English-Indonesian Neural Machine Translation for Spoken Language Domains

Meisyarah Dwiastuti
Charles University, Faculty of Mathematics and Physics
Prague, Czech Republic
meisyarah.dwiastuti@gmail.com

Abstract

In this work, we conduct a study on Neural Machine Translation (NMT) for English-Indonesian (EN-ID) and Indonesian-English (ID-EN). We focus on spoken language domains, namely colloquial and speech languages. We build NMT systems using the Transformer model for both translation directions and implement domain adaptation, in which we train our pre-trained NMT systems on speech language (in-domain) data. Moreover, we conduct an evaluation on how the domain-adaptation method in our EN-ID system can result in more formal translation outputs.

1 Introduction

Neural machine translation (NMT) has become the state-of-the-art method in the research area of machine translation (MT) in the past few years (Bojar et al., 2018). As a data-driven method, NMT suffers from the need of a big amount of data to build a robust translation model (Koehn and Knowles, 2017). The lack of parallel corpora for some languages is one of the reasons why the research of NMT for these languages has not grown. Indonesian is one of the examples of such under-researched language. Despite the huge number of speakers (more than 200 million people), there have been only a few works on Indonesian MT, even towards the heavily researched language like English. While the lack of data was an issue for NMT research in this language (Trieu et al., 2017; Adiputra and Arase, 2017), the recent release of OpenSubtitles2018 corpus (Lison et al., 2018) containing more than 9 millions Indonesian-English sentence pairs gives us an opportunity to broaden the study of Indonesian NMT systems.

One of interesting linguistic problems is language style, which is the way a language is used depending on some circumstances, such as when and where it is spoken, who is speaking, or to whom it is addressed. We are interested in studying the formality of MT output, focusing on spoken language domains. Given a small dataset of speech-styled language and a significantly larger dataset of less formal language, we would like to investigate the effect of domain adaptation method in learning the formality level of MT output. Learning formality level through domain adaptation will help MT systems generate formality-specific translations.

In this paper, we conduct a study of NMT for English-Indonesian (EN-ID) and Indonesian-English (ID-EN) directions. This study has the following objectives:

1. to present a set of baseline results for EN-ID and ID-EN NMT systems on spoken language domains.
2. to examine the effectiveness of domain adaptation in:
   (a) boosting the performance of the NMT systems for both directions.
   (b) learning the formality change in spoken language EN-ID NMT systems.

To accomplish both objectives, we build the NMT systems for both EN-ID and ID-EN directions using the Transformer model (Vaswani et al., 2017). This model relies on self-attention to compute the representation of the sequence. To the best of our knowledge, there has not been any work on building NMT for those language pairs using the Transformer model.

We perform experiments using domain adaptation. We consider formal speech language as our in-domain data, and colloquial dialogue-styled language from movie subtitles as our out-of-domain data. We adopt the domain-adaptation

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 309–314
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

method used by Luong and Manning (2015) to fine-tune the trained model using in-domain data. For each translation direction, we run five experiments: three in which we do not perform domain adaptation and two when we do. We evaluate the effectiveness of the domain adaptation method using automatic evaluation, BLEU (Papineni et al., 2002), and report the score obtained from each experiment on in-domain test set. Moreover, we analyze how domain adaptation affects formality change in the translations of EN-ID NMT systems by performing a human evaluation.

2 Background

In this section, we provide background information on Indonesian language and the approaches used in our experiments.

2.1 Indonesian language

Similarly to English, Indonesian's writing system uses the Latin alphabet without any diacritics. The typical word order in Indonesian is Subject-Verb-Object (SVO). The language does not make use of any grammatical case nor gender. The grammatical tenses do not change the form of the verbs. Most of the word constructions are derivational morphology. The complexity of its morphology includes affixation, clitics, and reduplication.

