

Language Resources and Tools Development for Indonesian Languages

Budi Irmawati^{1*}, Arik Aranta¹, Wirarama Wedhaswara¹

M. Iqbal D. Putra¹, Siti Oryza Khairunnisa^{2**}

¹Universitas Mataram, Jl. Majapahit 62 Mataram, Indonesia

²Tokyo Metropolitan University, 1 Chome-1 Minamiosawa, Hachioji, Tokyo 192-0397, Japan

{budi-i, arik, wirarama}@unram.ac.id, iqbalwinfor@gmail.com, siti-oryza-khairunnisa@ed.tmu.ac.jp

Abstract

These works generated resources for languages in Indonesia. We started our works on Indonesian and Balinese and will continue with languages in the west part of Lesser Sunda Islands. We collected parallel learner sentences, documents with different age levels, and scientific papers. We use those resources to solve some problems such as preposition error correction, identify words for different age levels, and mapping reviewers who best match with a submitted paper in a referred publication. Then, to preserve ancient documents, we defined an input mechanism to write Balinese scripts recognized as the *Bali Simbar* font.

Keywords: Age level word list, Balinese transliteration, Dependency annotation scheme, Indonesia L2 resources

Résumé

Tulisan ini menjelaskan kegiatan-kegiatan untuk mengumpulkan data dan membuat aplikasi dalam bahasa-bahasa di Indonesia untuk membantu masyarakat dalam mempelajari bahasa-bahasa tersebut. Kami mulai dari bahasa Indonesia dan bahasa Bali dan akan melanjutkan dengan bahasa-bahasa di Nusa Tenggara Barat. Kami mengumpulkan kalimat paralel yang terdiri dari kalimat yang ditulis oleh pelajar bahasa, dokumen dari berbagai tingkatan usia, dan paper akademik. Kami menggunakan sumber daya tersebut untuk menyelesaikan beberapa permasalahan seperti perbaikan kata depan, mengenerate kata-kata yang digunakan pada tingkat usia tertentu, dan menentukan reviewer yang bidang keahliannya sesuai dengan paper yang disubmit ke sebuah seminar. Selanjutnya untuk menjaga keberlanjutan bahasa daerah, kami membuat metode untuk menuliskan huruf dalam bahasa daerah, dalam hal ini huruf Bali yang telah dikenal sebagai Simbar Bali dalam format UTF.

1. Introduction

Many people agreed that producing language resources are time consuming and highly cost. In the case of under-resource languages, large raw texts are also difficult to find. The problems are even harder because rare institutions that want to financially support those productions. Many people do not understand that language resources are really worth. Not many people know that recent methods may generate automatic decision extracted from a large sized of text data especially ones that have already been annotated, as those happen in the languages that have many resources.

In the production of a language resource, many linguists understand that it has already taken time to collect or record the raw data. The next efforts are to analyze what kind of treatments that fit with the data such as finding features to be extracted to useful information. To collect informative features, people need to build language tools and know how to extract those features. Then, to extract those features we also need data that have been annotated. Therefore, we have to decide what annotation schemes are appropriate with the data, that will support our efforts to mine the features, based on the analysis of previous experimental results. After applying proposed methods, we still need to validate whether the annotation really benefits to the data extraction and whether the extracted features appropriate with the goals.

We have worked to develop language resources on the sentences written by L2 learners for Bahasa Indonesia (Irmawati et al., 2016a). We also defined a dependency relation annotation scheme for Bahasa Indonesia (Irmawati et al., 2017a) by considering the language characteristic. Then we, annotated sentences with the annotation and trained the MST Parser to build a dependency relation parser. The parser was used to annotated a large sized of Indonesian corpus to extract their dependency relations. The dependency relation features has been proofed successfully improved a preposition error correction performance (Irmawati et al., 2016b; Irmawati et al., 2017b).

We realized that language learners will use different vocabulary from ones used by the natives. Moreover, the development of languages for people in Indonesia is influenced by their local, indigenous, languages, which may more than one languages. Therefore, the term ‘*second language learners*’ in Indonesia are not only fit with foreigners, but also for the Indonesian people. For that reason, we are currently working to generate specific word lists based on an age level (e.g. children and teenagers). To find the specific words for each age level, we calculate a word that has higher occurrence in a document but lower occurrence in the documents of different age level. Later, we will build a game for children so they may develop a simple sentence using the word list from their age level (easy words). The game may also be used by language learners in beginner level.

*Corresponding author.

