Khmer POS Tagging Using Conditional Random Fields

Sokunsatya Sangvat and Charnyote Pluempitiwiriyawej(✉)

Faculty of Information and Communication Technology, Mahidol University,
Salaya, Nakhon Pathom, Thailand
sangvat.sok@student.mahidol.ac.th, charnyote.plu@mahidol.ac.th

Abstract. The transformation-based approach with hybrid of rule-based and tri-gram have already been introduced for Khmer part-of-speech (POS) tagging. In this study, in order to further explore this topic, we present an alternative approach to Khmer POS tagging using Conditional Random Fields (CRFs). Since the features greatly affect the tagging accuracy, we investigate five groups of features and use them with the CRF model. First, we study different contextual information and use it as our baseline model. We then analyze the characteristics of Khmer and come up with three additional groups of language-related features including morphemes, word-shapes and name-entities. We also explore the use of lexicon as features to further improve the accuracy of our tagger. Our proposed approach has been evaluated on a corpus of 41,058 words and 27 POS tags. The comparative study has shown that our proposed approach produces a competitive accuracy compared to other Khmer POS tagging approaches.

Keywords: Khmer · Part-of-speech tagging · POS tagging
Conditional Random Fields

1 Introduction

Part-of-speech (POS) tagging, a process of assigning each word in a text with the most suitable part-of-speech or lexical category, is an important task in natural language processing. It is used as a pre-processing step in many applications such as information retrieval/extraction, machine translation and speech recognition.

Many work have been published on POS tagging for many different languages. However, the research of POS tagging for Khmer, the official language of Cambodia, is currently in early stage. Based on our study, there is only one published work that discusses the topic [1]. The development of Khmer POS tagging is necessary as it would have a significant impact on others NLP researches and applications of Khmer. Therefore, the problem of Khmer POS tagging should be further explored.

In this paper, we present a new approach for automatic Khmer POS tagging using Condition Random Fields (CRFs). Because the features have great impacts on the tagging accuracy, we introduce four groups of features for the CRF model. First, we investigate different templates of contextual information and use it as our baseline model. Then, we investigate some characteristics of Khmer and come up with three

K. Hasida and W. P. Pa (Eds.): PACLING 2017, CCIS 781, pp. 169–178, 2018.
https://doi.org/10.1007/978-981-10-8438-6_14

additional groups of language-related features including morphemes, word-shapes and name-entities. In addition, we also explore the use of lexicon to further improve the accuracy of our tagger. Our POS tagger has been developed and evaluated on a corpus of 41,058 words and 27 POS tags that have been borrowed from the previous work in Khmer POS tagging conducted by Nou and Kameyama [1]. Our comparative study has shown that our proposed approach produces a competitive accuracy compared to Nou's and Kameyama's [1] approach.

2 Related Work

A number of statistical models have been proposed for English POS tagging. A set of top performing models includes transformation-based approach, support vector machine, maximum entropy and decision tree. These models achieve from 96% to 97% accuracy [2–5].

An elastic-input neuro tagger [6] and a hybrid tagger, combined with a neural network and Brill's error-driven learning [7] have been proposed for Thai, a language which shares a considerable number of vocabularies and grammar rules with Khmer. The two models achieve 94.4% and 95.5% accuracies respectively. Another study [8] has developed a more accurate tagger using support vector machine which increased the precision up to 96.1%.

A genetic-algorithm-based POS tagger [9] has been introduced for Chinese, which has overlapping grammar structures with Khmer. The proposed model has been shown to be more flexible to incorporate both statistical and rule information than other models such as recurrent neural net, hidden markov model and rule-based model. As a result, genetic algorithm achieves 95.8% accuracy which is higher than the other models. The maximum entropy model has also been proposed for Chinese POS tagging [10]. The model has been built using a large annotated corpus and produces 96.8% accuracy.