In spoken language, while formal speech is similar to written language, people tend to use non-standard spelling in colloquial language by changing the word forms or simply using informal words. For example, '*bagaimana*' (how) → '*gimana*' or '*tidak*' (no) → '*nggak*'. Although the measure of formality level can be relative to some people depending on their culture, there are words that are only used in formal situation. For example, the use of pronouns like *saya*' (I), '*Anda*' (you), or certain words like '*dapat*' (can) or '*mengkehendaki*' (would like).

2.2 Neural Machine Translation

Neural machine translation (NMT) uses an encoder-decoder architecture, in which the encoder encodes the source sentence $\boldsymbol{x} = (x_1, ..., x_n)$ to a continuous representation sequence $\boldsymbol{z} = (z_1, ..., z_k)$ and the decoder translates the representation $\boldsymbol{z}$ into a sentence $\boldsymbol{y} = (y_1, ..., y_m)$ in the target language.

Several previous works implemented recurrent neural networks (RNN) in their encoder-decoder architecture (Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2015; Luong and Manning, 2015). While Bahdanau et al. (2015) used bidirectional RNN for their encoder and Luong and Manning (2015) used multilayer RNN in their architecture, both works implemented an attention mechanism in their decoder, which was able to handle the problem in translating long sentences.

The model that we use in this paper is Transformer model (Vaswani et al., 2017), which gets rid of all the recurrent operations found in the previous approach. Instead, it relies on self-attention mechanism to compute the continuous representation on both the encoder and the decoder. In order to keep track of the token order within the sequence, the model appends *positional-encoding* of the tokens to the input and output embeddings. Both of the encoder and the decoder are composed of stacked *multi-head* self-attention and fully-connected layers.

Our choice of the Transformer model is motivated by its good performance reported recently in various translation tasks, such as bilingual translation of various language directions (Bojar et al., 2018), multilingual translation (Lakew et al., 2018), and also for the low-resource with multi-source setting (Tubay and Costa-jussà, 2018). While some works empirically compare the performance of Transformer and RNN-based models (Vaswani et al., 2017; Lakew et al., 2018; Tang et al., 2018), this is not the aim of this paper. We leave the comparison of both methods for EN-ID and ID-EN NMT as future research.

2.3 Domain-adaptation

One of the challenges in translation is that words can be translated differently depending on the context or domain (Koehn and Knowles, 2017). While in-domain data is limited, we expect using available large amounts of out-of-domain data to train our model and implementing a domain-adaptation method will give the model a robust performance. Therefore, we implement the method of Luong and Manning (2015). First, we train our model on general domain data consisting of around 9 millions parallel sentences. After that we fine-tune the model using in-domain data, which means the model training is continued on only in-domain data for a few more steps.

3 Experimental Setup

We run experiments on both EN-ID and ID-EN pairs with different training scenarios as follows:

1. **IN (baseline)**: using only small in-domain data (speech language)

2. **OUT**: using only large out-of-domain data (colloquial language)

3. **OUT+DA**: using only large out-of-domain data, then fine-tune the model using only in-domain data

4. **MIX**: using a mixture of in-domain and out-of-domain data

5. **MIX+DA**: using a mixture of in-domain and out-of-domain data, then fine-tune the model using only in-domain data

3.1 Dataset

We use OpenSubtitles2018 (Lison et al., 2018) parallel corpus as our out-of-domain data and TEDtalk (Cettolo et al., 2012) as in-domain data. OpenSubtitles2018 corpus contains movie subtitles which can represent colloquial language in dialogue style. On the other hand, TEDtalk corpus contains speech language which has higher level of formality than colloquial language. The details of the dataset setting is shown in Table 1. As the training data, we use all the sentences from OpenSubtitles2018 and the train set of TEDtalk. For training the baseline system (IN), we use only TEDtalk train set. For OUT and MIX, we use OpenSubtitles2018 train set and both sets in the first phase of training, respectively. Then, for the second phase of training (fine-tuning), we use TEDtalk train set while keeping the vocabulary from the first phase train set.

