УДК 81`32

NAMED ENTITY RECOGNITION FOR KAZAKH USING CONDITIONAL RANDOM FIELDS

Tolegen Gulmira, Kazakhstan, 100000, c. Astana, *National Laboratory Astana*, gtolegen13@fudan.edu.cn
Toleu Alymzhan, Kazakhstan, 100000, c. Astana, *National Laboratory Astana*, atoleu13@fudan.edu.cn
Xiaoqing Zheng, China, 200433, c. Shanghai, *Fudan University*, zhengxq@fudan.edu.cn

We addressed the Named Entity Recognition (NER) problem for the Kazakh language by using conditional random fields. Kazakh is a typical agglutinative language in which thousands of words could be generated by adding prefixes and suffixes to the same root, which arises a serious data sparsity problem for many NLP tasks. To reduce the data sparsity problem, a necessary preprocessing step is to split the words into their roots and morphemes by morphological analysis. In this study, we designed a CRF-based NER system for Kazakh, which leveraged the features derived from the results of a new-developed morphological analyzer, and found that the performance can be boosted by introducing such derived features. Moreover, we assembled a NER corpus which was manually annotated with location, organization and person names.

Keywords: Kazakh language, agglutinative language, named entity, NER, CRF

ИЗВЛЕЧЕНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ ИЗ ТЕКСТА НА КАЗАХСКОМ ЯЗЫКЕ С ИСПОЛЬЗОВАНИЕМ УСЛОВНЫХ СЛУЧАЙНЫХ ПОЛЕЙ

Толеген Гулмира, Казахстан, 100000, г. Астана, *Национальная Лаборатория*, gtolegen13@fudan.edu.cn
Толеу Алымжан, Казахстан, 100000, г. Астана, *Национальная Лаборатория*, atoleu13@fudan.edu.cn
Xiaoqing Zheng, КНР, 200433, г. Шанхай, *Фудан Университет*, zhengxq@fudan.edu.cn

Мы обратились к проблеме извлечения именованных сущностей (Named Entity Recognition, NER) для казахского языка, используя условные случайные поля (Conditional Random Fields, CRF). Казахский – типичный агглютинативный язык, в котором тысячи слов могли бы быть получены путем добавления префиксов и суффиксов к тому же корню, что создает серьезную проблему разреженности данных для решения многих задач NLP. Чтобы уменьшить проблему разреженности данных, необходим шаг предварительной обработки для разделения слов в их корни и морфемы путем морфологического анализа. В данном исследовании мы разработали CRF-систему NER для

казахского языка, которая использовала возможности, полученные из результатов новоразвитого морфологического анализатора, и обнаружили, что производительность может быть повышена путем внедрения таких производных функций. Кроме того, мы собрали NER корпус, который был вручную аннотированный с названиями мест, организаций и лиц.

Ключевые слова: Казахский язык, агглютинативный язык, именованные сущности, NER, CRF

1 Introduction

Named Entity Recognition is an important task in many Natural Language Processing (NLP) applications nowadays. State-of-the-art NER systems have been produced for several languages, but despite these recent improvements, developing an NER system for Kazakh is still a challenging task due to the structure of the language. Kazakh is the official state language of Kazakhstan. It is an agglutinative and highly inflected language. To illustrate this, consider the following English phrase "*people who live in the city*" which can be translated into Kazakh with only one word "*қаладағылардың*" which can decompose into the root and additional suffixes: "*қала+да+ғы+лар+дың*".

This productive nature of Kazakh results in a single root which may produce hundreds or thousands of word forms, thus causing the data sparsity problem. In order to prevent this behavior in our NER system, we build a Finite State Transducer (FST)-based morphological analyzer, which is able to decompose complex words into their constituent root and morphemes. We also developed a perceptron algorithm based morphological disambiguation system.

In this paper, a CRF-based system used to extract named entities from Kazakh text is presented. In order to obtain accurate generalization, we used several syntactic and semantic features of the text, including: more fine-grained morphological features and word type information. The morphological features are important for Kazakh; these features are extracted by morphological analyzer, and effectively used in this study to increase the recognition performance. We evaluated our models on Kazakh NER corpus (KNC) that we have annotated with location, person and organization names.

The rest of the paper is organized as follows: Section 2 gives brief information about FST-based morphological analyzer and describes features used in this work. Section 3 gives detailed information about the corpus of named entity and reports the results of experiments. Section 4 reviews the existing work. Section 5 concludes with possible future work.