The most related work to our work was conducted by Nou and Kameyama [1]. Nou and Kameyama [1] have proposed a transformation-based approach with hybrid of rule-based and tri-gram for Khmer POS tagging. The problem is separated into two tasks: known words and unknown words handling. The known word handling adopted the transformation-based learning approach [2]. In this approach, the tagger first assigns the most frequent POS to each word by looking up in the lexicon (dictionary) and then applies the learned rules, obtained from the training data, to reduce the error. However, the tagger often encounters a number of unknown words, which do not appear in the lexicon and the training data. Thus, the tagging accuracy can be low because the rules of those words cannot be obtained. Therefore, the unknown words are handled in the second task which uses a hybrid of rule-based approach and tri-grams. The rule-based approach tags the words based on their internal structure while the tri-grams relies on contextual information. The hybrid of both models combines the strength of each other and thus produces a higher accuracy. Overall, the proposed approach achieves 95.27% and 91.96% of recall on the training and the testing set respectively.

3 Overview of Conditional Random Fields

Conditional Random Fields (CRFs), introduced by [11], are conditional probabilistic distribution models, designed for the problem of sequence labelling. They are a type of undirected graphical model that computes a single log-linear distribution over a state sequence (e.g. POS tags sequence) given an observation sequence (e.g. words in a sentence). The conditional probability of a state sequence $S = \left(s_1, s_2, s_3, \ldots, s_T\right)$ given an observation sequence $O = \left(o_1, o_2, o_3, \ldots, o_T\right)$ is defined as:

$$P(S|O) = \frac{1}{Z_0}\exp\left(\sum\nolimits_{t=1}^{T}\sum\nolimits_{k=1}^{K}\lambda_k f_k\left(s_{t-1}, s_t, o, t\right)\right) \tag{1}$$

where $f_k\left(s_{t-1}, s_t, o, t\right)$ is a feature function whose values could have a huge range, but they are typically binary. There are totally K feature functions. Each feature function is associated with λ, a learned weight computed using the training data. Z_0 is the normalization factor which is used to make all conditional probability of all candidate paths sum up to 1. Thus, the normalization factor is calculated as:

$$Z_0 = \sum\nolimits_{s}\exp\left(\sum\nolimits_{t=1}^{T}\sum\nolimits_{k=1}^{K}\lambda_k f_k\left(s_{t-1}, s_t, o, t\right)\right) \tag{2}$$

We are given training data $D = \left\{s^{(i)}, o^{(i)}\right\}_{i=1}^{N}$, where each $o^{(i)} = \left\{o_1^{(i)}, o_2^{(i)}, \ldots, o_T^{(i)}\right\}$ is an observation sequence, and each $s^{(i)} = \left\{s_1^{(i)}, s_2^{(i)}, \ldots, s_T^{(i)}\right\}$ is a state sequence. The CRF model is trained by using the standard maximum log-likelihood estimation method. In order to avoid overfitting, we use regularization which penalizes the weight vectors whose norm is too large. Thus, the log-likelihood L of a given training data D is calculated as:

$$L = \sum\nolimits_{i=1}^{N}\log\left(P(s^{(i)}|o^{(i)})\right) - \sum\nolimits_{k=1}^{K}\frac{\lambda_k^2}{2\sigma^2} \tag{3}$$

In the second sum, the parameter λ is set to maximize the log-likelihood L by using various training algorithms. The parameter σ^2 is a free parameter which controls how much the weights are penalized.

We use CRFsuite [12], an open-source implementation of CRFs, for our experiment. The CRF model is trained using Gradient Descent algorithm with the L-BFGS method. We try this training method with different parameter values and find the best result by setting all parameters with the default values.

4 Khmer and Features Analysis

Khmer is the official language of Cambodia. It is spoken by approximately 16 million people of Cambodia, Khmer ethnicity in southern Vietnam and northeastern Thailand. Khmer is largely influenced by Pali and Sanskrit through Hinduism and Buddhism.

