American Journal of Applied Sciences

Technical Reports

Comparison of Stochastic and Rule-Based POS Tagging on Malay Online Text

Kalaiarasi Sonai Muthu Anbananthen, Jaya Kumar Krishnan,
Mohd. Shohel Sayeed and Praviny Muniapan

Department of Information Science and Technology, Multimedia University, Melaka, Malaysia

Article history
Received: 28-09-2016
Revised: 21-12-2016
Accepted: 30-03-2017

Corresponding Author:
Kalaiarasi Sonai Muthu Anbananthen
Department of Information Science and Technology, Multimedia University, Melaka, Malaysia
Email: Kalaiarasi@mmu.edu.my

Abstract: Extensive development of web 2.0 has led to production of gigantic amount of user generated data. These data consist of many useful information. Manual analyzing these data and classifying sentiment in them, is an exhausting task, thus opinion mining method is needed. Opinion mining approach uses natural language processing where Part-of-Speech (POS) Tagging is a crucial part. The performance of any NLP system depends on the accuracy of a POS tagger. Two main issues that affect the accuracy of POS tagger are unknown words and ambiguity. Although research on POS tagging has been back dated few decades ago, yet they have been mostly focused on English. Research on Malay language is still in the early stage. Also, online Malay Text differs from proper Malay text, in the sense of structure and also grammar. Online users tend use a lot of abbreviations and short forms in their text. Besides this, the “BahasaRojak” phenomena complicate tagging process even further. Thus taking all these into consideration, in this study, we will review stochastic and rule-based POS tagging methodologies to deal with ambiguous and unknown words on online Malay text.

Keywords: Opinion Mining, Part-of-Speech Tagging, Malay Language, Malay Online Text, Rule Based Approach, Stochastic Approach

Introduction

Growth of web 2.0 has led to various range of development in the usage of social network and media. Social networks such as Twitter, Facebook, Blogs and Instagram are rapidly increasing on daily basis and this has led to the extensive generation of data and text. This large amount of data consists of daily conversations between users, opinions, dissatisfactions and comments. Such data can be efficiently used by government agencies to understand how far the policies and services are received by the citizens, healthcare providers to provide better services, or for people to be aware of others’ opinions and to benefit from their experiences to make informed decisions. Although these data are highly informational for various tasks, but it is simply exhausting for a person or organization to process these ever evolving data manually. Therefore, it becomes a real challenge to track, understand and extract the meaningful opinion from these texts. This has raised the demand to develop automatic methods to analyze and summarize these opinions. This leads to opinion mining.

Opinion mining is the process for tracking and identifying the opinion expressed by a person about a particular topic in natural language by analyzing the syntactic structure of the sentence. Opinion mining approach uses Natural Language Processing (NLP) to automatically extract and classify people’s “opinion” or “sentiment” from text. NLP is a process where machines are programed to understand human readable language. Machines are programed to understand sentence, its semantic values and also the syntactic structure of it. With this, machines are able to identify and determine the sentiments expressed in a sentence and classify the polarity of the text.

One of the most crucial processes in NLP is Part of Speech (POS) Tagging. POS tagging is the process of assigning the correct grammatical category for a word within a sentence taking in account to the syntactic structure of the sentence and also the composition of the word within the sentence itself.

A POS tagger consists of three main elements (Manning *et al*., 2014). The first part of a POS tagging is annotation. Annotation is the process of assigning the accurate POS tag for a word within a sentence, where this annotated word will then be used in the creation of corpus. Before the annotation process, tag set for the annotation process have to be determined. Every language consist of various tag set based on its morphology thus a tag set is chosen normally based on the language application for which the POS tags are used.

The second part of a POS tagging is to assign an accurate grammatical position for each word within the sentence. This tagging technique can be classified into two types which are stochastic tagging and rule based tagging. Stochastic taggers work based on mathematical equation taking account the probability distribution of occurrence of one POS followed by another in a sentence sequence (Samuelsson and Voutilainen, 1997). Whereas the rule based taggers work fully based on linguistic method, where a set of handcrafted linguistic rules are assigned and the tagger tags each sentence based on these handcrafted rules (Brill, 1992).

The last element of a POS tagger is the parser. Parsing is the process of analyzing a string of symbols whether it is confined within the rules of grammar and the analysis is then presented in a diagram (tree) form to give a clear visible structure of the sentence construction (Manning, 2003). Generally parsers are used to split the subject and the predicate of a sentence. With the usage of a parser it is able to identify the exact meaning of a sentence (Aranzabe *et al*., 2012).

There are two main methodologies for automatic POS tagging: Rule-based and stochastic approach (Kupiec, 1992). Rule-based POS approach work by the rules constructed by linguist experts. This methodology is non-automatic, costly and time-consuming. On the other hand, stochastic taggers are based on statistic approach where training consists of learning lexical and contextual probabilities. Stochastic models are robust and can be automatically trained but less accurate than rule-based models.

