Implementation of Kadazan Tagger Based on Brill's Method

Marylyn Alex & Lailatul Qadri Zakaria

CAIT Research Group, Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, Jalan Tun Ismail Ali, 43600 Bangi, Malaysia
Email: alexmarylyn@gmail.com

Abstract. We present and evaluate the implementation of Part of Speech (POS) Tagging for the Kadazan language by using the Transformation-based approach. The main purpose of this study is to develop an automatic POS tagging for the Kadazan language, which had never, been developed before. POS tagging can tag the Kadazan corpus automatically and can help reduce the disambiguation problem of this language. The implementation of this approach in this study is to achieve a better and higher accuracy or at least similar to that of the other tagging approaches such as the statistical and the original rule-based approach. This approach can transform the tags based on the prescribed set of rules. A number of objectives were set in order to achieve the main purpose of this study. Firstly, to apply the lexical and contextual rules for this language. Secondly, to implement the Brill's algorithm based on the set of rules and finally to determine the effectiveness of the Kadazan Part of Speech by using this approach. The tagging system had been trained using four Kadazan corpuses containing 5663 words in all. Based on the evaluation results, the tagging system had achieved around 93% accuracy.

Keywords: *brill's tagger; kadazan language; Part of Speech tagger; rule-based; statistical; transformation-based.*

1 Introduction

Part of Speech (POS) tagging is a system that read the text in some languages and assign POS such as noun, verb, adjective, adverb, pronoun, *etc*. to every word in the text (corpus). The tagging process could be linked with morphological process such as the formation of adjective from verb. For example, the word *'charm'*, which is tagged as a verb, could be transformed into an adjective when *'ing'* is suffixed to that word which would then become *'charming'*. In Natural Language Processing, POS tagging is important in order to show how the words could be related to each other and how the ordered structures of the sentence could help resolve the ambiguity problem in different kinds of analysis levels. POS tagging had been used in many applications such as machine translation, speech recognition, information retrieval, dictionary (Wordnet) and so on. Hence, the importance of POS tagging cannot be ignored

Received September 23rd, 2013, Revised October 1st, 2013.
 ISSN: 2337-5787, DOI: 10.5614/itbj.ict.res.appl.2013.7.3.1

at all. POS tagging is quite difficult because of the ambiguity problem. Ambiguity is defined as the existence of two or more possible meaning in one word or in a sentence that can confuse the reader. However, the ambiguity problem could be reduced through the tagging process where the words could be tagged based on their meaning. A few different approaches could be applied to the POS tagging. The first method that was known to have implemented POS tagging was the rule-based method [1]. Then, statistical technique came into existence after 1980's and was found to have obtained more popularity. Then came the Brill rule-based system that was used in 1992 and quite different from the original rule-based method [2]. This tagger had been trained in tagging for the English words by using the Transformation-Based approach and was found to have achieved 97% accuracy as a result [2].

Rule-based approach acts by assigning tags to a word using contextual information according to the rules developed by human. Statistical approach is known as a stochastic tagger when it disambiguates the word based on probability and where the word occurs with a specific tag. The tag occurs frequently in the training set is the one to be assigned to the ambiguous instance of that word. However, statistical approach requires complex computation.

Over the past few years, statistical approach were thought to be the most successful method compared to the original rule-based method [3], until after the introduction of the Brill's tagger approach in 1992. Brill's tagger approach is the advance version of rule-based method. After 1992, most of the researches who used the original rule-based method referred to Brill's method [1]. Brill's method uses the rule templates and is easier to implement compared to the statistical approach because complex computation is not required.

Brill's approach was originally developed for the English language and was found to have achieved high accuracy compared to the original rule-based and statistical approach. There were few other languages using the Brill's approach for POS tagging and the accuracy obtained were also high. In this paper, we are trying to implement the Brill's approach to see the average of accuracy in applying this method for the Kadazan language. It is done by evaluating the tagging accuracy and the tagging performance. This study was carried out to develop the Kadazan POS tagger for the Kadazan text automatically and at the same time to observe and determine the effectiveness of the Brill's approach for the Kadazan language.