In this study, the POS tagger has been built by considering a number of Khmer characteristics which are the following:

- Khmer text are written continuously without explicit word boundary. Text need to be segmented before assigning POS.
- A word in Khmer can have more than one possible part-of-speeches depending on the meaning and the context of the word. For example, the word “មុន” /m′un/ can be used as an adverb with the meaning of “before” or as a noun with the meaning of “acne”.
- Most of the words that are borrowed from Pali and Sanskrit have more than one acceptable spellings. They can be written in both Pali/Sanskrit or modern simplified way. e.g. The Khmer word for “value” can be written as “តម្លៃ” /t′aṃl′ai/(Pali/Sanskrit) or “តំលៃ” /t′aṃl′ai/ (modern simplified).
- Adding prefixes or suffixes to a word can change the POS of the word. For example, adding the prefix “អ្នក” /ʿ′n′a′k/ to the word “បើកបរ” /b′oekbr/ Drive [Verb] will produce the word “អ្នកបើកបរ” /na:kb′oekbr/ Driver [Noun]·
- A considerable number of words are compound words, which are the combinations of multiple root words. The POS of the compound word is dependent on the POS of the root words. For example, the compound noun “ក្រុមហ៊ុន” /kr′umh′un/ (Company) [Noun] is a combination of the word “ក្រុម” /kr′um/ (Group) [Noun] and the word “ហ៊ុន” /h′un/ (Investment) [Noun].

We analyze the characteristics of Khmer and come up with features that are necessary for Khmer POS tagging. The features are categorized into four groups: contextual, morphological, word-shape and lexical features.

4.1 Contextual Features

Contextual features of a particular word in a sentence are defined with respect to a window of its surrounding words (proceeding words and succeeding words) and their combinations. The contextual features play an important role in producing high tagging accuracy because the POS tag of a word is dependent on the word itself and the surrounding context. However, it is challenging to come up with the template of contextual features that work best for Khmer POS tagging. Therefore, we conduct an experiment on three pre-defined templates of contextual features as shown in Table 1. The first template is an example that comes with in CRFsuite tool [12]. It contains five word-features and two two-words-transition features. The second template extends the first template by adding three three-words-transition features. The third template extends the second template by adding a five-words-transition feature. The objective of the experiment is find the most suitable template of contextual features for Khmer POS tagging.

Table 1. Templates of contextual features where *w[i]* (*w[−i]*) is the *i*-th word to the right (left) of the current word *w[0]*

Template	Features
1	The word features: *w[−2], w[−1], w[0], w[1], w[2]* The words-transition features: *(w[−1]/w[0]), (w[0]/w[1])*
2	The word features: *w[−2], w[−1], w[0], w[1], w[2]* The words-transition features: *(w[−1]/w[0]), (w[0]/w[1]),* *(w[−2]/w[−1]/w[0]), (w[−1]/w[0]/w[1]), (w[0]/w[1]/w[2])*
3	The word features: *w[−2], w[−1], w[0], w[1], w[2]* The words-transition features: *(w[−1]/w[0]), (w[0]/w[1]),* *(w[−2]/w[−1]/w[0]), (w[−1]/w[0]/w[1]), (w[0]/w[1]/w[2]),* *(w[−2]/w[−1]/w[0]/w[1]/w[2])*

4.2 Morphological Features

A significant number of Khmer words are formed using morphological processes including prefixes, suffixes and compounding. With respect to the morphological rules of Khmer, we come up with the following morphological features that are useful for predicting POS.