As development set, we use TEDtalk tst2013 and tst2014. As test set, we use TEDtalk tst2015-16 and tst2017-plus. We notice that the test set tst2017-plus provided at the website[1] contains a small part of the train data. Therefore, we remove these common sentences from the test set and obtain tst2017-plus-mod with 1035 sentences not overlapping with the training data.

Part	Dataset	#sentences
Train	OpenSubtitles2018	9,273,809
	TEDtalk train	107,329
Dev	TEDtalk tst2013	1034
	TEDtalk tst2014	878
Test	TEDtalk tst2015-16	980
	TEDtalk tst2017-plus-mod	1035

Table 1: Dataset used in our experiments

3.2 Training details

We run our experiments using Tensor2Tensor (T2T) (Vaswani et al., 2018) on a GeForce GTX 1080 machine using a single GPU. We use Transformer model with hyperparameter set `transformer_base` (Vaswani et al., 2017). Some hyperparameters follow the suggestion of Popel and Bojar (2018): maximum sequence length=150, batch size=1500, learning rate=0.2, learning rate warmup steps=8000. We optimize our model using the Adafactor optimizer (Shazeer and Stern, 2018). For the vocabulary, we use the default subword units implemented in T2T, SubwordTextEncoder (STE), which is shared between source and target languages with approximate size of 32,678 units. Our data is not tokenized.

We run the baseline and the first phase of our domain-adaptation experiments training for 300,000 and 500,000 steps, respectively, and save the checkpoint hourly. However, we find an overfit on baseline systems during the training so we stop early and select the model from the checkpoint resulting the highest BLEU score on development set. For the second phase of training in domain-adaptation experiments, we set the steps to 50,000 in order to avoid overfit to the in-domain data and save the checkpoint every 10 minutes. We use the last value of learning rate in the first training phase for the second training phase.

During decoding, we use beam search with beam size of 4 and alpha value (length normalization penalty) of 0.6. We evaluate our model on the development set during the training and the test set after the model selection using case-sensitive BLEU score computed by the built-in command `t2t-bleu`.

3.3 Formality level evaluation

We conduct a manual evaluation for the formality level of translations resulted from our best EN-ID system. The purpose of this evaluation is to

[1] https://wit3.fbk.eu/mt.php?release=2017-01-more, accessed on 25th February 2019

System	EN-ID	ID-EN
IN	22.03	23.06
OUT	20.75	22.81
OUT+DA	27.47	26.93
MIX	24.84	25.18
MIX+DA	**29.10**	**28.18**

Table 2: BLEU scores of our English-Indonesian (EN-ID) and Indonesian-English (ID-EN) NMT systems on test set. The bold texts mark the best scores.

see whether the domain-adapted system generates more formal translation based on human evaluation. The evaluation is inspired by human assessment of Niu et al. (2017). We randomly select 50 translation pairs from the test set generated by the first and second phases of our EN-ID system. We make sure each pair does not consist of the same sentences. Then 48 Indonesian native speakers vote which sentence is more formal between two of them. An option of ”neutral or difficult to distinguish” is also available. The voters are not aware that the sentence pairs are generated by MT systems in order to keep the purity of the evaluation based on formality level and not biased to the translation quality.

4 Result

4.1 NMT performance

Table 2 shows the BLEU evaluation of our systems. For both EN-ID and ID-EN directions, the result shows similar patterns: (1) System trained with only in-domain data (IN) works better than with only out-of-domain data (OUT) although the training-data sizes are significantly different. (2) Domain adaptation (fine-tuning) helps to improve the BLEU score in both cases when the model is first trained without and with in-domain data (OUT and MIX respectively). Despite the best performance of our mixture system, the domain adaptation method has higher impact on the out-of-domain system.

While IN systems suffer from overfit, both training and evaluation loss in OUT and MIX systems still slightly decrease in the end of the training which indicates the training steps still can be increased.

