Improving NER Models by exploiting Named Entity Gazetteer as External Knowledge

Atefeh Zafarian, Habibollah Asghari
ICT Research Institute
Academic Center for Education,
Culture and Research (ACECR)
Tehran, Iran
{Zafarian, habib.asghari }@ictrc.ac.ir

Abstract

This paper proposes a supervised NER model based on gazetteer for NSURL-2019 Task 7: Named Entity Recognition (NER) in Farsi. Supervised methods generate acceptable results in many Natural Language Processing tasks such as Named Entity Recognition (NER). Since these methods are domain-based, so their quality is related to the volume and the domain of training data. External knowledge can help the supervised methods to compensate this deficiency. In this paper, we use an unlabeled corpus as external knowledge to extract a named entity gazetteer for improving the performance of NER systems. We apply a supervised NER model on the unlabeled corpus to extract named entities with high probability. Finally we train a new NER model by using the gazetteer as a new feature to be employed with other features. The results show that the performance of NER model exploiting the gazetteer outperforms the ordinary models.

1 Introduction

This paper proposes a NER model for (Taghizadeh et al., 2019). NER systems extract important names from text such as person, location and organization. Some NER systems may cover other tags such as time, date, money etc. due to the type of the information that we expect to extract from the text.

Many of the previous works use supervised methods for constructing high performance NER models. They generate good results but only on their specific domain, but if the domain changes, they won't work efficiently. For compensation, some of the researches used external knowledge such as entity dictionaries or gazetteers. Gazetteers contain named entities that researchers add them as an external knowledge for improving the performance of NER model. However, generating and maintaining high-quality gazetteers is very time consuming. There are some methods that have been proposed for solving this problem by automatically extracting gazetteers from external knowledge for example Torisawa (2007). In a research investigated by Torisawa, (2007), they have extracted NEs from Wikipedia by automatic methods. Although extracted information of Wikipedia as a gazetteer is useful for training NER models, they don't cover all of the new entities because of rapid changes in the information content. Moreover, they cannot extract all of the tags and only focus on a limited set of tags such as person, location and organization names. However, many of the applications need more tags.

In this paper, we propose an automatic method for generating a named entity gazetteer from a big unlabeled corpus. At first, we use a supervised NER model to decode unlabeled corpus. At the second step, we extract a high-confidence named entity list from the unlabeled corpus as an entity gazetteer. Finally we add the gazetteer as a new feature to the model and retrain our NER model to generate a new one.

For generating the corpus for both the NER model and the gazetteer, we use the news data from Persian news agencies. This approach is beneficial and improves the performance of NER model because it adds the newest information from recently released news to our model. Moreover, the gazetteer is designed in such a way to be in similar domain with the NER corpus. So, it gives a better performance in comparison to Wikipedia resource because it contains most recent information from the news text.

The experiments in this paper are conducted in Persian and the data set is a large NER corpus, coming from the NSURL shared task. We show our results in phrase level and word level for a 3 classes and a 7 classes NER system. We also show the achieved results for all of the tags. Our results show an acceptable accuracy in F-score and a good result in precision. Also we got a good accuracy in new tags such as 'Time', 'Date', 'Money' and 'Percent'.

The paper is organized as follow: in the second section, an overview of the pervious works in exploiting gazetteers to enhance the performance of NER models will be presented. Section 3 comes with our proposed method. In section 4, the experimental setup is explained including the data set and evaluation measures. In section 5, our experiments and results are thoroughly described. Conclusion and future works are described in the last section.

2 Related Work

There are different approaches for generating NER models. Some of them use external knowledge as a feature to improve their model. For example, Torisawa (2007) retrieves the corresponding Wikipedia entry for each candidate word sequence and improves the NER system by the candidates. (Nothman et al., 2008) transforms the Wikipedia link into Named Entity Recognition by classifying the target Wikipedia pages into common entity types. Cucerzan (2007) have employed Wikipedia in order to support a Named Entity Recognition and disambiguate extracted named entities. (Bøhn and Nørvag, 2010) have applied Wikipedia contents to automatically generate an entity dictionary to connect the same named entity to the same tag. In a research investigated by (Nadeau et al., 2006) they proposed an unsupervised named entity Recognition by automatically extracting gazetteers from a large amounts of text. (Toral and Monachini, 2008) improved the performance of a named entity recognition by using external knowledge. (Etzioni et al., 2005) focused on automatic extraction from the Web for improving a Named Entity Recognition system. It should be noted that some researches have shown that larger NE lists do not necessarily correspond to increased NER performance (Mikheev et al., 1999).

