Samawa Language Part of Speech Tagging with Probabilistic Approach: Comparison of Unigram, HMM and TnT Models

Trienani Hariyanti¹,², Saori Aida¹, Hiroyuki Kameda¹

¹ School of Computer Science, Tokyo University of Technology, 1404-1 Katakuramachi, Hachioji, Tokyo
² Department of Computer Science, Sumbawa University of Technology, Batu Alang, Sumbawa, Indonesia

E-mail: trienanihariyanti@gmail.com

Abstract. Samawa language is one of living languages with more than 500K native speakers in Sumbawa island, Indonesia. There are, however, extremely small amounts of available resources and efforts to develop tools in the Natural Language Processing (NLP) discipline area. In this paper, we observe and evaluate three models of probabilistic approach which are Unigram, Hidden Markov Model (HMM) and Trigram’n’Tags (TnT) models for part of speech tagging problem, which is a process to label either word or punctuation in a sentence. We used k-fold cross-validation (with k = 5 and 10) and tagged corpus around 20K tokens with 24 tags. TnT model gives the best performance reached 96.18% compared with the other models. This result shows that TnT model could be considered and used to extend Samawa corpora and help some NLP tasks in the future.

1. Introduction

Indonesia is well-known as a country which has many local languages. The Ethnologue declared that there are 719 local languages, where 707 of them are living, and the remaining are extinct [1]. Samawa language is one of living languages with more than 500K native speakers in Sumbawa island. There are, however, extremely small amounts of available resources and efforts to develop tools in the Natural Language Processing (NLP) discipline area. Nonetheless, in this year, we have started Samawa language project in [2] and have been creating Samawa tagged corpus from the various domain. We also developed a tagset and manually tagged around 20K of words.

There are many interesting researches in NLP such as speech synthesis and recognition, information retrieval, question answering, and machine translation. They principally need a fundamental task which is called Part of Speech (PoS) tagging. PoS tagging is a process to label either word or punctuation (token) in a sentence with grammatical category, i.e. verb, noun, personal pronoun, adjective, adverb, preposition and so on. Labelling part of speech manually is a time-consuming and quite exhausting process. Automatic tagging is widely used to classify and predict token tags as well as overcome these obstacles. However, the main problem in this approach is ambiguous words or lexical ambiguity.

In general, automatic tagging could be accomplished by three approaches, i.e. probabilistic approach, rule-based approach and the combination of them. Probabilistic approaches promise to find the most likely tag for a token on probability obtained from tagged corpus as training data. In this paper, we observe and evaluate three models of probabilistic approach, i.e. Unigram, Hidden Markov Model (HMM) and Trigrams’n’Tags (TnT) by using our Samawa tagged corpus. They are modules in Natural
Language Toolkit (NLTK) [3], a platform for working with human language data and Python programs related to NLP.

We organized the rest of this paper by providing related works with the probabilistic approach in section 2. Furthermore, in section 3 we provide our methodologies for PoS tagging based on Unigram, HMM and TnT models. We present our experiment, discuss and compare their performance among each other in section 4 and 5, respectively. In the last section, we summarize our conclusions and present future works.

2. Related Work
There are several works that have been done in PoS tagging problems with the probabilistic approach. For Bahasa Indonesia, Wicaksono et al. in [4] presented PoS tagger based on HMM model. Their experiment showed that affix tree and additional lexicon are effective in increasing the HMM PoS tagger accuracy up to 96.50% with 15% Out-of-Vocabulary (OOV) word. Also, HMM model has adopted for Nepali language and achieves over 96% accuracy for known words [5]. On the other research, Ahmad et al. in [6] conducted a comparison between probabilistic and rule-based approach, i.e. Unigram, Hidden Markov Model (HMM) and Brill Tagger. They showed Unigram provides a higher accuracy compared to HMM and Brill Tagger with 88.37% on their tagged corpus. TnT as a part of probabilistic approach also has been used to train Spanish in [7]. Carrasco et al. showed that TnT gives the best performance with accuracy over 95.84%.

3. Methodology
As mentioned above, we used modules in Natural Language Toolkit (NLTK), a suite of library for text analytics and NLP written in Python. It provides many datasets including corpora from various languages and tutorials to conduct processing tasks such as tokenization, stemming, tagging, chunking and many more. In the following subsection, we describe three models of probability approach in general and our Samawa tagset.

