Fifteenth International Conference On Computer Applications (ICCA2017)

Exploring Features for Myanmar Named Entity Recognition

Hsu Myat Mo, Khin Thandar Nwet, Khin Mar Soe
Natural Language Processing Lab, University of Computer Studies, Yangon, Myanmar
hsumyatmo@ucsy.edu.mm, khinthandarnwet@ucsy.edu.mm, khinmarsoe@ucsy.edu.mm

Abstract

Named Entity Recognition (NER) is the task of classifying or labeling atomic elements in the text into predefined sets of named entity categories such as Person, Location, Organization or Number. NER is also a crucial piece of Information Extraction System. Robust handling of proper names is essential for many applications in natural language processing and key knowledge acquisition infrastructure for the Semantic Web. For Myanmar Language, exploring features for NER with machine learning approach is a still challenging task because of the complex nature of the language. This paper describes our effort on feature exploring using local and external information for Myanmar NER that applied Conditional Random Fields (CRFs). The experimental results show that the best result is obtained by combining the local feature, such as neighboring words, and the external information such as clue words and name lists.

Keywords: Named Entity Recognition, Feature Exploring, Myanmar Language

1. Introduction

Recognition of named entities (e.g. people, organizations, location, etc.) is an essential task in many natural language processing applications nowadays. NER could be used in various NLP systems to improve their performance such as sematic parsers, question and answering system, summarization, document categorization, document and news searching, etc. and also in machine translation. A well performing NER is important because it can have an impact on other NLP applications and can be applied for future level of NLP work.

Named Entity Recognition (NER) is the process of automatically tagging, identifying or labeling different named entities in text in accordance with the predefined set of types.

Most of the NER systems use three approaches: rule-based, machine learning (statistical) approach and hybrid approach. There are two types of machine learning models that are used for NER called Supervised and Unsupervised machine learning model. Several Statistical approaches with supervised and unsupervised learning have been applied in English, European and some other Asian

languages successfully. Myanmar language is very complex compared to English and other European languages. On the other hand, it has both complex and rich morphology as well as ambiguity. Moreover, there is no capitalization and Myanmar texts are written without spaces. Therefore, how to identify named entities written in Myanmar text automatically is intriguing.

A statistical approach to NER for the Myanmar Language using one of the supervised machine learning approaches called Conditional Random Fields (CRF) had been proposed in [1]. In statistical machine learning based approach, annotated corpus is used as training data and a probabilistic model is built with features of the data which is similar to the rules that are used in ruled-based approaches. The corpora with correctly labeled name entities are learned to produce the features of the data. The model then uses the features to calculate and identify the most probable named entities. It is need to discover relevant features and construct the features that increase conditional likelihood.

In this paper, various combinations of features have been explored and compare their impact on recognition performance of the CRF-based Named Entity Recognition for Myanmar Language.

The remainder of the paper is structured in the following way. In section 2 related works will be described. In section 3 a brief introduction of the nature of Myanmar language and challenges of Myanmar language to NER is going to be presented. In section 4 Conditional Random Fields (CRFs) and various features used are discussed followed by proposed Myanmar NER work flow in section 5. Summary of the performed work and future plans are described in section 6.

2. Related Works

Previous effort on Myanmar NER had been done by two research works. A method for Myanmar Named Entity Identification using a hybrid method has been presented by (Thi Thi Swe and Hla Hla Htay, 2010). This method is a combination of ruled based and statistical N-grams based method which use name database. They classified Myanmar NEs into three classes, namely person name (PER), organization name (ORG) and location name (LOC).[2]

Moreover, (Thida Myint and Aye Thida, 2014) proposed Myanmar Named Identification algorithm. In the algorithm, the system defines the names by using some of the POS information, Name entity identification rules and clue words in the left and/or the right contexts of NEs carry information for NE identification.[3]

Analysis of the impact of various features from the construction of a CRF-based supervised NER system on the CoNLL 2003 dataset and the OntoNotes version 4 CNN dataset is performed by (Maksim Tkachenko and Andery Simanovsky, 2012).[4]

For the work on identification and classification of named entities in Tamil, the authors of [5] discussed about the linguistic features used in CRFs for Tamil language. In their work, the features are designed based on the theoretical linguistic information such as morphological and syntactic information and features are classified into three: context based, word based and structure based.

3. Myanmar Language

Myanmar language is the official language of the Republic of the Union of Myanmar. Myanmar Language Commission (MLC) standardized that it is composed of nine parts of speech in Myanmar grammar such as noun, pronoun, adjective, verb, adverb, post-positional marker, particle, conjunction and interjection. It

is written from left to right and usually with no space between words. Myanmar language is mainly characterized as a SOV (subject, object and verb) language; would probably defined as postpositional language and it is also regarded as a free order of word language which means that the part of speech of the word in the text can vary according to its position in the sentence.