This research is focused on the development POS tagging software for Malay language. For Malay language, POS Tagging is still in the early stages of development. Also, limited resources and tools available for Malay language. POS taggers are very much language dependent. Most of the existing POS tagging approaches focus mainly on English. Malay Language on the other hand differs entirely in terms of composition, structure and complexity compared to English. Therefore, the available taggers cannot be directly used to tag Malay texts. To this date there is no available annotated text created for Malay Language.

Also, most Malaysians use abbreviations, improper grammar or "BahasaRojak" (mixture of English and Malay) online. New words are coined every day. Unknown words are non-negligible in Malay POS tagging. Therefore, in order to analyze these Malay online opinions, the POS tagger must be incurred with some knowledge of suggesting tag for unknown words and overcome ambiguous problem. The term "Unknown words" means the words that appear in sentences, but do not exist within the corpus. Ambiguity means that a word in a sentence can be categorized into more than one category; i.e., it can be noun or adjective. The performance of any NLP system depends on the accuracy of POS tagger.

In this study, we will review the rule-based and stochastic methodologies discussed above to find the best approach to deal with ambiguous and unknown words. Based on the review, a Malay POS tagging framework will be developed.

The paper is organized as follows. Section 2 gives a comparison of Bahasa Malaysia and English language. The overview of POS tagging is described in section 3. Section 4 describes the basic framework of a stochastic tagger and rule based tagger and the processes involved, while section 5 is where the discussion on both the taggers methodology is done, as for section 6 is where the evaluation whereby the performance of both the tagger across Malay online text is shown. Lastly section 7 is where the conclusion and future work are presented.

Difference in English and Malay Language

English is a global language and widely spoken language among many countries in the world (Crystal, 2004). Given the vast usage of this language, lots of researches have been carried out. Also, many resources and tools have been developed to analyze English language compared to Malay language. Malay language is the national language in Malaysia, Brunei and Indonesia and it is one of four official languages of Singapore. It has more than 270 million speakers across the Malacca Straights (Frawley, 2003). Although it is a widely spoken language, the research and resources for this language is still limited. Also, the available resources or tools for English language cannot be directly utilized to analyse Malay language due to the difference in morphology and syntactic structure in both of these languages. Morphology and syntactic structure affect tagging of the words.

Malay Morphology

Morphology is the study of way, words form sentences, from small meaning-bearing units which are referred to as morphemes (See, 1980; Bakar *et al*., 2013). Malay is an agglutinative language which is extraordinarily rich in morphology (Hassan, 1974). The basic morphologies in Malay are divided into three which are:

i. Affixation
ii. Reduplication
iii. Compounding

Affixation

An affix is a morpheme that is attached before, after or within to a word stem to form a new word. Affixation has the power of changing a tense, meaning and also part-of-speech category of a word entirely. Besides being the most important morphology category, affixation is also the most commonly used morphological type in Malay language. Affixes are divided into: Prefixes, suffixes, infixes and circumfixes (Hassan, 1974).

Prefix is a word that is present in beginning of a word. Common prefixes include *me-, pe-, be.* Suffixes are words that are present at the end of another word. Common suffixes are *–i, -kan, -nya.* Infixes are words that are present in the middle of a word. Common infixes are *–el-, -em- and –er-*. Circumfixes are words that appear at the beginning and end of a sentence. Common circumfixes are *peN-..-an, pe-.*

Reduplicaiton

Reduplication is a morphological process in which the word is repeated exactly. In other word, reduplication is duplicating a word into two, giving a different meaning to the word. There are three types of reduplication in Malay: Full duplication, partial duplication and rhyming and chiming. Full duplication is when a word is fully replicated for example *langit-langit "*palate/ceiling" (from *langit* "sky"). Partial reduplication is when a root word is partially replicated for example *jejari* "radius" (from *jari* "finger"). For rhyming and chiming is when phonological changes occur, this reduplication suits the sound when it is pronounced, example *lauk- pauk* "variety of dishes" (from *lauk* "dishes").

Compounding

Compounding is a process of combining two words into one and producing a new compound word. Example in Malay is the word *adat-istiadat* "culture and traditions" this word is made of two words which are *adat* "culture" and *istiadat*"traditions".

English morphology varies slightly from Malay. In English, morphology can be categorized into three which are derivational, inflection and compounding. Derivational are as similar as affixation in Malay morphology. The difference between the both are that derivation can only occur either as a suffix or affix not both together in a word whereas for Malay affixation, it may occur as prefix, suffix, circumfix and also as an infix. Inflection on the other hand, is suffixes which are added to a root word. This inflection does not change the meaning or grammar position, only changes past/present tense and plural/singular form of the root word. Whereas in Malay, inflection falls under affixation in general. Compounding in English Language is the same as in Malay morphology.