**This work was done when the author was a student in Institut Teknologi Sepuluh Nopember, Indonesia.

topic modelling to map scientific papers with prospective committee members who better review them (Pradina and Khairunnisa, 2018). We also built a mobile application to write a word in Balinese script. User types the word in alphabet then the application will convert it to a unicode so the related word will be printed in *Bali Simbar* fonts (Aranta et al., 2018).

2. Language Resources

These works resulted parallel learner sentences (sentences written by L2 learner with their correction sentences) and a dependency relation annotation scheme.

2.1. Language Learner (L2) Resources

L2 learner resources contained mistakes made by second language (L2) learners that were collected from mistaken sentences written by the L2 learners taken from lang-8 website¹. The mistaken sentences have their corrections (corrected by their native speakers) so the pairs are considered as parallel data (one contained mistakes and one is its correction).

The size of the original mistaken sentences were limited. It is only 6,559 learners' sentences with the vocabulary sized of 8,673 words. We firstly experimented on preposition errors. To increase its size, we produced artificial error sentences from formal sentences that were taken from Leipzig corpus (Quasthoff et al., 2006). We proposed two methods to inject the correct sentences. The first method, *embeddings*, used dependency-based word embeddings to find a normal sentence that its verb and its preposition object were similar to the verb and the preposition object of the learner sentence that has incorrect preposition. Then, we replaced the preposition in the normal sentence with the incorrect preposition taken from the learners' sentence (Irmawati et al., 2016b). The second method, *selection*, selected randomly injected sentences in which the mistakes highly resembled the mistaken sentences written by L2 learners (Irmawati et al., 2017b).

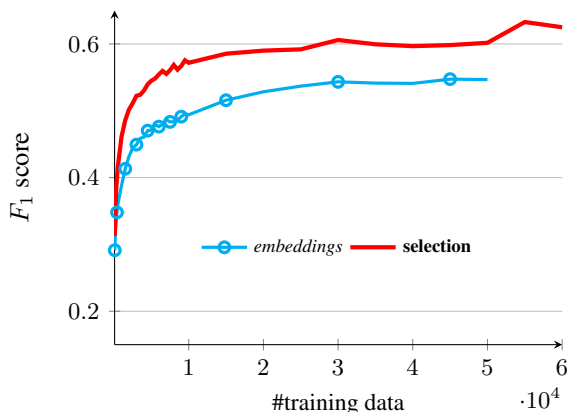


Figure 1: The *selection* method presents better results than the *embeddings* method.

We evaluated the resulted artificial sentences by using them as a training data to correct the preposition errors. The *em-*

¹<http://lang-8.com>

beddings method was promising because it took lesser time to generate artificial L2 learners' sentences. However, our experiments resulted that the *selection* method performed better as showed in Figure 1. The results also shows that training using more data resulted better but the best results were obtained by the data resemble to the data contained mistake. Both experiments used dependency features, so we may conclude that the dependency relation annotation scheme really benefits to preposition error correction for Indonesian.

2.2. Dependency Relation Annotation Scheme

As explained in Subsection 2.1., the verb related to a preposition is important as well as the object of the preposition. Therefore, dependency relation annotation scheme is important to train a dependency relation parser. We defined a dependency annotation scheme (Irmawati et al., 2017a) and annotated 1,132 sentences to build the dependency parser model with accuracy of 81.2% for UAS.

Our work is an adaptation of the Stanford annotation scheme for English proposed by de Marneffe and Manning (2013). This annotation scheme was really useful to our task in Section 2.1. because without that dependency relation features, we obtained less than 50% F₁-score in the preliminary experiments.

2.3. Word List Based on Age Level

In the context of language learning, word choices are critical and be a foolishness for learners. Many researchers tried to help them by developing a simple word list (Coster and Kauchak, 2011) to help learners understand documents not in their spoken language, in the side of vocabulary and structure.

Our goal is to generate vocabularies used by children and teenagers and to justify whether the two different age levels use similar vocabulary in Indonesian. We took data by crowdsourcing from the internet (<https://hai.grid.id/> and <https://bobo.grid.id/>). As the comparison document, we also crowdsourced a national newspaper (detik.com) to comply that the two targeted vocabulary were different from one used in the formal articles.

