bookChapter

by Budi Irmawati

Submission date: 09-Apr-2023 12:56AM (UTC-0500)
Submission ID: 2059390004
File name: 027_Irmawati_Komachi_Matsumoto.pdf (1.52M)
Word count: 5660
Character count: 30753

CHAPTER TWENTY-SEVEN

TOWARDS CONSTRUCTION OF AN ERROR-CORRECTED CORPUS OF INDONESIAN SECOND-LANGUAGE LEARNERS

BUDI IRMAWATI[1], MAMORU KOMACHI[2] AND YUJI MATSUMOTO[3]

1. Introduction

Recently, much natural language processing (NLP) research has benefited from using learner corpora as resources (Ng et al. 2013). While some grammatical errors made by second-language (L2) learners overlap with the errors made by native speakers, other learner errors do not. In writing, L2 learners also make errors that occur comparatively infrequently in native speakers' production (Leacock et al. 2014) because the learners' first language (L1) influences how they develop a sentence. Therefore, working on native speaker corpora is not a highly suitable basis for the error detection and correction of L2 learner writings.

In English, many people have worked on preparing corpora (Dahlmeier et al. 2013; Leacock et al. 2014; Marcus et al. 1993; Rozovskaya and Roth 2010). Other learner corpora are also available for languages such as Korean (Lee et al. 2012), Arabic (Abuhakema et al. 2008), German (Boyd 2010), and Czech (Hana et al. 2010). However, for some underdeveloped languages such as Indonesian, such a corpus is unavailable. Hence, tailoring L2 learning resources of Indonesian for the development of a language learning support system is especially useful. This type of language resource helps language teachers understand the types of learner

[1] Corresponding author. Nara Institute of Science and Technology. Email: budi-i@is.naist.jp. On leave from University of Mataram, Indonesia.
[2] Tokyo Metropolitan University. Email: komachi@tmu.ac.jp
[3] Nara Institute of Science and Technology. Email: matsu@is.naist.jp

problems, design course materials, provide feedback about mistaken grammar or word choices, construct a confusion matrix for L2 learners, and emphasize the use of Indonesian words instead of borrowed words from other languages.

Developing a learner corpus manually is laborious and error-prone while rule-based machine annotation is too coarse and inaccurate, so in an initial effort to develop an error-annotated learner corpus of Indonesian language, we combined manual and automated-based techniques. We extracted learners' writings from a language learning Social Networking Service (SNS), Lang-8[4], as raw data. Lang-8 is a website where learners write journals and native speakers manually highlight and correct the errors sentence by sentence.

After some preprocessing cleaning, we aligned words in the learners' sentences with words in the native-corrected sentences to identify error positions and the incorrect words in a two-step procedure. In the first step, we aligned the words with dynamic programming and some heuristic rules and then asked a native speaker to correct the results. Then, we extracted a confusion matrix from the aligned corrected sentences. In the second step, an automatic procedure aligned sentence pairs and assigned error tags based on the confusion matrix. We called these procedures ***rule-based*** and ***hybrid*** (*rule-based improved by the confusion matrix*), respectively. Our automatic work focused on a word-to-word error alignment because we do not have a gold standard for a phrase-based alignment of manually annotated data.

This alignment covers spelling errors, unnecessary words, omitted words, affixation errors, and replacement error types. The precision of the alignment increased from 70.4 per cent in the *rule-based* to 89.4 per cent in the *hybrid* procedure. We also carried out an experiment to show that human beings have difficulty doing semantic and syntactic analysis using a short window size.

We performed a preliminary experiment to identify the error types using the alignment data. We noticed that some error types show poor accuracy.

The next section briefly reviews related work on statistical alignment in monolingual corpora. Section 3 explains how we prepared and pre-processed the data. Section 4 explains the two main experiments conducted to show how the re-correction improved alignment precision and the need for human judgement of the error position and error type. This also describes a preliminary automated error-detection system using

[4] http://lang-8.com

this corpus. In Section 5, we describe the experimental results and in Section 6, we point out the importance of this corpus and future directions for our work.

2. Related Works

Many learner corpora, especially in English, are publicly available as language resources. Starting in 2011, the NLP community has come together to share the task of grammatical error correction. The advantage is all the participants use the same training and test sets, and the same evaluation metrics (Leacock et al. 2014). Instead of the training data and the test data provided by the organizer, the participants can exploit other language resources as long as they are publicly available.

Related research for annotated learners' sentences has been done in some languages such as English (Fraser and Marcu 2006; Izumi et al. 2005), Korean (Lee et al. 2012), and German (Boyd 2010). For instance, Hana et al. (2010) worked on developing a learner corpus for Czech from handwritten documents that were transcribed into HyperText Markup Language (HTML). They converted these documents into Prague Markup Language (PML) format and then an annotator manually corrected the document.

Nagata et al. (2011) did other work on a learner error corpus. They created an English learner corpus that was manually error-tagged and shallow-parsed. The error annotation was given using Extensible Markup Language (XML) syntax by tagging a word or phrase that contained errors while a missing word was inserted in the missing word position.

