IIT Bombay's English-Indonesian submission at WAT: Integrating Neural Language Models with SMT

Sandhya Singh Anoop Kunchukuttan Pushpak Bhattacharyya
Center for Indian Language Technology
Department of Computer Science & Engineering
Indian Institute of Technology Bombay
{sandhya, anoopk, pb}@cse.iitb.ac.in

Abstract

This paper describes the IIT Bombay's submission as a part of the shared task in WAT 2016 for English–Indonesian language pair. The results reported here are for both the direction of the language pair. Among the various approaches experimented, Operation Sequence Model (OSM) and Neural Language Model have been submitted for WAT. The OSM approach integrates translation and reordering process resulting in relatively improved translation. Similarly the neural experiment integrates Neural Language Model with Statistical Machine Translation (SMT) as a feature for translation. The Neural Probabilistic Language Model (NPLM) gave relatively high BLEU points for Indonesian to English translation system while the Neural Network Joint Model (NNJM) performed better for English to Indonesian direction of translation system. The results indicate improvement over the baseline Phrase-based SMT by 0.61 BLEU points for English-Indonesian system and 0.55 BLEU points for Indonesian-English translation system.

1 Introduction

This paper describes IIT Bombay's submission for the English-Indonesian and Indonesian-English language pairs for the shared task in the 3rd Workshop on Asian Translation[1] (WAT) (Nakazawa et al., 2016).

Every language pair in machine translation brings in new challenges in the form of their linguistic features. The Indonesian language, also known as Bahasa(Indonesia) is the official language of Indonesia. It is the fourth most populous country[2] in the world with approximately 190 million[3] people speaking this language. The language belongs to the Austronesian language family and has a lot of influence from Dutch language. It is also considered mutually intelligible with the Malay language. The script used is Roman/Latin script. The sentence structure followed is similar to English language i.e. Subject Verb Object (SVO). But it is highly agglutinative and morphologically rich as compared to English language. Hence, English-Indonesian is a very important language pair for translation studies.

There is very limited work related to Indonesian language machine translation. Some of the previous work done is discussed here. Yulianti et al. (2011) experimented with a hybrid MT system (HMT) for Indonesian-English translation. They created a pipeline system where the input is first translated using a rule based MT system (RBMT) and the output is further processed with statistical MT system (SMT) to improve the translation quality. The results indicate that a pure SMT system outperforms HMT system in all cases. Larasati (2012) focused on resources and tool preparation for Indonesian-English SMT system as the author described this language pair as under-resourced and

[2] http://www.infoplease.com/world/statistics/most-populous-countries.html

[3] https://en.wikipedia.org/wiki/Indonesian_language

under-studied. MorphInd, a morphanalyzer was developed as a part of the experiment. The tool could give more morphological information at a word level compared to its previous versions. The author also developed a standard parallel corpus IDENTIC, which could be used by the research community for MT related task. The experiment with preprocessed Indonesian data resulted in an improved SMT system output. Mantoro et al. (2013) attempted to find the optimal parameter for English-Indonesian SMT system by varying the weights of translation model, language model, distortion (reordering) and word penalty. And the optimally tuned SMT system is able to give a BLEU score of 22.14. Above discussed work clearly indicate that there is a lot of scope for experimentation for this language pair.

Recently, Hermanto et al.(2015) performed an experimental study with RNN language model for English-Indonesian MT system. The experiment was done on a very small set of data for neural LM and the output was compared with SMT system trained on same data. The perplexity analysis of both the systems show that RNN model system outperforms SMT system with n-gram LM.

The results of Hermanto et al.(2015) and various other research outcomes on different language pair using neural language model motivated our approach of experimentation using NLM and NNJM as a feature in SMT.

2 System Description

For our participation in WAT 2016 shared task for English ←→ Indonesian language pair, we experimented with the following systems –

1. *Phrase-Based SMT system* : This was our baseline system for the WMT shared task. The standard Moses Toolkit (Koehn et al., 2007) was used with MGIZA++ (Gao and Vogel, 2008) for word alignment on training corpus followed by *grow-diag-final-and* symmetrization heuristics for extracting phrases and lexicalized reordering. Tuning was done using Batch MIRA (Cherry and Foster, 2012) with the default 60 passes over the data and *–return-best-dev* flag to get the highest scoring run into the final moses.ini file. A 5-gram language model using SRILM toolkit (Stolcke, 2002) with Kneser-Ney smoothing was trained.