English to Malay Translation

Every language is unique on its own. Language composition, sentence construction, grammatical reasoning and syntactic structures vary from one to another. Therefore, translating one language to another, especially from different language category is a problem. Translation of a word or a sentence from English to Malay causes two problems which are grammatical error and also changing in the meaning of the word or a sentence in whole. Example of a word or a sentence where its grammatical position is changed is as below:

That watch is sold in Jusco

When this sentence is directly translated to Malay it is has below:

Itu jam adajual di Jusco
(Jusco sold that watch)

Even though the Malay translated sentence comes up to almost similar meaning with the English sentence, but if it is analyzed in Malay Morphology or grammar context, the translated sentence is grammatically wrong. The accurate translation for the sentence would be as below:

Jam tanganitudijual di Jusco
(That watch is sold in Jusco)

The second major problem in direct translation from English to Malay is that the translation may change the entire meaning of a sentence. Example for the stated problem is as below:

Please don't beat around the bush, come to a proper conclusion soon

When the sentence is directly translated the result shows the whole meaning of the sentences changes as shown below:

Silajanganmenewaskansekitarbelukar,
membuatkesimpulan yang betultidak lama lagi
(The term "beat around the bush" is in literal referred as the action of beating the bush)

Due to difference in the grammar and structural composition, direct translation from English to Malay is not appropriate. It is quite difficult to utilize the resources in English to Malay. From the examples above, we can see that the direct translation may also change the meaning of the whole sentence. Therefore, it is important to develop a POS tagger to tag Malay text.

Problems in Malay POS Tagging

Two main issues which are encountered in any POS Tagging development is the tagging of unknown words and ambiguous words (Gungor, 2010). In Malay Language, this problem is more serious due to lack of resources and word corpus. Problem of unknown words occurs when a word appear in sentences, but is not in the corpus. Especially online, new words are coined every day. It is impossible to train the tagger for every possible word in the language. Therefore, in order to build a complete POS tagger for Malay language, the tagger must be equipped with some intelligent or knowledge to suggest the tag for unknown word.

The second issue in POS tagging is ambiguous word. Ambiguous words are words with more than one sense (meaning) to it. There are many words in Malay that can be used in multiple ways, which means it can be tagged with more than one part of speech. If we take the word *jalan* (walk), in Malay this word brings two meanings one is walk another one is road. Therefore when we tag this word, there will be a conflict whether to tag this word is a verb or a noun.

Thus in this research we develop a Malay tagger which has the capability to overcome the above stated two problems and achieve a decent accuracy in the process of tagging Malay online text.

Related Works

Two main issues encountered in any POS Tagging application is appearance of unknown words and ambiguous word (Kumawat and Jain, 2015; Kumar *et al.*, 2015; Weischedel *et al.*, 1993). Currently the available approaches for POS tagging are based on either stochastic or ruled based methods (Silva *et al.*, 2013) to tag unknown word or disambiguate word. Stochastic taggers utilized maximum likelihood of a word in the sentence. It depends on the probability model that is determined based on the rules of the language (Nand and Perera, 2015). Whereas rule based taggers works fully based on linguistic method, where a set of handcrafted linguistic rules are assigned and the tagger tags each sentence based on these handcrafted rules. Although there are many researches and available tools to perform POS tagging for English text, only very limited applications have been developed to tag Malay language. Also, the available developed Malay POS taggers only work on grammatically correct structure.

In 2011, Mohammed, H., developed a stochastic approach based POS tagger to tag Malay text (Mohamed *et al.*, 2011). This approach consists of three parts. The first part is to manually tag the words in a sentence based on the corpus, followed by prediction of unknown words based on affixation. The last part was to check and manually correct the wrongly tagged word based on Malay Morphology. The Malay tagged corpora have been developed consisting of the tag set that had been defined in bilingual dictionary with some minor changes to suit the purpose of tagging Malay texts (Hock, 2009). Since these corpora are manually tagged, an extensive need of labor is required to perform the annotation process. Affixation method is used to tag the unknown words in the second part. In Malay language each affixation contributes to a particular POS category. With this characteristic, in this approach affixation in an unknown word is utilized to identify the corresponding POS category. While a probability distribution formula of HMM model is used to check whether it is the accurate POS category for the given word. One of the main disadvantages of this method is in the third part where input from linguistic experts are needed to correct the wrongly tagged words.

