POS-Tagging for informal language (study in Indonesian tweets)

Endang Suryawati, Devi Munandar, Dianadewi Riswantini, Achmad Fatchuttamam Abka, Andria Arisal
Research Center for Informatics, Indonesian Institute of Sciences

endang@informatika.lipi.go.id

Abstract. This paper evaluates Part-of-Speech Tagging for the formal Indonesian language can be used for the tagging process of Indonesian tweets. In this study, we add five additional tags which reflect to social media attributes to the existing original tagset. Automatic POS tagging process is done by stratified training process with 1000, 1600, and 1800 of annotated tweets. It shows that the process can achieve up to 66.36% accuracy. The experiment with original tagset gives slightly better accuracy (67.39%) than the experiment with five additional tags, but will lose important informations which given by the five additional tagset.

1. Introduction

Twitter is one of the most popular information sharing media. This microblog is continuously producing data. This data are in the form of tweets, messages sent using Twitter, which has a limit of the number of characters for each tweet. It is an interesting task for enthusiasts in the field of computational linguistic to obtain useful information from this huge data. The information are usually about public opinion on social, politics, health, education, disaster, etc. In an effort to extract information derived from tweets, it is necessary to understand the pattern of tweets.

Tweets are known to have informal language styles. This requires special attention in the process of recognizing every word compared to formal language. The main problems encountered in informal language, such as in tweets, are the use of non-standard words and abbreviation as the effect of character limitation for each message. Sometimes, tweet’s writer also use ellipsis and emoticon to show expression. Tweets also can be a mixing between different local languages and dialects with the formal/official language.

One of the fundamental process when extracting information from twitter data is part-of-speech (POS) tagging. POS tagging determine tags of words in tweets based on the syntactic analysis. The studies about POS tagging for twitter data have been performed using both rule-based and statistical-based approaches. The addition of Twitter data that tends to be informal to a formal language corpus can decrease the performance of a model. It becomes a challenge to be able to handle the problem of extracting Twitter data pattern, in addition to building the tweets corpus.

Gimpel et al. [1] developed POS tagging for English using conditional random field (CRF) method by training 1827 manually annotated tweets. The results shows better performance than Stanford POS tagger. The research by Gimpel et al. inspired another group of researcher to develop POS tagging for tweets in Dutch [2]. Avontuur et al. [2] use SoNaR's standard twitter collection added with five special
tags to form a golden corpus. The tagger used is existing Frog tagger combined with post-processing module. In addition to English, POS tagging for Twitter data has been developed for several other languages such as Myanmar [3], Japanese [4], Irish [5], Indian [6] and Arabic [7]. In these studies, various treatments for informal language have been conducted to provide better results.

Indonesian tweets that have variety of language styles also requires special handling. Indonesian tweets is strongly influenced by foreign languages and local languages that have diverse dialects. Although there is no specific research about Indonesian tweets corpus for POS tagging, however some research have taken advantage of POS tagging for Indonesian tweets to support natural language processing (NLP) applications, namely event extraction [8][9]. The research shows an attempt to correct deficiencies when handling out of vocabulary (OOV) problem.

In order to achieve good performance of NLP applications, the handling of problem in tweets should be done in pre-processing stage. In an NLP application that utilizes POS tagger and named-entity recognition (NER) for extracting information, improper OOV handling will worsen the performance. The poor performance of POS tagger will decrease NER accuracy and further reduce overall system performance [10]. Therefore, special handling in developing POS tagger for informal language, such as in twitter data, is needed.

In general, there are two ways to develop POS tagging for Twitter data. The first way is to build a tweets corpus by manually annotating each word, then building a tagger model based on the corpus [8]. The second way is to automatically annotate each word in twitter data collection using existing formal language tagger and make corrections if needed, then the tagger is retrained to get new model for tweets data [1][2]. The retraining process will recur in accordance with the addition of training data that has been corrected, so that the performance continue to increase.

In this paper, POS tagger for Indonesian tweets is developed using the second way, that is utilizing existing POS tagger for formal language to automatically perform annotations on twitter data collection. The goal is to prove that POS tagger for the formal language can be used to annotate twitter data collection, with the addition of five new tags related to twitter attribute and the use of informal words. Word embeddings generated by using word2vec [11] and glove [12] are used to represent word in addition to random method. The random word embeddings is used as the baseline.