3.1. Challenges of Myanmar Language to NER

The task of identifying names in Myanmar text automatically is complex compared to other languages for many reasons. One of the reasons is the lack of resources such as annotated corpus, name lists, name dictionaries, etc. which means that Myanmar is resource-constrained language. Moreover, Myanmar language has distinct characteristics and having no capitalized letter which is the main indicator of proper names for some other languages like English and which is of the free order, makes the NER a complex process. Myanmar names also take all morphological inflections which can lead to ambiguity. This ambiguity of named entities may lead to problem in classifying name entities into predefined types.

4. Conditional Random Fields (CRFs)

Models for many natural language tasks benefit from the flexibility to use overlapping, non-independent features. While it is difficult to capture inter-dependent features such as word features, part-of-speech tags, character n-grams, capitalization patterns for like English language and some other orthographic features with a generative probabilistic model, conditionally-trained models, can handle them well.

Conditional models are used to label the observation sequence x_* by selecting the label sequence y_* that maximizes the conditional probability $p(y_*|x_*)$. The conditional nature of such models means that no effort is wasted on modeling the observation and one is free from having to make unwanted independence assumptions about these sequences; arbitrary attributes of the observation data may be captured by the model, without the modeler having to worry about how these attributes are related.[7] Additionally, CRFs avoid the label bias problem, a weakness exhibited by maximum entropy Markov models (MEMMs) and other conditional Markov models based on directed graphical models.[6][7] Conditional Random Field is an example of discriminative models.

CRFs are undirected graphical models used to calculate the conditional probability of values on designated output nodes given values on other designated input nodes.

Lafferty [6] defined the probability of a particular label sequence y given observation sequence x to be a normalized product of potential functions, each of the form as:

$$p(y|x,\lambda) = \frac{1}{Z(x)} \exp\left(\sum_j \lambda j F j(y,x)\right) (1)$$

where $F_j(y,x)$ is either a state function $s_j(y_{i-1},y_i,x,i)$ or transition function $t_j(y_{i-1},y_i,x,i)$, λj is the weight of indicating the precision of feature f_j, $Z(x)$ is a normalization factor.

Feature function F_j whose weight λj is to be learned via training. Feature functions could ask arbitrary questions about two consecutive states, any part of the observation sequence and current position.

A set of real-valued features b(x,i) of the observation that should hold the model distribution is constructed to expresses some characteristic of the empirical distribution of the training data.

ဖြူ သည် ရန်ကုန် တွင် နေသည် ။

For example, in the above sentence which means “Phyu lives in Yangon”, the feature

observation of word ရန်ကုန် at position 3 is constructed as follow:

$$b(\text{ရန်ကုန်}, 3) = \begin{cases} 1 \text{ if the observation at position 3 is the word "ရန်ကုန်"}, \\ 0 \text{ otherwise.} \end{cases}$$

Each feature function takes on the value of one of these real-valued observation features b(x, i) if the current state (in the case of a state function) or previous and current states (in the case of a transition function) take on particular values.

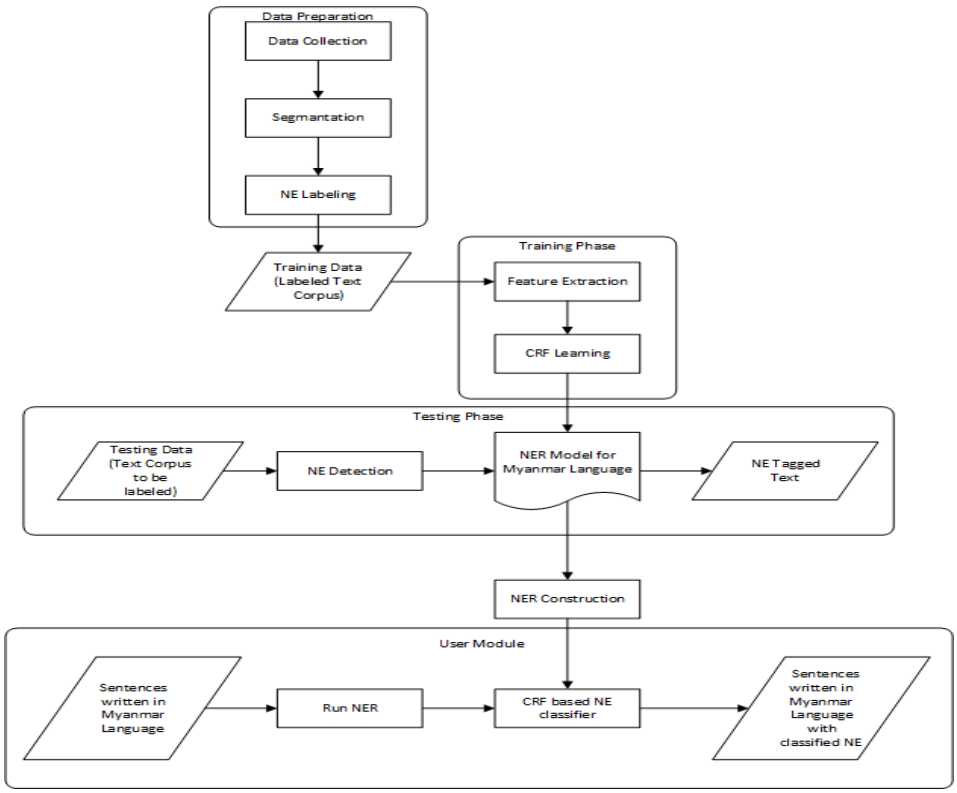

Figure 1. Work Flow of NER

5. Myanmar NER Work Flow

The work flow of CRF-based approach to Myanmar NER is shown in Figure 1.