Alfred developed Rule Based POS (RPOS) tagger to tag Malay texts (Alfred *et al.*, 2013). This method has two parts, where the first one checks the word in the existing dictionary. The dictionary for this RPOS tagger is extracted from Thesaurus Bahasa Malaysia (Thesaurus, 2008). When the tagger gets a sentence as an input, it checks with the dictionary. When the word is present in the dictionary its designated tag as per in the dictionary will be given to the word. If a word does not exist inside the dictionary then the word will go through the affixation rules to identify its proper POS category (Karim *et al.*, 2008). Once all the words in the sentence is tagged, then the tagged sentence will be crosschecked with the word relation rules. Here with the word relation rules ambiguity will be distinguished. This tagger has obtained an overall accuracy of 89.2% for news articles and 86% for Biomedical Articles which is in proper grammar structure.

Based on review, both stochastic and rule based methodologies have its method of tagging and solving unknown and ambiguous problem. In term of performance, both of the methods perform well in proper Malay text. In this study we are comparing rule based approach by Rayner and stochastic approach by Muhammad to see which method is suitable for online text which contains abbreviation, improper grammar or “BahasaRojak”.

System Description

This section describe the proposed Malay POS tagging framework. We have developed two frameworks. The first framework is based on stochastic approach and the second utilizing rule based methodology.

Stochastic Tagger

Stochastic tagger framework as shown in Fig. 1 is the enhanced framework of Mohammad’s stochastic tagger. The framework is divided into four modules which are pre-processing, frequent labelling module, feature templates and transition probability module. Input and output of the four modules are shown in Fig. 2. All the text used in this research has more than 70% words in Malay. The text extracted online will be process in pre-processing module. This module will clean the text extracted. Each word in the sentence will be checked and corrected. Texts are checked for abbreviation and spelling errors. Abbreviation words are expanded and spelling errors are corrected as shown in Fig. 2, part a. The word "Nape" is an abbreviation word. It will be expanded to "Kenapa". Besides spelling and abbreviation correction, this module also checks for unwanted symbols and repeated letters. These unwanted symbols and words will also be removed from the sentence.

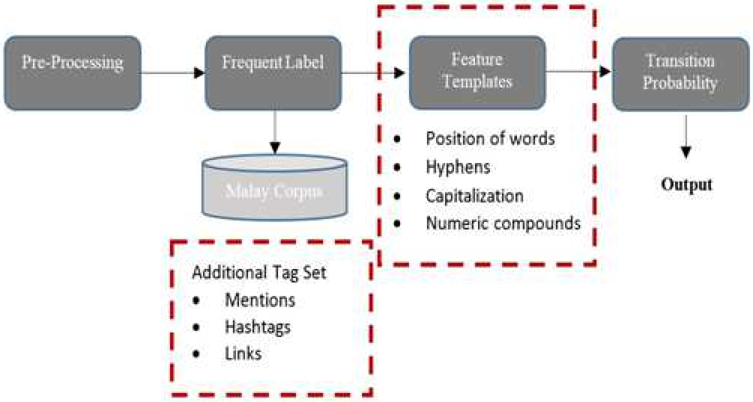
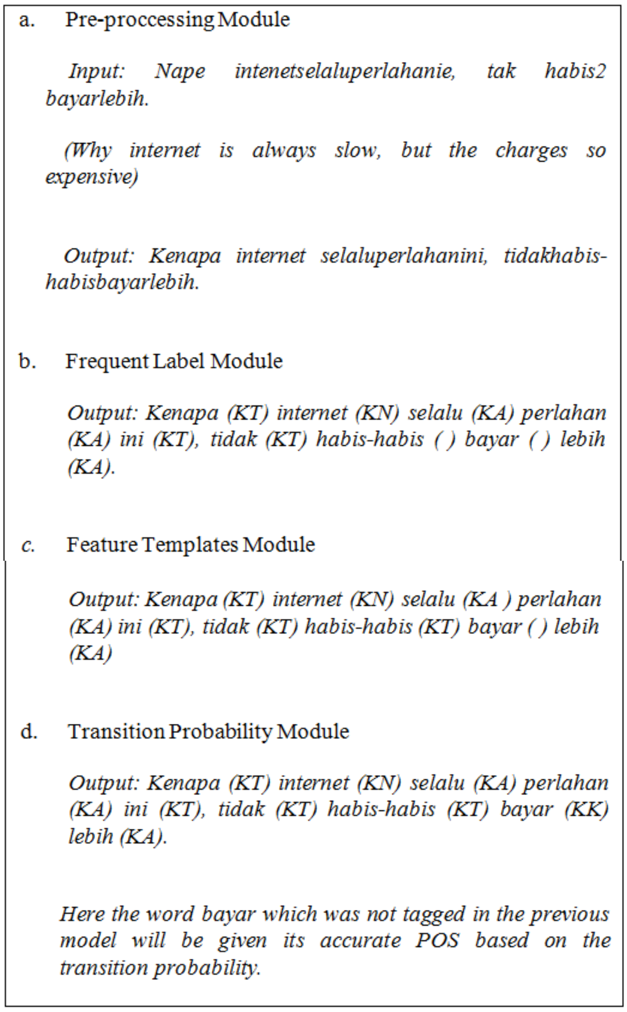


Fig. 1. Stochastic tagging framework



a. Pre-proccessing Module

Input: Nape intenetselaluperlahanie, tak habis2 bayarlebih.