2 The CRF-based NER system

The CRF-based state-of-art NER systems for agglutinative languages often use a large feature set to improve system performance (Yeniterzi, 2011; Şeker and Eryiğit, 2012). As described before, Kazakh words may contain different morphemes to determine their meaning, and these morphemes can be viewed as a set of morphological features. In order to extract such features and prevent data

sparseness problems, developing a morphological analyzer and disambiguator will be a crucial step for Kazakh NER.

2.1 Morphological analysis and disambiguation

Morphological analysis is a problem of breaking word such as "*болмесін*" (*don't let him split sth.*) into the stem and morphemes, *бол* (*split*) and *-ме -сін*. These features specify the additional information about the stem. There are mainly two approaches, FST-based and data-driven approaches. Finite state techniques (Oflazer and Güzey, 1994) are the most suitable technique to represent the morphosyntactic rule of agglutinative languages.

In order to develop a FST-based morphological analyzer, we need three components: 1) a lexicon of roots annotated with some information (part of speech etc.), to determine which morphological rules apply to them, 2) a morphotactic component that describes the word formation by specifying the ordering of morphemes, and 3) Morphophonemics rules description.

For the lexicon list, we collected a new lexicon of 151,463 roots. To compile this lexicon and to ensure the correct spelling of the words, we manually checked and annotated with POS tags. We have considered almost all inflectional suffixes of nouns, verbs, adjectives, pronouns, adverbs, e.g. plural, case, possession, predication, tense, mood etc. We edited 3,039 word-formation rules and developed an FST-based morphological analyzer for Kazakh. The morphological analyzer achieved 95% coverage on news corpus (800K words), and 91% on Kazakh's Wikipedia (10M words).

In order to select the most probable analysis of the words, depending on the context, we developed a perceptron algorithms (Collins, 2002; Sak et al., 2008) based morphological disambiguator and manually disambiguated a corpus (15K words) for model training, and the best model achieved 90.54% accuracy on the test data. At the next stage, we use the information coming from the raw data and the disambiguated morphological analysis to extract named entity features.

2.2 Feature categories

We generalize named entities (NEs) by using a set of features that are capable of describing various properties of the text. In this study, these features can be grouped into two categories: morphological and word type information. In morphological features, since the part-of-speech of NEs will always be a noun, instead of using morphological features as sequence morpheme, we extracted more useful morpheme tags that attached to the noun from inflectional and derivational morphemes, and used theses as atomic units, and then each feature can be combined with other features effectively.

Morphological features:

- **Root Feature (RF)**: the root of word as a feature. Although the lexicon of morphological analyzer may not contain all the proper nouns and cannot analyze some proper nouns, but the root of the surrounding words of the proper nouns has an influence on the entity recognition.

- **Part of speech (POS)**: the Part-of-speech tag of the root as a feature.
- **Inflectional suffixes (NC, PL, PS, and PR)**: these four features that are extracted from all inflectional suffixes, the feature *NC* includes: Nominative, Genitive, Locative, Accusative, Dative, Ablative, and Instrumental suffixes; PL - plural; *PS* - possessive; *PR* - predication.
- **Derivational suffixes (SP, SN)**: these two features extracted from derivational suffixes.
- **Proper Noun (NP)**: this feature is stated as "1" when the selected morphological analysis includes tag "np".
- **Name suffixes (NS)**: some suffixes most used in the Kazakh surname formations such as (*+ев, +ов, +ин, +ева, +ова, +ина, +улы, +қызы*).

Word type features:

- **Latin words (LW)**: Latin spelling words.
- **Acronym (AC)**: this should help to identify organizations and persons abbreviated as acronyms.
- **Case Feature (CF)**: the information about lowercase and uppercase letters used in the current word.
- **Start of the Sentence (SS)**: this feature indicates whether or not the current token represents the beginning of a sentence.

In this work, we provided atomic features within a window of {-4, 4}, and the feature template are designed by using wrapper methods (Kohavi, 1997). NER is typically treated as a sequence labeling task. We utilized CRF (Lafferty et al., 2001), which provides advantages over other statistical approaches such as the Hidden Markov Model and enables the use of any number of features.

2.3 The CRF model

Let $X = \{x_1, x_2, \dots, x_n\}$ be the sequence of words in sentence, and $Y = \{y_1, y_2, \dots, y_n\}$ the hidden labels. A linear chain Conditional Random Field defines a conditional probability

$$P(Y \mid X) = \frac{1}{Z} exp\left(\sum_{i}^{N}\sum_{j}^{K} \lambda_j f_j(y_{i-1}, y_i, X, i)\right) \quad (1)$$

where λ_j is the weight for feature f_j, and must be learned, the scalar Z is the normalization factor. For NER task, each y_i is named entity label and f_j is feature function which produces a real binary value. Training involves finding the λ_j and maximize the conditional log-likelihood of the training data $\mathcal{D}$

$$\sum_{(X,Y) \in \mathcal{D}} \log p(Y|X, \lambda) \quad (2)$$

In this work, we used CRF++[1], which is an open source CRF sequence

[1] CRF++: Yet Another CRF toolkit.

labeling toolkit.