As for an Indonesian corpus, in general, some Indonesian corpora are available online such as the bilingual Indonesian-English parallel corpus called the *Identic Corpus* (Larasati 2012), the *1 Million POS-Tagged Corpus*[5] that contains about 39 thousand sentences, and an Indonesian corpora repository (Manurung et al. 2010). To the best of our knowledge, no error-corrected annotated learner corpus for Indonesian language is currently available.

3. Data Preparation

We used raw data from Lang-8 written by Indonesian learners[6] gathered in 2011.[7] Lang-8 is a multilingual language learning and

[5] http://panl10n.net/english/OutputsIndonesia2.htm

[6] Language is initially identified via the author profile as metadata and then checked by human processors.

language exchange SNS where learners from 180 countries write a topic journal that consists of some sentences in their target language, which are then corrected by native speakers of that language, who have likewise written a journal in the language they are learning. Each learner is encouraged to correct others' entries written in their first language. It is possible to have more than one correction for one sentence because more than one native speaker can make corrections to one journal entry.

Learner's sentence	Saya benar-benar tertalik dari manganya selama panjang. (I was really attracted to the manga for a long time.)		
Native speakers' corrections	1	❐	Selama ini saya benar-benar tertarik dengan manga ini.
		✍	**[f-blue]**Selama ini saya benar-benar tertarik dengan manga ini**[/f-blue]**.
	2	❐	Saya benar-benar tertarik ~~dari~~ dengan manga~~nya~~ ini selama ~~panjang~~ ini.
		✍	Saya benar-benar tertarik **[sline]**dari**[/sline] [f-blue]**dengan**[/f-blue]** manga**[sline]**nya**[/sline] [f-red]**ini**[/f-red]** selama **[sline]**panjang**[/sline] [f-blue]**ini**[/f-blue]**.
	3	❐	Saya benar-benar tertalik dari manganya selama panjang. Saya benar-benar tertarik dengan manga ini.
		✍	Saya benar-benar**[f-red]** tertalik dari**[/f-red] [f-red]**manganya selama panjang**[/f-red]**. Saya benar-benar **[f-blue]**tertarik dengan**[/f-blue] [f-blue]**manga ini**[/f-blue].**

Table 1. Examples of learner's sentences.

These data have extraneous content, noises, such as a native's tags, comments, and sentences in the learner's first language. A native's tags are the tags produced when a native speaker highlights and corrects the learner's sentences. Table 1 shows an example of these noises, along with how the native speakers highlighted and corrected the sentence. The native's tags are inside square brackets and the comment or L1 sentence is in the parentheses in the first row. The symbol ❐ indicates the website view of the native-corrected sentence, while the symbol ✍ indicates a sentence in the raw data with native's tags. In example 1, the native speaker only highlighted the corrected sentence. In example 2, the native speaker crossed out the incorrect words and highlighted the corrected

[7] http://cl.naist.jp/nldata/lang-8

words with different colours. With a different style of correction, in example 3, the native speaker rewrote and highlighted the sentence then wrote the correction in the second line. He/she highlighted the original words and the words corrected with a different colour.

We cleaned these noises automatically by deleting the native's tags then we excluded the sentences written in other languages. Originally, the data consisted of 783 journals written by 107 learners from fifteen different countries. These journals contained 6,559 learners' sentences or 77,201 words with 8,673 word types. Because some sentences had two or more corrections, after the cleaning process, this resulted in 7,420 sentence pairs.

To produce high-quality native-corrected sentences, we asked a native speaker to check the sentence pairs manually. First, a native speaker re-corrected spelling and punctuation errors; re-corrected words that violated the multi-word expression rules (a multi-word expression is written as one word if it is surrounded by both a prefix and a suffix); and replaced out-of-vocabulary words of corrected sentences. In Lang-8, some native speakers wrote corrected sentences in an informal writing style. Thus, we were required to rewrite them and discard sentence pairs that could not be rewritten. These re-correction processes referred to Indonesian grammar books (Alwi 2000; Sneddon et al. 2010) yet the learners' sentences were not affected.

To keep the information about the writer for later functionality, each error-corrected sentence pair has a journal ID that refers to the topic, learner ID and L1. The result is aligned sentences with their word types, POS tag, word correction, and their error annotations.

4. Experiments

We conducted two experiments with this corpus. The first experiment was to show that re-corrected alignments help the automatic alignment. After cleaning noises, we split the data into two parts: 658 sentence pairs for rule-based alignment and 5,557 sentence pairs for hybrid alignment.

The second experiment had two goals. The first goal was to evaluate the difficulty of error identification of Indonesian learners' sentences. The second goal was to report the quality of human annotators in evaluating the automatic alignment results. Finally, in this section, we show how we used the corpus for preliminary error identification.