3.1. Unigram
Unigram is a type of the simplest n-gram (with n = 1), which is a probabilistic language model for estimating next word in a sequence of n words. Without considering preceding word, Unigram calculates the probability of the next word and also assigns the probabilities to entire sequences. The probabilities can be estimated from a training corpus. When Unigram encounters a word that does not exist in the training corpus, Unigram assigns it a default tag ‘None’. Furthermore, given a string $W = w_1, w_2, \ldots, w_N$, where N is the number of word tokens, a unigram model can be represented by the following formula:

$$P(w_1 \cdots w_N) = \prod_{i=2}^{N} p(w_i)$$

3.2. Hidden Markov Model
The way Hidden Markov Model (HMM) works in PoS tagging problem is to assign all the possible sequences of PoS tags and choose the sequence with overall highest score. HMM generates probability over tags and then combines to calculate the probability scores for tag sequences. Tag sequences that have the highest probability will be chosen [8]. The estimation of tag sequences $T = t_1, \ldots, t_n$ underlying observed sequence of words $w_1, w_2, \ldots, w_n$ are formulated by the following equation:

$$\hat{t}^n_1 = \arg \max_{t^n_1} \prod_{i=1}^{n} P(w_i|t_i) \cdot P(t_i|t_{i-1})$$

3.3. Trigrams’n Tags (TnT)
This model was first introduced by Brants in [9] and tried to assign an appropriate tag based on the calculation of probabilities of possible tags for each word. It is a type of second-order Markov models
which combines all the \( n \)-gram models (Unigram, Bigram and Trigram) together to choose the best tag for a word [10]. Given sequence of words \( w_1, \ldots, w_n \) and \( t_1, \ldots, t_n \) as their corresponding PoS tags, the prediction of tags sequence can be calculated by the following formula:

\[
\arg \max_{t_1, \ldots, t_T} \prod_{i=1}^{T} P(t_i | t_{i-1}, t_{i-2}) P(w_i | t_i) P(t_{T+1} | t_T)
\]  

(3)

3.4. Samawa Tagset

In Samawa language, Sumarsono et al. in [11] classified Samawa tagset into 11 categories, i.e. noun, pronoun, verb, adjective, particle, number, question, coordinate conjunction, negate and modal. We proposed our original tagset that we have mentioned in [12] and present it in Table 1.

| No. | Category  | Description             | No. | Category | Description             |
|-----|-----------|-------------------------|-----|----------|-------------------------|
| 1   | CC        | Coordinate Conjunction  | 13  | PR       | Demonstrative           |
| 2   | CD        | Cardinal Number         | 14  | PRP      | Personal Pronoun        |
| 3   | DT        | Determiner              | 15  | RB       | Adverb                  |
| 4   | FW        | Foreign Word            | 16  | RP       | Particle                |
| 5   | IN        | Preposition             | 17  | SC       | Subordinate Conjunction |
| 6   | JJ        | Adjective               | 18  | SYM      | Symbol                  |
| 7   | MD        | Modal / Auxiliary       | 19  | UH       | Interjection            |
| 8   | NEG       | Negate                  | 20  | VB       | Verb                    |
| 9   | NN        | Noun                    | 21  | VBT      | Transitive Verb         |
| 10  | NNP       | Proper Noun             | 22  | VBI      | Intransitive Verb       |
| 11  | NND       | Measure Word            | 23  | WH       | Question                |
| 12  | OD        | Ordinal Number          | 24  | Z        | Punctuation             |

4. Experiment

In this section, we present our experiment using our small tagged corpus which consists of 20,246 tokens arranged into 1,190 sentences. The corpus contains folklore and history articles which collected from humanists and archive office of Sumbawa. We manually tagged the corpus by following the guidance of Samawa morphological and grammatical rules in [10] [11] [12].

In this experiment, we used k-fold cross-validation (with \( k = 5 \) and \( k = 10 \)) in our training and testing process. It means the corpus will be separated into \( k \) partitions. Then, each of these will be used for testing and the rest (the other \( k - 1 \) partitions) as for training process.

4.1. Samawa Tagged Corpus

In Samawa corpus, we applied the simplest format of the form word/tag. We present a Samawa tagged corpus example in Figure 1 below.
Figure 1. Samawa tagged corpus example.

