Ensemble(zh)), Weighted Voting (Zhou, 2012) is used to weight the different taggers. The ensemble based pre-annotators learn the weights incrementally after each sentence in a dataset is annotated by human.

[1] `http://www.datatang.com/data/44067/`

[2] `http://ictclas.nlpir.org/` (version 5.0).

[3] `http://nlp.fudan.edu.cn/` (version 2.1).

[4] `http://nlp.stanford.edu/software/CRF-NER.shtml` (version 3.6.0).

Table 1: Assisted annotation experiments. Annotators are assigned to annotate sentences under various pre-annotation conditions.

Dataset	H_1	H_2	H_3
AKSW-News	Stanford(en)	Ensemble(en)	Stanford(en)+Illinois
CoNLL-Test	Stanford(en)+Illinois	Stanford(en)	Ensemble(en)
Reuters-128	Ensemble(en)	Stanford(en)+Illinois	Stanford(en)
CTB5	ICTCLAS	Ensemble(zh)	ICTCLAS+Stanford(zh)
ITNLP	ICTCLAS+Stanford(zh)	ICTCLAS	Ensemble(zh)
People's Daily	Ensemble(zh)	ICTCLAS+Stanford(zh)	ICTCLAS

Table 2: Results of assisted annotation experiments.

Pre-annotator	Language	N_{add}	N_{modify}	Precision	Recall	F_1	Time
Stanford(en)	English	0.47	**0.16**	0.947	0.923	0.935	18.8
Stanford(en)+Illinois	English	0.37	0.46	0.942	0.924	0.933	18.7
Ensemble(en)	English	**0.11**	0.67	0.950	0.930	0.940	**18.2**
ICTCLAS	Chinese	1.32	**0.08**	0.948	0.936	0.942	14.6
ICTCLAS+Stanford(zh)	Chinese	0.82	0.91	0.949	0.947	0.948	14.6
Ensemble(zh)	Chinese	**0.59**	0.84	0.951	0.951	0.951	**13.3**

3.4 Assisted Annotation Experiments

Three human annotators (H_1, H_2, and H_3) participate in our assisted annotation experiments. They are graduate students in our school, and major in NLP study. After they have annotated some sentences in a training dataset to get familiar with the Web based UI, each of them has to annotate all of the sentences in the six datasets. The human annotators are presented with the sentences in the same order (Table 1), but for different human annotators, each sentence is pre-annotated by different pre-annotators. We carefully design the experiments, to ensure that each sentence will be pre-annotated by all the pre-annotators, and will be annotated by all the human annotators.

4 Results and Analysis

The results of assisted annotation experiments under various pre-annotation conditions are presented in Table 2. After the sentences are pre-annotated by ensemble based pre-annotators (Ensemble(en) and Ensemble(zh)), human annotators take less annotation time per sentence without loss in annotation performance (in terms of F_1-measure), on both English and Chinese datasets.

The Web based UI automaticly records the number of *adding actions* (N_{add}) and number of *modifying actions* (N_{modify}) when human annotators are annotating. As presented in Table 2, ensemble based pre-annotation approach significantly reduces the number of *adding actions*. However, ensemble based pre-annotation approach introduces more *modifying actions*, compared to single taggers (Stanford(en) and ICTCLAS). We will analyze the results in the following.

We utilize linear regression to model the annotation time, where T_{total} is the total time in seconds spent on a sentence, N_{token} is the number of English tokens or Chinese characters in the sentence, T_{token} is the time taken on reading an English token or a Chinese character, N_{add} is the number of *adding actions*, T_{add} is the time taken on performing an *adding action*, N_{modify} is the number of *modifying actions*, T_{modify} is the time taken on performing a *modifying action*, and additionally, there is T_c seconds of overhead per sentence.

$$T_{Total} = N_{token} \cdot T_{token} + N_{add} \cdot T_{add} + N_{modify} \cdot T_{modify} + T_c$$

Table 3: Estimated time spent on reading a token, adding a new annotation, modifying an existed annotation, etc.

Language	T_{token}	T_{add}	T_{modify}	T_c
English	0.25	4.79	1.95	7.45
Chinese	0.11	4.22	1.94	1.64

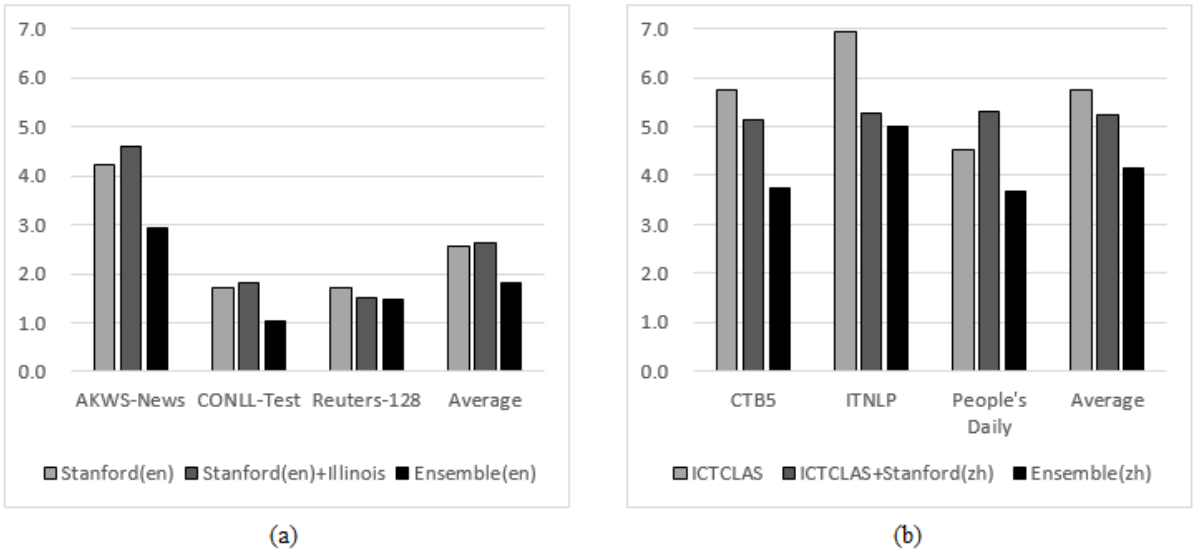

Figure 1: Estimated time taken by human annotators on performing adding and modifying actions on English (a) and Chinese (b) datasets.