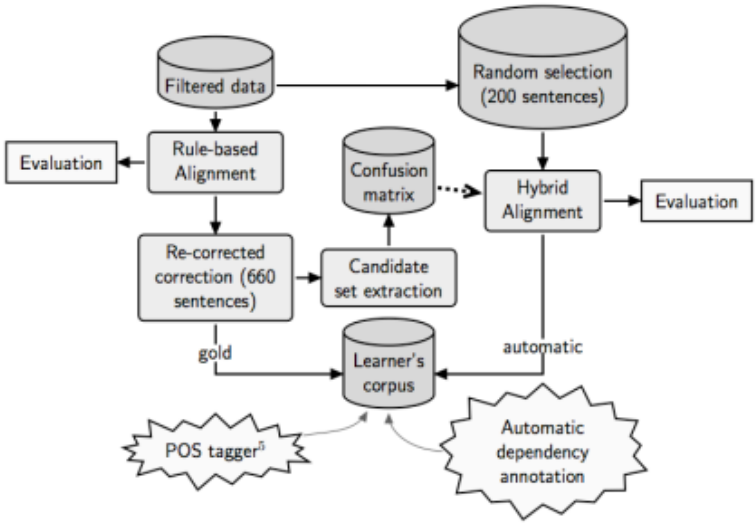


Figure 1. Semi-automated alignment process.

4.1. Semi-Automatic Alignment

Figure 1 illustrates the semi-automatic alignment generally. It was performed in two steps after filtering out the Lang-8 data as explained in Section 3. First, we did *rule-based* alignment by constructing an *edit distance* (ED) matrix (Jurafsky and Martin 2009) between a learner's sentence and a native-corrected sentence using dynamic programming. Because the learner's sentence and its correction are in the same language, we defined several heuristic rules to align them. Based on the ED matrix, we aligned a word from the learner's sentence and a word from the native-corrected sentence if they matched each other. If the words did not match, we extracted the stem word (lemma) from morphology information obtained from *Morphind* (Larasati et al. 2011). If the stem words matched, we assigned the word pair as an *affixation error* (A). The second step, *hybrid* alignment, is described later.

For pairs that did not have the same stem word, we aligned them as a *replacement error* (R) if their predecessor and successor were also aligned. We called this the *neighboring dependency rule* as explained in Wang and Zhou (2004) and shown in Equation 1. Let the triple (w_1,R,w_2) be a syntactic dependency relating two words w_1 and w_2 and a dependency relation R between them. Given a pair of learner sentences L and a corrected sentence C, where C is a native correction of L, L contains a triple l_1,R_L,l_2, and C contains a triple c_1,R_C,c_2.

[8] http://septinalarasati.com/work/morphind/

$$align(R_L, R_C) \Leftrightarrow \begin{cases} align(l_1, c_1) \\ align(l_2, c_2) \end{cases}$$

Equation 1.

The method that was used in computing *spelling error* (S) in the *rule-based* alignment is the Python built-in library SequenceMatcher.ratio[9] from the difflib package. If the word pair follows the rule in Equation 2, it is annotated as a spelling error.

$$abs(x_i.|vowel| - x_j.|vowel|) \leq 2 \wedge abs(x_i.|consonant| - y_j.|consonant|) \leq 2 \wedge ratio \geq 0.7 \Rightarrow (x_i, y_j)$$

Equation 2.

where $x_i.|vowel|$ and $x_i.|consonant|$ are the number of vowels and consonants in the i^{th} word of sentence x, respectively.

Depending on what rule it followed, an alignment was tagged as a spelling, affixation or replacement error. Next, the system annotated unaligned words in the learner's sentence as unnecessary words and unaligned words in the corrected sentence as omitted words. After that, we asked a native speaker to evaluate and re-correct the alignment results again to create a gold standard alignment. Furthermore, we extracted a confusion matrix and computed the confusion matrix probability.

Attribute	Description
id	The id of the learner-corrected sentence
jid	The id of the journal which the sentence belongs to
uId	The id of the learner related to learner's metadata
nId	The id of learner's L1
oText	Sentence written by learner
oPOS	POS tag of learner's sentence
oDeps	Dependency relation of learner's sentence
cText	Sentence corrected by native speaker
cPOS	POS tag of corrected sentence
cDeps	Dependency relation of corrected sentence
Token list	List of words that were corrected Each token has index position in learner's sentence and corrected sentence, an error type, and a corrected word

Table 2. XML file structure.

[9] http://docs.python.org/2/library/difflib.html

Second, we chose a candidate error from the confusion matrix in assigning a spelling, affixation or replacement error in the *hybrid* process based on Equation 3 to improve the precision before applying the same rules as in the *rule-based* procedure to align the remaining words.

$$\widehat{y_j} = \underset{y_j}{arg\ max}\ M(y_j|x_i)$$

Equation 3.

where *M* is the confusion matrix.

In a different process, we performed POS tagging on every word using *Morphind* and merged them into the aligned sentences. We also manually assigned dependency relations for the *rule-based* results, trained them using an MST parser (McDonald et al. 2006) and ran an automatic parser for *hybrid* results using the in-house training parser, with 81.17 per cent accuracy. This alignment was saved in XML format to achieve readability and easy extensibility. Table 2 shows the XML format for the annotation schema.

