Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

Named Entity Recognition for the Mongolian Language

Zoljargal Munkhjargal, Gábor Bella,
Altangerel Chagnaa, and Fausto Giunchiglia

2015

Technical Report # DISI-17-013

Proceedings of the 18th International Conference on
Text, Speech, and Dialogue, Pilsen, Czech Republic

Named Entity Recognition for Mongolian Language

Zoljargal Munkhjargal[1], Gabor Bella[2], Altangerel Chagnaa[1], and Fausto Giunchiglia[2]

[1] DICS, National University of Mongolia, 14200, Mongolia
{zoljargal,altangerel}@num.edu.mn
[2] DISI, University of Trento, 38100, Italy
gabor.bella@unitn.it, fausto@disi.unitn.it

Abstract. This paper presents a pioneering work on building a Named Entity Recognition system for the Mongolian language, with an agglutinative morphology and a subject-object-verb word order. Our work explores the fittest feature set from a wide range of features and a method that refines machine learning approach using gazetteers with approximate string matching, in an effort for robust handling of out-of-vocabulary words. As well as we tried to apply various existing machine learning methods and find optimal ensemble of classifiers based on genetic algorithm. The classifiers uses different feature representations. The resulting system constitutes the first-ever usable software package for Mongolian NER, while our experimental evaluation will also serve as a much-needed basis of comparison for further research.

Keywords: Mongolian Named Entity Recognition, Genetic Algorithm, Machine Learning, String Matching

1 Introduction

The volume of textual information made available every day exceeds by far the human ability to understand and process it. As a consequence, automated information extraction has become an essential and pervasive task in computing. One particular component of such systems is Named Entity Recognition (NER) that consists of identifying personal names, organization names, and location names within sentences of natural language text. NER is an extensively researched topic. However, In less-studied and resource lack languages such as Mongolian, there is not enough research.

While the general problem of NER has been approached from widely varying perspectives, methods tend to follow 1) dictionary-based, 2) rule-based, 3) stochastic machine learning-based approaches or 4) combinations of these. If applied directly to text, these approaches are not considered robust as they cannot tackle spelling mistakes, orthographic variations, or out-of-vocabulary names, the latter being a very common phenomenon as the set of commonly used named entities (people and things) is open and constantly evolving. It is in great part for these reasons that statistical machine learning-based methods, based on manually pre-annotated text corpora, have become the basis of most NER systems. State-of-the-art results have been obtained using Maximum Entropy [1], Hidden Markov Models, Support Vector Machines [2], and Conditional Random Fields [3].

However, for agglutinative languages such as Mongolian, supervised learning methods tend to produce weaker results. This is due to the morphology, characterized by an almost unbound number of word forms. As a result, machine learning is hindered by frequent occurrences of word forms rarely or never seen during training.

Ensembling several NER classifiers that each one is based on different feature representation and different classification approach improves general performance accuracy [4] [5] [6]. This general improvement depends on a diversity of classifiers that see the NER task from different aspects. However, exploring the fittest-feature set and selecting appropriate classifier for constructing an ensemble classifier are a difficult problem.

This paper describes what we believe to be the first serious attempt at designing a supervised NER system for Mongolian. The system implements an ensemble approach consisting of supervised machine learners with a corresponding new annotated corpus, newly created gazetteers, as well as a simple rule-based matcher. A genetic algorithm is applied to find optimal classifier ensemble. Furthermore, to tackle the out-of-vocabulary word form problem, we took inspiration from studies showing how approximate string distance metrics can be used for robust name-matching tasks [7] [8], for taking into account inflectional variations of names.

2 Mongolian names

Mongolian is an agglutinative language that a word is inflected by rich suffix chains in the verbal and nominal domains. It is often considered part of Altaic language family that includes Turkic languages, Korean and Japanese. As Hungarian is also agglutinative, from a computational linguistic point of view very similar problems need to be solved in the two languages [9].