4.2. Tagging Example

In NLTK, the result of tagging process will be converted from a list of tokens to a list of tuples, and the form of each tuple is \((\text{word}, \text{tag})\). Table 2 shows some results of tagging using Unigram, HMM and TnT models.

| Table 2. Tagging example. |
|---------------------------|
| **Untagged Text**         |
| Kele menan, Bonong ta kena bilin teris. Satelu dengan Bonong ta dunung dapat orong nya Karim. Rupa-rupa sate ya sajerang leng dengan nya Bonong ta. Tau telu dengan nya Bonong leha-leha mo nguling-nguling pang bale orong nya Karim, nanta nya Bonong bongkok bangkang pongo sedo dengan. Kele menan sabar nya Bonong bele. |
| **Unigram**               |
| ['(\'Kele\', \'SC\'), \('menan\', \'RB\'), \('\', \'Z\'), \('Bonong\', \'NNP\'), \(\'ta\', \'PR\'), \(\'kena\', \'VB\'), \('bilin\', None), \(\'teris\', \'RB\'), \('\', \'Z\'), \('Satelu\', None), \(\'dengan\', \'NN\'), \('Bonong\', \'NNP\'), \(\'ta\', \'PR\'), \(\'dunung\', \'JJ\'), \(\'dapat\', \'VB\'), \('orong\', \'NN\'), \(\'nya\', \'DT\'), \('Karim\', None), \('\', \'Z\'), \(\'Rupa-rupa\', None), \(\'sate\', \'JJ\'), \(\'ya\', \'MD\'), \(\'sajerang\', None), \(\'leng\', \'IN\'), \(\'dengan\', \'NN\'), \(\'nya\', \'DT\'), \('Bonong\', \'NNP\'), \(\'ta\', \'PR\'), \('\', \'Z\'), \(\'Tau\', \'NN\'), \(\'telu\', \'CD\'), \(\'dengan\', \'NN\'), \(\'nya\', \'DT\'), \('Bonong\', \'NNP\'), \(\'leha-leha\', None), \(\'mo\', \'RP\'), \(\'nguling-nguling\', None), \(\'pang\', \'IN\'), \(\'bale\', \'NN\'), \(\'orong\', \'NN\'), \(\'nya\', \'DT\'), \('Karim\', None), \(\'\', \'Z\'), \(\'nanta\', \'JJ\'), \(\'nya\', \'DT\'), \('Bonong\', \'NNP\'), \(\'bongkok\', None), \(\'bangkang\', \'NN\'), \(\'pongo\', \'VB\'), \(\'sedo\', None), \(\'dengan\', \'NN\'), \(\'\', \'Z\'), \('Kele\', \'SC\'), \(\'menan\', \'RB\'), \(\'sabar\', \'JJ\'), \(\'nya\', \'DT\'), \('Bonong\', \'NNP\'), \(\'bele\', \'UH\'), \('\', \'Z\')] |
| **HMM**                   |
| ['(\'Kele\', \'SC\'), \('menan\', \'RB\'), \('\', \'Z\'), \('Bonong\', \'NNP\'), \(\'ta\', \'PR\'), \(\'kena\', \'VB\'), \('bilin\', \'VBT\'), \(\'teris\', \'RB\'), \('\', \'Z\'), \('Satelu\', \'IN\'), \(\'dengan\', \'IN\'), \('Bonong\', \'IN\'), \(\'ya\', \'MD\'), \(\'sajerang\', \'IN\'), \(\'leng\', \'IN\'), \(\'dengan\', \'IN\'), \(\'nya\', \'IN\'), \(\'Bonong\', \'IN\'), \(\'\', \'IN\'), \(\'Tau\', \'IN\'), \(\'telu\', \'IN\'), \(\'dengan\', \'IN\'), \(\'nya\', \'IN\'), \(\'Bonong\', \'IN\'), \(\'leha-leha\', \'IN\'), \(\'mo\', \'IN\'), \(\'nguling-nguling\', \'IN\'), \(\'pang\', \'IN\'), \(\'bale\', \'IN\'), \(\'orong\', \'IN\'), \(\'Karim\', \'IN\'), \(\'\', \'IN\'), \(\'nanta\', \'IN\'), \(\'nya\', \'IN\'), \(\'Bonong\', \'IN\'), \(\'bongkok\', \'IN\'), \(\'bangkang\', \'IN\'), \(\'pongo\', \'IN\'), \(\'sedo\', \'IN\'), \(\'dengan\', \'IN\'), \(\'\', \'IN\'), \(\'Kele\', \'IN\'), \(\'menan\', \'IN\'), \(\'sabar\', \'IN\'), \(\'nya\', \'IN\'), \(\'Bonong\', \'IN\'), \(\'bele\', \'IN\'), \(\'\', \'IN\')] |
5. Performance and Result
Using our corpus, we compared the performance of three probabilistic models to each other with Samawa tagset. The result is given in Table 3 below.