Figure 2, below, is an example of two alignment pairs. For each example 2a and 2b, the upper sentence is a learner's sentence and the lower sentence is the native-corrected sentence. The vertical lines show alignments between two sentences; the curves represent dependency relations. We provided two examples to show how one learner's sentence was corrected in two different ways. The *replacement* in Figure (2a) and Figure (2b) was corrected in the same way. Before we analyze the error, we will explain the clitic. The *clitic* '-nya' can be written together as one word after a transitive verb as a direct object, after a ditransitive verb as an indirect object, after a noun as a determiner or third person possessive pronoun, or after a preposition as a preposition object. In this example, *Morphind* assigned 'nyanyinya' with VSA_PS3, a verb with a third person pronoun. However, our heuristic rule found that 'nyanyi' is an intransitive verb (VSA is changed into VSAI) that never has a direct object. The conclusion is that the word 'nyanyinya' is an incorrect construction.

For Figures 2a and 2b, the *relative pronoun* 'yang' is never followed by a preposition; the word 'yang' is followed by a verb or an adjective so the phrase 'yang di' is an incorrect form. We use the symbol λ for a null string. In these figures, 'yang di' can be corrected in two different ways. The first correction (2a) is to add a verb, 'bermain', so the preposition 'di' follows the verb and is labelled as a *prep*. The second correction (2b) is to delete the relative pronoun 'yang' so the verb 'mendengarkan' is directly

followed by the *preposition* 'di' and is labelled as a *prep*. Therefore, in Figure 2a, the object 'adik' is in the room while in Figure 2b, it is the subject 'Saya' that is in the room.

4.2. Human Judgement

In the second experiment, we chose 100 sentences randomly and extracted sequences of five and seven words (partial sentences) randomly. We chose such limited contexts for humans to judge because those are common window sizes used as features in machine learning. Then, we asked two native speakers to report on three tasks: (TASK_A) is whether a partial sentence had an error, (TASK_B) is which word was incorrect, and (TASK_C) is their suggested correction. Finally, we computed the inter-annotator agreement using the Kappa Statistic (Chodorow et al. 2012), as in Equation 4 for TASK_A and TASK_B.

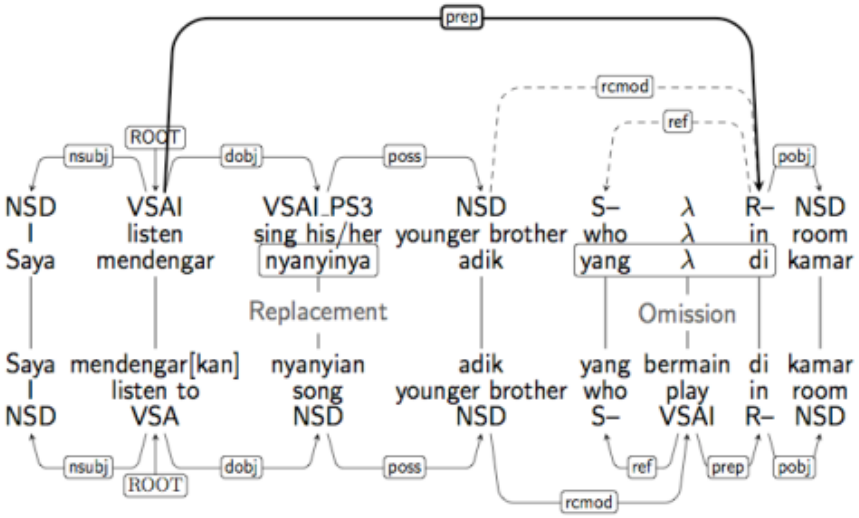
$$\kappa = \frac{P_a - P_e}{1 - P_e}$$

Equation 4.

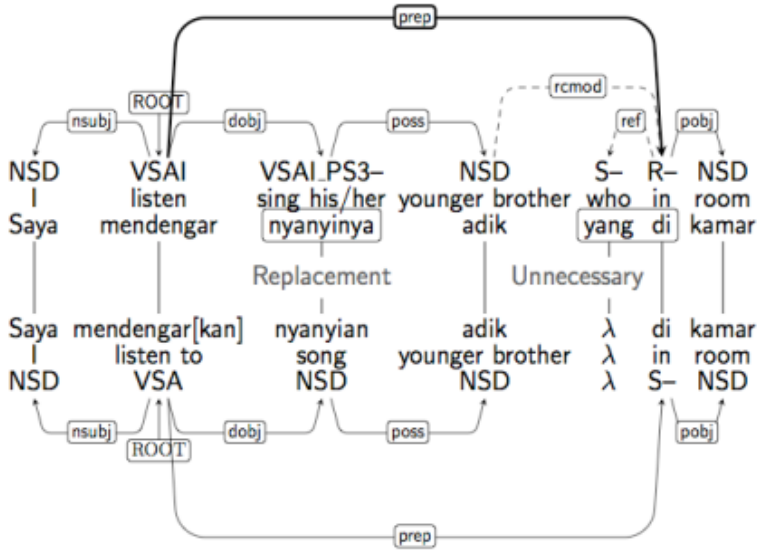
4.3. Preliminary Error Identification

We utilized the corpus for a preliminary experiment on learner's error identification using LibSVM[10] (Chang and Lin, 2011) with 10-fold cross validations that run on the manual re-corrected alignment data. We used surface words, POS tags, lemmas, prefixes, suffixes, bigrams, and trigrams as the features. We plan to use the corpus to develop an automatic system that gives feedback to the learners about the mistakes they make as well as the position of the mistaken words.