2 Brief Overview of Kadazan Language

Kadazan language is a language which is spoken by the Kadazan race in Borneo along the region from the Nosoob-Kepayan area through Penampang-Putatan

and to Papar, Sabah. Same as any other languages, the Kadazan language also has its own characteristics, grammatical structures and rules. The morphology of the Kadazan language could be formed from the affixes of a word. For example, in the Kadazan language, the formation of noun from an adjective involves all three affixes either by prefixing, infixing, suffixing or the combination of prefix and suffix. For example, the word 'avasi' (good) => '**k**avasi**an'** (goodness). The word 'avasi', which is an adjective, is changed to noun after the 'k' is prefixed and the 'an' is suffixed to that word. Another example is the word 'poit' (bitter) => 'p**in**oit' (bitterness). The word 'poit', which is an adjective, is changed to a noun after the *'in'* is infixed after the first character of the word 'poit'. The next example is by prefixing 'mang' at the beginning of the word. For example, 'ajal' (teach) => '**mang**ajal' (teaching). The word 'ajal', which is a verb, is changed to noun after the word 'mang' is prefixed to that word, which is 'mangajal'. The contextual rules is based on 'previous' word and 'next' word. It requires condition. For example, 'change noun to verb if the previous tag is AISO'. If the sentence is tagged wrongly such as 'aiso **louti (adjective)**', which means 'no bread', by applying the rule as mentioned before, the sentence will be tagged as 'aiso **louti (noun)**'.

3 Related Work

There are few different approaches for POS tagger besides the Brill's Tagger. The two well known approaches are the rule-based approach and the statistical approach. The rule-based approach was the first method introduced by [4]. In general, rule-based approach are rules written by humans based on linguistic knowledge. They were done by generating the input sentence to the output text in the basic morphological, syntactical and semantic analysis from both sources and the target languages that were involved in the tagging or translation. This approach usually depends on the dictionary and human to tag the words. The related work [5] had developed a POS Tagger for the Pashto language using the rule-based approach and had achieved a 88% accuracy. To use this POS Tagger approach, the first step is to input all the raw text into the system and the tokenization process would then take part. The lexicons taken from newspapers, books or from Internet are then used to tag the text. All the words taken from the sources would be extracted and tag manually. Any new words not in lexicon would be tagged using the rules written by human. The output would be checked manually and correction made for any wrong tags found in the text.

The second method that became popular after the original Rule-based approach was the statistical approach. This approach disambiguated the words depending on the probability of the word that occurred with a particular tag. The most frequented tag that occurred in the training set was the one assigned to the ambiguous instance of that word. The probability of the given sequence of tags

occurred would be calculated. One of the statistical methods that had been used widely was the Hidden Markov Model (HMM). In this method, the lexical and the probabilities were used to search a tag for a word. Hence, the statistical approach requires a complex computation. One of the related works that used the statistical approach for tagging is [6]. The Hidden Markov Model (HMM) were used to tag the Manipuri language and had achieved a 92% accuracy as a result. First of all, the input text would be split into words and the stemmer process would then be applied to separate the affixes from each word. The statistical analyzer were used to extract the unigram and the bigram probability from the tag corpus. The most likely tag for every word in the text would be determined and the tags having the highest probability for each word would be chosen.

After the statistical method, the Brill's Tagger method was then introduced and it was found to have a similar or even better accuracies than those of the two above mentioned methods introduced earlier. The transformation-based approach also known as the Brill's approach acted by transforming the tags based on the applied rules. In this approach, a tag would be assigned to each word and then transformed by using a set of rules. The rules would be applied repeatedly to transform the tags until no more rules were left to be applied. Furthermore, this method was also known as self learning where a comprehensive technique called TEL and the rule templates were used instead of the pure n-gram. In terms of its linguistic accessibility and flexibility, it is defined by linguistic knowledge to be statistically investigated. The Brill's Tagger would first set the rule templates before they were allowed to change the rule files so as to help analyze the results and to highlight the remaining corpus. One of the related works that used the Brill's approach was [7]. The Brill's approach were used to tag the Greek language and were found to have achieved 95% accuracy. At first, all the words in the corpus were tag to their tag based on the most likely tags in the lexicon. The lexicon was taken from the Penn Tree Bank of the Wall Street Journal and the Brown corpus. After all the words were tagged to their initial tag. The lexical and the contextual rules were then applied to transform and to correct the wrong tags. The explanation of the approach in detail could be found in the following section. Besides the English and the Greek languages, other languages also implemented this method for tagging. For example, the Brill's Tagger approach had been used for the German language [8]. It showed how this approach could improve the tagging performance by increasing the size of the corpus. The error rates could be reduced by adding more rules and by trying to search for unknown words in the lexicon. Furthermore, this approach had also been used for the Hungarian language [2], which has a rich morphology, agglutinative with free word orders as compared to the English language. The accuracy achieved was around 85% - 88%. Brill's Tagger had also been used to train the Danish language [9] for

checking grammar errors, incorrect commas and incorrect article noun agreement. Besides tagging, this approach could also be used to identify scores of grammar errors in the Danish and other languages. Based on the overall discussion, Brill's approach was chosen to develop a Kadazan POS tagger due to better performance as shown in the results for tagging different languages.