3 Proposed Method

Our method includes five steps as follow:

- Preprocessing of the text.
- Training a CRF-based NER model.
- Crawling a large amount of news from Persian news agencies for generating an unlabeled corpus.
- Applying NER model on the unlabeled corpus and extracting high-confidence named entities as a gazetteer.
- Adding the gazetteer to CRF-based model and training the new model.
- In the following subsections we will describe the above mentioned steps in detail.

3.1 Preprocess

At the first step, we preprocess the NSURL corpus. We use Parsivar tools for text preprocessing (Mohtaj et al., 2018). There are some problems in the corpus; for example the whole of some sentences in the corpus were tagged as a single named entity. We remove the sentences because it increases the runtime and has negative effect on the results. Furthermore, we apply a normalizer on the corpus to unify the character codes.

3.2 Training the NER Model

We use CRF algorithm for training the model. Because of the supervised algorithm we used, it gives a high performance model. The tool that has been employed is CRF-based Stanford Named Entity Tagger. It presents good facilities for define NER features.

We checked different features for NER model and identified a series of n-gram features such as the assigned class of the word, the word itself and the previous and next words as best features for training the model. Table 1 shows the feature set used for our proposed model.

Description	Feature
Current Word	W3
Left Word	W2
Right Word	W4
Left Tag	T2
Two Left Words	W1W2
Two Left Tags	T1T2

Table 1: Feature Set.

3.3 Generate Unlabeled Data

As mentioned before, since the domain of NSURL corpus is from Persian news, so we use the text from Persian news for making unlabeled corpus. We crawl some popular news agencies and extract news from different categories. We focus on the domain of training corpus; for example if the training corpus contains only the text in sport domain, we crawl only sports news. Then, we apply a preprocessing tool on unlabeled Corpus and tokenize and normalize the sentences and remove very short and very long sentences.

3.4 Generate Named Entity Gazetteer

For generating named entity gazetteer, we decode unlabeled corpus with our NER model and extract words with high probability. For extracting these entities, we also consider sentence confidence using following the equation (Zafarian et al., 2015).

$$sentence\ confidence = \frac{\sum_{word \in i} word\ confidence}{\max(10, sentenc\ length)}$$

If the word confidence and sentence confidence are both reliable, we extract entities from that sentence.

3.5 Retrain NER Model

Finally, we add the named entity gazetteer as a new feature to our proposed NER model and re-train the model with this new feature.

4 Experimental Setup

4.1 Dataset

We used NSURL corpus as training data. It is a Persian NER corpus with more than 900 thousand words that is manually labeled for NER tasks. This corpus was published by NSURL-2019 Workshop for Farsi (Persian) NER Task.

4.2 Evaluation Measure

For Evaluation of NER systems, most of the researches use precision, recall, and F-score as performance measures. Precision is the number of NEs a system correctly detected divided by the total number of NEs identified by the system. Recall is the number of NEs a system correctly detected divided by the total number of NEs contained in the input text. F-Score combines these two into a single score and is defined with the following equation (Tsai et al., 2006).

$$F - score = \frac{2 * precision * recall}{\text{precision} + \text{recall}}$$

5 Experiments and Result

We participated in NER resolution Shared Task for Farsi under the NSURL-2019 Workshop as Team-4. Our results in the workshop are shown in Tables 2 to 6. As we mentioned in section 4, the NSURL corpus is prepared as a training NER corpus but we used only 57 percent of corpus because of the limitations in hardware and computation platform. We expect that the performance of our system be improved if all of the dataset is used for the training phase of the system. To reduce the computational complexity, we removed long sentences with less than two tags. Table 7 shows the characteristics of NSURL and our shortened training corpus.

Corpus	Sentence	Word	Tag
NSURL	23,321	912,032	100,118
Sh_NSURL	10,388	502,989	85,265

Table 2: The characteristics of Corpus.

The results of phrase level and word level NER model for 3 classes (Person, Location and Organization) are shown Table 3 and 4. Moreover, the results in phrase level and word level NER model for 7 classes (Person, Location, Organization, Time, Date, Currency and Percent) are shown in Table 5 and 6.

