Probabilistic Part Of Speech Tagging for Bahasa Indonesia

Femphy Pisceldo
Faculty of Computer Science
University of Indonesia
Depok 16424, Indonesia
femphy.pisceldo@gmail.com

Mirna Adriani
Faculty of Computer Science
University of Indonesia
Depok 16424, Indonesia
mirna@cs.ui.ac.id

Ruli Manurung
Faculty of Computer Science
University of Indonesia
Depok 16424, Indonesia
maruli@cs.ui.ac.id

Abstract

In this paper we report our work in developing Part of Speech Tagging for Bahasa Indonesia using probabilistic approaches. We use Condtional Random Fields (CRF) and Maximum Entropy methods in assigning the tag to a word. We use two tagsets containing 37 and 25 part-of-speech tags for Bahasa Indonesia. In this work we compared both methods using using two different corpora. The results of the experiments show that the Maximum Entropy method gives the best result.

1. Introduction

Part-of-speech (POS) tagging, also called grammatical tagging, is the process of assigning part-of-speech tags to words in a text. A part-of-speech is a grammatical category, commonly including verbs, nouns, adjectives, adverbs, and so on. Part-of-speech tagging is an essential tool in many natural language processing applications such as word sense disambiguation, parsing, question answering, and machine translation.

Manually assigning part-of-speech tags [6] to words is an expensive, laborious, and time-consuming task, hence the widespread interest in automating the process. The key issue that must be addressed is that of ambiguity: in most languages, words behave differently given different contexts, and thus the challenge is to correctly identify the POS tag of a token appearing in a particular context.

There are several approaches to automatic POS tagging, i.e. rule based, probabilistic, and transformational-based approaches. Rule-based POS taggers [6] assign a tag to a word based on several manually created linguistic rules, e.g. a word is assigned a noun tag if it follows an adjective. Probabilistic approaches [2, 3, 7, 10] determine the most probable tag of a token given its surrounding context, based on probability values obtained from a manually tagged corpus. The transformational-based approach combines rule-based and probabilistic approaches to automatically derive symbolic rules from a corpus. One of the most widely used transformational-based approaches is the Brill tagger [4].

The Brill tagger has been adapted for German [11], Hungarian [8], Swedish [9] and other languages. POS tagging for English has been extensively studied, and its accuracy comfortably exceeds 90% in identifying the correct tag from a sophisticated tagset [3]. Currently, however, there is relatively little work on POS tagging for Bahasa Indonesia [5, 12]. Bahasa Indonesia is the official language of Indonesia with its inhabitants of more than 250 million people [15]. Even though Bahasa Indonesia is spoken by most of the people in the country, the availability of language tools and resources for research related to Bahasa Indonesia is still limited. One language tool that is not yet commonly available for Bahasa Indonesia is a POS tagger. In this study, we report our attempt in developing an Indonesian part-of-speech tagger based on two probabilistic methods, namely Conditional Random Fields (CRF) and Maximum Entropy.

2. Probabilistic Part of Speech Tagging

As mentioned above, probabilistic approaches determine POS tags based on conditional probabilities given surrounding context features, where these probability values are obtained from a manually tagged corpus. Given the large amount of corpora that is available nowadays, and the language-independence of the tagging methods and algorithms, probabilistic methods are often favoured by the NLP community. In this paper we investigate two particular methods,

namely Conditional Random Fields (CRF) and Maximum Entropy (ME) models.

2.1 Conditional Random Fields

A conditional random field (CRF) is a framework for building discriminative probabilistic models for segmenting and labeling sequential data [2, 7]. A CRF is an undirected graphical model in which each vertex represents a random variable Y_i whose distribution is conditioned on some observation sequence X, and each edge represents a dependency between two random variables. The conditional dependency of each Y_i on X is defined through a set of *feature functions* of the form $f(Y_{i-1}, Y_i, X, i)$, which can be thought of as measurements on the observation sequence X that partially determine the likelihood of each possible value for Y_i.

