Mongolian Named Entity Recognition System with Rich Features

Weihua Wang, Feilong Bao*, Guanglai Gao
College of Computer Science, Inner Mongolia University
Huhhot, China, 010021
Email: *wangweihuacs@163.com*;{csfeilong,csggl}@imu.edu.cn

Abstract

In this paper, we first build a manually annotated named entity corpus of Mongolian. Then, we propose three morphological processing methods and study comprehensive features, including syllable features, lexical features, context features, morphological features and semantic features in Mongolian named entity recognition. Moreover, we also evaluate the influence of word cluster features on the system and combine all features together eventually. The experimental result shows that segmenting each suffix into an individual token achieves better results than deleting suffixes or using the suffixes as feature. The system based on segmenting suffixes with all proposed features yields benchmark result of F-measure=84.65 on this corpus.

1 Introduction

Named Entity Recognition (NER) is a natural language processing (NLP) task that consists of finding names in an open domain text and classifying them among several predefined categories such as person, organization and location. It is an important tool in almost all NLP application areas, such as Question Answering, Machine Translation (Chen et al., 2013), Social Media Analysis, Semantic Search or Automatic Summarization.

Since the MUC (Sundheim, 1995) and CoNLL (Sang, 2002) conferences, NER has drawn more and more attention in NLP community. Many NER systems have been developed for English and other language (Ratinov and Roth, 2009; Benajiba et al., 2010; Kravalová and Žabokrtskỳ, 2009). Machine learning based approach have been the predominant in these systems to achieve state-of-the-art results (Radford et al., 2015). As one of them, Conditional Random Fields (CRF) (Lafferty et al., 2001) was proved to an efficient classifier for NER.

Recently, there are more and more concern about how to incorporate more latent sematic features into the NER system (Konkol et al., 2015). Therefore, word cluster IDs function as a non local feature to improve the performance of NER system (Turian et al., 2010; Zirikly and Diab, 2015).

Mongolian is a widely spread language in the world. It is called classical Mongolian in China and called Cyrillic Mongolian in Mongolia and Russia. The classical Mongolian uses Uighur-script, while Cyrillic Mongolian uses Cyrillic-script. In this paper, we address the problem of NER for classical Mongolian.

Compared with other languages, the research on Mongolian NER is still at its initial stage and many issues in Mongolian NER remain unsolved. As far as we know, there has been very little work in the area of NER in Mongolian. Tong (2013) only investigated Mongolian person name recognition. There is still no work publicly reported on recognition of Mongolian location and organization name. Moreover, there are no public available resources and tools for Mongolian NER.

However, proper identification and classification of named entities are very crucial in Mongolian information processing. Therefore, we propose a framework to develop resources and several methods for

*Corresponding author

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 505–512, Osaka, Japan, December 11-17 2016.

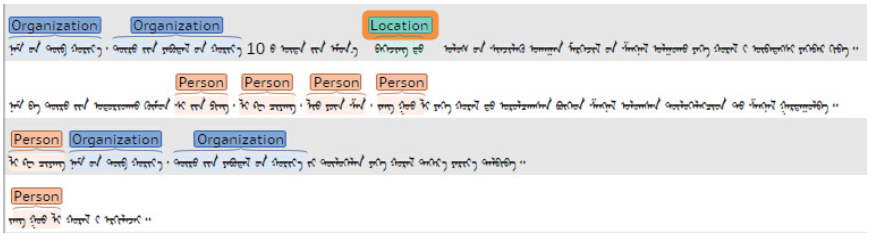

Figure 1: The annotation platform for Mongolian NER.

Mongolian named entity recognition. This paper introduces the work on Mongolian NER that is still in progress.

As one of agglutinative languages, Mongolian has complex morphological structures. We explore different morphological processing methods to alleviate data spareness. Different from the work (Şeker and şen Eryiğit, 2012), we separates Mongolian suffixes from the words and even delete the suffixes. We also propose rich features by exploiting Mongolian orthographic feature, morphological feature, syntactic feature and semantic feature.

The remainder of this paper is organized as follows: Section 2 introduces the construction of Mongolian NER resources; Section 3 presents three morphological processing methods; Section 4 introduces the language independent and language specific features we used; Section 5 describes the results of experiments; Section 6 concludes the paper and summaries some future work.