Data for training and testing are collected from Myanmar news articles websites. Training and testing data are prepared through the process of segmentation, manually label the name entities tags. For the segmentation work, Myanmar online Word Segmentation application that is available on University of Computer Studies, Yangon official website[1] is applied. For the training process different kinds of features are explored to train the model. In this paper, some particular types of features are considered and compare their impact on the F score value.

[1] http://www.nlpresearch-ucsy.edu.mm/NLP_UCSY/wsandpos.html

6. Experiment and Evaluation

Since CRFs are log-linear models, and high accuracy may require complex decision boundaries that are non-linear in the space of original features, the expressive power of the models is often increased by adding new features that are conjunctions of the original features. For example, a conjunction feature might ask if the current word is a lexicon of location names and the next word is "မြို့နယ်".

Firstly, we start with local knowledge features which can be extracted from a token (word) being labeled and its surrounding context. Next, evaluation with external knowledge features (name lists and clue word lists) is carried out.

In this work, ten types of named entities types including person name, location name, organization name, government organization name, names of season, year, month and date, number and percent value are predefined to label the name entities in the text.

Training data and testing data are prepared as follow.

Table1.Specification of training data

Total No. of Tokens	Total No. of named entity Tokens	Total No. of not named entity Tokens
18001	1110	16891

Table 2.Specification of testing data

Total No. of Tokens	Total No. of named entities Tokens	Total No. of Not named entity Tokens
2017	78	1939

6.1. Local Context Features to NER

Firstly, only local information that is only current token is used to deal with the NER. It is not surprising that a token-based CRF perform poorly. Next, the previous and next tokens and the current token are considered. Moreover, the information of the two previous and two next tokens are used to train the model. The impact of each experiment is shown in Table 3.

Table 3. Evaluation of context as feature in NER

Features	Precision	Recall	F score
W0	36.81	22.57	32.6
W-1,W0,W1	65.52	70.21	69.31
W-2,W-1,W0,W1,W2	51.36	57.89	54.69

6.2. Combining External Features to NER

Most NER systems of other language such as English use additional features like part-of-speech (POS) tags, shallow parsing, gazetteers, etc,.

For this work clue word list for person names and name lists are used as external features. Myanmar person names are usually started with ဦး ၊ ကို ၊ မောင် for male and မ ၊ ဒေါ် for female. Likewise, some of the person names are started with salutation words such as ဒေါက်တာ၊ ပါမောက္ခ.

Such possible clue words for person names, location names and organization names are prepared first and use these as external features for the NER task. With these features the result show that

Besides the clue word lists, name lists are also used as external features and compare the result.

Table 4. Evaluation with external feature in NER

Features	Precision	Recall	F score
W-1,W0,W1+clue	64.33	81.12	73.85

word lists			
W-1,W0,W1+ clue word lists+ name lists	65.99	89.09	75.82

According to the experiment, the training process with the information of previous and next word and current word and with the help of clue words list and name lists information can provide the best rest to the given training and testing data.

7. Conclusion

In this paper, a comprehensive set of features that are used in supervised NER has been explored and their impact on the performance of the NER for Myanmar language has been described. As future work, more and more researches will be carried out focusing on how to optimize the statistical NER for Myanmar Language.

References

[1] Hsu Myat Mo, Khin Thadnar Nwet and Khin Mar Soe, "CRF-based Named Entity Recognition for Myanmar Language", Genetic and Evolutionary Computing, Vol(536), pp 204-211.

[2] Thi Thi Swe, Hla Hla Htay, "A Hybrid Methods for Myanmar Named Entity Identification and Transliteration into English", http://www.lrec-conf.org/proceedings/lrec2010/workshops/W16.pdf

[3] Thida Myint, Aye Thida, "Named Entity Recognition and Transliteration in Myanmar Text", PhD Research, University of Computer Studies, Mandalay, May, 2014.

[4] M.Tkachenko and A.Simanovsky, "Named Entity Recognition: Exploring Features", Proceedings of KONVENS 2012, Vienna, September 20, 2012.

[5] V.S. Ram R, Akilandeswari A and S.L. Devi, "Linguistic Features for Named Entity Recognition Using CRFs", International Conference on Asian Language Processing, 2010.

[6] J. Lafferty, A. McCallum, F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning, Pages 282-289, 2001.

[7] H. M. Wallach, 'Conditional Random Fields: An Introduction', University of Pennsylvania CIS Technical Report MS-CIS-04-21, February 24, 2004.

[8] C. Sutton and A. McCallum, "An Introduction to Conditional Random Fields" Foundations and Trends in Machine Learning, Vol.4, No. 4(2011)267373, 2012.