The next sections of this paper is managed as follows. Chapter 2 describes issues on twitter data for Indonesian and foreign languages (non-Indonesian) and how to handle them. Chapter 3 describes the experiments (corpus and method used). The results and discussion are outlined in Chapter 4 and Chapter 5 is the conclusion of this paper.

2. Twitter Sentence Pattern

2.1. The Problem of Twitter Data
The informal pattern often used in tweets are abbreviation, misspelled word, use of emoticons, use of interjection words, dialect variation and noisy data. Tweets also have orthographic forms with many variants. Indonesian tweets has more irregular pattern. In addition to have such an informal pattern, Indonesian tweets is influenced by foreign languages, local languages, and social languages. Social languages are often referred as the alay language. The alay language contributes to the addition of OOV lists that degrade model performance.

To deal with the informal pattern of Twitter data, A Le et al. [13] grouped informal language pattern into six groups of patterns, namely abbreviation, interjection, foreign word, blending, emoji, and emoticons. In the foreign word group, sometimes also found the form of abbreviation, such as the word "ht" which means "hot topic". In the blending group, there is an example of mixing between Indonesian tweets and English such as “gagal move on”, which means failed to move on. Although this approach is intended for sentiment analysis applications, it can be adopted for POS tagging for the normalization process at the pre-processing stage.

In twitter data also found words derived from local languages, such as nak, ora, piye, and kabare. In addition, there is an unusual form of the loop like tetembak which means firing a shot. The
influence of *alay* language is also strong enough to affect the pattern of Indonesian tweets, word such as *cucok* which means fit or suitable and a number of other *alay* words.

### 2.2. Handling The Problem of Twitter Data

In general the problem in Twitter data can be addressed either through the selection of appropriate tagging methods, the selection of appropriate word representation (feature), or the process of normalizing the informal pattern into a formal pattern. According to Avontuur et al. [2], there are two approaches to handle problem in twitter data pattern. The first approach is to revise or normalize the text of tweets from its informal form to formal form. The second approach attempt to adopt NLP tools that address twitter data issues. In addition to the choice of methods, features, and normalization, it is also necessary to have a large number of data.

Owoputi et al. [14] tried to improve POS tagging performance developed by Gimple et al. [1] by addressing the issue of informal pattern in Twitter data. This is done by the addition of cluster-based features and the level-token name lists feature. Word clustering groups informal words based on their variants, for example, the words "lololol", "lool", and "lmbo" will be in different groups with words "gonna", "gunna", "gonne" and "goona".

Other researchers propose POS tagging with bootstrapping techniques to reduce sparsity data [15]. It contribute in a module that is part of GATE TwitE toolkit for processing social media text. Kaji et al. [4] proposed a combination of model with lexical normalization to overcome orthographic diversity. Gosh et al. [6] used various pre-processing modules (word lists and some word features) and post-processing with logical reasoning to annotate words that fail to be annotated by the model.

Another approach is by choosing the right method of tagging. Some of these methods have been used in POS tagging for Twitter data research such as Conditional Random Field (CRF) [1] [16] [17] and Neural Network [17]. Some tools that have been used in POS tagging for Twitter data research are Stanford [5] [6] [17], GATE [17], AMIRA [7] and MADA [7], NLTK, OpenNLP, CoreNLP, Pattern, TweetNLP, TwitterNLP and TwitIE [18].

### 3. Experimental Setup

First we performed an automatic annotation using existing POS tagger for formal Indonesia [19], with addition of five new tags for Twitter data. The data annotation is done semi-automatically, where the tagger annotates new data automatically and the result of the annotation is corrected manually. The resulting data is added to the training data to rebuild the model. The model is trained multiple times, with data amounts of 1000, 1600 and 1800. This technique is referred to as active learning. The normalization process does not use normalized word lists but utilizes the Kateglo dictionary and Kamus Besar Bahasa Indonesia, to find the equivalent word of the informal language manually.