2. *Use of Neural Language Model* : A neural probabilistic language model was trained and integrated as a feature for the phrase-based translation model. For this, the default NPLM[4] implementation in Moses which is similar to the method described in Vaswani et al. (2013) was used. The goal was to examine if neural language models can improve the fluency for Indonesian-English translation and English-Indonesian translation by making use of distributed representations. We experimented with various word embedding sizes of 700, 750 and 800 for the first hidden layer in the network to get the optimal parameter while decoding.

3. *Use of Bilingual Neural Joint Language Model* : Devlin et al. (2014) have shown that including source side context information in the neural language model can lead to substantial improvement in translation quality. We experimented with Devlin's method which uses NPLM[3] in the back-end to train a neural network joint language model (NNJM) using parallel data and integrated it as a feature for the phrase-based translation as implemented in Moses. A 5-gram language model augmented with 9 source context words and single hidden layer required for fast decoding was used as a parameter to train the joint model.

4. *Use of Operational Sequence Model* : Operation sequence model was trained as it integrates N-gram-based reordering and translation in a single generative process which can result in relatively improved translation over phrase based system. OSM approach as suggested in Durrani et al. (2013) considers both source and target information for generating a translation. It deals with minimum translation units i.e. words, along with context information of source and target sentence which spans across phrasal boundries. A 5-gram OSM was used for the experimentation here.

[4] http://nlg.isi.edu/software/nplm/

These 4 systems were trained for both directions of language pair and the test data was decoded and evaluated with BLEU points, RIBES scores, AMFM scores, Pairwise crowdsourcing scores and Adequacy scores for comparative performance evaluation.

3 Experimental Setup

The data provided for the WAT 2016 shared task experiment for English-Indonesian language pair comprised of news domain data with a good mix of finance, international, science, national and sports news. The data was prepared using the scripts available with moses. After extracting the data in individual files for training, tuning and testing purpose, it was tokenized and truecased using the learnt truecased model. The training data was further cleaned for the maximum sentence length of 80 words.

For training the neural language model (Vaswani et al., 2013), additional monolingual data was used for each direction of language pair. For Indonesian-English, additional 2 million sentences of English Europarl data[5] was used for the experimentation. The data was tokenized and truecased for the experiment. For English-Indonesian direction, additional 2 million Indonesian sentences from Commoncrawl[6] was used for experiment. Since Commoncrawl provides raw data by web scraping, the Indonesian data obtained was cleaned for noisy sentences and then tokenized and truecased for training the language model. Table – 1 gives the statistics of the data used for experimentation.

Language	Training Set	Tuning Set	Test Set	For LM
English	44939 sentences	400 sentences	400 sentences	50000 sentences + 2M sentences (Europarl)
Indonesian	44939 sentences	400 sentences	400 sentences	50000 sentences + 2M sentences (Commoncrawl)

Table 1. Data used for the experiments

For training the joint neural language model (Devlin et al., 2014), the parallel data used for training the SMT system was used to train the bilingual neural language model.

4 Results & Analysis

4.1 Indonesian to English MT system

A comparative performance of baseline phrase based system, OSM system and neural LM and with joint neural LM for Indonesian-English MT system have been shown in Table-2. The translated output of all the three systems trained are evaluated for Bilingual Evaluation Understudy (BLEU), Rank-based Intuitive Bilingual Evaluation Score (RIBES) and Adequacy-Fluency Metric (AMFM).

For OSM experiment, a 5-gram operation sequence model was trained with the default settings of phrase based system as discussed in section 2. The BLEU scores shows a relative improvement of 0.21 points over the baseline phrase based system. The output of this system was submitted for human evaluation process for this direction of language pair.

For neural LM system, a 5-gram model with a vocabulary size of 100K and word embedding

[5] http://www.statmt.org/europarl/

[6] http://commoncrawl.org/

dimensions of 150 units in second hidden layer was trained with 3 different first hidden layer parameter i.e. 700 units, 750 units , 800 units. The aim was to use the most fitting model for decoding.

The model was optimized for only 5 epochs of stochastic gradient ascent due to time constraint with small batch sizes of 1000 words. The neural model obtained was added to moses.ini file as a feature with a default weight of 0.5. The translation model was tuned further to get better weights for all the parameters required of the translation system.

Approach Used	BLEU score	RIBES score	AMFM score
Phrase based SMT	22.03	0.78032	0.564580
Operation Sequence Model*	22.24	0.781430	0.566950
Neural LM with OE= 700	**22.58**	0.781983	**0.569330**
Neural LM with OE = 750	21.99	0.780901	0.56340
Neural LM with OE = 800	22.15	**0.782302**	0.566470
Joint Neural LM	22.05	0.781268	0.565860

Table 2. Experiment Results for Indonesian-English MT system
(OE – Output Embeddin; * : submitted to WAT)

Similarly, the joint neural LM using the bilingual data was also trained with the source and target vocabulary size of 100K and total n-gram size of 14 comprising of 5-gram target language model and 9-gram source context window with word embedding dimension of 750 units for the single hidden layer. The neural model obtained was included in the moses.ini file as a feature with default weight as 0.5. This decoding model was tuned further to learn the new weights with added feature and then used for translation.

