WDS 12 Proceedings of Contributed Papers, Part I, 123–129, 2012. ISBN 978-80-7378-224-5 © MATFYZPRESS

Towards an Indonesian-English SMT System: A Case Study of an Under-Studied and Under-Resourced Language, Indonesian

S. D. Larasati

Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic.

Abstract. This paper describes a work on preparing an Indonesian-English Statistical Machine Translation (SMT) System. It includes the creation of Indonesian morphological analyzer, MorphInd, and the composing of an Indonesian-English parallel corpus, IDENTIC. We build an SMT system using the state-of-the-art phrase-based SMT system, MOSES. We show several scenarios where the morphological tool is used to incorporate morphological information in the SMT system trained with the composed parallel corpus.

Introduction

Statistical Machine Translation (SMT) is one of the major research topics in Computational Linguistics. Many approaches and methodologies have been applied to solve translation problems between a language pair. Among recent developments, the statistical phrase-based approach [Koehn et al., 2003] is the one that is mostly researched.

Many kinds of linguistic information, such as morphology, syntax, and semantics, have been incorporated to solve translation problems for many different language pairs that introduce new and different challenges in the SMT field. The languages in SMT are also varied, such as agglutinative languages (e.g., Turkish, Finnish), highly inflected languages (e.g., Czech), languages without word segmentation (e.g., Chinese), languages with different word order or sentence structure, etc. Those languages come with their individual unique linguistic properties and some interesting phenomena differences when they are paired.

Unfortunately, although SMT is one of the popular research topics, many competitions, shared tasks, and research focused on only several heavily researched languages. While on the other hand, under-resourced languages are still struggling on creating more language resources and tools. These languages often encounter an issue of data sparseness when a phrase-based approach is applied. Lack of interest in these languages is also unfortunate since some of these languages may offer different and interesting language phenomena that may have been neglected or not seen yet.

This paper focuses on the resource and tool preparation to build an Indonesian-English SMT system, with several scenarios on incorporating Indonesian morphological information into the system. Unlike English or other major European languages, Indonesian is not rich in language resources and tools, which makes it difficult to catch up with research in other languages. Indonesian has a complex morphology, such as affixation, cliticization, reduplication, etc. Most of the words constructions are derivational morphology. Pairing it to English in an SMT scenario introduces new challenges in solving different language phenomena.

MorphInd, an Indonesian Morphological Analyzer

We improve SMT translation quality by utilizing the language pair's morphological features. This is done by encoding (and transforming if necessary) the morphological features of the source language, that bring the most suitable linguistic information, to get a better translation in the target language.

In order to get morphological features of Indonesian surface words, we developed an Indone-

sian morphological analyzer, MorphInd [Larasati et al., 2011]. MorphInd was developed based on a previous work by Pisceldo et al. [2008]. MorphInd delivers more robust morphological information in its output, such as the word's morphemic segmentations, lemma tag, and three positional morphological tags, which are not provided by the previous work. To understand more on how Indonesian words are constructed, we examine Table 1 which gives several Indonesian morphological operation examples and their corresponding MorphInd output. MorphInd is a Finite State Automata tool developed on FOMA [Hulden, 2009] and wrapped in a *perl* script. The tool is in *Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported* license and publicly available for download.

Table 1. Several Indonesian morphological operation examples for the verb '*ajar*' (teach) and their corresponding MorphInd output.