Most personal names are compounds of two or more simple names or common words (that can themselves serve as names). For example, Ганболд (Ganbold) joins two common nouns: Ган (Gan-steel) and Болд (Bold-alloy).

The full personal name is written in the reverse order with respect to the Western convention: either a *patronymic* or a *matronymic* (roughly equivalent to the surname) comes first and a given name (equivalent to the first name) comes second. In general, the surname is inflected in a genitive case, depending on vowel harmony and on several other morphological factors. Because of the usage of patronymic/matronymic as surnames instead of distinct family names, the order of the surname and the given name is never inversed. The full name also consists of abbreviation of surname, which ends with period, and given name. This is commonly used in newswire domain.

In general, organization and location names are capitalized. For names of international organizations, countries, councils, ministries and associations consisting of several words, all of the words have to be capitalized. For other typical multi-word names—such as industry branches, provinces, districts, scientific and cultural organizations—the head word is capitalized and the others remain in lowercase.

If a proper name of any kind is a composite of two names and the second name starts with a vowel, a hyphen is placed between two names and the second word is capitalized: Баруун-Урт (Baruun-Urt: city name).

Another feature that makes NER more difficult for Mongolian is the subject-object-verb word order: boundaries between named entities are easy to miss when the subject and the object are both proper names.

3 The NER system

A range of machine learning algorithms are successfully applied to the task of NER. To effectively solve classification problem, we hypothesized that ensemble of a diverse set of classifiers would benefit the performance, assuming different methods lies in tackling the problem from different angles. Depending on the various feature combinations, a classifier produces various results, too. Thus we tried to find the most optimal combination of features sets using a brute force method. The building blocks of our system are shown in Fig 1.

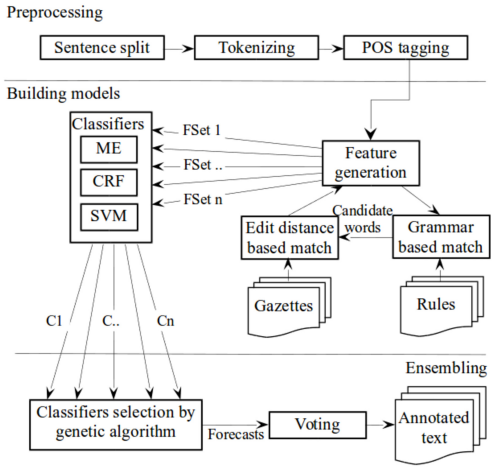

Fig. 1. Outline of building complex NER model. FSet: feature set that is a subset of 5 group feature sets; C: classifier that trained on particular subset.

3.1 Preprocessing

From the point of NER, feature of context/trigger words and sentence position is one of common clues to determine NEs. Thus we should involve a sentence detector. The Mongolian proper name writing rules such as surname abbreviation and hyphenation is leading to consider one word as two or more tokens. Further, we hope that tokenization before classification is more suitable for the sequence of instance (feature vector of token) tracking algorithms (ME, SVM, CRF, HMM). As well as part-of-speech (POS) tag is commonly used for a statistical NE classifier as a feature [10] [11].

3.2 Feature generation

We experimented with a rich set of features, describing the characteristic of the token with its context (a moving window of size five), many of which are used in related works [5], [6], [12] and [9].

1. **Orthographic properties of the word form**: is first letter capitalized, is entire word uppercased, is entire word lowercased, does it contain any hyphen, does token only consist of punctuation marks, does token contain at least one punctuation mark except hyphen and word length.
2. **Word Shape**: long pattern (maps all uppercase characters to "X", all lowercase ones to "x", and the rest to "_"), and short pattern (consecutive character types are not repeated: "X_x") and a symbolic feature outputting one of following labels: allLowerCase, allCaps, firstCap, capPeriod, onlyDigit, onlyPunct, hyphened or other.
3. **Affix information**: the first 4 characters of token and character 3-grams. We validated 3-, 4- and 5-character prefixes one by one, the 4-character prefix reaches best result, as well as 2-, 3- and 4-grams are tested, 3-gram gives the best result.
4. **Morphological and Contextual information**: full part-of-speech (POS) tag, high and low-level POS tag (with and without information about inflection, resp.), position in the sentence (start, mid or end) and is the word between quotes.
5. **Gazetteer information**: if the token is included in one of the gazetteers, it receives a feature containing the name of the category.