There are 180 sentences in the three English datasets. Each of them is pre-annotated by three pre-annotators, and then annotated by three human annotators. Finally we get 540 instances. Similarly, from the experimental results on the Chinese datasets, we get 540 instances. Based on the time model, we get T_{token}, T_{add}, T_{modify}, and T_c, as listed in Table 3. As we can see, adding a new annotation takes twice more time than modifying an existing annotation, both on English and Chinese datasets.

Given a sentence under different pre-annotation conditions, $N_{token} \cdot T_{token} + T_c$ is constant, while the *adding* and *modifying action* time $T_{a+m} = N_{add} \cdot T_{add} + N_{modify} \cdot T_{modify}$ varies. Now, we can estimate the *adding* and *modifying action* time for the different pre-annotation approaches on all the datasets. From Figure 1, we can see that after the sentences are pre-annotated by ensemble based pre-annotators, human annotators take less time on performing *adding* and *modifying actions* than the two baseline approaches on all datasets.

5 Conclusion

In this paper, we evaluate the effects of ensemble based pre-annotation which combines multiple existing NE taggers on three English datasets and three Chinese datasets. The experimental results show that the ensemble based pre-annotation approach reduces the number of *adding actions* and the total human annotation time, without loss in annotation performance (in terms of F_1-measure). Based on a linear regression model, we estimate the time taken on performing *adding* and *modifying actions* by human annotators, and conclude that ensemble based pre-annotation approach reduces the human annotation time on all datasets. In future work, we will study how different ensemble algorithms affect the performance, and will try to apply ensemble based pre-annotation approach to other NLP tasks, such as Entity Linking, Relation Extraction, etc.

Acknowledgements

This work is partially funded by the National Science Foundation of China under Grant 61170165, 61602260, 61502095. We would like to thank all the anonymous reviewers for their helpful comments.

References

Bart Desmet, and Vronique Hoste 2010. Dutch named entity recognition using classifier ensembles. *LOT Occasional Series*, vol 16, pp. 29–41.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL*, vol 4. Association for Computational Linguistics.

Guohong Fu, and Kang-Kwong Luke. 2005. Chinese named entity recognition using lexicalized HMMs. *ACM SIGKDD Explorations Newsletter*, vol. 7, no. 1, pp.19–25.

Kuzman Ganchev, Fernando Pereira, and Mark Mandel. 2007. Semi-automated named entity annotation. *Proceedings of the linguistic annotation workshop,* pp. 53–56. Association for Computational Linguistics.

Nancy Ide, James Pustejovsky, Christopher Cieri, Eric Nyberg, Denise DiPersio, Chunqi Shi, Keith Suderman, Marc Verhagen, Di Wang, and Jonathan Wright. 2015. The language application grid. *International Workshop on Worldwide Language Service Infrastructure*, (pp. 51–70). Springer International Publishing.

Todd Lingren, Louise Deleger, Katalin Molnar, Haijun Zhai, Jareen Meinzen-Derr, Megan Kaiser, Laura Stoutenborough, Qi Li, and Imre Solti. 2014. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. *Journal of the American Medical Informatics Association,* 21(3), 406-413.

Qun Liu, Huaping Zhang, Hongkui Yu, and Xueqi Cheng. 2004. Chinese lexical analysis using cascaded hidden Markov model. *Journal of Computer Research and Development*, vol. 41, no. 8, pp. 1421–1429.

Tingming Lu, Man Zhu, and Zhiqiang Gao. (under publication) 2016. Reducing Human Effort in Named Entity Corpus Construction Based on Ensemble Learning and Annotation Categorization.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60.

Philip V. Ogren, Guergana K. Savova, and Christopher G. Chute. 2008. Constructing evaluation corpora for automated clinical named entity recognition. *Proceedings of the Language Resources and Evaluation Conference (LREC),* pp. 28–30.

Stelios Piperidis, Dimitrios Galanis, Juli Bakagianni, and Sokratis Sofianopoulos. 2015. Combining and extending data infrastructures with linguistic annotation services. *International Workshop on Worldwide Language Service Infrastructure*, (pp. 3–17). Springer International Publishing.

Xipeng Qiu, Qi Zhang and Xuanjing Huang. 2013. FudanNLP: A Toolkit for Chinese Natural Language Processing. *Proceedings of the 51nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Ren Speck, and Axel C. N. Ngomo. 2014. Ensemble learning for named entity recognition. *Semantic Web–ISWC 2014. LNCS*, vol 8796, pp. 519–534. Springer, Heidelberg.

Decai Wu, Grace Ngai, and Marine Carpuat. 2003. A stacked, voted, stacked model for named entity recognition. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, CONLL 2003*, vol. 4, pp. 200–203. Association for Computational Linguistics, Stroudsburg

Naiwen Xue, Feixia Fu, Dong Chiou, and Marta Palmer. 2005. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, vol. 11, no. 2, pp.207–238.

Zhihua Zhou. 2012. Ensemble methods: foundations and algorithms. *CRC Press*, pp 74–75.