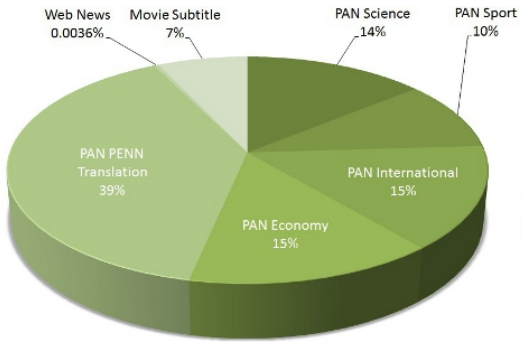
Example		**English Translation**	**MorphInd Output**
Affixation			
meN+ajar	→ *mengajar*	teach	`meN+ajar<v>_VSA`
di+ajar	→ *diajar*	to be taught	`di+ajar<v>_VSP`
peN+ajar	→ *pengajar*	teacher	`peN+ajar<v>_NSD`
peL+ajar	→ *pelajar*	student	`pelajar<n>_NSD`
peL+ajar+an	→ *pelajaran*	lesson	`pelajar<n>+an_NSD`
ajar+an	→ *ajaran*	teaching	`ajar<v>+an_NSD`
Cliticization			
peL+ajar+an+ku	→ *pelajaranku*	my lesson	`pelajar<n>+an_NSD +aku<p>_PS1`
ajar+an+nya	→ *ajarannya*	his/her teaching	`ajar<v>+an_NSD +dia<p>_PS3`
ku+ajar	→ *kuajar*	I teach	`aku<p>_PS1+ajar<v>_VSA`
Reduplication			
peN+ajar + peN+ajar	→ *pengajar-pengajar*	teachers	`peN+ajar<v>_NPD`
peL+ajar + peL+ajar	→ *pelajar-pelajar*	students	`pelajar<n>_NPD`

IDENTIC, an Indonesian-English Parallel Corpus

We composed an Indonesian-English parallel corpus and named it as IDENTIC, [Larasati, 2012], since there was no official Indonesian-English parallel corpus available. We use this corpus to train, tune, and test the SMT system.

The corpus contains texts coming from different sources which makes it vary in genres and language styles. The corpus source statistics with the number of sentences, number of tokens, and average sentence length is provided in Figure 1. The corpus was manually preprocessed which includes operations such as spelling correction, sentence segmentation, and sentence alignment. IDENTIC can be downloaded in plain text format and also in a morphologically enriched format provided by MorphInd. The morphologically enriched format is stored as a CoNLL format [Buchholz and Marsi, 2006]. Example on how the linguistic information is stored can be seen in Figure 2.

As an additional work, we also manually annotated the dependency structure for 100 sentences, with Part-of-Speech tags provided by MorphInd. Currently we do not label the dependency relations.



	#sentences	#id token	#en token	avg. id length	avg. en length
Web News	164	3208	3566	19.56	21.74
Movie Subtitles	3161	24292	28544	7.68	9.03
PAN Science	6355	111095	123223	17.48	19.39
PAN Sport	4465	112462	114155	25.19	25.57
PAN International	6641	168010	177164	25.29	26.68
PAN Economy	6540	169236	183425	25.88	28.05
PAN PENN Translation	17674	407854	429447	23.08	24.29
TOTAL	**45000**	**996157**	**1059524**		

Figure 1. IDENTIC corpus source statistics.

Indonesian

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS
1	ku	aku	aku<p>_PS1	PS1	p|P|S|1|aku|1|1
2	mencintai	cinta	meN+cinta<n>+i_VSA	VSA	n|V|S|A|meN+cinta+i|1|1
3	mu	kamu	kamu<p>_PS2	PS2	p|P|S|2|kamu|1|1
4	.	.	.<z>_Z–	Z–	z|Z|-|-|.|1|1

English

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS
1	I	I	I_PRP	PRP	PRP
2	love	love	love_VBP	VBP	VBP
3	you	you	you_PRP	PRP	PRP
4	.	.	.	.	.

Figure 2. The snippets of IDENTIC 'morphologically enriched' type, stored in 2006 CoNLL Shared Task Data Format. The fields HEAD, DEPREL, PHEAD, and PDEPREL are omitted since the values will always be set to '0', 'ROOT', '_', and '_' respectively.

