Handling Indonesian Clitics: A Dataset Comparison for an Indonesian-English Statistical Machine Translation System

Septina Dian Larasati
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
SIA TILDE
Riga, Latvia
larasati@ufal.mff.cuni.cz, septina@tilde.lv

Abstract

In this paper, we study the effect of incorporating morphological information on an Indonesian (id) to English (en) Statistical Machine Translation (SMT) system as part of a preprocessing module. The linguistic phenomenon that is being addressed here is Indonesian cliticized words. The approach is to transform the text by separating the correct clitics from a cliticized word to simplify the word alignment. We also study the effect of applying the preprocessing on different SMT systems trained on different kinds of text, such as spoken language text. The system is built using the state-of-the-art SMT tool, MOSES. The Indonesian morphological information is provided by MorphInd. Overall the preprocessing improves the translation quality, especially for the Indonesian spoken language text, where it gains 1.78 BLEU score points of increase.

1 Introduction

Incorporating linguistic information into statistical Natural Language Processing (NLP) applications usually helps to improve a particular NLP. Simplifying the problem beforehand, for languages with complex language constructions, is one of the approaches that is usually applied, especially when the constructions cannot be represented by a statistical model.

Incorporating morphological information as part of a preprocessing module in the SMT pipeline has been long studied, for instance in rich morphology languages such as Arabic (Habash and Sadat, 2006) or agglutinative languages such as Turkish (Bisazza and Federico, 2009) (Yeniterzi and Oflazer, 2010), and many more. This paper shows an example on how to use Indonesian morphological information on an Indonesian-English SMT system by preprocessing to gain better translation quality.

Indonesian has a complex morphology system, including affixation, reduplication, and cliticization. Here we address the problem of cliticized phrase constructions in Indonesian that occur more frequent in spoken language and social media text than in the formal written text. Having more cliticized phrases in a text makes a spoken dialogue text difficult to translate. Here we also evaluate the effect of the preprocessing on other different types of text.

2 Related Work

Indonesian or *Bahasa Indonesia* ("language of Indonesia"), is the official language of the country. Indonesian is the fourth most spoken language in the world with approximately 230 million speakers including its 30 million native speakers. In spite of that fact, Indonesian is an under-resourced language within the Austronesian language family. There is still a lot of work that is needed to be done to collect language resources or to build language tools for this language. Given the lack of language resources, the research on Indonesian Machine Translation (MT) is not so prolific, although MT is one of the major research topics in NLP.

Related MT research is mostly done for Malay, a mutually intelligible language to Indonesian, which has richer parallel language resources. Although Indonesian and Malay share a similar morphological mechanism, they mostly differ in vocabulary and in

26th Pacific Asia Conference on Language,Information and Computation pages 146–152

having several false friends.

There was a work done by (Nakov and Ng, 2009) for translating a resource-poor language, Indonesian, to English by using Malay, the related resource-rich language, as a pivot. There was another related work on incorporating morphological information for Malay-English SMT (Nakov and Ng, 2011), that focused on the pairwise relationship between morphologically related words for potential paraphrasing candidates. Unlike their previous research that focused on word inflection and concatenation, here they focused on derivational morphology. They used Malay Lemmatizer (Baldwin, 2006) and an in-house re-implementation of Indonesian Stemmer (Adriani et al., 2007) to get the paraphrasing candidates.

3 Indonesian Clitic

"A clitic is a morpheme that has syntactic characteristics of a word, but shows evidence of being phonologically bound to another word."[1]. In this paper, we focus on the Pronoun and Determiner clitics which are mainly bound to Indonesian Verbs and Nouns. Figure 1 shows examples on how these clitics are bounded.

(1) *kumengirimkanmu*
ku+ *mengirimkan* *+mu*
I send you
"I send you"

(2a) *bukunya*
buku *+nya*
book his/her
"his/her book"

(2b) *bukunya*
buku *+nya*
book the
"the book"

Figure 1: Indonesian cliticized phrase examples. The suffix '*-nya*' is ambiguously translated to English, which can be either a Possessive Pronoun or a Determiner depending on the context (2a and 2b).