[10] http://www.csie.ntu.edu.tw/~cjlin/libsvm



(a) Replacement and omission example



(b) Replacement and Unnecessary Example

Figure 2. Examples of error-corrected sentence alignment.

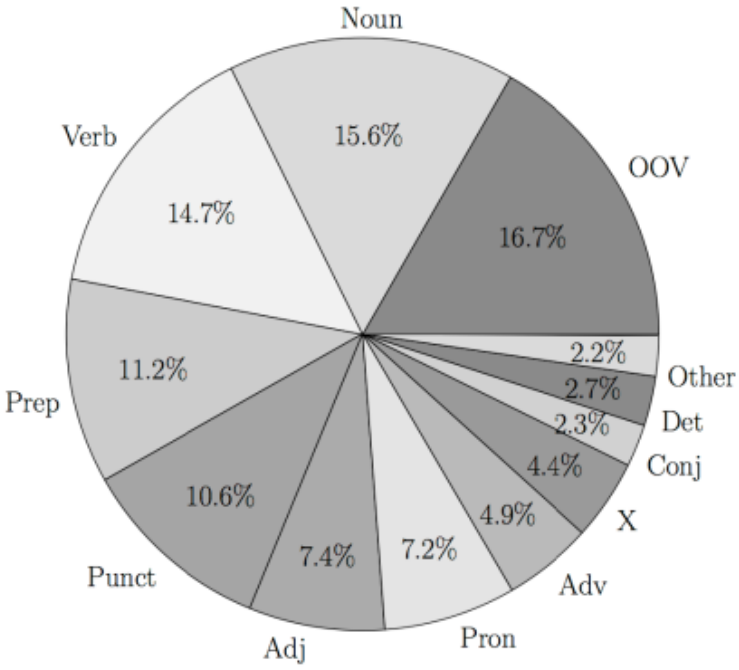


Figure 3. Error distribution.

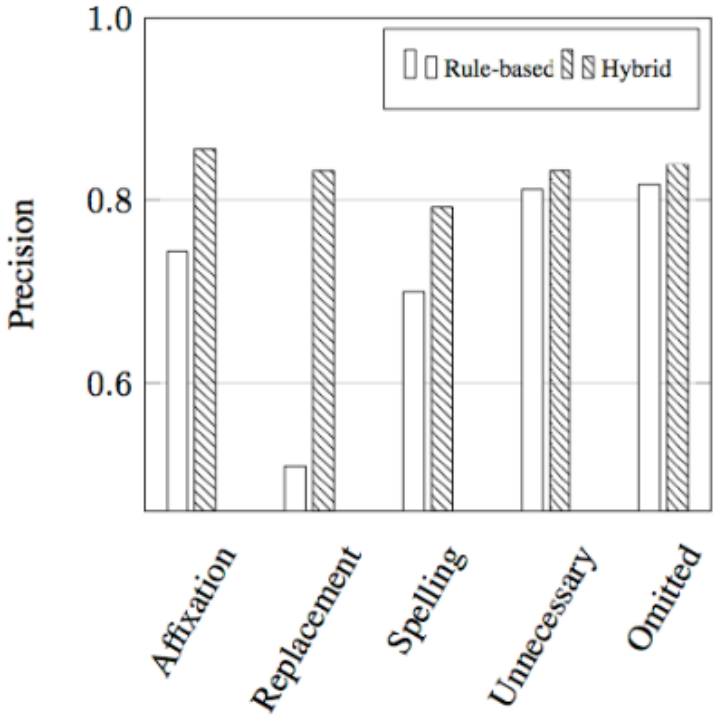


Figure 4. Precisions of rule-based and hybrid alignments.

Machine Annotation		Corrected Annotation					Multi-Alignment
		A	R	S	U	M	
Affixation	A	6	0	0	1	0	1
Replacement	R	1	88	1	14	4	0
Spelling	S	♣17	3	80	0	0	1
Unnecessary words	U	1	*16	3	114	0	3
Omitted words	M	0	**24	0	0	126	0

Table 3. Confusion matrix of the error-corrected sentences from 202 randomly selected sentences.

5. Evaluation

5.1. Evaluation of Semi-Automatic Alignment

Based on the alignment results, Figure 4 shows that seven error types dominated the error distribution, with more than 7 per cent in each type. We differentiated between OOV and X; OOV is all unknown words caused by a spelling error as well as words that are not used in writing while X is all words that are not recognized by the POS tagger. Then we compared the *rule-based* alignment results with the native speaker re-corrected results.