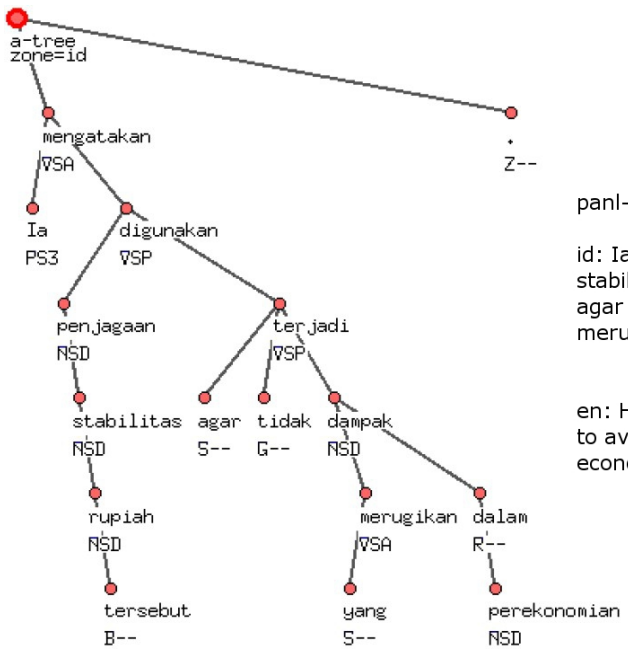


Figure 3. Dependency tree example.

SMT Experiment

We use a state-of-the-art phrase-based SMT system, MOSES [Koehn et al., 2007], to build our Indonesian-English SMT system. We use GIZA++ [Och and Ney, 2003] for the word alignment and SRILM [Stolcke, 2002] for language modeling. We will build our Baseline system and we will create several early experimental systems that incorporate the morphological information into the SMT pipeline. The translation quality metric that will be used is BLEU [Papineni et al., 2002], although the options are also open for any other metrics such as METEOR, TER, etc.

Related Work

The lack of research activities in the Indonesian-English SMT field is mostly caused by the lack of language resources and tools to support it. One simple Indonesian-English Machine Translation (MT) experiment focusing on the effect of different training data sizes has been described in BPPT [2009]. Beside the experiment's result, the report also includes a how-to installation guide for the tools that were involved.

Another MT research project was done for Malay, a mutually intelligible language to Indonesian, which has richer parallel language resources. Indonesian and Malay share similar morphological mechanism but mostly differ in vocabulary and having several false friends. There was a work done by Nakov and Ng [2011] for Malay-English SMT 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 focus on derivational morphology. They utilize a Malay Lemmatizer [Baldwin, 2006] and an inhouse reimplementation of an Indonesian Stemmer [Adriani et al., 2007] to get the paraphrasing candidates.

Proposed Early Work

There is still a lot to explore on this Indonesian-English translation direction. With the morphological analyzer tool, we explore the possibility incorporating the morphological features into the SMT system. Preprocessing in SMT and incorporating morphological information has been applied on different languages. Here we give several scenario examples to see the effect of preprocessing Indonesian text, by making its morphological structure similar to English. The preprocessed text will be the input of the SMT system and go through the whole SMT system pipeline. The particular phenomena that we will look into are the case of Indonesian cliticized nouns and plural noun phrases.

Cliticized Nouns - A clitic is a morpheme that has the syntactic characteristics of a word, but is phonologically bound to another word. Clitics in Indonesian, when it is attached to a verb, represent pronouns that are involved in the event. It is also a possessive or an explicit definiteness marker when it is attached to a Noun. We will separate the clitic(s) from its main word by detecting them using MorphInd. The examples can be found in Figure 4.

in:	(1)	*kumengirimkanmu*			(2a)	*bukunya*		(2b)	*bukunya*	
		ku+	*mengirimkan*	*+mu*		*buku*	*+nya*		*buku*	*+nya*
		I	send	you		book	his/her		book	the
		"I send you"				"his/her book"			"the book"	
out:		*ku*	*mengirimkan*	*mu*		*buku*	*nya*		*buku*	*nya*

Figure 4. Indonesian phrase with clitic examples. The suffix '-nya' is ambiguously translated to English, which can be a Possessive Pronoun or a Determiner.

Plural Noun Phrases - Plural forms in Indonesian are marked by reduplicating the word form and separating them with a hyphen. The plurality occurs not only on Nouns but

also on other POS, such as verbs, to marks a repeated action or habitual manner. For example '*buku-buku*' (books) or '*menari-nari*' (dancing over and over).

Noun Phrases with an explicit plurality marking, such as plural Numerals and plural Determiners, have a different construction compared to English. In that construction the Nouns are not in a plural form, since the plurality is already represented by the plural marking.