(Why internet is always slow, but the charges so expensive)

Output: Kenapa internet selaluperlahanini, tidakhabis-habisbayarlebih.

b. Frequent Label Module

Output: Kenapa (KT) internet (KN) selalu (KA) perlahan (KA) ini (KT), tidak (KT) habis-habis ( ) bayar ( ) lebih (KA).

c. Feature Templates Module

Output: Kenapa (KT) internet (KN) selalu (KA) perlahan (KA) ini (KT), tidak (KT) habis-habis (KT) bayar ( ) lebih (KA)

d. Transition Probability Module

Output: Kenapa (KT) internet (KN) selalu (KA) perlahan (KA) ini (KT), tidak (KT) habis-habis (KT) bayar (KK) lebih (KA).

Here the word bayar which was not tagged in the previous model will be given its accurate POS based on the transition probability.

Fig. 2. Example processing of a sentence

Table 1. Sample entries of corpus

Word	Probability 1	Probability 2
Kenapa (*Why*)	KT (*Adverb*)	KN (*Noun*)
lebih (*more*)	KA (*Adjective*)	KT (*Adverb*)

The cleaned sentence from pre-processing module will be passed to the frequent labelling module. In this module, each word in the sentence will be tagged. That is, each word will be cross checked with the Malay corpus and assigned with the most frequent tag for that particular word. Corpus is collection of words labelled with their most frequent tags. Each word in the Malay corpus may have more than one probable tag. If a word has more than one tag, the probability equation (Equation 1 and 2) will be used to determine the highest probable tag. Table 1 shows the example of the Malay corpus. The word "Kenapa" can be tagged as KT (Adverbs) or KN (noun). Based on the equation, the word "Kenapa" is tagged KT as shown in Fig. 2, part b. If a particular word does not exist in the corpus (unknown words) or special characteristics words it will not be given any label, this word will be later tagged in feature template or transition probability module.

The next module is feature template module which contains tag for special characteristics words like hyphens, capitalization, numeric compounds. All the words that is not tagged in frequent labelling module will be passed to this module. This module will find all the special characteristics words and tagged these words. For instance the word habis-habis which contain hyphen, will be tagged as KT as shown in Fig. 2 part c. The tagged sentences will then be passed to transition probability module.

Transition probability module contains rules of tagging. Therefore it is able to tag unknown words and also recheck the overall tagging in the sentences. Unknown words which were not tagged in the previous 2 modules, will be tagged based on rules the position of words in the sentences. For example the word "bayar" is between of "KT and KA", therefore it is tagged as "KK" as shown in Fig. 2 part d. Also, in this module, the tagging of whole sentence is rechecked based on the rules. If the word was wrongly tagged, the word will be corrected based on the sequence of tag. Therefore the accuracy of tagging can be increased and at the same time ambiguity of the words can be reduced. The output from this module acts as the output for the whole tagger as shown in Fig. 2:

$$\frac{P\left(most\ freq\ tag\ for\ word\right)}{Total\ number\ of\ words} \tag{1}$$

$$\left(next\ tag\right) \leftarrow P\ \left[\left(current\ tag\right)\right] \tag{2}$$

Rule Based Tagger

Rule based tagger framework as shown in Fig. 3 is the enhanced framework of Alfred Rayner's RPOS tagger. The framework is divided into pre-processing, frequent labelling, unknown word identification and rule check module. The functionality of first and the second modules are the same as in stochastic approach. In pre-processing module, the texts will be cleaned and in frequent label module, each word in the sentence is crossed checked with the corpus and assigned with the most frequent tag. The only difference this module with stochastic approach is the structure of corpus. Here the corpus has only one designated tag for each word whereas in stochastic approach each word can have more than one tag as shown in Table 2. In rule based approach the tagging depends on the rule set that have been created where each word have only one tag assigned to it. On the other hand, in stochastic approach, tagging of the words depends on the probability condition set where it can have more than one tag for a word. The output of frequent labelling module is as shown in Fig. 4 part b. If a word does not exist inside the corpus then it will go through the unknown word identification.

In the unknown word identification module, unknown words in the sentence will be tagged based on the affixation and word ending rules. For example if an affix is present in the beginning of a word then that particular word will be tagged as a verb (KK) and if the affix is present at the end of the word then the word will be tagged as an adverb (KT). In Fig. 4 part b, the word "bayar" does not have affixation in the beginning or end of the word; therefore it will be tagged as noun (KN). Besides that, this module will also tag special characteristics word. The output of this module is shown in Fig. 4 part c.