2 Construction of Mongolian NER resources

2.1 Characteristics of Mongolian

Mongolian writes from top to bottom, and the same letter has different presentation forms decided by position in the word. The NER task of Mongolian was difficult due to the following reasons:

-*Mongolian has large scale vocabulary*: Mongolian has complex morphological structures that each root can be followed by several suffixes to formulate new words. So the larger vocabulary decreased the performance of Mongolian NER system.

-*Absence of capital letters in the orthography*: In English and other Latin language, the proper names always appear with capitalized letter, but there is no concept of capitalization in Mongolian.

-*Multi-category word is very common to named entities*: many common nouns, adjectives and verbs can act as person names or location names, such as an adjective word "" (means "clever") is a very common person name in Mongolian.

-*Subject-Object-Verb word order*: boundaries between named entities are easy to confuse when the subject and object are both proper names.

2.2 Corpus

Nowadays, there is no public annotated corpus about Mongolian named entities. In this paper, we firstly created a new corpus gathered from several Mongolian news web site. We extract mainly content for every web page by analysing the character of each web page html tags. The content of this corpus includes political news, economic news, cultural news and daily news.

This corpus contains 33209 sentences, 59562 named entities and 119M tokens. It annotated manually with person, location and organization by a Mongolian native speaker under the open source platform "Brat" (Stenetorp et al., 2012). The interface of this platform shows in Figure 1. The annotation task cost about three months. At beginning, we discuss almost every sentence to guarantee the quality of annotation. It is time consuming but worth to be done. After twenty days then the annotation become more quick and more unambiguous.

This corpus was converted into BIO2 (Kudo and Matsumoto, 2001) label format. A token is labeled as "B-label" if the token is the beginning of a named entity, and labeled as "I-label" if it is inside a named entity but not the first token within the named entity, others will "O". So there are seven types, that is

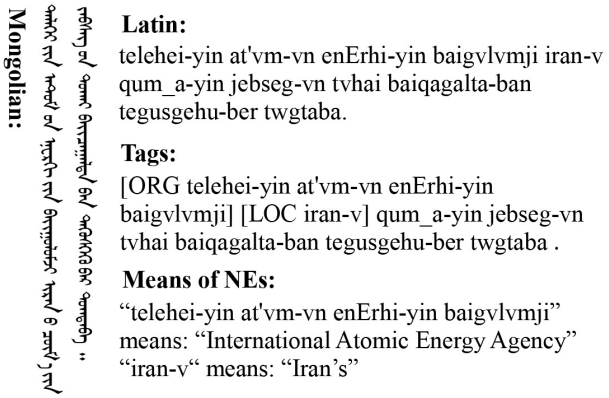

Figure 2: Example of Mongolian NNBS suffixes and named entity tags

"B-PER", "I-PER", "B-LOC", "I-LOC", "B-ORG", "I-ORG" and "O", will be classified by learning algorithm.

The average length of named entities is 2.87 words in our corpus. The person, location and organization entities account for 20.74%, 47.62% and 31.64%, respectively. There are 55% organization entities length are above three words. Mongolian person name always express in one word. However, when transliterating Chinese person into Mongolian, the person name length unchanged. So about 39% person names are three words and 33% person names are only one word.

3 Approach

3.1 Model

CRF is a probabilistic framework that suitable for labeling input sequence data (Lafferty et al., 2001). For an input sequence $X = x_1, x_2...x_n$, CRF model aims to find the best named entity label sequence $Y = y_1, y_2...y_n$ that maximizes the conditional probability $p(y|x)$ among all possible tag sequences. The $p(y|x)$ can be expressed as:

$$p(y|x) = Z(x)^{-1} exp(\sum_t \sum_k \lambda_k f_k(y_{t-1}, y_t, x) \quad (1)$$

where λ_i represents the weight assigned to different features and $Z(x)$ is the normalizing function, it can be defined as:

$$Z(x) = \sum_{y \epsilon Y} exp(\sum_t \sum_k \lambda_k f_k(y_{t-1}, y_t, x) \quad (2)$$

$f_k(y_{t-1}, y_t, x)$ is the binary feature function, such as

$$f_k(y_{t-1}, y_t, x) = 1(y_{t-1} = y', y_t = y, x_t = x) \quad (3)$$

3.2 Morphological processing

For the morphological structure of Mongolian, a Mongolian words can be decomposed into roots, derivational suffixes and inflectional suffixes. As for nouns, the inflectional suffixes contain case suffixes, reflexive suffixes and plural suffixes. All the case suffixes, reflexive suffixes and partly plural suffixes connected to stem through a Narrow Non-Break Space (NNBS) (U+202F, Latin:"-"), so we called them NNBS suffixes. For example, in Figure 2, there are 8 NNBS suffixes in one sentence, and the suffixes appeared inside or beside the named entities.