A clitic can occur before its main words (proclitic) or after (enclitic). Figure 2 shows some of the patterns on how the clitics (proclitics and enclitics) are usually bounded to Verbs and Nouns as their main word.

[1]http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsACliticGrammar.htm

(1)	(I) *ku+*		*+ku* (I)
	(you) *kau+*	[Verbs]	*+mu* (you))
			+nya (him/her/it)
			+-nya (him/her/it)
(2)			*+ku* (my)
		[Nouns]	*+mu* (your)
			+nya (his/her/the/a)
			+-nya (his/her/the/a)

Figure 2: Examples of Indonesian clitic patterns on Verbs (1) and Nouns (2).

Clitics can also be bound to other Parts-of-Speech (PoS) as well, such as Adjectives, in a more complex Verb Phrase or Noun Phrase constructions.

4 Data

We want to observe the different kinds text that gain the most benefit, in terms of translation quality, from applying a preprocessing on Indonesian clitics. In order to do that, first we split the data into several different datasets that contain different kinds of text.

4.1 Data Source

The corpus we use in this work is the IDENTIC (Larasati, 2012) Indonesian-English parallel corpus. We chose this corpus because it consists of various types of text. We categorized the text in two categories by how it was produced, i.e. *en-to-id translated* text and *id-to-en translated* text. This corpus consists of $\pm$45K sentences or $\pm$1M words. In those categories, we also found different types or genres of text that we exploit. Given below are the two text categories, by how they were produced, and the types of text they consist of.

- **en-to-id translated text**: the text that was produced by translating English text to Indonesian and it consists of
  - (**p**) the Indonesian text that was translated from PENN Treebank sentences (Marcus et al., 1993)
  - (**a**) a small portion of comparable Indonesian-English international articles taken from the web
  - (**s**) English movie subtitles in which the texts are mainly in a spoken dialogue style

- **id-to-en translated text**: the text that was produced by translating Indonesian text to English and it consists of articles in Science (**c**), Sport (**o**), International (**t**), and Economy (**e**) genres.

The statistic of the text based on the sources are given in Table 1.

source	#sentences	id#token	en#token
p	17626	404540	424974
a	164	3208	3566
s	3161	24274	28544
c	6355	111065	123205
o	4465	112451	114155
t	6641	167839	177164
e	6532	168611	182795
Total	44944	991988	1054403

Table 1: Text source statistics in terms of number of sentences and number of tokens on Indonesian and English side.

4.2 Dataset

For our dataset comparison, we divide the text into five different datasets (F,H,S,E,I) to be compared in section 6. The division of the text for the datasets is shown in Figure 3.

- **F**: a dataset with proportional mixed texts for training, tuning, and testing data
- **H**: a dataset with proportional mixed texts for training, tuning, and testing data, but with a smaller training data compared to **F**
- **S**: a dataset with proportional mixed texts for training data (excluding the subtitles) and subtitles text as the tuning and the testing data
- **E**: a dataset with *en-to-id translated* text as the training data, and *id-to-en translated* text as the tuning and the testing data
- **I**: a dataset with *id-to-en translated* text as the training data, and *en-to-id translated* text as the tuning and the testing data

For each datasets, the sentences are chosen randomly without replacement, but keeping them in the same proportion as to the original text source. We keep the tuning and the testing data size similar (1K sentences), while the training data varies depending on the rest of the text available. We make the same tuning data for **F** and **H** dataset and for their testing data as well.

distribution	**training**	**tuning**	**testing**
	pas-cote	pas-cote	pas-cote
F	●●●-●●●●	●●●-●●●●	●●●-●●●●
H *	●●●-●●●●	●●●-●●●●	●●●-●●●●
S	●●○-●●●●	○○●-○○○○	○○●-○○○○
E *	●●●-○○○○	○○○-●●●●	○○○-●●●●
I *	○○○-●●●●	●●●-○○○○	●●●-○○○○

● *: included in the dataset*

○ *: excluded from the dataset*

size	**training**	**tuning**	**testing**
F	42944	1000	1000
H *	20951	1000	1000
S	41783	1000	1000
E *	20951	1000	1000
I *	23993	1000	1000

Figure 3: Division of the text for the datasets. Datasets marked with * are dataset with much smaller training data (±21-24K sentences) compare to the full size ones (±41-43K sentences). **p**,**a**,**s** text type are *en-to-id translated* text, while **c**,**o**,**t**,**e** are *id-to-en translated* text.