4 Brill's Tagger for Kadazan

Figure 1 shows the overall model for Kadazan POS Tagger based on Brill's approach.

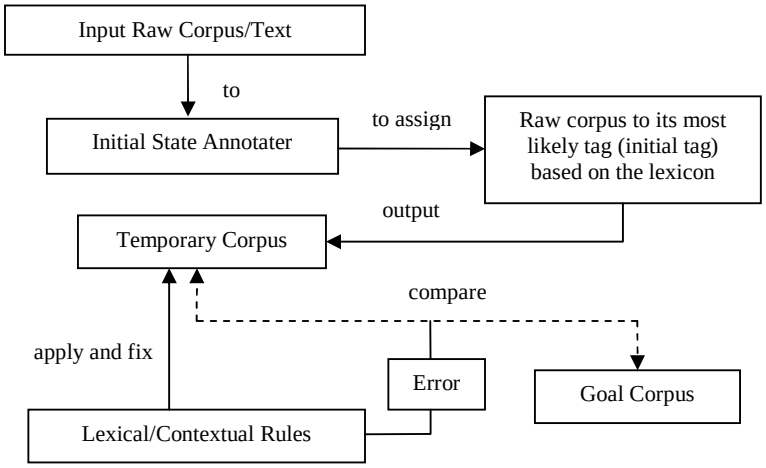

Figure 1 Kadazan POS tagger model based on Brill's approach.

The Kadazan POS tagger is divided into four phases. The first phase of tagging begins by inputting raw text into the system. The second phase continues when the corpus is going through the initial state annotator to tag all the words to its most likely tag based on the lexicon. The output of this process is the temporary corpus. The other possible tags act as the second tag if and only the initial tag is wrong. The rules are then be applied to change the initial tag (most likely tag) to one of the other possible tags. The word with possible tags shows that it is ambiguous because it has more than one meaning. For example, the word 'kalaja' can be classified into a verb or a noun. For example, if the sentence goes like this, 'mamaso isido monoodo **kalaja** do kabaahan', which means, 'he is in the middle of doing an artisan work.' From here, the correct tag for 'kalaja' is a verb. If the sentence goes like this, 'onu oh **kalaja** diozu?', which means, 'what is your job'?, the correct tag for the word 'kalaja' in this sentence is a noun. However, not every word has other possible tags as shown in Table 1.

4.1 The First Phase

The first phase of tagging begins by inputting an annotated text into the system. Figure 2 shows the diagram of the first phase.

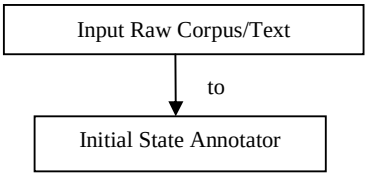

Figure 2 First phase of tagging.

4.2 The Second Phase

The second phase of tagging begins when the input text go through the initial state annotator to tag all the words inside the corpus to its most likely tag which is given in a lexicon. The words not in the lexicon, are considered as unknown words. The unknown words would be tagged automatically as noun (N). Table 1 shows the example of the lexicon for most likely tag.

Table 1 Examples of the lexicon.

Word in Kadazan	English Translation	Most Likely Tag	Other Possible Tags
Aanangaan	Like	V	R
Kalaja	Work/Working	V	N
Ahasu	Warm/ Hot	J	-
Kampil	Bag	N	-

Based on Table 1, the word 'Aanangaan' is usually tagged as a verb (V) but may also be tagged as an adverb (R) depending on the sentence. The word 'Kalaja' is usually tagged as a verb (V) but can also be tagged as a noun (N). The word 'Ahasu' is usually tagged as an adjective (J) and there is no other possible tags for that word. The word 'Kampil' is usually tagged as a noun (N) and there is no other possible tags for that word too. All these depend on the structure of the sentence. The same thing applies to other words in the lexicon. The output of this phase is temporary corpus in every single word in the corpus had been tagged to their initial tags. Figure 3 shows the diagram of the second phase.