Reference Sentence	Translated Sentence	Error Analysis
Moreover, syariah banking has yet to become a national agenda, Riawan said.	In addition, the banking industry had not so national agenda, said Riawan who also director of the main BMI.	Phrase insertion
Of course, we will adhere to the rules, Bimo said.	We will certainly *patuhi* regulations, Bimo said.	All words not translated
The Indonesian government last year canceled 11 foreign-funded projects across the country for various reasons, the Finance Ministry said.	The government has cancel foreign loans from various creditors to 11 projects in 2006 because various reasons.	Phrase dropped
As the second largest Islamic bank with a 29% market share of the Islamic banking industry's total assets at end-2007 albeit only 0.5% of overall banking industry's total assets, net financing margin NFM on Muamalat's financing operations increased to 7.9% in 2007 from 6.4% in 2004 due to better funding structure.	As the second largest bank of the market by 29 percent of the total assets syariah banking loans at the end of December 2007 although the market only 0.5 percent of the total assets banking industry as a whole, financing profit margin Muamalat rose to 7.9 percent in 2007 from 6.4 percent in 2004 thanks to funding structure.	Phrase dropped

Table 3. Indonesian-English NPLM based MT system output

The scores clearly indicate that both the approaches of LM i.e. neural LM generated from much bigger monolingual corpus or joint neural LM outperforms the baseline phrase-based SMT system. For WAT, the neural LM with word embedding dimensions of 700 units for the first hidden layer is submitted for participation. The BLEU score shows an improvement of 0.55 points over our baseline system. These scores may be improved with further tuning of the neural parameters.

Some translation outputs of relatively better performing NPLM system compared against the reference sentences have been given in Table-3. An analysis of the translation output was done for NPLM based Indonesian-English MT system. The output sentences were adequate and fluent to some extent. The major error found was of dropping and insertion of phrases. In some instances, the Indonesian words could not be translated to English due to lack of vocabulary learnt. Though, OOV word percentage was found to be 5% of the total words in the test set. Another major pattern error was in the choice of function words used for English language. This error might require some linguistic insight on the Indonesian side of the language pair to understand the usage of function words in the source language.

4.2 English to Indonesian MT system

For the reverse direction of language pair i.e. English-Indonesian, similar set of experiments were performed with same parameters as mentioned in section 4.1. The results obtained for the baseline phrase-based system, OSM based system, neural LM with additional monolingual data from commoncrawl with 3 different parameter variations and joint neural LM system have been given in Table-4. Since the authors do not know the Indonesian language, the translated output could not be manualy evaluated for error analysis at authors' end.

For this direction of language pair, the scores of OSM experiment is comaparable to baseline phrase based system with a score of 21.70 BLEU points. However, the joint neural language model has outperformed the neural LM and the baseline system by 0.61 BLEU scores. Joint neural LM output was submitted for manual evaluation.

Approach Used	BLEU score	RIBES score	AMFM score
Phrase based SMT	21.74	0.804986	0.55095
Operation Sequence Model	21.70	0.806182	0.552480
Neural LM with OE = 700	22.12	0.804933	0.5528
Neural LM with OE =750	21.64	0.806033	0.555
Neural LM with OE = 800	22.08	0.806697	0.55188
Joint neural LM*	**22.35**	**0.808943**	**0.55597**

Table 4. Experiment Results for English-Indonesian MT system

(OE – Output Embedding; * : submitted to WAT)

4.3 Human Evaluation Result Analysis

As a part of shared task evaluation process, the translation system performance was human evaluated using two methods: pairwise crowdsourcing evaluation compared against the baseline system and JPO adequacy evaluation for content transmission.

For Indonesian-English system, human evaluation was done on OSM system output. The crowdsourcing results show that 20% of the translations were better than the baseline system, 34% translations were comparable and 46% were worse than the baseline system. The system scored -26.00 in the crowdsourcing evaluation and 2.98 in adequacy evaluation. Table-5 shows the adequacy score

distribution as received in JPO adequacy evaluation. However, the automatic evaluation scores are found to be comparable to the baseline system.