#### 3.1. Word Embeddings

The construction of word embeddings started with preparation of unlabeled corpus to build a vocabulary. Unlabeled corpus is built from 101,800 Indonesian tweets. Pre-processing of the unlabeled text corpus produces vocabulary of 178,736 words. Each method of word embeddings (random, CBOW, Skip-Gram, and Glove) are trained using the vocabulary. The training process will produce a word vector for every word that is in the vocabulary. This word vector is used as input in the tagging process. Some important parameters used in this experiment are vocabulary size, context window size and vector dimension. We use vector dimension of 500 and the context window size is 2.

#### 3.2. Tagset

The tagset used are an available 23 tags [19] with addition of five new tags related to the twitter attributes (bold texts in Table 1). These five tags (twitter attributes) are HASH, DISC, AT, URL and EMO. HASH is a tag used for word starts with "#". DISC is a tag used for “RT” and “:” in a re-tweets sentence. AT is a tag used for word starts with the “@”. Tag of URL is used for word that indicate an URL. EMO is used for word that contain emotion expressions. The addition of these five new tags refers to the five twitter tags used by Avontuur et al. [2]. These five new tags are used to accommodate
five twitter attributes. These new tags (classes) can be used as the keyword for the information searching process from twitter data. Meanwhile, informal words will be kept classified as the original twenty-three class, even though these informal words will be automatically classified as the NNP class before they are manually corrected for the retraining process.

| Table 1. Indonesian tagset for social media |
|---|---|
| Tag | Description |
| CC | Coordinating Conjunction |
| CD | Cardinal Number |
| OD | Ordinal Number |
| DT | Determiner / article |
| FW | Foreign Word |
| IN | Preposition |
| JJ | Adjective |
| MD | Modal and auxiliary verb |
| NEG | Negation |
| NN | Noun |
| NNP | Proper Noun |
| NND | Classifier, partitive, and measurement noun |
| PR | Demonstrative Pronoun |
| PRP | Personal pronoun |
| RB | Adverb |
| RP | Particle |
| SC | Subordinating Conjunction |
| SYM | Symbol |
| UH | Interjection |
| VB | Verb |
| WH | Question |
| X | Unknown |
| Z | Punctuation |
| AT | Mention (‘@’) |
| DISC | Discourse Maker (‘RT’) |
| HASH | Hastag (‘#’) |
| URL | Uniform Resources Locator (‘http’) |
| EMO | Emoticon (‘☺’, ‘☻’) |

3.3. POS Tagging with Neural Network
The model used to perform the tagging process is Feedfroward Neural Network with one hidden layer. This model has been built on the previous research for the formal Indonesian language trained on online media corpus [19]. In our research, the model is trained using semi-automatic labeled corpus, that created from 1000, 1600 and 1800 Indonesian tweets. The labeled corpus used in this training process is divided into 64% for training, 16% for validation and 20% for testing. The use of original tagset (without five additional tags) is intended to build a tagger model which used as the baseline.

4. Result and Discussion
From previous research, it has been concluded that almost all methods of word embedding (CBOW, Skip-Gram, and Glove) give good performance (roughly 93 %) for tagging process of Indonesian language sentences from online media, except random method that gives a fairly low performance [19]. Table 2 shows the accuracy for each embeddings method which produced from 1000 tweets as training data. Tables 3 and Table 4 show the same measurement which produced from 1600 tweets and 1800 tweets as training data.
Table 2. Tagging accuracy (%) with 1000 tweets

| Method     | Testing | Validation | Testing | Validation | Testing | Validation | Testing | Validation |
|------------|---------|------------|---------|------------|---------|------------|---------|------------|
| Random     | 64.87   | 65.37      | 43.19   | 42.08      | 43.05   | 42.85      | 46.05   | 46.22      |
| CBOW       |         |            |         |            |         |            |         |            |
| Skip-gram  |         |            |         |            |         |            |         |            |
| Glove      |         |            |         |            |         |            |         |            |