- Prefixes and Suffixes: The POS of a Khmer word can be changed by adding a prefix (e.g., អ្នក /ʿ′n′a′k/, ភាព /bh′āb/, ការ /k′ār/, …) or a suffix (e.g., ជន /jn/, កិច្ច /k′icc/, ធម៌ /dhmr/, …).. However, it is challenging to define all prefixes and suffixes in Khmer. Therefore, we extract fixed numbers (from 2 to 6) of characters from each word and use them as prefix and suffix features.
- Compound Words: The POS of a compound word can be predicted from the POS of the root words. We propose five features corresponding to the five most common types of compound words: noun-noun, noun-verb, verb-verb, adjective-adjective and adverb-adverb. Each feature check whether or not a word is one of the compounding types. For example, the word “ក្រុមហ៊ុន” /kr′umh′un/ (Company) [Noun], a combination of the word “ក្រុម” /kr′um/ (Group) [Noun] and the word “ហ៊ុន” /h′un/ (Investment) [Noun], is a noun-noun-compounding type and not the others.

4.3 Word-Shape Features

Word-shape features check whether or not a word matches a particular pattern. For example, abbreviation in English is often written with all capital letters. We have studied the patterns of Khmer words and introduce four different patterns that could predict four different POS tags: cardinal number (ប.សំ.),, ordinal number (បូ.សំ.),, abbreviation (អ.ក.) and foreign words (ប.រ.).

- Cardinal number: It is identified by having each character as Khmer number 0 to 9 (០, ១, ២, ៣, ៤, ៥, ៦, ៧, ៨, ៩). For example, "១២" /dɔ:b pi:/ (12) is a cardinal number.
- Ordinal number: It is identified by having the word "ទី" /d′ī/ (at) followed by Khmer number 0 to 9 (០, ១, ២, ៣, ៤, ៥, ៦, ៧, ៨, ៩). For example, "ទី១២" /d′ī dɔ:b pi:/ (12th) is an ordinal number.
- Abbreviation: It is identified by having two to five consonants separated by period ".". For example, "អ.ស.ប." is the abbreviation for "អង្គការ សហ ប្រជាជាតិ" /′a′ngk′ā s′aḥh′aḥ br′a′jr′āj′āt/ (United Nations).
- Foreign words: It is identified if it has a non-Khmer character.

4.4 Name-Entity Features

In Khmer, name entities such as names of people, places and organizations are difficult to tag. Unlike English, Khmer does not have capitalized structure to identify name-entities. Moreover, there a significant number of people names that overlap with Khmer words. To help the model identify name-entities, we propose two features which are described as follows:

- Gazetteer-lookup feature: We have built 3 gazetteer lists (lists of name entities): people, places and organization lists. The three lists are extracted from KCorpus released by PAN Localization of Cambodia [13]. However, the names that also belong to other POS tags are excluded from the lists. For example, "ហេង" is a very popular Khmer name, but it is also an adjective meaning "lucky". The word "ហេង″ is excluded from the gazetteer lists. After the gazetteer lists are built, the feature checks whether or not a particular word appears in the gazetteer lists.
- Tittle feature: Most name-entities in Khmer are preceded by some common titles such as លោក /l′ok/ (Mr), កញ្ញា /kññ′ā/ (Miss), ខេត្ត /kh′ett/ (Province), ទីក្រុង /t′īkr′ug/ (City), ក្រុមហ៊ុន /kr′umh″un/ (Company), …, etc. This feature checks whether or not a particular word is followed by one of the title.

4.5 Lexical Features

Lexical features of a particular word are defined with respect to all of its possible POS(es) obtained from a lexicon (dictionary). In this work, we use the lexicon built by [1]. The lexicon contains approximately 32,000 words extracted from Royal Academy wordlist which has been approved by the Royal government of Cambodia. The lexicon includes most of the Khmer root words and their corresponding POS(es). The lexicon is used in two ways. First, all possible POS(es) of each word are used as features to CRFs. This will help the model predicts the POS of the words which do not appear in training set. Second, the words in the lexicon are used as root words in the compounding features described in Sect. 4.2.

5 Experimental Setup and Results

5.1 Experimental Setup

To be able to compare the performance with [1], we have set up the experiment that uses the same corpus. The corpus totally contains 41,058 words (1,298 sentences) and is tagged with 27 POS classes.