5 Experiment

For the SMT experiment, we built five *baseline* SMT systems each trained using different datasets (**F**,**H**,**S**,**E**, and **I**) and compare each of them against another system (*unclitic*) trained using its preprocessed dataset version.

5.1 *baseline* system

The *baseline* SMT system is in lowercased-to-lowercased Indonesian-to-English translation direction. We use the state-of-the-art phrase-based SMT system MOSES (Koehn et al., 2007) and GIZA++ tool (Och and Ney, 2003) for the word alignment.

We build our Language Models (LMs) from the seven English monolingual LM data provided by the Seventh Workshop on Statistical Machine Translation (WMT 2012) translation task[2]. Those monolin-

[2]http://www.statmt.org/wmt12/translation-task.html

input	*kumengirimkanmu*			*bukuku*	
	ku+	*mengirimkan*	*+mu*	*buku*	*+ku*
analysis	aku<p>_PS1+	meN+kirim<v>+kan_VSA	+kamu_PS2	buku<n>_NSD	+aku<p>_PS1
gloss	I	send	you	book	I
english	I send you			my book	
output	*ku*	*mengirimkan*	*mu*	*buku*	*ku*
input	*buku kecilku*			*buku-bukunya*	
	buku	*kecil*	*+ku*	*REDP.buku*	*+nya*
analysis	buku<n>_NSD	kecil<a>_ASP	+aku_PS1	buku<n>_NPD	+dia<p>_PS3
gloss	book	small	I	books	he/she/the
english	my small book			his/her/the books	
output	*buku*	*kecil*	*ku*	*buku-buku*	*nya*
input	*buku resepku*			*kukirim*	
	buku	*resep*	*+ku*	*ku+*	*kirim*
analysis	buku<n>_NSD	resep<n>_NSP	+aku_PS1	aku<p>_PS1+	kirim<v>_VSA
gloss	book	recipe	I	I	send
english	my recipe book			I send	
output	*buku*	*resep*	*ku*	*ku*	*kirim*

Figure 4: MorphInd analysis examples for Indonesian phrases that contain cliticized word and the preprocessing output after separating the clitic(s). The Verb Phrase's clitics are the Subject or Object of the Verb, while the enclitic on the Noun Phrase is a Possessive Pronoun of the Noun.

gual data are:

- Europarl Corpus
- News Commentary Corpus
- News Crawl Corpus (2007-2011)

We treat them as seven separate LMs, which correspond to seven LM features in MOSES configuration file. We use SRILM (Stolcke, 2002) to build the LMs. The quality of the translation result is measured using the BLEU score metric (Papineni et al., 2002).

5.2 *unclitic* system

As we have seen in Figure 2, Indonesian clitics have a fairly simple pattern and each is aligned to a different individual word in English. We use a finite state Indonesian morphological analyzer tool, MorphInd (Larasati et al., 2011) to find the correct clitics instead of just using a simple pattern matching with regular expression. This is to make sure that we do not cut the word in a wrong morpheme segmentation.

We preprocess the text by separating the clitics given the Indonesian clitics schema and MorphInd correct clitics detection, to make the alignment model simpler. Figure 4 shows several MorphInd analysis examples. The *input* shows the original words in Indonesian and the *output* shows the new text after we apply the preprocessing.

The preprocessing is applied on the training, the tuning, and the testing data. Then we build another SMT system (*unclitic*) with the same setting as the *baseline* system but using the new preprocessed data.