The NNBS suffixes are used very flexible that each stem can add several NNBS suffixes to change the word form. What's more, the NNBS suffixes in Mongolian can be located unambiguously, while other suffixes segmented may lead to some letters insertion, lost and substitution. Therefore, we proposed three methods to process NNBS suffixes.

RE:Remove all NNBS suffixes in text. After this processing, the sentence in Figure 2 will be "*telehei at'vm enErhi baigvlvmji iran qum_a jebseg tvhai baiqagalta tegusgehu twgtab*."

FE:Take NNBS suffixes as a new feature and replace word with stem. After this processing, sentence length remain unchanged but the feature dimension will add one.

SE:Segment NNBS suffixes as a new token. After then, the sentence will longer, for example, the sentence in Figure 2 will be turned into "*telehei -yin at'vm -vn enErhi -yin baigvlvmji iran -v qum_a -yin jebseg -vn tvhai baiqagalta -ban tegusgehu -ber twgtaba*".

If the suffixes are the last tokens in the entity, we will remove the suffixes from the entity. Because this kind of suffixes only add the syntax function for previous stem. For example, the tag of "iran-v" will change to "*[LOC iran] -v*".

4 Features

Supervised NER is sensitive to the selection of features, we consider the following feature sets for Mongolian. In the following experiment, we fixed all features window at [-1,1], that means take the previous feature, current feature and next feature into consideration, except the contextual feature.

Contextual Feature (CXT): this feature was automatically generated, and mean to combine the current and previous output tokens.

Orthographic Feature (ORT): this feature defined the lexical orthographic nature of the tokens in the text, which means the n-gram of tokens. If the suffixes were split, the n-gram tokens will include suffixes directly.

Syllable Feature (SYN): this feature contained syllable count, first and end syllable of the current token.

Syllable count: we concluded 28 rules about counting Mongolian syllables for the first time, according to Mongolian grammar. In general, too many syllables might not be names.

First and end syllable: some first syllables or end syllables occur frequently in many Mongolian person names.

Look up feature: defined as binary features and matched exactly with the lookup table.

Gazetteers (GAZ): this collected gazetteer consist of 8735 location names and 2731 person names. We extracted location names from Mongolian Chinese dictionary mainly contained Inner Mongolian location names manually. The person names list found in few Mongolian blogging web sites, and mainly contained Mongolian names.

Transliteration table (TRS): this table contained 564 Mongolian borrowed words from Chinese. For example, a very common surname in Chinese "王" ("wang").

Person title and job title list(TIT): this list contained 373 person title entries and 582 job title entries.

Morphological Feature: this feature explored rich morphological structure of Mongolian.

Part-of-speech (POS): we employed a rule and dictionary based POS tagger to produce this features. This top level POS marking set include 15 classes which according with (China Standard, 2011). When SE method applied, the POS feature of NNBS suffixes will be denoted by "F".

NNBS suffixes: used as feature only when the FE method applied. If a word contains NNBS suffixes, the suffixes themselves will be referred as features.

Word Clusters IDs: this feature gained from massive unlabeled corpus after the same SE method preprocessed. The corpus used also crawled from web sites in a more wider range. It contained 337M sentences and its token size and vocabulary showed in Table 1. From Table 1, we found the vocabulary decrease 27% while the token number growth 22%.

Word2vec clusters IDs (W2V): this feature achieved by performing K-means clustering on word2vec vectors in (Mikolov et al., 2013) and directly used the cluster IDs as features. The vectors' dimension in our experiments are 200, the minimum occurrences number of token is 3 and the context window fixed at 8. We retrained word vectors with negative sampling used skip-gram model. A new cluster number assigned to the test token without trained cluster ID.

