View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by UUM Repository

MAXENTTAGGER FOR MALAY JAWI POS-TAGS

[1]Juhaida Abu Bakar, [2]Khairuddin Omar, [3]Mohammad Faidzul Nasrudin & [4]Mohd Zamri Murah

[1]School of Computing, Universiti Utara Malaysia,

[2,3&4]Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia

juhaida.ab@uum.edu.my

ko;mfn;zamri@ukm.edu.my

ABSTRACT

Purpose - Malay is a major language of the Austronesian family spoken in many countries. Malay Jawi is lacking in annotated resources and tools. In addition, Part-of-speech (POS) ambiguity in Natural Language Processing (NLP) is a vague important phenomenon that needs to be solved immediately. Since POS is an important feature of the word, and is the link between the words and syntax, POS tagging (POST) needs to provide intermediate results showing superior performance to the next NLP tasks. POS ambiguity is a main problem in increasing POST performance. POST performance is often measured with accuracy and precision of a tag and it was considered critical to NLP application. Some of the standard package POS tagging provided in Natural Language ToolKit (NLTK) are Brill tagger, HMM tagger, and CRF Tagger. In this paper, POST Malay Jawi implemented NLP tools, NLTK for the *state-of-the-art* methods tagger; maximum entropy models. NLTK is used as the implementation tool for Jawi tagging, as syntax and semantics of the language is transparent, and it has the good functionality of NLP-operator. The tool also uses Python as the implementation language.

Methodology - In this model, feature set will be used for tagging learning using Classify package in the NLTK module. Feature selection method is an important task which will determine the classification performance increase (Tang, Alelyani, & Liu, 2014). Feature selection improves the performance in terms of speed and effectiveness of learning. Feature selection also reduces the number of data dimensions and discards irrelevant data, repetitive, and noisy data. NLTK has a number of feature set based classifiers built-in; these operate on a variety of algorithms, including decision-tree models, Naïve Bayesian models, the Mallet and Weka machine-learning package, and maximum-entropy models (Malecha & Smith, 2010). Some works have already been done to create a part-of-speech tagger in NLTK using maximum entropy models (Ratnaparkhi, 1996) and megaM package (Daume III, 2004). Based on previous work (Hassan, Nazlia, & Mohd Juzaiddin, 2015; Malecha & Smith, 2010), some features are included, which is expected to correspond to the Malay Jawi as well as appropriate to different language and writing. The simplest type of tag feature is affix features. These features are based on the prefixes and suffixes of a word. Construction of these features is done automatically from the training corpus by recording all

prefixes and suffixes up to a certain length, together with their neighborhood information. In addition to using the current word, the tags of surrounding words can also be used as features. A common example in Malay Jawi might be that the word following a cardinal is often a noun and sometimes a verb, but rarely an adjective or preposition. These features are expected to be useful for classification when languages use modifiers and word positioning to convey meaning. This paper based on the experimental study achieved in Juhaida et. al (2017). We have conducted five experiments on features using NLTK parameters for selecting the best features that maximize accuracy.

Findings - The best model for the Malay Corpus is used in classifying the non-annotated Quran corpus. Table 1 shows the result of the words with ambiguity classes in the test corpus. According to the Malay Corpus tagset, for the word "ةلبق" (*kiblat*), the maximum amount of ambiguous word is the sum of four words that are tagged as Direction (KAR), Adjective (ADJ), Noun (KN) and Symbol (SYM). The word "ةلبق" (*kiblat*) is not in the training data and gives a variety of results. There are also prefix and suffix features that do not reflect the meaning of word affixation for words such as words that start with "me", such as "***me****reka* [them]", "***me****rah* [red]", or words ending in "an", such as "*adeg****an*** [scene]", and "*kawas****an*** [place]". Examples of words mentioned are not word affixation for words. This indicates features information for the whole words needs to be taken into account.

Table 1: Ten highest words with ambiguity class in the test corpus (Quran Jawi Translation Corpus)

	Buckwalter format	Translation		Occurrences	Ambiguity
		Malay Jawi	English		
1.	lyht	تهيل	see	4	KK, KN
2.	jAenlh	هلنڠاج	do not	3	KG#, KN
3.	AntArA	اراتنا	between	6	KSN, KN
4.	hAdVknlh	هلنکڤداه	face it	3	ADJ, KN
5.	lAlw	ولال	then	3	KK, ADV
6.	brAymAn	نامياراب	believer	18	KK, KT
7.	sQAlA	الڬس	everything	5	ADV, KN
8.	kVdAX	ثادڤک	to him	36	KSN, ADJ
9.	tAhw	وهات	know	13	KK, KN
10.	kAmw	وماک	you	41	ADJ, KG

Keywords: features selection, machine learning algorithm, part-of-speech, malay jawi

CONCLUSIONS

In this paper, experiments have been conducted on Jawi MaxEntTagger POS-Tags. The comparison is made for the results of the development and implementation of algorithms by calculating the accuracy of the state-of-the-art word tagging to identify the best models. The average accuracy is calculated based on the k-fold cross-validation, k = 10. The best model with useful features obtained with the highest accuracy, is displayed in percentage accuracy, precision, F-measure and confusion matrix. This paper covers part of the methodology of testing, experiments of the best tagging, and the results of each involved sub-corpus. This corpus is unique due to the Buckwalter code applied. This corpus will serve as a benchmarking corpus for the development and evaluation systems in word tokenization, as well as further language processing in Malay Jawi. This study is focusing on ambiguity classes and out-of-vocabulary (OOV) problem in the Jawi POS tagging. The findings in this study are comparable to previous studies of words not found in the dictionary (OOV). This is because the model used does not add a special literal feature on the words found in the corpus.

REFERENCES

Daume III, H. (2004). Notes on CG and LM-BFGS Optimization of Logistic Regression. Retrieved November 12, 2015, from http://www.umiacs.umd.edu/~hal/docs/daume04cg-bfgs.pdf

Hassan, M., Nazlia, O., & Mohd Juzaiddin, A. A. (2015). Malay Part of Speech Tagger: A Comparative Study on Tagging Tools. *Asia-Pacific Journal of Information Technology and Multimedia*, *4*(1), 11–23.

Juhaida, A. B., Khairuddin, O., Mohammad Faidzul, N., & Mohd Zamri, M. (2017). POS-Taggging Malay Corpus: A novel approach based on maximum entropy [Accepted]. *Journal of Engineering and Applied Sciences*.

Malecha, G., & Smith, I. (2010). Maximum Entropy Part-of-Speech Tagging in NLTK (pp. 1–10). unpublished course-related report.

Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-Of-Speech Tagging. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 133–142).

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, 37.