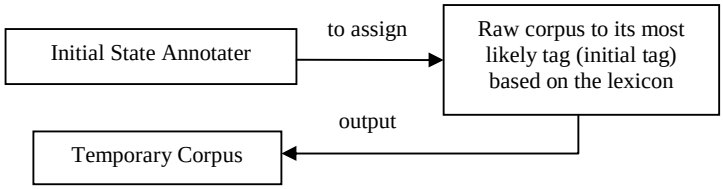

Figure 3 Second phase of tagging.

4.3 The Third Phase

The third phase continues by comparing the temporary corpus with the goal corpus to detect if there is any error (wrong tags) that occurs in the temporary corpus. The goal corpus is the manually tagged corpus. For example, if the sentence in Kadazan goes like this 'onu oh **kalaja** diozu?' meaning 'what is your job?'. In temporary corpus, the word 'kalaja' is tagged as a verb as an initial tag based on the lexicon for its most likely tag as shown in Table 1. However, in the goal corpus, the correct tag for the word 'kalaja' is a noun. So, there is an error detected after both sentences or corpuses were compared to each other. Figure 4 shows the diagram of third phase of tagging.

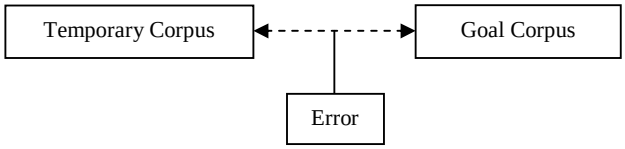

Figure 4 Third phase of tagging.

4.4 The Fourth Phase

The last phase continues when the lexical and the contextual rules are applied to fix the errors, which occurred from the third phase. The lexical rules are based on prefixes, infixes and suffixes of the word. Usually, the lexical rules only affect the unknown words. However, the contextual rules are the rules that transform and correct the wrong tags based on the 'next' or 'previous' word in the sentence. Figure 5 shows the tagging diagram of phase four.

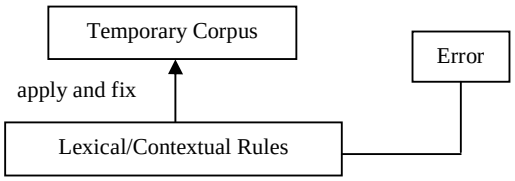

Figure 5 Fourth phase of tagging.

Table 2 shows the examples of lexical rules for Kadazan language.

Table 2 Examples of the Lexical Rules.

Rule	Current Tag	New Tag	Condition
L1	J	N	Prefix ‘in’
L2	J	N	Infix 'in'
L3	J	N	Prefix ‘k’ and Suffix ‘an’

Rule L1 stated that if the current tag of a word is an adjective (J), after prefixing ‘in’ to that word, the word is transformed into a noun (N). For example, agang (J) => prefix 'in' => **in**agang (N). The word 'agang (red)', which is an adjective, is changed into '**in**agang (redness)', which is a noun, after prefixing 'in' into it. Rule L2 stated that if the current tag of the word is an adjective (J), after infixing 'in' into it, it will be tagged as a noun (N). For example, vasi (J) => infix 'in' => v**in**asi (N). The word 'vasi (good)', which is an adjective, is changed into 'v**in**asi (goodness)', which is a noun, after infixing 'in' into it. Rule L3 stated if the current tag of the word is an adjective (J), after prefixing ‘k’ and suffixing ‘an’ to that word, it would be transformed into a noun (N).

For example, avasi (J) => prefix 'k', suffix 'an' => **k**avasi**an** (N). The word 'avasi (good)', which is an adjective, is changed into '**k**avasi**an** (goodness)' which is a noun after prefixing 'k' and suffixing 'an' into it. The same case applies to other words based on their lexical rules. Next is by learning the contextual rules. Table 3 shows the examples of contextual rules.

Table 3 Examples of Contextual Rules.