To compare the precision, because we do not have a gold standard for the test data, we chose 202 random sentences and ran them through the *rule-based* and *hybrid* processes, and then we asked a native speaker to evaluate them manually. We computed the alignment precision by dividing the correct alignments by the number of automatic alignment results. In Figure 4, the precisions are drawn as a white bar and a diagonal line bar. On average, the precision increased from 70.4% to 89.4%, and the sentence alignment completeness increased from 14.8% to 56.9%. A sentence was counted as complete if alignments and error types in one sentence are all correct. Unfortunately, the replacement precision for the *rule-based* result is quite low because the system proposed more than one error annotation for each alignment to help manual evaluation, which means the number of errors increased. We provided multiple annotations to help the annotator easily choose which error tag is appropriate for each alignment.

This experiment showed that the confusion matrix from manual error annotation improved the alignment quality of the learner's error-corrected corpus. Table 3 above represents an analysis of the confusion matrix. Figure 4 indicates that the automated system aligns word pairs well but Table 3 shows that the system cannot assign a correct error type for some cases. For example, (1) affixation errors were identified as spelling errors (♣) (if the correction words were only one or two characters, similar to one of the affixes, and were unavailable in the confusion matrix); (2) some replacements were identified as unnecessary words (*) or omitted words (**); and (3) incorrect replacements happen because the *neighboring dependency rule* failed to identify them. Besides, the multi-alignment is a phrase that could not be assigned as a word-to-word alignment.

```
<s> id="5" jId="598062" nId="5" uId="179855"
oText="Saya mendengar nyanyinya adik yang di kamar"
oPOS="PS1 VSA VSAI_PS3 S-- R—NSD" oAffixes="-|- meN|- -
|- -|- -|- -|- -|-" oStem="saya dengar nyanyi_dia adik yang di
kamar" oDeps="1:nsubj -1:root 1:dobj 2:poss 5:ref 3:rcmod
5:pobj"
cText="Saya mendengarkan nyanyian adik yang bermain di
kamar" oPOS="PS1 VSA NSA NSD S—VSA R—NSD"
cAffixes="-|- meN|kan -|an -|- -|- ber|- -|- -|-" cStem="saya
dengar nyanyi adik yang main di kamar" cDeps="1:nsubj -1:root
1dobj 2:poss 5:ref 3:rcmod 5:prep 6:pobj"
<t i="2" j="2" p="nyanyinya;nyanyian" eT="RV">nyanyinya</t>
<t i="5" j="5" eT="OV" c="bermain" >
</s>

<s> id="6" jId="598062" nId="5" uId="179855"
oText="Saya mendengar nyanyinya adik yang di kamar"
oPOS="PS1 VSA VSAI_PS3 S-- R—NSD" oAffixes="-|- meN|- -
|- -|- -|- -|- -|- -|-" oStem="saya dengar nyanyi_dia adik yang di
kamar" oDeps="1:nsubj -1:root 1:dobj 2:poss 5:ref 3:rcmod
```

Figure 5. Alignment result in XML format.

In Table 3, some spelling errors were identified as affixation errors because the system cannot differentiate spelling and affixation errors if the word pair satisfies both spelling and affixation rules. The morphology analyzer *Morphind* cannot produce a correct affixation if the word is spelled incorrectly. This misspelled word is tagged as OOV or other POS. This table also shows some unnecessary words and omitted words were identified as replacements. The unnecessary and omitted words were misidentified because a native speaker made a correction that changed the order of the sentence, so they satisfied the *neighboring dependency rule*.

Figure 2 shows the result of the alignment examples from Figure 2 in XML format. Figures 2a and 2b are represented as XML with sentence id "5" and "6" respectively for the same journal id (jId), learner L1 (nId), and user id (uId). We use same variables listed in Table 2. The tag s represents a sentence where tag t represents a token. We specifiy the omitted error type differently, which is tagged as I, which means 'insert'. The dependency relation for each word is written as head:label while the affix is written as prefix|suffix. A word that does not have a prefix or suffix is written as -|-. The omitted word still has i variable although the corresponding word is not available in the learner's sentence to identify where the word is inserted. For the same reason, for an unnecessary word, the j variable has the value of the index of the next word after deletion. The error tags for these examples are translated into *verb replacement* (RV), *verb omitted* (OV) for sentence id "5"; and *verb replacement* (RV) and *conjunction insertion* (IS) for sentence id "6". The POS, the stem word, and the dependency relation are written as sentence elements because we only save the token that has a correction. For ease of use by any researcher, the automatic result of the corpus, without the dependency construction, can be accessed freely from http://sourceforge.net.[11]

5.2. Evaluation of Human Judgement

Table 4 shows the κ score for TASK_A and TASK_B defined in Section 4.2. In general, a 7-word context provides better agreement while the 5-word partial sentences are more likely to be mistaken. For TASK_A the native annotators better agree when identifying whether the partial sentence has an error compared to TASK_B, identifying which word is grammatically erroneous. TASK_B shows that native speakers mostly disagree about which word is incorrect, typically in grammatical cases. This disagreement is possibly because people pick a different word as the error in the same sentence. This indicates the lowest agreement score in Table 4 for a 5-word column.

Agreement	5-words	7-words
TASK_A (κ)	0.599	0.655
TASK_B (κ)	0.196	0.634

Table 4. Inter-annotator agreement of partial sentence error identification.