Feature set selection. To measure the strength of the above five groups of features we trained all of classifiers, which we involved, for all possible sub set of the five groups (31 models per classifier). Between 15 and 20 best performing models achieved very similar results better than the others, in each classifier. We used the top 5 models of each classifier, and then recombined the models in a voting scheme.

Grammar- and Edit Distance-Based Matching. To complement the statistical classifier by making use of existing Mongolian name resources, we implemented a simple pattern-based and a gazetteer-based recognizer. The former is a regex-based matcher using simple grammatical rules while the latter uses fuzzy string matching on name lists.

A simple rule set representing an intentionally rough (and, by consequence, easy-to-implement and fast-to-apply) grammatical model of Mongolian proper names is created. The rules, given below, are optimized for recall rather than precision, since their main purpose is to extract candidate named entities that will be further verified using string distance metrics and gazetteers.

1. Personal name: a) capitalized word in genitive + capitalized word; b) capitalized word + period + capitalized word.
2. Organization name: a) sequence of capitalized words + organization designator (e.g., ХХК “Co. Ltd.”, холбоо “association”, and компани “company”); b) capitalized word + sequence of lowercased words + organization designator; all-capital word (e.g., МУИС-NUM).

3. Location name: a) capitalized word + location designator (e.g., гудамж “street”, талбай “square”, хот “city”).

Edit Distance Metrics. In order to robustly match inflectional form or naming variations, we validated using the main string similarity measures used in [7] and [8]. A token based metrics are Levenshtein, Smith-Waterman (SW) [13] (gap cost is *1*, the mismatch cost is *-1* and the match cost is *2*), Jaro and Jaro-Winkler (JW)[14]. The multi-token based metrics are Jaccard similarity, Fleggi-Shunter [15], Monge-Elkan [16] and SoftTFIDF [7].

In [8], considering the declension paradigm of Polish, approved a basic and time efficient metric based on the longest common prefix information, which would intuitively perform well in the case of single-token names. We have been inspired by this method and modified the declension paradigm to agglutinative feature. It is defined as $CP_b = (|lcp(s,t)| + b)^2/(|s| * |t|)$. b is set to 0 . If s ends in common suffix inflection such as “-ын” (genitive case), “-ынд” (genitive case + locative case), b is set to 1.

3.3 Classifiers

For the statistical classifer, we used the ME (OpenNLP v1.5.3), CRF (CRF++ v0.53) and SVM (Yamcha v0.33).

3.4 Ensembling

A Genetic algorithm (GA) [17] is a search method of an optimal solution, inspired by biological evaluation processes.

In our experiments, the genome for optimal ensembles from a set of $n = 15$ classifiers (5 models for each 3 classifiers), can be a binary string of length n, in which every bit represents a classifier. A bit value 1 means the classifier is selected and 0 is vice versa. For instance the chromosome 101010101010101 represents an ensemble in which every first classifier is used. Our implementation follows specific parameters and methods described in [6] and [5].

To combine the outputs of classifiers we use majority voting mechanism. We implemented and tested the following decision function: each classifier’s output tag votes for a NE class for each token, and the tag that has highest score wins.

4 Resource building

We manually annotated a Mongolian POS-tagged corpus [11] from the Newswire domain with NE tags. The corpus consists of 310 articles, about 277,000 tokens, 14,837 sentences, 4,382 personal names, 4,932 location names, and 3,366 organization names.

This corpus is equivalently created to English corpora used on the CoNLL02 and CoNLL03 conferences, in format, annotation style and the number of NEs.