Rule	Current Tag	New Tag	Condition
C1	J/R/V	N	Previous word is 'aiso'
C2	J/R/V	N	Previous word is 'i'
C3	J/R/V	N	Next word is 'togumu'

Rule C1 stated that if the previous word is ‘aiso’, then the word after 'aiso' where the current tag is either a verb (V) or an adjective (J) or an adverb (R) would be transformed into a noun (N). Rule C2 stated that if the previous word is 'i', then the word next to that word where the current tag is either a verb (V), an adjective (J) or an adverb (R) would be transformed into a noun (N). Rule C3 stated that if the next word is ‘togumu’, the word before 'togumu' where the current tag is either a verb (V), an adjective (J) or an adverb (R) would be transformed into a noun (N). The same thing applies to other words based on their contextual rules.

5 Results and Evaluation

The evaluation of the Kadazan POS Tagging had been carried out and the results are shown here for further discussion in this section. The tagger had been evaluated in two ways. Firstly, by getting the accuracy and error rate by applying the lexical rules and contextual rules and by combining both rules without using any other rules. Secondly, by calculating the tagging time and the evaluation time to calculate the accuracy.

5.1 Training the Kadazan Tagger

The Kadazan children's story books were used to train the tagger. There are four corpuses that were used where the first corpus entitled *'Mobubuvat, Laja' Om Raani'* contains 741 words (corpus 1), the second corpus entitled *'I Duok Om I Vuhan'* contains 901 words (corpus 2), the third corpus entitled *'Zi Ombong-Ombong'* contains 1328 words (corpus 3) and the fourth corpus entitled *'Zi Osong Om I Vuhanut'* contains 2693 words (corpus 4).

Table 4 Tagging results for manually tag compared to system tag by applying lexicon, lexical, contextual and both rules (for corpus 1).

Tags	Lexicon only	Lexical Rules + Lexicon (%)	Contextual Rules + Lexicon (%)	Lexicon + Contextual + Lexical Rules (%)
Correct	670	675	693	695
Wrong	71	66	48	46
Accuracy (%)	90.42	91.09	93.52	93.79
Error Rate (%)	9.58	8.91	6.48	6.21

Based on Table 4 (corpus 1), which contains 741 words, by using the lexicon and without applying any rules, we obtained 670 words for correct tags, 71 wrong tags and overall we obtained 90.42% accuracy and 9.58% error rate. By using the lexicon and applying the lexical rules only, we obtained 675 words for correct tags, 66 wrong tags and overall we obtained 91.09% accuracy and 8.91% error rate. By using the lexicon and applying the contextual rules only, we obtained 693 words for correct tags, 48 wrong tags and overall we obtained 93.52% accuracy and 6.48% error rate. Lastly, by using the lexicon and applying both lexical and contextual rules, we obtained 695 words for correct tags, 46 wrong tags and overall we obtained 93.79% accuracy and 6.21% error rate.

Based on Table 5 (corpus 2), which contains 901 words, by using the lexicon and without applying any rules, we obtained 839 words for correct tags, 62 wrong tags and overall we obtained 93.12% accuracy and 6.88% error rate. By using the lexicon and applying the lexical rules only, we obtained 834 words for

correct tags, 67 wrong tags and overall we obtained 92.56% accuracy and 7.44% error rate. By using the lexicon and applying the contextual rules only, we obtained 855 words for correct tags, 46 wrong tags and overall we obtained 94.89% accuracy and 5.11% error rate. Lastly, by using the lexicon and applying both lexical and contextual rules, we obtained 848 words for correct tags, 53 wrong tags and overall we obtained 94.12% accuracy and 5.88% error rate.

Table 5 Tagging results for manually tag compared to system tag by applying lexicon, lexical, contextual and both rules (for corpus 2).

Tags	Lexicon only	Lexical Rules + Lexicon (%)	Contextual Rules + Lexicon (%)	Lexicon + Contextual + Lexical Rules (%)
Correct	839	834	855	848
Wrong	62	67	46	53
Accuracy (%)	93.12	92.56	94.89	94.12
Error Rate (%)	6.88	7.44	5.11	5.88

Table 6 Tagging results for manually tag compared to system tag by applying lexicon, lexical, contextual and both rules (for corpus 3).