Thus, the probability of a label sequence Y given an observation sequence X and CRF model λ can be written as

$$p(Y \mid X, \lambda) = \frac{1}{Z(X)} \exp\left(\sum_j \lambda_j \sum_{i=1}^{n} f_j(Y_{i-1}, Y_i, X, i)\right)$$

where $Z(X)$ is a normalization factor, and the λ_j values are weights learned from training data.

Compared to Hidden Markov Models (HMM), CRFs offer several advantages such as the ability to relax strong independence assumptions which are commonly made in other models to ensure tractability. In fact, a CRF can be viewed as a generalization of a HMM where the transition probabilities are not restricted to constant values, but can be arbitrary feature functions.

Given their ability to label sequential data, CRFs have successfully been applied to the task of POS tagging. It has been applied to English with an accuracy of up to 97% [16] and on languages such as Hindi and Telugu with accuracies of 78.66% and 77.37% respectively [2].

2.2 Maximum Entropy Models

Another widely used probabilistic method for POS tagging is the maximum entropy model [10, 14]. This method allows high flexibility in utilizing contextual information, and assigns an appropriate tag based on a probability distribution that has the highest entropy values found on the training corpus according to certain conditional values. Maximum entropy models the POS tagging task as

$$p(h,t) = \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h,t)}$$

where h is a 'history' of observation and tag sequences and t is a tag, π is a normalization constant, and $f_j(h,t)$ are feature functions with binary values of 0 and 1, and μ and $\alpha_1, \ldots, \alpha_k$ are model parameters.

The model parameters must be set so as to maximize the entropy of the probability distribution subject to the constraints imposed by the value of the f_j feature functions observed from the training data. These parameters are commonly trained using the Generalized Iterative Scaling (GIS) algorithm [10]. The underlying philosophy is to choose the model that makes the fewest assumptions about the data whilst still remaining consistent with it. The accuracy obtained on experiments with English is 96.6% [10].

2.3 POS Tagging for Bahasa Indonesia

There has been some initial study conducted on POS tagging for Bahasa Indonesia. Sari et al. applied Brill's transformational rule based approach in developing a POS tagger for Bahasa Indonesia on a limited tagset trained on a small manually annotated corpus [13]. Utilizing various lexical and contextual features, they showed that the method obtained an accuracy of 88% in assigning the correct tags to Indonesian words.

2.4 Tagset for Bahasa Indonesia

Since a standardized Indonesian annotated corpus is not available, we consulted what is widely accepted to be the official Indonesian grammar reference [1] to define a tagset for Bahasa Indonesia to be used for annotating an Indonesian corpus.

Bahasa Indonesia has 5 main parts of speech: *kata kerja (*verb), *kata sifat* (adjective), *kata keterangan (*adverb), *kata benda (*noun), and *kata tugas* (function words). Nouns can be further divided into subcategories, such as countable and uncountable common nouns, genitive common nouns, proper nouns, and various pronouns. Function words can be further divided into subcategories, such as preposition, conjunction, interjection, article, and particle.

The above description of Indonesian language is far from complete and only highlights some of the major characteristics of Bahasa Indonesia. Based on the above description, and from observation of data, we defined 37 tags for Indonesian words including punctuation. This tagset can be seen in Table 1. Additionally, we created a smaller subset consisting of 25 tags by conflating several different categories into one, e.g. the various common nouns NNC (countable), NNU (uncountable), and NNG (genitive) are replaced with a single NN category. The transitive (VBT) and

intransitive (VBI) verb categories are represented by the single VB category, and the various cardinals (CDP, CDO, CDI, CDC) are conflated into a single CD tag.