Experiment	Approach Followed	Adequacy distribution					Adequacy Score
		5	4	3	2	1	
Indonesian-English	OSM approach	12%	18.75%	31.75%	30.5%	7%	2.98
English-Indonesian	NNJM	17.75%	25.25%	23.25%	16.5%	17.25%	3.10

Table 5. JPO Adequacy scores for English↔Indonesian

The joint neural LM approach for English-Indonesian system was submitted for human evaluation. The human evaluation scores shows that 23% of the translation were better than the baseline system, 44.75% were in tie with baseline system and 32.25% were worse than the baseline system. The crowdsourcing evaluation score is -9.250 and adequacy evaluation score is 3.10. For the JPO adequacy score, we observed that 33% sentences have at least 3 point difference between the annotator scores. The scores received have been given in Table-5.

5 Conclusion and future work

In our research group, we have been working on a usecase related to English-Indonesian Machine Translation. This motivated us to participate in this shared task despite of having no exposure to Indonesian language. Since no member of the team had any previous experience with Indonesian language, not much of the linguistic insight was used in performing the experiments. This was an enriching experience in the terms of using computational ability for machine translation with minimum linguistic insight of one of the language in pair for translation. The BLEU scores show that using neural LM helps in improving the translation quality .

In future , we would like to investigate the hyperparameters for the neural language model. We also plan to look at pure neural machine translation approaches for the English-Indoneian language pair.

Reference

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. "*Neural machine translation by jointly learning to align and translate*." In ICLR.

Cherry, Colin, and George Foster. 2012. "*Batch tuning strategies for statistical machine translation."* Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics.

Devlin, Jacob, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard M. Schwartz, and John Makhoul. 2014. "*Fast and Robust Neural Network Joint Models for Statistical Machine Translation."* In conference of the Association of Computational Linguistics.

Durrani, Nadir, Helmut Schmid, and Alexander Fraser. 2011. *"A joint sequence translation model with integrated reordering."* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics.

Durrani, Nadir, Alexander M. Fraser, and Helmut Schmid. 2013. *"Model With Minimal Translation Units, But Decode With Phrases."* HLT-NAACL.

Durrani, N., Fraser, A. M., Schmid, H., Hoang, H., & Koehn, P. 2013. *“ Can markov models over minimal translation units help phrase-based smt?.”* In conference of the Association of Computational Linguistics .

Gao, Qin, and Stephan Vogel. 2008. *"Parallel implementations of word alignment tool."* Software Engineering, Testing, and Quality Assurance for Natural Language Processing. Association for Computational Linguistics.

Hermanto, Andi, Teguh Bharata Adji, and Noor Akhmad Setiawan. 2015. *"Recurrent neural network language model for English-Indonesian Machine Translation: Experimental study."* 2015 International Conference on Science in Information Technology (ICSITech). IEEE.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan. 2007. *"Moses: Open source toolkit for statistical machine translation."* In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. Association for Computational Linguistics.

Larasati, Septina Dian. 2012. "*Towards an Indonesian-English SMT system: A case study of an understudied and under-resourced language, Indonesian."* WDS’12 Proceedings of Contributed Papers 1.

Mantoro, Teddy, Jelita Asian, Riza Octavian, and Media Anugerah Ayu. 2013. *"Optimal translation of English to Bahasa Indonesia using statistical machine translation system."* In Information and Communication Technology for the Muslim World (ICT4M), 2013 5th International Conference of IEEE.

Nakazawa, Toshiaki and Mino, Hideya and Ding, Chenchen and Goto, Isao and Neubig, Graham and Kurohashi, Sadao and Sumita, Eiichiro. 2016. *“Overview of the 3rd Workshop on Asian Translation.”* Proceedings of the 3rd Workshop on Asian Translation (WAT2016), October.

Niehues, Jan, Teresa Herrmann, Stephan Vogel, and Alex Waibel. 2011. *"Wider context by using bilingual language models in machine translation."* InProceedings of the Sixth Workshop on Statistical Machine Translation. Association for Computational Linguistics.

Schwenk, Holger. 2012. *"Continuous Space Translation Models for Phrase-Based Statistical Machine Translation."* COLING (Posters).

Stolcke, Andreas. 2002. *"SRILM-an extensible language modeling toolkit."* Interspeech. Vol. 2002.

Vaswani, Ashish, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. *"Decoding with Large-Scale Neural Language Models Improves Translation."* In *EMNLP*.

Yulianti, Evi, Indra Budi, Achmad N. Hidayanto, Hisar M. Manurung, and Mirna Adriani. 2011. *"Developing Indonesian-English Hybrid Machine Translation System."* In Advanced Computer Science and Information System (ICACSIS), 2011 International Conference of IEEE.