6 Result and Discussion

For this study, we make three combinations of dataset comparison (F-H, E-I-H, and F-S) to see how is the translation quality differs by using different datasets. Then we also observe the gain or loss caused by the preprocessing on the Indonesian clitics. The translation evaluation as a whole can be seen in Figure 5.

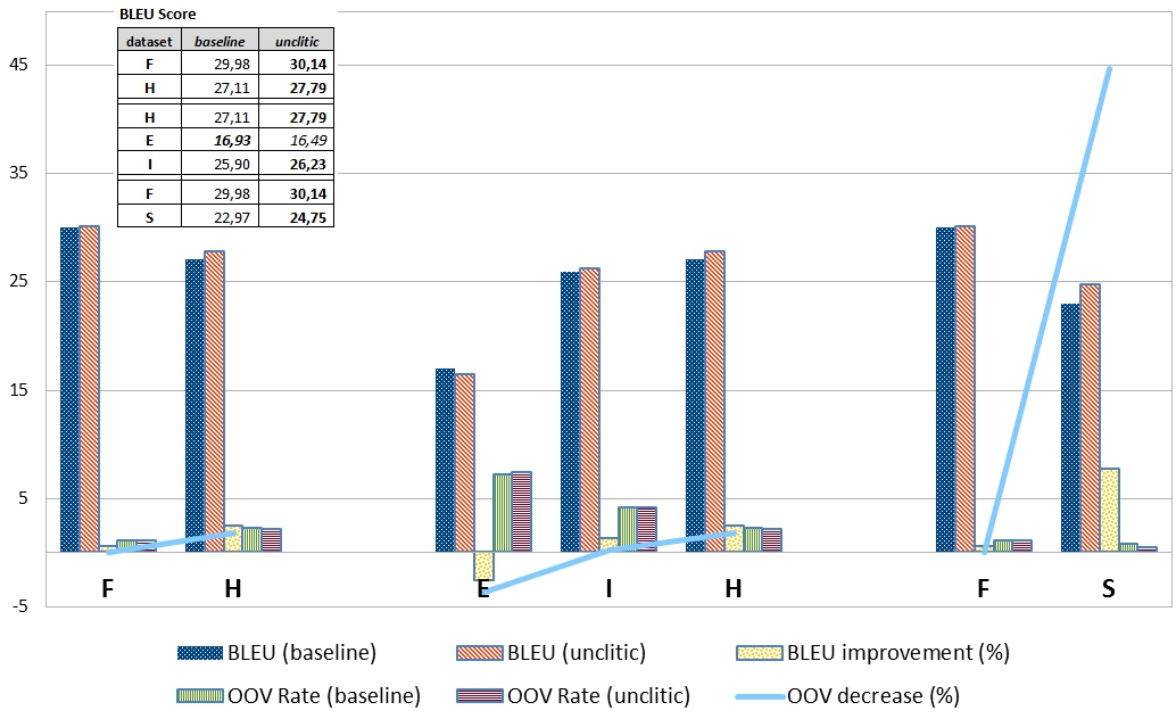

Figure 5: The *baseline* and *unclitic* SMT systems translation quality in terms of BLEU Score and their corresponding OOV Rate (%) on different datasets (F-H-S-E-I).

6.1 Working with Smaller Training Data (F-H)

The Indonesian-English parallel data is relatively small to begin with (±45K sentences or ±1M words). Here we try to push it even further to train an SMT system with only half of the training data that we have and observe the effect of applying the preprocessing on the clitics.

In this experiment, we compare the systems that are trained on **F** and **H** datasets, where the training data is in the same type but differ in size. Considering the small number of the training data that **H** has, having more data at this stage still helps to get a better translation quality. Here we also see that the smaller system gain more improvement by applying the preprocessing.

6.2 Different Text Categories (E-I-H)

Here we compare three different systems trained on three different smaller training data (21K-24K sentences), i.e. **E**,**I**, and **H** datasets. Here we see that the **E** dataset has a very high Out-of-Vocabulary (OOV) rate, which makes a poor translation result, and even the clitic preprocessing cannot help to improve the translation. In spite of that, the system trained on **H** and **I** datasets gain a better translation quality by applying the preprocessing.