3 Experiments

We have conducted several sets of experiments to explore the effectiveness of features on NER task for Kazakh. We used the CRF and examined the cumulative contribution of the features. For evaluation, we use the Kazakh NER corpus (Section 3.1). This corpus is divided into training (80%), development (10%) and test (10%) set. The development set was used for choosing hyper-parameters and model selection. We adopted IOB tagging scheme (Tjong Kim Sang, 2002) for all experiments. For evaluation, we used the *conlleval*[2] evaluation script to report the F1-score, precision and recall values.

3.1 Corpus Construction

A major obstacle to Kazakh NER is the scarcity of publicly available annotated corpora. We created a corpus by manually annotating the text of 2500 articles from general news media[3]. These articles were randomly selected from all articles. Only body text was extracted from the chosen articles for inclusion in the corpus. The annotations have been executed manually by native speakers of Kazakh. In order to assist the annotation process and to improve the correctness of the annotation, we have developed a web-based annotation tool. We followed the MUC-7 NE task definition (Chinchor, 1998) as a guideline for annotations.

The current corpus was annotated with person, location and organization names and the annotations were double-checked. We have measured the inter-annotator agreement (IAA) that was calculated using Fleiss' kappa. We randomly sampled 500 articles for this purpose. Each article was annotated by two annotators. The IAA is 0.93 without following guidelines or discussing difficult cases with each other. After discussions, the IAA achieved over 0.98.

Web text often contains errors. In this corpus, we did not correct grammatical and typographical errors, as we wished that the corpus remains as similar as possible to the source text. The final corpus was created by correcting annotation mistakes. This corpus consists of 18,054 sentences and 270,306 words with 4,292 person names, 7,391 location names, and 2,560 organization names.

3.2 Experimental Result

In CRF model, we adopted wrapper methods (Kohavi, 1997) to feature selection and tuned the feature template on the development set[4]. After find the best feature template, we grouped these features combinations into different classes, and each class is related to each feature category as described in Section 2.2.

In order to explore the contribution of each feature, we first evaluated the baseline (Word) performance of a CRF model, in which the tokens are only used in their surface form, and then added each feature combination to it. The results are evaluated with respect to the CoNLL metric and shown in Table 1. A plus (+) sign

[2] www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt
[3] Available at: http://www.inform.kz
[4] These experiments are not added to the paper due to the space constraints.

before the feature name indicates that these feature combinations[5] are added on top of the rest with suitable feature templates.

The system achieved a 69.91% F1 using only word surface form. The F1 is improved by +4.57% when using the root feature *(+RF)* with Word. This result indicates that the root of the surrounding words of the proper nouns have an enormously positive impact on system performance. The F1 is improved by+6.34% when using *+POS* and improved by +3.51% when using morphological features such as *+NC, +PL, +PS, +PR*, which are extracted from all inflectional suffixes. As we can see, the inflectional suffix features not bring a significant improvement for organization names due to the complex of structure organization names, such as long NEs, mixed with other language and abbreviation etc. Then the feature *(+PS)* may hurt system performance. The morphological features such as *+SP, +SN* are extracted from derivational suffixes, these features also improves (+0.54%) system performance. These results show that using morphological features can significantly improves the NER from Kazakh texts due to the agglutinative nature of the language. This also indicates there is still room for improvement using more fine grained usage of these inflectional and derivational suffixes. Using name suffixes *(+NS)*, the person F1 improved by +2.94%. As shown, the feature *(+LW)* may hurt system performance (-0.31%). The F1 improved by +0.18% using *+AC*. Since only the proper nouns and the initial words of the sentences start with a capital letter, the case features *(+CF)* improved the system significantly for all labels.