LDA word classes (LDA): we followed the work in (Chrupala, 2011) to induce LDA to produce different word clusters with the minimum occurrences number of token is 3. If the test tokens are out of

Table 1: Vocabulary decrease after processing by SE method in cluster training corpus (mincount=3)

	Vocabulary	**Tokens number**
Word model	395511	72051575
SE model	285063	88021157

Table 2: Results of different morphological processing methods

Feats.	**Baseline**			**RE**			**FE**			**SE**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
+ORT	84.50	77.05	80.65	84.57	79.50	81.96	84.57	79.50	81.96	84.61	80.07	82.28
+POS	84.70	78.71	81.59	85.09	81.10	83.05	85.11	81.25	83.13	85.11	81.67	**83.35**
+GAZ	84.88	79.49	82.10	85.20	81.95	83.55	85.20	82.28	83.62	85.21	82.35	83.75
+TRS	84.97	79.60	82.20	85.56	82.25	83.87	85.57	82.15	83.83	85.30	82.34	83.79
+SYN	85.02	81.04	82.98	85.11	82.63	83.81	85.18	82.63	83.88	85.07	83.16	84.10
+TIT	85.01	81.14	83.03	85.01	82.42	83.69	85.30	82.71	83.99	85.28	83.32	**84.29**

vocabulary of trained LDA word, we also assigned a new LDA classes number for them.

5 Experiment

In our experiments, we analyzed the impact of various morphological processing and various categories features under an CRF framework with the same parameters. All the experiments carried on 5-fold cross-validation, the proportion of train and test set is 80% , 20%. We evaluated the results by the CoNLL metrics of precision, recall and F-measure.

Precision, means the percentage of corrected named entities (NEs) found by the classifier. It can be expressed as:

$$precsion = \frac{Num(correct\ NEs\ predicted)}{Num(NEs\ predicted)} \tag{4}$$

Recall is the percentage of NEs existing in the corpus and which were found by the system. It can be expressed as:

$$recall = \frac{Num(correct\ NEs\ predicted)}{Num(all\ NEs)} \tag{5}$$

F_1 is the harmonic mean of precision and recall. It can be expressed as:

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{6}$$

5.1 Impact of morphological processing

Firstly, we incrementally added features to the three methods mentioned above and each feature window fixed at [-1,1]. The results show in Table 2, the "Baseline" means without any morphological processing. From Table 2 we found that all the three methods can improve the overall performance. When incorporated all features, the SE method achieved the best and improved F-measure by 1.26. Because segmenting by NNBS can decrease the percentage of unknown word in sentences and do help to detect named entities. The features played more effect roles when the suffixes took apart.

Each feature improves the performance based on the former one. The F-measure improvement caused by POS feature is obvious. The contribution of GAZ feature lies in improving both the precision and recall of person names and location names. The TRS feature improves the F-measure because the corpus contains amount of Chinese person and location names. The SYN feature slightly indicates some cues for named entities, the first and end syllable act on prefixes or suffixes. The TIT feature benefits the person names recognition to improve overall F-measure.

Table 3: Results (in F-measure) of different semantic space of LDA word classes and word2vec clusters

Clusters numbers	LDA	Word2vec
50	83.16	83.27
100	83.20	**83.49**
200	**83.26**	83.26
500	83.24	–

Table 4: Results of different LDA word classes and word2vec clusters combination

Clusters IDs combination	F1
LDA200+W2V100	**83.53**
LDA500+W2V200	83.29
LDA100+LDA200	83.13
W2V100+W2V200	83.29
All Cluster IDs	83.24
SE+ORT	82.28

5.2 Impact of word cluster features

Secondly, we evaluate the impact of semantic features, that is, only adding LDA word cluster or word2vec cluster features onto the SE method, Table 3 shows the results. The F-measure varies with cluster number and cluster type, the more word classes does not always mean better performance. The best cluster number is 200 for LDA word cluster and 100 for word2vec cluster. Word2vec clusters outperform the LDA word cluster because that it can induce more context to cluster.

We then combined the best and second performance in Table 3 without other features to produce the best word cluster combination. Table 4 shows the results. In Table 4, SE+ORT means only using context feature with SE method, as baseline system. The F-measure reaches 83.53 when coupled with LDA200 and W2V100. This best F-measure even surpassed the performance of POS and ORT feature combination about F_1 is 83.35 under SE method in the same condition. However, the overall performance reduced when added all type clusters features to the feature set.

5.3 Final system

Finally, we integrate all features including traditional features and word cluster features in SE method, Table 5 shows the final system performance.