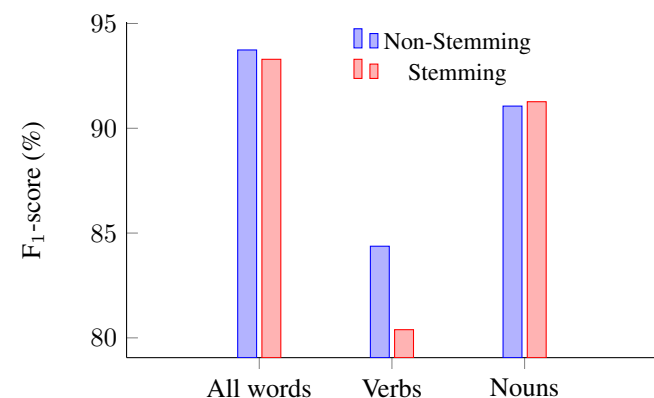


Figure 2: The results of classification on documents consumed by different age levels show that Non-Stemming words performs better.

Like other text processing, we also preprocessed the data with tokenization, case folding, stop words filtering, and stemming. We then calculated the TF-IDF for each word related to each document class. Then we used *Multinomial Naïve Bayes* as the classifier. We did some experiments by differentiated the dataset of using all word types, only verbs, and only nouns. For each experiment, we did with stemming and without stemming the words. The experimental results show that all words features performed better than individual word type experiments, which means that the POS of the words did not contribute well.

Figure 2 shows that the dataset of all words and all verbs without stemming performed better. This phenomenon indicates that the existent of affixes influences the complexities of a word. However, the experiment need to be extended to justify what affixes mostly used in the lower age level articles. In Indonesian, every time an affixes is added to a word, it may derived the word to different POS tag. Therefore, it is necessary to evaluate whether several times word derivations may increase word complexity.

The experiment also listed some words contained in specific class as shown in Table 1. However, we have not find a specific pattern that may be concluded. Therefore, we need to work further to find some interesting pattern from the data.

Affixes Words	Categories		
	Children	Teenager	Newspaper
antara	antaranya	antaranya	di antara, antaranya
buruk	buruk	buruknya	memperburuk, memburuk
budaya	kebudayaan kebudayaan-nya	berbudaya	budayanya,
negara	negara	senegara-nya	kenegaraan

Table 1: Some affixes words based on age level category

2.4. Assigning Prospective Reviewers to Scientific Papers

As a conference attracts high number of participants, it would be time consuming for program committee to assign reviewers to bunch of submitted papers. Moreover, they need very wide knowledge to pair varied paper topics and reviewers' expertise to make sure that each paper will obtain the best valuable comments.

We employee *author-topic modelling* to solve this situation. The method is very simple. We assumed that authors who wrote a published paper are ones with expertise-related to the topic of the paper. Then, we trained a model for those relations. Next time a program committee need to assign reviewers to a submitted paper, they may apply the trained model to the submitted paper to find the best reviewers who match with the paper.

In this experiment, we collected 422 scientific papers from various conferences in Indonesian, from 2013 to 2016,

from <http://is.its.ac.id/pubs/oajis/>. We did three scenarios, said *without-stemming*, *with-stemming*, and *only nouns*. The experimental results showed that the *with-stemming* scenario got the lowest perplexity and the highest topic coherence with 100 number of topics as shown in Table 2. The *with-stemming* scenario even obtained -1.528 for the 50 number of topics.

	Without Stemming	With Stemming	Only Nouns
Number of topics	50	100	150
Mean of perplexity	171.18	127.46	85.17
Std. Dev. of perplexity	1.158	1.040	0.667
Topic coherence	-1.765	-1.615	-1.741

Table 2: The results of three scenarios

We analyzed that a problem that may affect the results are the number of papers written by an authors, which is very few during a year. It is usually only about two to three papers when he/she wrote as the first author. Moreover, there are possibility of an author to write little bit different topic from his/her main area though the area may have some similarities in context.

3. A Transliteration for Balinese

As other languages in Asia, some parts in Indonesia used scripts to write their ancient documents. The script contains some rules because one letter represents one syllable with inherent vowel /a/. Each letter covers the consonant, vowel, and some accent speech. The languages are still be used in the daily conversation with some simplification in vocabularies and language levels. It also differentiates the vocabulary spoken to elderly and honour people.

On the other hand, to face of the diversity, Indonesian tend to use their national language that is used as a unity language. The use of alphabet to write the documents also supports the development of the national language. Therefore, the local language speakers are decreasing. In the spirit of preserving endanger languages in Indonesia, we developed an application as an input method for Balinese script (Aranta et al., 2018). Balinese is a language used in Bali, Northern Nusa Penida, Western Lombok and Eastern Java. It is a *Malayo-Polynesian* language spoken by 3.3 million people (as of 2000). Balinese is not mutually intelligible with Indonesian. Some words in the higher level are almost similar to Javanese but ones used in the daily conversation have different in meaning². In the case of Balinese, the script can be written with 18 consonants and 9 vowels³. The rules are *Gantugan*, *Gempelan*, *PasangPageh*, numbers, and punctuations.