	Development set				**Test set**			
Feature	**LOC**	**ORG**	**PER**	**Overall**	**LOC**	**ORG**	**PER**	**Overall**
Word	77.86	61.26	65.19	71.73	77.56	66.67	57.45	69.91
+RF	85.26	66.34	69.21	77.95	82.97	69.52	61.19	74.48
+POS	88.03	70.26	74.64	**81.57**	87.99	76.79	69.65	**80.82**
+NC	88.19	73.10	78.79	83.16	88.58	79.21	74.36	82.81
+PL	88.47	73.02	79.95	83.58	89.55	79.65	76.92	**84.10**
+PS	88.93	74.77	79.51	**83.99**	89.75	77.66	76.62	**83.76**
+PR	89.01	73.42	80.32	84.03	89.94	77.75	78.21	84.33
+SP	89.79	74.89	80.64	84.79	90.47	78.71	77.48	84.57
+SN	89.81	74.67	81.06	84.87	90.84	77.59	78.52	84.87
+NP	90.77	74.27	83.03	85.93	90.53	77.49	80.66	85.38
+NS	90.65	74.94	**84.11**	86.26	90.85	77.42	**83.60**	86.37
+LW	90.83	74.89	84.55	86.46	90.35	77.16	83.60	86.06
+AC	90.70	75.16	84.86	86.51	90.64	77.49	83.46	86.24
+CF	93.15	78.49	91.89	**90.46**	91.45	82.38	89.83	**89.43**
+SS	93.15	78.59	91.91	**90.47**	**91.71**	**83.40**	**90.06**	**89.81**

Table 1 F1-score of the CRF when each feature is added cumulatively.

4 Related Work

There are large number of studies have been performed on NER for many languages. Here we review the literatures most relevant to this work.

[5] These feature combinations are selected by wrapper methods and related to each feature category.

For Turkish, (Tatar and Cicekli, 2011) proposed an automatic rule learning method for Turkish and achieved an averaged F1-score of 91.08% on the data-set, the experiment result show that morphological features can significantly improve the NER performance. (Yeniterzi, 2011) obtained an F1-score performance of 88.94% by using CRF and exploiting the effect of morphology used inflectional features as tokens. In the same direction (Şeker and Eryiğit, 2012) proposed a successful CRF model for Turkish NER and extracted Proper Noun and Noun Case two features from all inflection suffixes, they report the result (89.55% in CoNLL metric) without using gazetteers on general news text. Few papers have been published in relation to Kazakh NER and this is one of the first systems to perform NER for Kazakh.

5 Conclusion

In this study, we have developed a CRF-based NER system for Kazakh language. Through a set of extensive experiments, the features of the CRF model have been carefully optimized for Kazakh NER. The experimental result shows that the features derived from the results of morphological analysis significantly improve the system's performance (from 69.91% to 89.81% in F1) by alleviating the data sparsity problem brought by the properties of agglutinative languages. Moreover, we created and manually annotated a Kazakh NER corpus (KNC). In the future, we plan to use deep learning approaches for Kazakh NER task.

Acknowledgments

The authors would like to thank Akmaral T., Olzhas S., Nazigul M., Galymzhan T., Duman S. for their help with data annotation. The authors would also like to thank Kakesh Kaiyrzhan for providing Kazakh language resources.

References

1. Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of CoNLL-2002, pages 155-158.
2. John D. Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, pages 282-289.
3. J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, pages 378-382.
4. Kessikbayeva Gulshat and Cicekli Ilyas. 2014. Rule Based Morphological Analyzer of Kazakh Language. In Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM, pages 46-54.
5. Michael Collins. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In Proceedings of EMNLP 2002.
6. N. Chinchor. 1998. MUC-7 named entity task definition, version 3.5 In

Proceedings of the Seventh Message Understanding Conference.
7. Oflazer K., Güzey C. 1994. Spelling correction in agglutinative languages. In Proceedings of the Fourth Conference on Applied Natural Language Processing, ANLP '94, pages 194-195.
8. Rabiner Lawremce R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. In Proceedings of the IEEE, 1989, pages 257-286.
9. Ron Kohavi and George H. John. 1997. Wrappers for Feature Subset Selection. Artificial Intelligence Archive, 97:273C324, 1997.
10. Seker Göhan Akın and Eryiğit Gülsen. Initial Explorations on using CRFs for Turkish Named Entity Recognition. In: Proceedings of COLING. 2012, pp. 2459–2474.
11. Serhan Tatar and Ilyas Cicekli. Automatic rule learning exploiting morphological features for named entity recognition in Turkish. In: Proceedings of the 28th International Symposium on Computer and Information Sciences. 2013, pp. 137–151.
12. Sak, H., Güngör, T., Saraçlar, M. A stochastic finite-state morphological parser for Turkish. In: Proceedings of the ACL-IJCNLP. Conference Short Papers, 2009, pp. 273–276.
13. Tür, G., Hakkani-T¨ur, D., and Oflazer, K. 2003. A statistical information extraction system for turkish. Natural Language Engineering, 9:181C210.
14. Yeniterzi Reyyan. 2011. Exploiting Morphology in Turkish Named Entity Recognition System. In Proceedings of the ACL 2011 Student Session, pages 105–110.