To guarantee the accuracy of tagging we set up a three-stage annotation procedure: first, linguists manually labeled the corpus with NE tags using clearly established guidelines; secondly, a reviewer validated the annotations, and finally, the linguist and reviewer discussed borderline cases. The agreement of linguist and reviewer F-measure was 98.56% (using conlleval script).

For gazetteers, we had access to a free-to-use 170,000-entry list of Mongolian personal names and to a 80,000-entry list of Mongolian company names.

For locations, we used *Geonames.org* that contains 4,151 names of mountains, rivers, landmarks, and localities of Mongolia. However, these names were in various *Latin* transcriptions because of crowdsourcing; we therefore had to convert 2,410 names from *Latin* transcriptions to *Mongolian Cyrillic*. We also extracted about 3,500 location names from *Wikipedia*.

5 Experimental Results and Discussion

A randomly selected 10% segment of the NE corpus was used as a test set and rest of 90% was used as a training set. We used the CoNLL03 evaluation methodology [10].

5.1 Results in edit distance functions

To validate efficiency of each string distance metrics, we compared 4,382 real-world names of the corpus to the non-inflected 176,343 person names (comparison pairs are 770 million). Table 1 shows results for each metrics that we involved. The SW based metrics achieved the best recall score, whereas Levenshtein was the worst metric. However, the SW based metrics matched irrelevant pairs. For example, "Лувсаншарав-ваас" (Лувсаншарав (person name) + аас (ablative case)) is matched to "шарав" and scored those as 1.0. The JW and the combination of Monge-Elkan with JW achieved the best F1-score 94.1% and 91.6%. We therefore chose it as default string matching method.

Metrics	**Pre**	**Re**	**F1**	**Metrics**	**Pre**	**Re**	**F1**
On a token				**On multiple tokens**			
Levenshtein	73.4	71.9	72.6	MEl & SW	82.4	94.7	88.1
SmithWaterman (SW)	85.1	97.5	90.8	MEl & JW	89.3	93.8	91.6
Jaro	93.2	88.9	90.1	Jaccard & DM	79.3	67.5	72.9
JaroWinkler (JW)	95.5	92.8	94.1	Fleggi-Shunter	75.6	70.1	72.7
Common prefix	84.5	83.1	83.8	SoftTFIDF & JW	92.2	90.2	91.2
				SoftTFIDF & SW	86.8	94.8	90.6

Table 1. The experimental results for edit distance metrics.

5.2 Results in feature selection experiments

We built 31 models for each 3 classifier methods (totally 93) from the available NE features. Table 2 presents the top 15 models that achieved the best F-measure for each classifier. The best individual classifier shows F-measure values 86.94%. One interesting thing is that ME reached its best performance without orthographic feature. The

ME		**CRF**		**SVM**	
Fgroups	F1	Fgroups	F1	Fgroups	F1
2,4,5	83.69	1,2,3,4,5	86.94	1,2,3,4,5	86.66
2,4	83.04	1,3,4,5	86.91	1,2,3,4	86.57
1,2,4	82.92	1,2,3,4	86.75	1,3,4	86.56
1,2,4,5	82.82	2,3,4,5	86.75	2,3,4	86.16
1,2,3,4	82.81	2,3,4	86.62	1,2,4,5	86.11

Table 2. The top five feature sets for each classifier.

gazetteer feature was frequently involved in the best performing classifiers. The first observation that can be made about applying gazette feature with robust string matching method is its good impact to the fine grained classifier. It can be also observed that CRF achieves the best-performance for the Mongolian NER task.

5.3 Results in the classifier ensemble

We tested the ensemble of different classifiers that are trained using one kind of classification method and 5 different feature set. As shown in Table 3, the scores of the ensemble of the best 5 models of ME, the F-measure is decreased by 0.97% from best individual ME. For CRF ensemble; the F-measure is improved by 0.42%. Finally, for it also decreased from best individual performance. We gathered the optimal genome

Classifier	**Pre**	**Re**	**F1**
Ensemble of MEs	88.86	77.38	82.72
Ensemble of CRFs	90.83	84.14	87.36
Ensemble of SVMs	88.19	84.75	86.43

Table 3. Result of ensemble on feature set.