In Table 5, AFH represents all the handcraft features, including ORT, POS, GAZ, TRS, SYN and TIT. From Table 5, we find that the same word cluster feature works different when combine with traditional features. With only single word cluster, the effect is weak, but when we use the combination of AFH, LDA100 and LDA200, result reach the best. It outperforms the handcraft features 0.36 in F_1. As the results shown, combining more features does not mean higher performance.

6 Conclusion

In this paper, we built a Mongolian named entity recognition corpus and explored three morphological processing methods with different features combination under the CRF framework. This is the first corpus for Mongolian and we carry on experiment on this corpus. The experimental results show that the proposed methods can alleviate the sparseness of data and improve the performance of Mongolian NER system. In addition, the word cluster features represent the latent semantic of word can also benefit the system. Among the above three methods, treating NNBS suffixes as individual token perform best. It can reach F-measure at 84.65 when combined all features including handcraft features and word cluster features. This work can also provide benchmark system to promote the future Mongolian NER research community.

In the future, we will try our method to other agglutinative languages and expand the work on using word cluster feature. We will also try to use deep neural network to perform to Mongolian NER.

Table 5: Final performance combined all features

Features	F1
AHF	84.29
AHF+LDA50	84.21
AHF+LDA100	84.33
AHF+LDA200	84.36
AHF+W2V50	84.34
AHF+W2V100	84.48
AHF+W2V200	84.54
AHF+LDA100+LDA200	**84.65**
AHF+LDA200+W2V100	84.57
AHF+LDA200+W2V200	84.54
AHF+LDA100+LDA200+W2V100	84.57
AHF+LDA100+LDA200+W2V100+W2V100	84.36

Acknowledgements

This research is partially supported by the China National Nature Science Foundation (No.61263037, No.61303165 and No.61563040), Inner Mongolia Nature Science Foundation (No.2014BS0604 and No.2016ZD06) and the program of high-level talents of Inner Mongolia University. Finally, we thank the anonymous reviews for their many helpful comments.

References

Yassine Benajiba, Imed Zitouni, Mona Diab, and Paolo Rosso. 2010. Arabic named entity recognition: using features extracted from noisy data. In *Proceedings of the ACL 2010 conference short papers*, pages 281–285. Association for Computational Linguistics.

Yufeng Chen, Chengqing Zong, and Keh-Yih Su. 2013. A joint model to identify and align bilingual named entities. *Computational Linguistics*, 39(2):229–266.

GB26235-2010 China Standard. 2011. *GB26235-2010 Information technology-Mongolian word and expression marks for information processing*. China National Standardization Technical Committee.

Grzegorz Chrupala. 2011. Efficient induction of probabilistic word classes with LDA. In *Fifth International Joint Conference on Natural Language Processing, IJCNLP 2011, Chiang Mai, Thailand, November 8-13, 2011*, pages 363–372.

Gökhan Akın Şeker and Gül şen Eryiğit. 2012. Initial explorations on using crfs for turkish named entity recognition. In *Proceedings of the 24th International Conference on Computational Linguistics, COLING 2012.*, Mumbai, India, 8-15 December.

Michal Konkol, Tomas Brychcin, and Miloslav Konopík. 2015. Latent semantics in named entity recognition. *Expert Syst. Appl.*, 42(7):3470–3479.

Jana Kravalová and Zdeněk Žabokrtskỳ. 2009. Czech named entity corpus and svm-based recognizer. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, pages 194–201. Association for Computational Linguistics.

Taku Kudo and Yuji Matsumoto. 2001. Chunking with support vector machines. In *Proceedings of the 2001 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 746–751.

Will Radford, Xavier Carreras, and James Henderson. 2015. Named entity recognition with document-specific KB tag gazetteers. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 512–517.

Lev-Arie Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning, CoNLL 2009, Boulder, Colorado, USA, June 4-5, 2009*, pages 147–155.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning, CoNLL 2002, Held in cooperation with COLING 2002, Taipei, Taiwan, 2002*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

Beth Sundheim. 1995. Overview of results of the MUC-6 evaluation. In *Proceedings of the 6th Conference on Message Understanding, MUC 1995, Columbia, Maryland, USA, November 6-8, 1995*, pages 13–31.

Gala Tong. 2013. *Automatic Recognition of Mongolian Names Based on Corpus*. Ph.D. thesis, Ming Zu University of China.

Joseph P. Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 384–394.

Ayah Zirikly and Mona Diab. 2015. Named entity recognition for arabic social media. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 176–185, Denver, Colorado, June. Association for Computational Linguistics.