Table 3. The accuracy of the probabilistic model

| Probabilistic Model | Accuracy 5-fold | Accuracy 10-fold |
|--------------------|-----------------|------------------|
| Unigram            | 79.57%          | 84.04%           |
| HMM                | 35.27%          | 41.04%           |
| TnT                | 94.15%          | 96.18%           |

Unigram model gives good results. It calculated probability of a token that often appears in a sentence and then tries to predict the next PoS tag with the most frequent tag. In our experiment, 79.57% and 84.04% were obtained for k = 5 and k = 10, respectively. Unlike the experiment with Unigram and TnT models, HMM delivers poor accuracy. It is due to small training data of around 20K tokens. A huge training data such as English, Hasan et al. in [15] showed satisfactory performance of HMM up to 90.3% with 1 million tokens.

For TnT model, the accuracy level obtained was very satisfactory up to 94.15% for k = 5 and 96.18% for k = 10, respectively. TnT model works by using all n-gram models together to select the best tags. High accuracy is obtained even though the training data has a small amount [9].

6. Conclusion and Future Works
This paper describes our effort to evaluate three probabilistic models for Samawa corpus. The result is TnT model gives the best performance of up to 96% accuracy better than Unigram and HMM models. The limitation of the corpus size and variation are still a major constraint. In the future, we wish to extend its amount and try to compare with the rule-based approach.

Acknowledgement
Mrs. Trienani Hariyanti would thank the Indonesian Endowment Fund for Education (LPDP) for gratefully acknowledge funding for her study.

References
[1] Simons, Gary F. and Charles D. Fennig (eds.) 2018 Ethnologue: Languages of the World, Twenty-first edition Ethnologue
[2] Hariyanti T, Aida S and Kameda H 2018 Samawa Language Project: Its Policy and Approach 11th Symposium of the 21st Century Science and Human (Tokyo: Tokyo University of Technology) pp 06 (1-8)
[3] Bird S 2006 NLTK: the natural language toolkit *Proceedings of the COLING/ACL on Interactive presentation sessions* (Association for Computational Linguistics) pp 69–72

[4] Wicaksono A F and Purwarianti A 2010 HMM based part-of-speech tagger for Bahasa Indonesia *Fourth International MALINDO Workshop, Jakarta*

[5] Paul A, Purkayastha B S and Sarkar S 2015 Hidden Markov Model based Part of Speech Tagging for Nepali language (IEEE) pp 149–56

[6] Amrullah A Z, Hartanto R and Mustika I W 2017 A comparison of different part-of-speech tagging technique for text in Bahasa Indonesia *Engineering Seminar (InAES), 2017 7th International Annual* (IEEE) pp 1–5

[7] Carrasco R M and Gelbukh A 2003 Evaluation of TnT Tagger for Spanish (IEEE Comput. Soc) pp 18–25

[8] Bird S, Klein E and Loper E 2009 *Natural language processing with Python: analyzing text with the natural language toolkit* ( O’Reilly Media, Inc.)

[9] Brants T 2000 TnT: a statistical part-of-speech tagger *Proceedings of the sixth conference on Applied natural language processing* (Association for Computational Linguistics) pp 224–231

[10] Perkins J 2014 *Python 3 text processing with NLTK 3 cookbook: over 80 practical recipes on natural language processing techniques using Python’s NLTK 3.0* (Birmingham: Packt Publ)

[11] Sumarsono, Nadera M, Basuki S, Merdhana N and Mertha N 1986 *Morfologi dan Sintaksis Bahasa Sumbawa* (Pusat Pembinaan dan Pengembangan Bahasa, Departemen Pendidikan dan Kebudayaan)

[12] Hariyanti T, Aida S and Kamed H 2018 Samawa Language: Part of Speech Tagset and Tagged Corpus for NLP Resources *Journal of Physics: Conference Series* **1061** 012007

[13] Seken I K, Basuki S, Gosong I M and Putu Antara I G 1990 *Morfologi Bahasa Sumbawa* (Pusat Pembinaan dan Pengembangan Bahasa, Departemen Pendidikan dan Kebudayaan)

[14] Mahsun 1990 *Morfologi Bahasa Sumbawa Dialek Jereweh* (Yogyakarta: Universitas Gadjah Mada)

[15] Hasan F M, UzZaman N and Khan M 2007 Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill’s tagger) for Bangla *Advances and Innovations in Systems, Computing Sciences and Software Engineering* (Springer) pp 121–126