We adopt five-fold cross validation to evaluate our approach. The corpus is divided into five equal sized subsets. One of the five subset is used as the testing data and the remaining four subsets are used as the training data. The process is repeated for five iterations, where each of the five subsets is used once as the testing data. The final accuracy is the average accuracy across all five iterations. The accuracy is calculated as:

$$Accuracy = \frac{Number\ of\ correctly\ tagged\ words}{Total\ number\ of\ words} \tag{4}$$

In this study, we perform three experiments. The objective of the first experiment is to find the most suitable template of contextual features. The objective of the second experiment is to evaluate the accuracy of our proposed features. The objective of the third experiment is to compare our proposed approach to [1] in term of accuracy.

5.2 Finding the Most Suitable Template of Contextual Features

As described in Sect. 4.1, it is challenging to come up with the template of contextual features that works best for Khmer POS tagging. Therefore, we define three templates of contextual features (as shown in Table 1) and test the accuracy of each one. With the five-fold cross validation, the tagging accuracy on testing set performed by each template is shown in Table 2. We can see that Template 1, which is an example that come with the CRFsuite [12], produces the highest accuracy. Therefore, the contextual features defined in Template 1 will be used as the baseline features for the rest of the experiments.

Table 2. Accuracy of different templates of contextual features

Features	Testing accuracy (%)
Template 1	86.54
Template 2	85.76
Template 3	85.48

5.3 Features Evaluation

To improve the accuracy upon the contextual features (the baseline features), the morphological, the word-shape, the name-entity and the lexical features are added. Table 3 shows the accuracy of continuous adding the morphological (M), the word-shape (WS), the name-entity(NE) and the lexical (L) features to the contextual features (C). As we can see, adding the morphological and the lexical features to the model significantly improves the accuracy by 5.64% and 3.12% respectively while adding the word-shape and name-entity features slightly increases the accuracy by 0.23% and 0.86% respectively. In general, the testing accuracy improves as we keep on adding the features to the model. From this observation, we can conclude that the uses of the morphological, the word-shape, the name-entity and the lexical features are effective in improving the tagging accuracy together with the contextual features. Our proposed feature set (C + M + WS + NE + L) produces 96.39% accuracy.

Table 3. Tagging accuracy of accumulative features starting with the baseline

Features	Testing accuracy (%)	Δ Accuracy (%)
C (baseline)	86.54	–
C + M	92.18	+5.64
C + M + WS	92.41	+0.23
C + M + WS + NE	93.27	+0.86
C + M + WS + NE + L	96.39	+3.12

5.4 Comparison to the Previous Work

We evaluate the accuracy of our proposed approach and compare the results with Nou and Kameyama [1]. Our proposed approach is evaluated in two separated cases. In the first case, we adopt the same approach as [1] in order to fairly compare the accuracy. The corpus is divided into two sets: 32,088 words for training set and 8,970 words for testing set. All the training data are extracted from Kohsantepheap newspaper (a very popular newspaper in Cambodia). 60% of the testing data are extracted from the same newspaper and the other 40% are from letters, legends, and reading articles in high school student's textbook. We call this case hold-out validation. In the second case, using the same corpus, we evaluate our approach using five-fold cross validation. In each fold, 80% of the corpus is used for the training and the rest 20% is used for testing. Both training and testing data in each fold are extracted from mixed domains.

Table 4 compares testing accuracy between our proposed approach and Nou's and Kameyama's [1] approach. As can be seen, the accuracy of our proposed approach is competitive compared to Nou's and Kameyama's [1] approach when we adopt the hold-out validation. The accuracy is even higher when adopting the five-hold cross validation. In this comparison, we only consider the experimental results in which Nou and Kameyama [1] use the complete lexicon.