Tags	Lexicon only	Lexical Rules + Lexicon (%)	Contextual Rules + Lexicon (%)	Lexicon + Contextual + Lexical Rules (%)
Correct	1206	1200	1235	1224
Wrong	122	128	93	104
Accuracy (%)	90.81	90.36	93.00	92.17
Error Rate (%)	9.19	9.64	7.00	7.83

Based on Table 6 (corpus 3), which contains 1328 words, by using the lexicon and without applying any rules, we obtained 1206 words for correct tags, 122 wrong tags and overall we obtained 90.81% accuracy and 9.19% error rate. By using the lexicon and applying the lexical rules only, we obtained 1200 words for correct tags, 128 wrong tags and overall we obtained 90.36% accuracy and 9.64% error rate. By using the lexicon and applying the contextual rules only, we obtained 1235 words for correct tags, 93 wrong tags and overall we obtained 93.00% accuracy and 7.00% error rate. Lastly, by using the lexicon and applying both lexical and contextual rules, we obtained 1224 words for correct tags, 104 wrong tags and overall we obtained 92.17% accuracy and 7.83% error rate.

Based on Table 7 (corpus 4), which contains 2693 words, by using the lexicon and without applying any rules, we obtained 2284 words for correct tags, 409 wrong tags and overall we obtained 84.81% accuracy and 15.19% error rate.

By using the lexicon and applying the lexical rules only, we obtained 2283 words for correct tags, 410 wrong tags and overall we obtained 84.78% accuracy and 15.22% error rate. By using the lexicon and applying the contextual rules only, we obtained 2471 words for correct tags, 222 wrong tags and overall we obtained 91.76% accuracy and 8.24% error rate. Lastly, by using the lexicon and applying both lexical and contextual rules, we obtained 2459 words for correct tags, 234 wrong tags and overall we obtained 91.31% accuracy and 8.69% error rate.

Table 7 Tagging results for manually tag compared to system tag by applying lexicon, lexical, contextual and both rules (for corpus 4).

Tags	Lexicon only	Lexical Rules + Lexicon (%)	Contextual Rules + Lexicon (%)	Lexicon + Contextual + Lexical Rules (%)
Correct	2284	2283	2471	2459
Wrong	409	410	222	234
Accuracy (%)	84.81	84.78	91.76	91.31
Error Rate (%)	15.19	15.22	8.24	8.69

The second evaluation part is done by calculating the tagging time and the time to calculate the accuracy of the tagger. Table 8 shows the time for tagging and evaluation process for all the corpuses.

Table 8 Tag time and evaluation time for corpus 1, 2, 3 and 4.

Corpus	1	2	3	4
Size	741	901	1328	2693
Tag Time/s	1.0	1.1	1.4	1.6
Evaluation Time/s	44.6	55.1	124	301

Based on Table 8, the tagging time for corpus 1 is 1.0 seconds and the time taken to evaluate the accuracy is 44.6 seconds. The tagging time for corpus 2 is 1.1 seconds and the time taken to evaluate the accuracy is 55.1 seconds. The tagging time for corpus 3 is 1.4 seconds and the time taken to evaluate the accuracy is 124 seconds. The tagging time for corpus 4 is 1.6 seconds and the time taken to evaluate the accuracy is 301 seconds.

5.2 Results and Discussion

Based on the results in Table 4, Table 5, Table 6 and Table 7, we noticed that without applying any rule and by using the lexicon only, the accuracies of the tagging are lower compared to the accuracy after applying both rules. This proved that by applying both rules, the tagging accuracies could be improved. By applying the lexicon and lexical rules only, we could see that some of the

accuracies are higher or lower compared to the accuracies before the lexical rules were applied. Based on Table 4, by applying the lexical rules a higher accuracy was obtained than before the lexical rules were applied. However in Table 5, 6 and 7, the accuracies obtained by using the lexical rules are lower compared to the accuracy before the rules were applied. Lexical rules were applied for unknown words by prefixing, infixing or by suffixing to certain words. In case the accuracy decreases after applying the lexical rules then there could be a reason yet to be discovered.

Take for an example, if this rule is applied, *'if the current tag of a word is an adjective (J), after prefixing 'in' into that word, the word is transformed into a noun (N)'*. From what we have discovered, even though the rule stated as it is and if the word is tagged as an adverb in the first place, the rule could still be applied to that word. Meaning to say the rule is applicable as long as the letter 'in' at the beginning of every word is detected. This maybe the reason why the accuracy decreased when the lexical rules were applied.

By applying the lexicon and the contextual rules, we could see improvements in terms of accuracies based on Table 4, 5, 6 and 7. This shows that the contextual rules were able to transform and correct the wrong tags. By applying the lexicon and both rules, a better accuracy of above 90% for every corpus were obtained.