Table 1. Tagset I for Bahasa Indonesia

No	Tag	Description	Example
1.	(	Opening parenthesis	({ [
2.	)	Closing parenthesis	) }]
3.	,	Comma	,
4.	.	Sentence terminator	. ? !
5.	:	Colon or ellipsis	: ;
6.	--	Dash	--
7.	“	Opening quotation mark	‘ “
8.	“	Closing quotation mark	‘ “
9.	$	Dollar	$
10.	Rp	Rupiah	Rp
11.	SYM	Symbols	% & ' " ".)). * + ,. < = > @ A[fj] U.S U.S.S.R * ** ***
12.	NNC	Countable common nouns	Buku, rumah, karyawan
13.	NNU	Uncountable common nouns	Air, gula, nasi, hujan
14.	NNG	Genitive Common nouns	Idealnya, komposisinya, fungsinya, jayanya
15.	NNP	Proper nouns	Jakarta, Soekarno-Hatta, Australia, BCA
16.	PRP	Personal pronouns	Saya, aku, dia, kami
17.	PRN	Number pronouns	Kedua-duanya, ketiga-tiganya
18.	PRL	Locative pronouns	Sini, situ, sana
19.	WP	WH-pronouns	Apa, siapa, mengapa, bagaimana, berapa
20.	VBT	Transitive Verbs	Makan, tidur, menyanyi
21.	VBI	Intransitive Verbs	Bermain, terdiam, berputar-putar
22.	MD	Modal or auxiliaries verbs	Sudah, boleh, harus, mesti, perlu
23.	JJ	Adjectives	Mahal, kaya, besar, malas
24.	CDP	Primary cardinal numerals	Satu, juta, milyar
25.	CDO	Ordinal cardinal numerals	Pertama, kedua, ketiga
26.	CDI	Irregular cardinal numerals	Beberapa, segala, semua
27.	CDC	Collective cardinal numerals	Bertiga, bertujuh, berempat
28.	NEG	Negations	Bukan, tidak, belum, jangan
29.	IN	Prepositions	Di, ke, dari, pada, dengan
30.	CC	Coordinate conjunction	Dan, atau, tetapi
31.	SC	Subordinate conjunction	Yang, ketika, setelah
32.	RB	Adverbs	Sekarang, nanti, sementara, sebab, sehingga
33.	UH	Interjections	Wah, aduh, astaga, oh, hai
34.	DT	Determiners	Para, ini, masing-masing, itu
35.	WDT	WH-determiners	Apa, siapa, barangsiapa
36.	RP	Particles	Kan, kah, lah, pun
37.	FW	Foreign words	Absurd, deadline, list, science

Table 2. Tagset II for Bahasa Indonesia

No	Tag	Description	Example
1.	(	Opening parenthesis	({ [
2.	)	Closing parenthesis	) }]
3.	,	Comma	,
4.	.	Sentence terminator	. ? !
5.	:	Colon or ellipsis	: ;
6.	--	Dash	--
7.	“	Opening quotation mark	‘ “
8.	“	Closing quotation mark	‘ “
9.	SYM	Symbols	% & ' " ".)). * + ,. < = > @

			A[fj] U.S U.S.S.R * ** ***
10.	NN	Common nouns	Buku, rumah, karyawan, air, gula, rumahnya, kambingmu
11.	NNP	Proper nouns	Jakarta, Soekarno-Hatta, Australia, BCA
12.	PRP	Personal pronouns	Saya, aku, dia, kami
13.	PR	Common pronouns	Kedua-duanya, ketiga-tiganya, sini, situ, sana
14.	WH	WH	Apa, siapa, mengapa, bagaimana, berapa
15.	VB	Verbs	Makan, tidur, menyanyi, bermain, terdiam, berputar-putar
16.	MD	Modal or auxiliaries verbs	Sudah, boleh, harus, mesti, perlu
17.	JJ	Adjectives	Mahal, kaya, besar, malas
18.	CD	Cardinal numerals	Satu, juta, milyar, pertama, semua, bertiga
19.	NEG	Negations	Bukan, tidak, belum, jangan
20.	IN	Prepositions	Di, ke, dari, pada, dengan
21.	CC	Coordinate conjunction	Dan, atau, tetapi
22.	SC	Subordinate conjunction	Yang, ketika, setelah
23.	RB	Adverbs	Sekarang, nanti, sementara, sebab, sehingga
24.	WDT	WH-determiners	Apa, siapa, barangsiapa
25.	FW	Foreign words	Absurd, deadline, list, science