Table 3. Tagging accuracy (%) with 1600 tweets

| Method     | Testing | Validation | Testing | Validation | Testing | Validation | Testing | Validation |
|------------|---------|------------|---------|------------|---------|------------|---------|------------|
| Random     | 64.03   | 65.11      | 64.94   | 64.83      | 65.69   | 65.86      | 63.91   | 63.91      |
| CBOW       |         |            |         |            |         |            |         |            |
| Skip-gram  |         |            |         |            |         |            |         |            |
| Glove      |         |            |         |            |         |            |         |            |

Table 4. Tagging accuracy (%) with 1800 tweets

| Method     | Testing | Validation | Testing | Validation | Testing | Validation | Testing | Validation |
|------------|---------|------------|---------|------------|---------|------------|---------|------------|
| Random     | 63.41   | 64.01      | 65.61   | 64.63      | 66.36   | 65.81      | 64.14   | 63.54      |
| CBOW       |         |            |         |            |         |            |         |            |
| Skip-gram  |         |            |         |            |         |            |         |            |
| Glove      |         |            |         |            |         |            |         |            |

Table 5. Tagging Accuracy (%) with 1800 Tweets without the five additional tags

| Method     | Testing | Validation | Testing | Validation | Testing | Validation | Testing | Validation |
|------------|---------|------------|---------|------------|---------|------------|---------|------------|
| Random     | 64.46   | 65.10      | 67.19   | 66.78      | 67.39   | 68.05      | 66.87   | 67.06      |
| CBOW       |         |            |         |            |         |            |         |            |
| Skip-gram  |         |            |         |            |         |            |         |            |
| Glove      |         |            |         |            |         |            |         |            |

Based on the accuracy shown in Table 2, random method gives the best performance among the other embeddings methods with accuracy of 64.87%. CBOW, Skip-Gram, and Glove give lower accuracy (approximately 43% - 46%). Table 3 shows an increase in performance almost on all methods caused by an increase in the amount of training data from 1000 tweets to 1600 tweets. The same result occurs when the amount of training data is increased to 1800 (66.36%), where almost all methods give performance improvement, except for Random which has a slight decrease in performance (see Table 4). This proves that an increase in the amount of data accompanied by increasing data pattern can improve performance. Based on this experiment Skip-Gram outperformed other methods in 1600 and 1800 training data. This indicate that the use of word embeddings can support the improvement of tagger performance.

The resulting accuracy of POS tagger for twitter data is smaller than the accuracy of POS tagger for formal Indonesian Language [19]. The main reason is the amount of data pattern is still small to support the performance of tagger. The source of twitter data is also obtained randomly (common account), not derived from specific accounts. The obtained data is mainly a private conversation that contains more noisy data.

As can be seen in Table 5, our experiment in POS tagging without the five additional tags (the original tagset) gives slightly better accuracy than the experiment with the five additional tagset.
(Table. 4). It is can be inferred that classification with smaller number of category gives better accuracy. It is because the five additional tags to classify Twitter attributes (‘#’, ‘@’, ‘http’) which dominate in tweets are correctly classified as NNP class. Other Twitter attributes (‘RT’, ‘emoticons’) are correctly classified as SYM class. Even though they are correctly classified, in this case, removing those five additional tags leads to missing important informations that relate with social media case. This experiment further emphasizes that adding the five additional tags does not significantly reduce POS tagging accuracy, but also keep important features of social media messages.

5. Conclusion
Part-of-speech tagging for the formal Indonesian language can be used for tagging of Indonesian tweets with the additional of five new tags. Approach to handle informal language patterns by matching a word with Kateglo dictionary (a simple normalization) helps model to recognize the word. Several word embedding techniques have been evaluated for part-of-speech tagging of social media messages (tweets) in Indonesian language. The experiment shows that the using of word embedding is an alternative technique for POS tagging with a promising results (66.36 %). This is because the word embeddings can capture syntactic and semantic information of word based on nearby words. The characteristic of word embeddings is suitable to recognize and predict tweets patterns.

For the future, in order to improve the performance of this tagger, it is necessary to add a large number of training data which brings many new patterns. In the normalization process, it needs to add informal word list. Data sources should derive from the specific twitter accounts as needed to avoid irregular informal patterns and noisy data. Unknown words also need to be treated separately and included in model learning iteration (active learning). It is recommended to attempt another word embeddings methods, such as TweetNLP, TwitterNLP, and TwitIE to find the best word embeddings method for Indonesian tweets.