We noticed that the accuracies decreased when all the rules were applied as the size of the corpuses increased as shown in Table 4, 5, 6 and 7. Take for an example, by comparing corpus 3 and corpus 4, we found that the corpus 3 obtained a higher accuracy than that of the corpus 4. Though the size of the corpus 3 is smaller than the corpus 4, but the number of rules applied to both the corpuses were the same. Hence, the accuracy for the corpus 4 could be increased if more rules were added to the tagger. This could also be that bigger size of corpus has a more complicated morphological structure compared to smaller corpus. This comparison is in term of new words, new sentence structure and the like.

In Table 8, we noticed that the tagging time and the evaluation time for corpus 2 took much longer than corpus 1, corpus 3 took much longer than corpus 2 and corpus 4 took even longer than corpus 3. This is because bigger size of corpus need more time for tagging process. This also depends on the speed of the computer. If the computer is slow, then the tagging time and the evaluation time would take much longer. In all, we noticed that the tagging results for Kadazan language had achieved high accuracies by using the Brill's approach.

For better evaluation results it is always recommended to add more rules from time to time where necessary especially the rules that can help to increase the

accuracy. By increasing the size of the corpus using more rules, a higher accuracy could be achieved and by using a larger lexicon could help reduce the number of unknown words.

6 Conclusion

In a conclusion, we found that the Brill's approach could be trained towards many languages besides the English language and most of the results showed that the accuracy achieved by using this approach were higher or equivalent compared to the original rule-based and statistical methods. However, the Brill's approach is simpler compared to the statistical method. Complex computation is not required in Brill's method and only rule templates were used and at the same time it could achieve better accuracy. The performance of the Kadazan POS tagger using Brill's approach had been highlighted and discussed. The approach had been implemented into POS Tagger in four phases based on the Brill's approach and the tagger had also been evaluated in two ways. Firstly, by getting the accuracy and the error in comparing each rules and combining both lexical and contextual rules together and without using any other rules. Secondly, by calculating the tagging and evaluation time to calculate the accuracy. The results achieved was around 93% accuracy.

The tagging performance showed that Brill's Tagger approach could be trained successfully over a small and bigger size of corpuses and the accuracy achieved were slightly higher and could possibly be increased by adding more rules to it. The lexicon also helped to identify the most likely tag and to provide all other possible tags. However, there were still a few remaining errors left outside the scope of the tagger's observation. Hence, we conclude that there is a need to find more solution to solve the tagger's problems in reducing the errors.

References

[1] Gulen, A. & Saka, E., *Part of Speech Tagging*, Middle East Technical University, 2001.

[2] Megyesi, B., *Improving Brill's POS Tagger for An Agglutinative Language*, Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, ACL-99, University of Maryland, MD, USA, 1999.

[3] Anwar, W., Bajwa, U.I., Munir, E.U. & Fareena Naz, F., *Urdu Part of Speech Tagging Using Transformation Based Error Driven Learning*, World Applied Sciences Journal, **16** (3), pp. 437-448, 2012.

[4] Voutilainen, A., *A Syntax-Based Part of Speech Tagger*, Proceedings of the Seventh Conference of the European Chapter of the Association for

Computational Linguistics, Association for Computational Linguistics, Dublin, Germany, 1995.

[5] Rabbi, I., Khad, M.A. & Ali, R., *Rule-Based Part of Speech Tagger for Pashto Language*, Proceedings of the Conference on Language & Technology, CLT09, Lahore, Pakistan, 2009.

[6] Singha, K.R., Purkayastha, B.S. & Singha, K.D., *Part of Speech Tagging in Manipuri with Hidden Markov Model*, IJCSI International Journal of Computer Science Issues, **9**(6), pp. 146-149, 2012.

[7] Petasis, G., Palioras, G., Karkaletsis, V., Spyropoules, C.D. & Androutsopoulas, I., *Resolving Part-of-Speech Ambiguity in Greek Language Using Learning Techniques*, Proceeding of the ECCAI Advanced Course on Artificial Intelligence, ACAI'99, Chania, Greece, 1999.

[8] Schneider, G. & Volk, M., *Adding Manual Constraints and Lexical Look-up to a Brill-Tagger for German*, Proceedings of the ESSLLI-98 Workshop on Recent Advances in Corpus Annotation, ESSLLI-98, Saarbrucken, Germany, 1998.

[9] Hardt, D., *Transformation-Based Learning of Danish Grammar Correction*, Proceedings of RANLP 2001, CLPP-BAS, Tzigov Chark, Bulgaria, 2001.