00001 10011 11111. The first 5 bits represents the ME classifiers, ordered per the best F1-accuracy rank (according to Table 2), followed by 5 CRFs and 5 SVMs. The optimal ensemble reached to 90.59% precision, 85.88% recall and 88.17% F1 score.

We seriously tried to apply existing machine learning and string matching methods into NER, as well as created the NER corpus and gazettee for Mongolian language.

The experimental results confirm that genetic algorithm can be successfully applied to the task of finding a classifier ensemble that outperforms the best individual classifier and simple ensemble method. However, the performance improvement measured in our experiments is not satisfactory, since running many classifiers that are trained on different feature sets took more CPU time than one classifier.

We nevertheless consider these as promising first results, especially taking into account the difficulties of Mongolian grammar. They will also be useful as a quantitative basis for comparison for further research in Mongolian NER.

References

1. Bender, O., Och, F., J., H., N.: Maximum entropy models for named entity recognition. In: Proceedings of CoNLL-2003. (2003) 148–151
2. Isozaki, H., Kazawa, H.: Efficient support vector classifiers for named entity recognition. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002), Taipei, Taiwan (2002)
3. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models-for segmenting and labeling sequence data. In: Machine Learning International Workshop. (2001)
4. Florian, R., Ittycheriah, A., Jing, H., Zhang, T.: Named entity recognition through classifier combination. In: Proceedings of the seventh conference on Natural Language Learning at HLT-NAACL. (2003)
5. Desmet, B., Hoste, V.: Dutch named entity recognition using classifier ensembles. In: Preceedings of the 20th Meeting of Computational Linguistics, Netherlands (2010)
6. Ekbal, A., Saha, S.: Maximum entropy classifier ensembling using genetic algorithm for ner in bengali. In: Proceedings of the International Conference on Language Resource and Evaluation (LERC). (2010)
7. Cohen, W., Ravikumar, P., Fienberg, S.: A comparison of string distance metrics for name-matching tasks. In: IJCAI-3 Workshop on Information Integration on the Web (IIWeb-03), Acapulco, Mexico (2003) 73–78
8. Piskorski, J., Wieloch, K., Pikula, M., Sydow, M.: Towards person name matching for inflective languages. In: NLPIX, Beijing, China (2008)
9. Szarvas, G., Farkas, R., A., K.: A multilingual named entity recognition system using boosting c4.5 decision tree learning algorithms. In: Springer Berlin Heidelberg. (2006)
10. Tjong, Kim, S., Fien, De, M.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: CoNLL-2003, Canada (2003)
11. Purev, J., Odbayar, C.: Part of speech tagging for mongolian corpus. In: The 7th Workshop on Asian Language Resources, Singapore (2009)
12. Simon, E., Kornai, A.: Approaches to hungarian named entity recognition. In: PhD thesis, Budapest University of Technology and Economics. (2013)
13. Smith, T., Waterman, M.: Identification of common molecular subsequences. In: Journal of Molecular Biology, 147. (1981) 195–197
14. Winkler, W., E.: The state of record linkage and current research problems. In: Technical report, Statistical Research Division, U.S. Bureau of the Census, Washington, DC (1999)
15. Fleggi, I., P., Sunter, A., B.: A theory for record linkage. In: Journal of the American Statistical Society 64. (1969) 1183–1210
16. Monge, A., Elkan, C.: The field matching problem: Algorithms and applications. In: Proceedings of Knowledge Discovery and Data Mining 1996. (1996) 267–270
17. Goldberg, D., E.: Genetic algorithm in search, optimization, and machine learning, Boston, Ma, USA, Addison-Wesley Publishing Company (1989)