Although we used only 57 percent of training data, we got acceptable results in NSRUL workshop. In Tables 3 to 5, our results show a lower recall compared to some groups, but we got a better result in precision measure. Table 6 shows the details of phrase level evaluation for 7 classes. As we expected, we got a better result in new tags such as ‘time’, ‘date, ‘money’ and ‘percent’.

Test Data 1	P	R	F1
In Domain	**87.5**	76.0	81.3
Out Domain	**87.5**	76.0	81.3
Total	**86.8**	72.3	78.9

Table 3: Phrase -level evaluation for subtask A: 3-classes

Test Data 1	P	R	F1
In Domain	**90.1**	78.2	83.7
Out Domain	88.7	70.2	78.4
Total	**89.4**	73.5	80.7

Table 4: Word -level evaluation for subtask A: 3-classes

Test Data 1	P	R	F1
In Domain	**87.0**	76.1	81.2
Out Domain	**86.2**	70.2	77.4
Total	**86.5**	72.7	79.0

Table5: Phrase -level evaluation for subtask A: 7-classes

Test Data 1	P	R	F1
In Domain	89.2	83.1	86.1
Out Domain	89.8	76.5	82.6
Total	89.7	79.4	84.2

Table 6: Word-level evaluation for subtask A: 7-classes

Test Data 1	F1
Per	76.2
ORG	75.9
LOC	82.8
DAT	**76.0**
TIM	67.1
MON	**91.3**
PCT	**93.6**
Total F1	79.0

Table 7: Details of phrase-level evaluation for subtask B: 7-classes

6 Conclusion

Supervised methods are domain based so that they generate good results but only on their specific domain. External knowledge can help supervised methods especially if they have common information with test data. In this paper, we extracted useful information from a large unlabeled corpus that it is in the same domain with the test data, both of them are from Persian news, so we added the gazetteer as a new feature to our supervised model. Our results show that this new feature is effective in our named entity recognition model and outperforms the ordinary model.

Acknowledgments

This research is a part of News Dashboard project to be deployed for Islamic Republic of Iran Broadcasting (IRIB). The authors would like to thank all of the members of the above mentioned project. Special credit goes to Dr. Shirin Ghanbari for her warm support.

References

Andrei Mikheev, Marc Moens, and Claire Grover. 1999. *Named entity recognition without gazetteers. In Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, Association for Computational Linguistics pages 1-8.*

Antonio Toral and Monica Monachini. 2008. Named entity wordnet. *In Proceedings of the 6th International Conference on Language Resources and Evaluation.*

Atefeh Zafarian, Ali Rokni, Shahram Khadivi, and Sonia Ghiasifard. 2015. *Semi-supervised learning for named entity recognition using weakly labeled training data. In 2015 The International Symposium on Artificial Intelligence and Signal Processing (AISP), pages 129-135. IEEE.*

Christian Bøhn, and Kjetil Nørvåg. 2010. *Extracting named entities and synonyms from wikipedia. In 2010 24th IEEE International Conference on Advanced Information Networking and Applications, pages 1300-1307. IEEE.*

David Nadeau, Peter D. Turney, and Stan Matwin. 2006. *Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In Conference of the Canadian society for computational studies of intelligence, pages 266-277. Springer, Berlin, Heidelberg.*

Joel Nothman, James R. Curran, and Tara Murphy. 2008. Transforming Wikipedia into named entity training data. *In Proceedings of the Australasian Language Technology Association Workshop 2008, pages 124-132.*

Kentaro Torisawa. 2007. Exploiting Wikipedia as external knowledge for named entity recognition. *In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL),* pages 698-707.

Nasrin Taghizadeh, Zeinab Borhani-fard, Melika Golestani-Pour, Mojgan Farhoodi, Maryam

Mahmoudi, Masoumeh Azimzadeh and Heshaam Faili. 2019. *Named Entity Recognition (NER) in Farsi. In Proceedings of the first International Workshop on NLP Solutions for Under Resourced Languages, Trento, Italy.*

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. *Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. Artificial intelligence 165(1):91-134.*

Richard Tzong-Han Tsai, Shih-Hung Wu, Wen-Chi Chou, Yu-Chun Lin, Ding He, Jieh Hsiang, Ting-Yi Sung, and Wen-Lian Hsu. *2006. Various criteria in the evaluation of biomedical named entity recognition. BMC bioinformatics 7, no. 1: 92.*

Salar Mohtaj, Behnam Roshanfekr, Atefeh Zafarian, and Habibollah Asghari. 2018. Parsivar: A language processing toolkit for persian. *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 708-716.*