We used the official *Balinese dictionary* to generate rules and to represent letters in unicodes, known as *Bali Simbar*. Then we used test data taken from Balinese Galang Foundation, a preservation institution for Balinese culture.

²https://en.wikipedia.org/wiki/Balinese_language

³https://en.wikipedia.org/wiki/Balinese_script

The evaluation was done on 151 letters containing 13 letter types taken from Balinese dictionary. It obtained 92.72 % accuracy that means that there are some unrecognized letters including ones related to pronunciation. The difficulties to represent correct letter were because not all users knew well how to pronounce a letter. For example, they cannot differentiate ‘e’ and ‘ē’ in the word ‘pēkēn’.

4. Conclusion

We described the process of collecting a learner corpus and how to generate data artificially to extend the original learner corpus. We have tried two artificial generation data methods involving dependency-based word embeddings and selection methods. By increasing its size, we showed that the preposition error correction system developed from the data, resulted better performance. Moreover, we also found that the dependency relation features also improved the performance.

For the document classification based on age level, we concluded that the document classifications works better by involving all words without stemming. However, there is still remaining work to confirm whether the complexity of affixes might be used to identify whether a word is difficult for a low level age such as children. In our experiments, we list some words that only appear in one age level. Therefore our work to build a word list of vocabulary used in an age level may still find other possible results. On the other hand, our experiments on author-topic model gave different results as the stemming scenario performed the best.

Our work on the transliteration of Balinese will also be continued to reverse the process from the script to alphabet letter for easy readability. Also it may be continue to be applied to other scripts used in local, indigenous, especially in the west part of Lesser Sunda.

5. Acknowledgements

The **L2 language resources works** were supported by *Directorate General of Higher Education* through DIKTI scholarship, Indonesia and the Computational Linguistics Laboratory of Nara Institute of Science and Technology (NAIST), Japan.

The **Balinese transliteration** was partially supported by the Indonesian Ministry of Research, Technology and Higher Education grant number 113/UN48.15/LT/2018.

6. Bibliographical References

Aranta, A., Gunadi, I. G. A., and Indrawan, G. (2018). Utilization of Hexadecimal numbers in Optimization

of Baliness Transliteration String Replacement Method. In *Proceedings of the 11th AUN/SEED-Net Regional Conference on Computer and Information Engineering*, pages 3746–3753, Surabaya, Indonesia, Nov. IEEE.

Coster, W. and Kauchak, D. (2011). Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA, June. Association for Computational Linguistics.

de Marneffe, M. and Manning, C. D. (2013). *Stanford Typed Dependencies Manual: Revised for Stanford Parser v.3.3 in December 2013*. September. Revised for the Stanford Parser v.3.3 in December 2013.

Irmawati, B., Komachi, M., and Matsumoto, Y. (2016a). Towards Construction of an Error-Corrected Corpus of Indonesian Second Language Learners. In Francisco Alonso Almeida, et al., editors, *Input a Word, Analyse the World: Selected Approaches to Corpus Linguistics*, chapter 27, pages 425–443. Cambridge Scholars Publishing, Newcastle upon Tyne.

Irmawati, B., Shindo, H., and Matsumoto, Y. (2016b). Exploiting Syntactic Similarities for Preposition Error Corrections on Indonesian Sentences Written by Second Language Learner. In Sakriani Sakti, et al., editors, *SLTU-2016, 5th Workshop on Spoken Language Technologies for Under-resourced Languages*, volume 81 of *Procedia Computer Science*, pages 214–220, Yogyakarta, Indonesia. Elsevier.

Irmawati, B., Shindo, H., and Matsumoto, Y. (2017a). A dependency Annotation Scheme to Extract Syntactic Features in Indonesian Sentences. *International Journal of Technology (IJTech)*, 8(5):549–558, November.

Irmawati, B., Shindo, H., and Matsumoto, Y. (2017b). Generating Artificial Error Data for Indonesian Preposition Error Correction. *International Journal of Technology (IJTech)*, 8(3):957–967, April.

Pradina, R. and Khairunnisa, S. O. (2018). Author-Topic Modelling for Reviewer Assignment of Scientific Papers in Bahasa Indonesia. In *Proceedings of 2018 International Conference on Asian Language Processing (IALP)*, pages 351–356, 11.

Quasthoff, U., Richter, M., and Biemann, C. (2006). Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1799–1802, Genoa, Italy.