Table 4. Comparison between our proposed approach and Nou's and Kameyama's [1] approach

Method	Testing accuracy
Nou's and Kameyama's approach [1]	95.10%
Our proposed approach (hold-out validation)	95.52%
Our proposed approach (5-fold cross validation)	96.39%

6 Conclusion and Future Work

In this study, we have introduced a new approach to Khmer POS tagging using Condition Random Fields (CRFs). We have come up with 5 groups of features: contextual, morphological, word-shape, name-entity and lexical features which are used in the CRF model. These features are based on the example template that comes with CRFsuite and the study of Khmer characteristics. The experimental results have shown that they are effective in improving the tagging accuracy. Our proposed approach produces a competitive accuracy compared to earlier approaches in Khmer POS tagging.

There are several ways to increase the tagging accuracy in order to reach state-of-the-art level. One possible way is to train the tagger on a larger annotated corpus with diverse kinds of text. This would allow the model to obtain more statistical information and decrease the number of unknown words in the testing set. Another possible way is to use more advanced machine learning models such as support vector machine or deep learning. It might also be useful to combine statistical approach with rule-based approach to improve the accuracy. This work can be used to as a pre-processing step in other researches of Khmer natural language processing such as name-entity recognition, word sense disambiguation and parsing.

Acknowledgements. This research project was supported by Faculty of Information and Communication Technology, Mahidol University.

References

1. Nou, C., Kameyama, W.: Khmer POS tagger: a transformation-based approach with hybrid unknown word handling. In: International Conference on Semantic Computing (ICSC 2007), pp. 482–489 (2007). https://doi.org/10.1109/icosc.2007.4338385
2. Brill, E.: Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. Comput. Linguist. **21**, 543–565 (1995)
3. Giménez, J., Màrquez, L.: SVMTool: a general POS tagger generator based on support vector machines. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, pp. 43–46 (2004)
4. Ratnaparkhi, A.: A maximum entropy model for part-of-speech tagging. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, vol. 1, pp. 133–142 (1996)
5. Black, E., Jelinek, F., Lafferty, J., Mercer, R., Roukos, S.: Decision tree models applied to the labeling of text with parts-of-speech. In: Proceedings of the Workshop on Speech and Natural Language, pp. 117–121 (1992). https://doi.org/10.3115/1075527.1075554
6. Ma, Q., Uchimoto, K., Murata, M., Isahara, H.: Elastic neural networks for part of speech tagging. In: IJCNN99 International Joint Conference on Neural Networks Proceedings, vol. 5, pp. 2991–2996 (1999). https://doi.org/10.1109/ijcnn.1999.835997
7. Ma, Q., Murata, M., Uchimoto, K., Isahara, H.: Hybrid neuro and rule-based part of speech taggers. In: Proceedings of the 18th Conference on Computational Linguistics, pp. 509–515 (2000). https://doi.org/10.3115/990820.990894

8. Murata, M., Ma, Q., Isahara, H.: Part of speech tagging in Thai language using support vector machine. In: NLPRS 2001 Workshop, The Second Workshop on Natural Language Processing and Neural Networks (NLPNN2001) (2001)
9. Lua, K.T.: Part of Speech Tagging of Chinese Sentences Using Genetic Algorithm (1996). In: Proceedings of ICCC96, pp. 45–49. National University of Singapore
10. Zhao, J., Wang, X.-L.: Chinese POS tagging based on maximum entropy model. In: Proceedings International Conference on Machine Learning and Cybernetics, vol. 1, pp. 601–605 (2002). https://doi.org/10.1109/icmlc.2002.1174406
11. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), pp. 282–289 (2001)
12. Okazaki, N.: CRFsuite: a fast implementation of Conditional Random Fields (CRFs) (2007). http://www.chokkan.org/software/crfsuite/
13. Khmer Part-of-Speech Tagger (2008). http://www.panl10n.net/english/Outputs%20Phase%202/CCs/Cambodia/MoEYS/Papers/2008/KhmerPOSTaggingV1.0.pdf