[11] http://sourceforge.net/projects/indonesianlearnercorpus/files/IndLCorpus/IndLearnerCorpus.xml

Kinds of Error	5-words	7-words
(1) Unidentified	2	3
(2) Error detection mismatch	15	4
(3) Error correction mismatch	17	7

Table 5. Error classification based on human judgment.

Error Type	Classification	
	Precision	Recall
*Replacement	0.089	0.577
After splitting error type		
Noun	0.160	0.800
Verb	0.103	0.316
Preposition	0.340	0.674
Adjective	0.314	0.792
Average Replacement	**0.229**	**0.645**

Table 6. Classification based result for error detection system.

In addition, we defined three categories to classify human judgement: (1) Unidentified, (2) Error detection mismatch, and (3) Error correction mismatch. We organized the human judgements into these three categories. Table 5 shows the results. In category (1), unidentified, the annotator was confused about the meaning of the partial sentence and was not able to make a decision. In category (2), error detection mismatch, one annotator stated that the sentences had an error while the other annotator said there was no error. In fact, such sentences were acceptable because at least one native classified it as correct. Category (3), error correction mismatch, was identified if the sentence had an error but the annotators identified different words as the error so they proposed different corrections. The last category also supports the conclusions deduced from Table 4 that identifying the word containing the error sometimes depends on human perception. The improvement in inter-annotator agreement in Table 4 and the decrease in mismatches in Table 5 show that, in a longer context, humans agree better on the sequence of words. For the complete sentences, we got Kappa Statistic κ about 0.896.

As preliminary work on automatic error identification, our experiment showed that classifying the error types into replacements, affixations and omitted words was difficult. To get better accuracy, we classified these errors into a more fine-grained error type based on the POS information. We subdivided the affixation errors into inflection and derivation errors, which further split into noun and verb inflections and derivations. We

divided replacement errors into several errors based on their POS. We found that splitting the replacement errors into more fine-grained error types improved the accuracy. As shown in Table 6, for example, we report the results of the replacement error type detected by our model using the SVM classifier. This table also shows the score of more fine-grained error types as nouns, verbs, prepositions, and adjectives. *Average replacement* is computed from the average score of noun, verb, preposition, and adjective for each precision and recall. This table shows that for the four error types, the average precision improved by 14 per cent and the recall improved by 6.8 per cent compared with the single-run replacement (*Replacement, the first line in the table). We will use the replacement result as a baseline for our next experiment. The model indeed needs to be improved significantly by extracting more features such as the dependency features combined with information that can be generated from a large native corpus running on a more sophisticated method.

As a starting point of providing learner resources for the Indonesian language, for instance, learner sentences can be analyzed to conceive the more frequent errors that are not made by native speakers. Moreover, a larger confusion matrix can be extracted from this corpus to suggest multiple-choice question answers for the evaluation of learners' abilities. Hopefully, the limitation of learner language resources will be a great motivation to develop NLP tools for this language.

6. Conclusions

In this paper, we presented our work on developing a corpus of second-language learners with native corrections. Our work is the first effort in creating a learner corpus for the Indonesian language, one of the underdeveloped languages for NLP. We showed how we organized the error annotation information for each sentence in XML format and how we improved the automatic error alignment by extracting a confusion matrix that helped in assigning the error type for each alignment.

In the second experiment, we verified that even for humans, semantic errors in partial sentences are difficult to judge. In the Indonesian language, a sentence can be written in several ways because of its free word ordering. Specifically, it is natural to have more than one correction for one grammatically wrong sentence, which requires use of longer contexts to work with statistical methods or machine learning. Indeed, the annotator agreement for complete sentences also represented the quality of the manual check of the alignments.

Lastly, we also showed our preliminary stage of utilizing the corpus and found that this corpus can be used in the error identification of learner sentences. Our next work is to extend the error identification using syntactic information in combination with dependency rules extracted from normal sentences. We also want to utilize a large native corpus to get more information to enrich our feature set.

For a future direction, we will utilize the corpus to develop an error-detection system to show the error type and error position as feedback for second-language learners and as a resource for developing other NLP tools for Indonesian such as a dependency parser, name entity or phrase segmentation. This resource is also useful for language teachers as described briefly in the final paragraph of the previous section.

Acknowledgements

This study is supported in part by the Directorate General of Higher Education, Republic of Indonesia under BPPLN Scholarship Batch 7 fiscal year 2012-2015. We would like to thank Lis Kanashiro, Erlyn Manguilimotan, Kensuke Mitsuzawa, Kevin Duh, and Mike Barker for valuable discussions and comments.

References

Abuhakema, G., R. Faraj, A. Feldman, and E. Fitzpatrick. 2008. "Annotating an Arabic Learner Corpus for Error." In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Alwi, H. 2000. *Tata Bahasa Baku Bahasa Indonesia.* Balai Pustaka, Indonesia, third edition.

Boyd, A. 2010. "EAGLE: an Error-Annotated Corpus of Beginning Learner German." In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC10),* 1897-1902, Valletta, Malta.

Chang, C. and C. Lin. 2011. "LIBSVM: A Library for Support Vector Machines." *ACM Transactions on Intelligent Systems and Technology,* 2:27:1-27:27.