The last module is the rule check module. This module will recheck the overall tagging in the sentences with the linguistic rules. If the tagging is same with the linguistic rules no changes will be made. On the other hand if there is a difference, the tagging of the word will be evaluated based on the condition and linguistic rules. As shown in Fig. 4 part d, the word "bayar" was tagged as a noun in unknown word identification module. Based on condition and linguistics rules "bayar" was changed to "KK" in this module. This module helps to overcome or reduce the ambiguous problem and increase the overall accuracy of the tagging. The output of this module acts as the output for the whole system as shown in Fig. 4 part **d.**

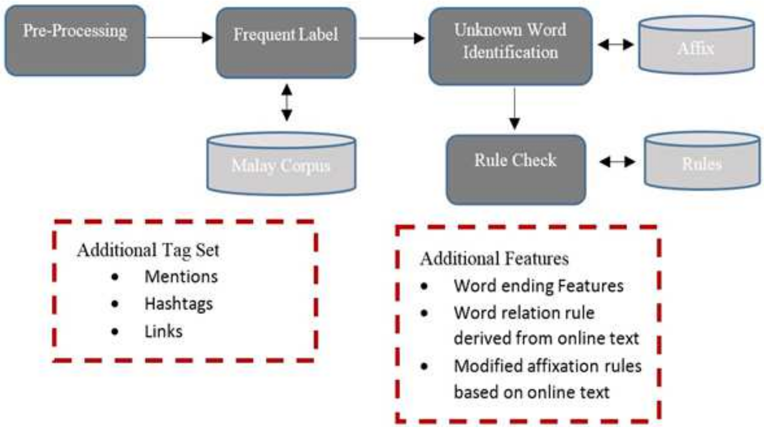
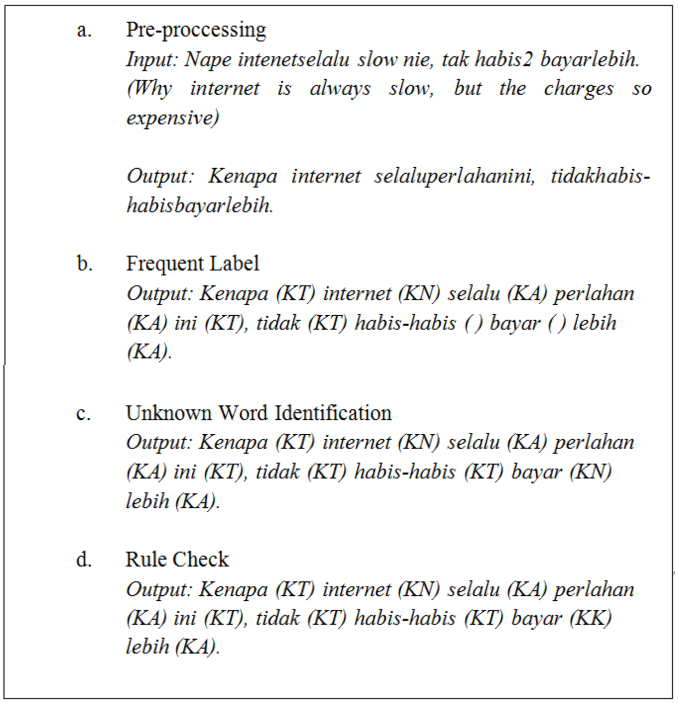


Fig. 3. Rule based tagging framework



a. Pre-proccessing
Input: Nape intenetselalu slow nie, tak habis2 bayarlebih. (Why internet is always slow, but the charges so expensive)

Output: Kenapa internet selaluperlahanini, tidakhabis-habisbayarlebih.

b. Frequent Label
Output: Kenapa (KT) internet (KN) selalu (KA) perlahan (KA) ini (KT), tidak (KT) habis-habis ( ) bayar ( ) lebih (KA).

c. Unknown Word Identification
Output: Kenapa (KT) internet (KN) selalu (KA) perlahan (KA) ini (KT), tidak (KT) habis-habis (KT) bayar (KN) lebih (KA).

d. Rule Check
Output: Kenapa (KT) internet (KN) selalu (KA) perlahan (KA) ini (KT), tidak (KT) habis-habis (KT) bayar (KK) lebih (KA).