3. Experiments

In this study we use the CRF++0.51 tool [7] and the Maximum Entropy tool [11]. CRF++ is an open source implementation of Conditional Random Field for segmenting and labeling sequential data. All the training and testing data in this research was ran on the CRF tools. We used the following template in our study:

```
# Unigram
    U01:%x[-2,0]
    U02:%x[-1,0]
    U03:%x[0,0]
    U04:%x[1,0]
    U05:%x[2,0]
    # Bigram
    B
```

For the Maximum Entropy method, we used the Stanford Postagger[1].

Table 2. The number of words in 5 folds.

Fold	No. of words (Corpus I)	No. of words (Corpus II)
1	2776	2776
2	2742	2742
3	2900	2900
4	2894	2894
5	2853	2853
Total	14.165	14.165

This study used two corpora: the first corpus (Corpus I) is a collection of local newspaper articles and the second corpus is a subset of Penn Treebank corpus (Corpus II). The first corpus is a small tagged-corpus that consists of 14,165 words. The second corpus is an Indonesian version of the English Penn TreeBank corpus. The translated Indonesian version of the Penn TreeBank corpus is part of an ongoing PANL10N1[2] (Pan Localization) project. PANL10N project is an initiative to build capacity in regional institutions for local language computing in South and South-East Asian countries. It is funded by the International Development Research Center (IDRC)-Canada. Currently there are eleven countries that participate in this project.

The corpus was annotated manually using the tagset shown in Table 1 and Table 2. We divided the corpora into 5 (Table 2) and 10 folds (Table 3). For each

[1] http://nlp.stanford.edu/software/tagger.shtml
[2] http://www.panl10n.net

experiment we used one fold for testing and the rest for training.

The evaluation was done by comparing the results of manual tagging and automatic tagging based on the CRF and the maximum entropy tagger.

Table 3. The number of words in 10 folds.

Fold	No. of words (Corpus I)	No. of words (Corpus II)
1	1399	2675
2	1377	2751
3	1382	2550
4	1360	2621
5	1460	2620
6	1440	2416
7	1436	2735
8	1458	2315
9	1464	2923
10	1389	2742
Total	14165	26348

4. Evaluation Results

We compare the performance of two POSTAG methods using two different tagsets in two different corpora. In our first study we used 37 tagsets on Corpus I. The best average accuracy was produced by the Maximum Entropy method (Table 4, 10 folds) which is 77.36%. On average, the Maximum Entropy method in 10 folds is 1.44% higher than in 5 folds. The average accuracy of CRF in 10 folds is 1.50% higher than in 5 folds. The performance of the Maximum Entropy method is about 8% higher than that of the CRF method.

Using 25 tagsets the Maximum Entropy also achieved the best average accuracy which is 85.02% (Table 4, 10 folds). The Maximum Entropy method in 10 folds is 6.19% higher than in 5 folds. The CRF method in 10 folds is 9.77% higher than in 5 folds.

Table 4. The Accuracy of the CRF and Maximum Entropy Methods Using Corpus I.

POSTAG Methods	Averg. Accuracy TAGSET=37	Averg. Accuracy TAGSET=25
CRF-5	67.98%	73.12%
CRF-10	69.48%	82.89%
Max Entropy-5	75.92%	78.83%
Max Entropy-10	77.36%	85.02%

In second study we used 37 tagsets on Corpus II (Table 5). The Maximum Entropy method achieved the highest accuracy, which is 95.19% if we 10 folds. On average, the Maximum Entropy method in 10 folds is 0.26% higher than in 5 folds. The average accuracy of CRF in 10 folds is 1.07% higher than in 5 folds. The performance of the Maximum Entropy method is about 12% higher than that of the CRF method.