Plural Determiners '*para*', that does not have any corresponding English translation, behave similarly to plural Numerals. It marks its following Noun to be in plural, for example "*para pelajar*" is translated into "(the) students", although the word '*pelajar*' (student) is in its singular form. We will transform Indonesian Plural Noun Phrases to have an English Noun Phrase structure. The examples can be found in Figure 5.

in:	(1)	*para*	*pelajar*	*mengirim*	*2*	*buku*
		DET-PL	student	send	2	book
		"students send 2 books"				
out:		*pelajar-pelajar*		*mengirim*	*2*	*buku-buku*
		students		send	2	books

Figure 5. Plural Noun Phrase Transformation examples.

Evaluation

We made two experiments to show the effects on separating clitics from their main word in an SMT system scenario. We trained the system using around 42K parallel sentences and using several language models (Europarl, News Commentary, and News Crawl corpus from WMT 2012 translation task[1]). We tuned and tested the system for two different types of data: a mix genre and subtitles. The result can be seen in Table 2. It shows that separating the clitics generally improves the translation quality, especially in spoken dialog where the clitics often occur.

Table 2. Experiment Results in terms of BLEU score.

System	**BLEU**
baseline (mix)	29.98
new (mix)	30.14
baseline (subtitles)	22.97
new (subtitles)	24.75

Conclusion

We showed the development of an Indonesian Morphological Analyzer and an Indonesian-English parallel corpus as a setup to build an Indonesian-English SMT system. We also include several simple scenarios where the morphological tool can be directly used to address some morphological structure differences between Indonesian and English.

Future Work

We mentioned some additional work on annotating the dependency structure on Indonesian sentences. This is done to encourage research in the supervised Dependency Parsing field for Indonesian or syntax-based SMT approaches. In the case of Indonesian-English SMT research, findings in that field will be valuable to improve SMT translation quality. It can be used

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

as a rule-based or statistical word preordering mechanism to order Indonesian Noun Phrases that have different Head-Modifier structure compared to English. In general improvements in Indonesian NLP research as an under-studied and under-resourced language are still widely open.

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

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

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

BPPT, Final report on Statistical Machine Translation for Bahasa Indonesia - English and English - Bahasa Indonesia, Tech. rep., Badan Pengkajian dan Penerapan Teknologi, URL `http://www.panl10n.net/english/outputs/Indonesia/BPPT/0902/SMTFinalReport.pdf`, 2009.

Buchholz, S. and Marsi, E., CoNLL-X shared task on multilingual dependency parsing, in *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pp. 149–164, Association for Computational Linguistics, 2006.

Hulden, M., Foma: a finite-state compiler and library, in *Proceedings of the Demonstrations Session at EACL 2009*, pp. 29–32, Association for Computational Linguistics, Athens, Greece, URL `http://www.aclweb.org/anthology/E09-2008`, 2009.

Koehn, P., Och, F. J., and Marcu, D., 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*, pp. 48–54, Association for Computational Linguistics, 2003.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E., 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*, pp. 177–180, Association for Computational Linguistics, Prague, Czech Republic, URL `http://www.aclweb.org/anthology/P07-2045`, 2007.

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

Larasati, S. D., Kuboň, V., and Zeman, D., Indonesian morphology tool (MorphInd): Towards an indonesian corpus, *Systems and Frameworks for Computational Morphology*, pp. 119–129, 2011.

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

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

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

Pisceldo, F., Mahendra, R., Manurung, R., and Arka, I. W., A two-level morphological analyser for the indonesian language, in *Proceedings of the Australasian Language Technology Association Workshop 2008*, pp. 142–150, Hobart, Australia, 2008.

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

Wang, P., Nakov, P., and Ng, H. T., Source language adaptation for resource-poor machine translation, in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 286–296, Association for Computational Linguistics, Jeju Island, Korea, URL `http://www.aclweb.org/anthology/D12-1027`, 2012.

Yusuf, H. R., An analysis of indonesian language for interlingual machine-translation system, in *Proceedings of the 14th conference on Computational linguistics-Volume 4*, pp. 1228–1232, Association for Computational Linguistics, 1992.