6.3 Translating Spoken Indonesian (F-S)

Indonesian speakers tend to use more clitics in Indonesian spoken language, than in a formal written text. Here we put the focus on the spoken language by comparing system trained on **S** dataset (subtitles as the tuning and testing data) and compared it with system trained on **F** dataset (the mixed types text).

The BLEU score for the *baseline* **S** is far below the *baseline* **F**, although their training data sizes only differ slightly (±43K (F) and ±42K (S) sentences). This happens because Indonesian spoken dialogue is more difficult to translate.

In spite of the score difference, here we see that translating the subtitle text gains the most improvement by applying the clitic preprocessing.

7 Conclusion

We showed one linguistically motivated example on how to incorporate morphological information into

an NLP application for Indonesian. We used the state-of-the art SMT tool, MOSES, and utilized the information provided from an Indonesian morphological analyzer, MorphInd.

We compared five different SMT systems in three different combinations, where we also applied a preprocessing on the datasets. We saw that the preprocessing overall improves the translation quality, except on the E dataset (with *en-to-id translated* text as the training data) where its OOV rate is too high. The S (subtitle text) dataset benefited the most from the preprocessing.

8 Future Work

There are still other straightforward Indonesian language constructions that can be exploited to improve Indonesian-English SMT system translation quality as part of a preprocessing.

Moving a step further from morphology, incorporating additional syntactical information will be an interesting approach to do. For example, since Indonesian and English have an opposite dependency for the Noun Phrase head-modifier construction, preordering Indonesian words in a Noun Phrase before the translation takes place will be a good approach to improve the translation quality.

Having more Indonesian-English parallel sentences for the training will hopefully improve the translation quality, since currently the parallel data is still very small. This will also increase the interest to do research in this language pair.

Acknowledgments

The research leading to these results has received funding from the European Commission's 7th Framework Program under grant agreement n° 238405 (CLARA), by the grant LC536 Centrum Komputační Lingvistiky of the Czech Ministry of Education, and this work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

References

M. Adriani, J. Asian, B. Nazief, SMM Tahaghoghi, and H.E. Williams. 2007. Stemming indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing (TALIP)*, 6(4):1–33.

T. Baldwin. 2006. Open source corpus analysis tools for malay. In *In Proc. of the 5th International Conference on Language Resources and Evaluation*. Citeseer.

A. Bisazza and M. Federico. 2009. Morphological preprocessing for turkish to english statistical machine translation. In *Proc. of the International Workshop on Spoken Language Translation*, pages 129–135.

BPPT. 2009. Final report on Statistical Machine Translation for Bahasa Indonesia - English and English - Bahasa Indonesia. Technical report, Badan Pengkajian dan Penerapan Teknologi.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 49–52, New York City, USA, June. Association for Computational Linguistics.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Septina Dian Larasati, Vladislav Kuboň, and Dan Zeman. 2011. Indonesian morphology tool (MorphInd): Towards an indonesian corpus. *Systems and Frameworks for Computational Morphology*, pages 119–129, august.

Septina Dian Larasati. 2012. IDENTIC corpus: Morphologically enriched indonesian-english parallel corpus. In *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

M.P. Marcus, M.A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Preslav Nakov and Hwee Tou Ng. 2009. Improved statistical machine translation for resource-poor languages

using related resource-rich languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1367, Singapore, August. Association for Computational Linguistics.

P. Nakov and H.T. Ng. 2011. Translating from morphologically complex languages: a paraphrase-based approach. In *Proceedings of the Meeting of the Association for Computational Linguistics (ACL2011), Portland, Oregon, USA*.

F.J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

James Neil Sneddon. 1996. *Indonesian Reference Grammar*. Allen & Unwin.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*.

Reyyan Yeniterzi and Kemal Oflazer. 2010. Syntax-to-morphology mapping in factored phrase-based statistical machine translation from english to turkish. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 454–464, Uppsala, Sweden, July. Association for Computational Linguistics.