Fig. 4. Example processing of a sentence

Table 2. Sample corpus entry

Word	Tag
Kenapa (*Why*)	KT (*Adverb*)
lebih (*more*)	KA (*Adjective*)

Evaluation

To evaluate the accuracy of stochastic and rule based tagger of unknown word, 8-fold cross-validation has been used. About 500 tweets are extracted from social media and pre-processed. About 70% of these tweets are in Malay and the rest are in "BahasaRojak". The tweets were split into 8 parts and 8 test sets were created from them. The first test set uses parts 2 through 8 for training and part 1 for tagging, the second test set is trained on parts 1 and 3 through 8 and is used for tagging part 2 and so on. Table 3 shows the testing average output over the 8 test sets for rule base with stochastic tagger.

The unknown word ratio denotes the average ratio of unknown words relative to the number of words in the tweets. The column tagging from corpus (known words) is the words in the corpus and unknown words are the words that are not in the corpus. Correctly tagged word means that the ambiguous words are given the right tag. Overall accuracy denotes the combined accuracy of known, unknown words and also disambiguated words over the whole tweets.

Table 3. Accuracy of taggers on test set

	Accuracy				
Taggers	Unknown word ratio	Tagging from corpus (%)	Unknown word (%)	Correctly tagged words (ambiguity) (%)	Overall accuracy (%)
Rule based tagger	6.8	100	89.90	93.40	92.90
Stochastic tagger	6.8	100	85.60	92.10	91.40

The total number of words from 500 tweets were 5850 words after preprocessing. Out of these 5850, around 397 words are unknown words. That is, around 6.80% words are unknown words. Both rule based and stochastic tagger achieved 100% accuracy for tagging known words. For unknown word, average accuracy achieved by rule based approach is 89.9% that is around 5% higher than stochastic tagger which achieved 85.6% accuracy. In tagging of ambiguous words, there is no significant difference between average accuracy of rule based (93.4%) and stochastic (92.1%) approaches. For the overall average accuracy, rule based tagger achieved 92.9%, which is slightly better (around 2.1%) than stochastic tagger which is 91.4%.

Discussion

In this study, we have not only compared the two classes of POS tagging approaches but also developed two POS taggers for Malay language. The developed taggers are enhancement of Muhammad's stochastic and Rayner's rule based tagger to tackle online texts. The structure and grammar of online Malay text is much different compared to proper Malay text. Also, these developed taggers are added with additional tag set such as mentions, hashtags, links, abbreviations and locations. The stochastic tagger is further enhanced with additional feature templates such as position of the word in a sentence, occurrence of hyphens, capitalization and numeric compounds. For rule based method, the affixation rules are extended and rule check module is added. With this enhancement, both of the developed taggers are capable of handling unknown word and ambiguity which are two undeniable issues faced in any POS tagging application.

Both stochastic tagging and rule based tagging method has different approaches in terms of handling unknown words and ambiguity. In stochastic approach, unknown words are tagged based on sequences of tags using transition probability. Whereas affix analysis is used to tackle unknown word issue in rule based approach. Based on the result in Table 3, rule based method is better in tackling unknown word problem compared to stochastic approach. To solve the ambiguity issue, stochastic tagger assigned all the probable tag for a words in the corpus. Besides corpus, transition probability also acts as a second phase checking for ambiguity. Whereas in rule based approach, ambiguity is tackled using the rule check module. Even though both methods perform well in solving ambiguity issue, for Malay language which is rich in morphology, rule base approach is a better solution. Rule based approach utilizes the linguistic rules which give more information to overcome ambiguity compared to stochastic module which solely depends on probability equation.

Conclusion

In this study we have developed and compared two types of POS tagging methods for dealing with unknown words and ambiguity in online Malays texts. Online texts contain lot of abbreviation and also most of Malaysian tends to use "BahasaRojak". To handle these problems, the existing stochastic and rule based tagger have been enhanced. Both of the enhanced taggers perform well on online Malay text. These algorithms are able to identify and tag unknown words and also disambiguate words well. Although both methods perform well, rule based method is better approach to tag Malay language which is rich in morphology compared to stochastic method. With further training and testing this algorithm can be further enhanced to achieve better accuracy. In future, combining both the said enhanced methods, developing a hybrid POS tagging module for online Malay text will also be looked into.

Acknowledgement

This work has been supported by TMR&D Grant No. MMUE/140065.

Author's Contributions

Kalaiarasi Sonai Muthu Anbananthen: Introduction, Methodology, Experiment and evaluation, Discussion, Conclusion.

Jaya Kumar and Mohd. Shohel Sayeed: Proof Reading and Literature review.

Praviny Muniapan: Introduction, Difference in English and Malay language, Literature review, System description, Experiment and evaluation.

Ethics

This article is unique and contains unpublished material. The corresponding author confirms that all of Co-authors have read and approved the manuscript and there are no ethical issues involved.