Chodorow, M., M. Dickinson, R. Israel and J. R.Tetreault, 2012. "Problems in Evaluating Grammatical Error Detection Systems." In *COLING,* 611-628. Indian Institute of Technology Bombay.

Dahlmeier, D., H. T. Ng and S. M. Wu. 2013. "Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English." In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications,* 22-31. Atlanta, Georgia: Association for Computational Linguistics.

Fraser, A. and D. Marcu. 2006. "Semi-Supervised Training for Statistical Word Alignment." In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 769-776. Sydney, Australia.

Hana, J., A. Rosen, S. Škodová and B. Štindlova. 2010. "Error-tagged Learner Corpus of Czech." In *Proceedings of the Fourth Linguistic Annotation Workshop, ACL,* 11-19. Upsala, Sweden: Association for Computational Linguistics.

Izumi, E., K. Uchimoto and H. Isahara. 2005. "Error Annotation for Corpus of Japanese Learner English." In *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora*, 71-80. Jeju Island, Korea: Association for Computational Linguistics.

Jurafsky, D. and J. H. Martin. 2009. Sp*eech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, New Jersey, second edition.

Larasati, S. D. 2012. "IDENTIC Corpus: Morphologically Enriched Indonesian-English Parallel Corpus." In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC12),* 902-906, Istanbul, Turkey.

Larasati, S. D., V. Kuboň and D. Zeman. 2011. "Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus." In *SFCM,* volume 100 of *Communications in Computer and Information Science,* 119-129. Zurich, Switzerland.

Leacock, C., M. Chodorow, M. Gamon and J. Tetreault. 2014. *Automated Grammatical Error Detection for Language Learners.* Morgan and Claypool Publishers.

Lee, S.-H., M. Dickinson and R. Israel. 2012. "Developing Learner Corpus Annotation for Korean Particle Errors." In *Proceedings of the Sixth Linguistic Annotation Workshop,* LAW VI '12, 129-133, Stroudsburg, PA, USA: Association for Computational Linguistics.

Manurung, R., B. Distiawan and D. D. Putra. 2010. "Developing an Online Indonesian Corpora Repository." In *Proceedings of the 24th Pacific Asia Conference on Language, Information, and Computation,* 243-249. Sendai, Japan.

Marcus, M. P., M. A. Marcinkiewicz and B. Santorini. 1993. "Building a Large Annotated Corpus of English: The Penn Treebank." *Comput. Linguist.* 19(2): 313-330.

McDonald, R., K. Lerman and F. Pereira. 2006. "Multilingual Dependency Analysis with a Two-stage Discriminative Parser." In *Proceedings of the Conference on Computational Natural Language Learning (CONLL),* 216-220.

Nagata, R., E. W. D Whittaker and V. Sheinman. 2011. "Creating a Manually Error-Tagged and Shallow-Parsed Learner Corpus." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, volume 1*, 1210-1219. Portland, Oregon, USA.

Ng, H. T., S. M. Wu, Y. Wu, C. Hadiwinoto and J. Tetreault. 2013. "The CoNLL-2013 Shared Task on Grammatical Error Correction." In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task,* 1-12. Sofia, Bulgaria: Association for Computational Linguistics.

Rozovskaya, A. and D. Roth. 2010. "Annotating ESL Errors: Challenges and Rewards." In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications,* IUNLPBEA '10, 28-36. Stroudsburg, PA, USA: Association for Computational Linguistics.

Sneddon, J. N., A. Adelaar, D. N. Djenar and M. C. Ewing. 2010. *Indonesian: A Comprehensive Grammar.* Australia: Routledge. Second edition.

Wang, W. and M. Zhou. 2004. "Improving Word Alignment Models Using Structured Monolingual Corpora." In *Proceedings of Conference on Empirical Methods on Natural Language Processing (EMNLP),* 198-205. Barcelona, Spain.

bookChapter

ORIGINALITY REPORT

5%	4%	3%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ijtech.eng.ui.ac.id Internet Source	1%
2	Claudia Leacock, Martin Chodorow, Michael Gamon, Joel Tetreault. "Automated Grammatical Error Detection for Language Learners, Second Edition", Springer Science and Business Media LLC, 2014 Publication	1%
3	aclweb.org Internet Source	1%
4	www.academicstar.us Internet Source	<1%
5	emmtee.net Internet Source	<1%
6	www.kaskus.co.id Internet Source	<1%
7	docplayer.net Internet Source	<1%
8	archive.org Internet Source	<1%
	vdoc.pub	

9	Internet Source	<1%
10	www.cl.cam.ac.uk Internet Source	<1%
11	www.apsce.net Internet Source	<1%
12	Lecture Notes in Computer Science, 2014. Publication	<1%
13	acl-bg.org Internet Source	<1%
14	citeseerx.ist.psu.edu Internet Source	<1%
15	docu.tips Internet Source	<1%
16	www.aclweb.org Internet Source	<1%

Exclude quotes	On	Exclude matches	< 3 words
Exclude bibliography	On		

bookChapter

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19