Using 25 tagsets the Maximum Entropy also achieved the best average accuracy which is 97.57% (Table 5, 10 folds). The Maximum Entropy method in 10 folds is 0.04% higher than in 5 folds. The CRF method in 10 folds is 0.69% higher than in 5 folds. Overall for 5 and 10 folds, the accuracy average using the Maximum Entropy method is higher than the using CRF method.

Table 5. The Accuracy of the CRF and Maximum Entropy Methods Using Corpus II.

POSTAG Methods	Accuracy TAGSET=37	Accuracy TAGSET=25
CRF-5fold	82.85%	90.46%
CRF-10	83.72%	91.15%
Max Entropy-5	94.93%	97.53%
Max Entropy-10	95.19%	97.57%

5. Conclusions

Our study focuses on evaluating the CRF and the Maximum Entropy methods for developing Part of Speech Tagger for Indonesian. The results show that the accuracy of the Maximum Entropy method is better than the CRF method.

In the future, we will conduct our study on POSTAG using a larger corpus and other methods so we can learn the best methods for Indonesian.

References

[1] Alwi, H., S. Dardjowidjojo, H. Lapoliwa, and A.M. Moeliono, *Tata Bahasa Baku Bahasa Indonesia, 3rd ed.*, Balai Pustaka, Jakarta, 2003.

[2] Avinesh, P., and Karthik, G. "Part of Speech Tagging and Chunking using Conditional Random Fields and Transformation Based Learning". *Proceedings of IJ-CAI Workshop on "Shallow Parsing for South Asian Languages.* 2007.

[3] Berger, Adam L., Stephen A. Della Pietra Y and Vincent J. Della Pietra Y. A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics (22): p. 39-71, 1996.

[4] Brill, E., "A Simple Rule-Based Part of Speech Tagger". In *Proceedings of the Third Conference on Applied Computational Linguistics (ACL)*, Trento, Italy, 1992.

[5] Chandrawati, Tutik. Indonesian Part of Speech Tagger Based on Conditiional Random Fields and Transformation Based Learning Methods. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2008.

[6] Cutting, D., Kupiec, J., Pederson, J., and Sibun, P. "A Practical Part of Speech Tagger". Proceedings of the Third Conference on Applied Natural Language Processing, Vol , 1992.

[7] Lafferty, J., McCallum, A., and Pereira, F., "Conditional Random Field: Probabilistic Model for Segmenting and Labeling Sequence Data". Proceedings of the Eighteenth International Conference on Machine Learning. 2001.

[8] Megyesi, B., "Brill's PoS Tagger with Extended Lexical Templates for Hungarian". In *Proceedings of the Workshop (W01) on Machine Learning in Human Language Technology, ACAI'99*, 1999, P.22-28.

[9] Prütz, K., "Part-of-speech tagging for Swedish". In *Parallel corpora, parallel worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22–23 April, 1999, edited by Lars Borin*. Rodopi, 2002.

[10] Ratnaparkhi, A. "A Maximum Entropy Model for Part-of Speech Tagging". Proceedings of the Conference on Empirical Methods in Natural Language Processing, Vol ,1996, p. 133–142.

[11] Ristad, Eric Sven. Maximum Entropy Modeling Toolkit, Release 1.6 beta. http://www.mnemonic.com/software/memt.

[12] Schneider, G. and M. Volk, "Adding Manual Constraints and Lexical Look-up to a Brill-Tagger for German". In *ESSLLI-98 Workshop on Recent Advances in Corpus Annotation.* Saarbrücken, 1998.

[13] Sari, Syandra, Herika Hayurani, Mirna Adriani, and Stéphane Bressan. Developing Part of Speech Tagger for Bahasa Indonesia Using Brill Tagger. *The International Second Malindo Workshop*, 2008.

[14] Toutanova, Kristina and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pp. 63-70.

[15] World FactBook, Indonesia. https://www.cia.gov/library/publications/the-world-factbook/print/id.html. 20 March 2008. [6 April 2008].

[16] Xuan-Hieu Phan, CRFTagger: CRF English POS Tagger, http://crftagger.sourceforge.net/, 2006.