References

Alfred, R., A. Mujat and J.H. Obit, 2013. A Ruled-Based Part of Speech (Rpos) Tagger for Malay Text Articles. In: Intelligent Information and Database Systems, Selamat, A., N.T. Nguyen and H. Haron (Eds.), Springer, Berlin, ISBN-10: 3642365434, pp: 50-59.

Aranzabe, M.J., A.D. De Ilarraza, N. Ezeiza, K. Bengoetxea and I. Goenaga *et al*., 2012. Combining rule-based and statistical syntactic analyzers. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, (ACL' 12), Jeju, Republic of Korea, pp: 48-54.

Bakar, J.A., K. Omar, M.F. Nasrudin and M.Z. Murah, 2013. Morphology analysis in malay analysis in malay pos prediction.

Brill, E., 1992. A simple rule-based part of speech tagger. Proceedings of the Workshop on Speech and Natural Language, Feb. 23-26, Harriman, New York, pp: 112-116. DOI: 10.3115/1075527.1075553

Crystal, D., 2004. The Cambridge Encyclopedia of the English Language. 1st Edn., Cambridge University Press, Stuttgart, ISBN-10: 312533862X, pp: 499.

Frawley, W., 2003. International Encyclopedia of Linguistics: AAVE-Esperanto. 1st Edn., Oxford University Press, Oxford, ISBN-10: 0195139771, pp: 525.

Gungor, T., 2010. Part of Speech Tagging. In: Handbook of Natural Language Processing, Indurkhya, N. and F.J. Damerau (Eds.), CRC Press, ISBN-10: 142008593X, pp: 704.

Hassan, A., 1974. The morphology of Malay. 1st Edn., Dewan Bahasa Dan Pustaka, Kementerian Pelajaran Malaysia, Kuala Lumpur, pp: 292.

Karim, N.S., F.M. Onn, H.H. Musa and A.H. Mahmood, 2008. Tatabahasa Dewan Edisi Ketiga. Dewan Bahasadan Pustaka, Kuala Lumpur.

Kupiec, J., 1992. Robust part-of-speech tagging using a hidden Markov model. Comput. Speech Language, 6: 225-242. DOI: 10.1016/0885-2308(92)90019-Z

Kumar, M., S.R. Anand and K. P. Soman, 2015. Cross-lingual preposition disambiguation for machine translation. Procedia Comput. Sci., 54: 291-300. DOI: 10.1016/j.procs.2015.06.034

Kumawat, D. and V. Jain, 2015. Pos tagging approaches: A comparison. Int. J. Comput. Applic., 118: 32-38.

Manning, C.D., M.J. Surdeanu, J. Finkel, S.J. Bethard and D. McClosky, 2014. The Stanford CoreNLP natural language processing toolkit. Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, (LSD' 14), At Baltimore, Maryland, pp: 55-60. DOI: 10.3115/v1/P14-5010

Manning, D.K.C.D., 2003. Natural Language Parsing. In: Advances in Neural Information Processing Systems, Becker, S., S. Thrun and K. Obermayer (Eds.), Morgan Kaufmann Publishers, San Mateo, ISBN-10: 0262025507, pp: 1687-1687.

Mohamed, H., N. Omar and M.J.A. Aziz, 2011. Statistical malay Part-Of-Speech (POS) tagger using Hidden Markov approach. Proceedings of the International Conference on Semantic Technology and Information Retrieval, Jun. 28-29, IEEE Xplore Press, pp: 231-236. DOI: 10.1109/STAIR.2011.5995794

Nand, P. and R. Perera, 2015. An evaluation of POS tagging for tweets using HMM modelling. Proceeding of the 38th Australasian Computer Science Conference (CSC' 15), Sydney, Australia, pp: 83-89.

Hock, O.Y., 2009. Kamus dwibahasa edisi kedua. Pearson Longman, Malaysia.

Samuelsson, C. and A. Voutilainen, 1997. Comparing a linguistic and a stochastic tagger. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, Jul. 07-12, Madrid, Spain, pp: 246-253. DOI: 10.3115/976909.979649

See, C.M., 1980. The morphological analysis of Bahasa Malaysia. Proceedings of the 8th Conference on Computational Linguistics, Sept. 30-Oct. 04, Tokyo, Japan, pp: 578-585. DOI: 10.3115/990174.990281

Silva, A.P., A. Silva and I. Rodrigues, 2013. A new approach to the POS tagging problem using evolutionary computation. Proceedings of Recent Advances in Natural Language Processing, Sept. 7-13, Hissar, Bulgaria, pp: 619-625.

Thesaurus, 2008. Thesaurus Bahasa Melayu, New Edition Kuala Lumpur, Dewan Bahasadan Pustaka.

Weischedel, R., R. Schwartz, J. Palmucci, M. Meteer and L. Ramshaw, 1993. Coping with ambiguity and unknown words through probabilistic models. Computat. Linguist., 19: 361-382.