This work is a study that aimed to enrich the Indonesian language corpus by creating a high quality informal language corpus. In addition to increasing amount of training data and selecting appropriate word embedding methods, it is possible to apply other types of neural networks such as CNN, RNN or other types of neural networks in order to improve the tagger performance.

References
[1] K. Gimpel, N. Schneider, B. O’Connor, D. Das, D. Mills, J. Einstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith 2011 Part-of-Speech for Twitter: Annotation, Features, and Experiments Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics Portland, Oregon pp. 42-47
[2] T. Avontuur, I. Balemans, L. Elshof, N. V. Noord, M. V. Zanen 2012 Developing a part-of-speech tagger for Dutch tweets Computational Linguistics in the Netherlands Journal Vol 2, pp. 34-51
[3] B. O’Connor, C. Dyer, K. Gimpel and N. A. Smith 2011 Normalization of Myanmar Grammatical Categories for Part-of-Speech Tagging International Journal of Computer Applications Vol 36 No. 1 pp. 10-17
[4] N. Kaji and M. Kitsuregawa 2014 Accurate Word Segmentation and POS Tagging for Japanese Microblogs: Corpus Annotation and Joint Modeling with Lexical Normalization Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) pp. 99-109
[5] T. Lynn, K. Scannel, and E. Maguire 2015 Minority Language Twitter: Part-of-Speech tagging and Analysis of Irish Tweets Proceedings of the ACL pp. 1-8
[6] S. Lynn, S. Ghosh, and D. Das 2016 Part-of-Speech Tagging of Code-Mixed Social Media Text Proceedings of the second Workshop on Computational Approaches to Code Switching, pp. 90-97
[7] F. Albogami, and A. Ramsay 2016 Fast and Robust POS Tagger for Arabic Tweets Using Agreement-based Botstrapping Proceedings of the Tenth International Conference on
Language Resources Workshop on Computational Approaches to Code Switching and Evaluation pp. 90-97

[8] F. Muhammad, and A. Purwanti 2016 Handling Out of Vocabulary in Supervised Event Extraction on Indonesian Tweets *International Conference on Data and Software Engineering*

[9] F. Muhammad, and M. L. Kodra 2015 Event Information Extraction from Indonesian Tweets using Conditional Random Field *Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*

[10] A. Purwarianti, A. Andhika, A. F. Wicaksono, I. Afif, F. Ferdiane 2016 InaNLP: Indonesia Natural Language Processing Toolkit *Proceedings of International Conference on Advanced Informatics: Concept, teori and Application*

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean 2013 Efficient estimation of word representations in vector space *arXiv* preprint arXiv:1301.3781

[12] J. Pennington, R. Socher, and C. D. Manning 2014 Glove: Global vectors for word representation *EMNLP* vol. 14, pp. 1532

[13] T. A. Le, D. Moeljadi, Y. Miura, and T. Ohkuma 2016 Sentiment Analysis for Low Resource Languages: A Study on Informal Indonesian Tweets *Proceedings of the 12th Workshop on Asian Language Resources*

[14] O. Owoputi, B. O’Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith 2013 Improved Part-of-Speech Tagging for Online Conversational Text with Word Cluster *Proceedings of NAACL*

[15] L. Derzynski, A. Ritter, S. Clark, and K. Bontcheva 2013 Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data

[16] C. Li, and Y. Liu 2015 Joint POS Tagging and Text Normalization for Informal Text *Proceedings of the 24th International Joint Conference on Artificial Intelligence*

[17] T. Gui, Q. Zhang, HM. Peng and X. Huang 2017 Part-of-Speech for Twitter with Adversarial Neural Network *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*

[18] A. Pinto, H. G. Oliveira, and A. O. Alves 2016 Comparing the Performance of Different NLP Toolkits in Formal and Social Media Text *SLATE*

[19] A. F. Abka 2016 Evaluating The Use of Word Embeddings for Part-of-Speech Tagging in Bahasa Indonesia *Proceedings of International Conference on Computer, Control and Informatics: Concept, Teori and Application*