4.2 Formality level

We use the evaluation approach described in Subsection 3.3 on translation output of MIX and MIX+DA from the test set. Out of 50 pairs, 35 MIX+DA sentences are voted by the majority (>50% of the voters) as more formal than their pairs. For the remaining pairs, the majority either select MIX sentences as more formal (12 pairs) or the MIX+DA sentences are still the most selected but the frequency is less than 50% of the voters (3 pairs). We consider the latter condition has no difference to being indistinct, although none of the pairs with ”difficult to distinguish” option are selected by the majority,

Sentence 1	
Source	*You* have to listen to one another.
MIX	*Kau* harus mendengarkan satu sama lain.
MIX+DA	*Anda* harus mendengarkan satu sama lain.
Sentence 2	
Source	*I* enjoy **fashion** magazines and pretty things.
MIX	*Aku* menikmati majalah **fashion** dan hal-hal cantik.
MIX+DA	*Saya* menikmati majalah **adibusana** dan hal-hal yang cukup.
Sentence 3	
Source	It *could* even **be disseminated** intentionally.
MIX	Itu bahkan *bisa* **dibubarkan** secara sengaja.
MIX+DA	Hal ini bahkan *dapat* **diabaikan** dengan sengaja.
Sentence 4	
Source	They *come* from these cells.
MIX	Mereka *datang* dari sel-sel ini.
MIX+DA	Mereka *berasal* dari sel-sel ini.

Figure 1: Sample outputs of our EN-ID non-adapted (MIX) and domain-adapted (MIX+DA) systems, in which more than 50% of human assessors vote translation by the domain-adapted system as more formal.

Among those 35 MIX+DA sentence pairs, we analyze 13 pairs that are voted by more than 85% voters to observe which segment of the sentences might trigger the voters to label them as more formal. Figure 1 shows sample output sentence pairs with such condition. Interestingly, 9 of those pairs show similar pattern, namely they contain the change of pronouns to the formal one. For instance, ”*kau*” → ”*Anda*” or ”*aku*” → ”*saya*” in Sentence 1 and 2, respectively, in the figure. Note that English does not use honorifics that can give such context change in the translation.

Among 2015 translation pairs from the test set, we find 316 translations which change the pronouns to be more formal, 448 translations which already use formal pronouns before domain-adapted thus do not change, and, surprisingly, no translation that still uses informal pronouns after being domain-adapted. This indicates the style of using honorifics is successfully transferred from speech styled language.

Sentence 3 and 4 are of the remaining pairs that do not have such pattern. In sentence 3, there are 3 different segments in the translations. Although native speaker might easily find that "*dapat*" is more formal than "*bisa*", just like the use of "could" and "can" in English, we cannot find how to measure each of the lexical differences affects the formality level. Meanwhile, in a pair that only has one word difference like in sentence 4, we can infer that the highlighted words are the trigger of the formality of the sentences, if we assume that the translation is correct (which is true in this sample). Nevertheless, the focus on finding segments that trigger the formality of the whole translation outputs can be an interesting future work.

5 Related Work

Most works on ID-EN or EN-ID MT were based on phrase-based SMT (Yulianti et al., 2011; Larasati, 2012; Sujaini et al., 2014), in which other approaches were incorporated to the basic SMT to enhance the performance, such as by combining SMT with rule-based system or adding linguistics information. Neural method was used as a language model to replace statistical n-gram language model in EN-ID SMT (Hermanto et al., 2015), not as an end-to-end MT system like our models.

While we can not find any previous work on end-to-end Indonesian NMT paired with English, such work has been performed with some Asian languages. Trieu et al. (2017) built NMT systems for Indonesian-Vietnamese and Adiputra and Arase (2017) for Japanese-Indonesian NMT. Those works used RNN-based encoder-decoder architecture, while we use self-attention based model.

Our analysis of formality level is related to politeness or formality control in NMT output (Sennrich et al., 2016; Niu et al., 2017). Both works added a mark on the source side as an expected formality level on the translation output. While the former focused only on the use of honorifics, the latter had a wider definition of formality based on the calculation of formality score. Although the finding of our work is similar to the expected output of Sennrich et al. (2016), it differs from both works as we use domain-adaptation method instead of a formality mark.

6 Conclusions and Future Research

We have presented the use of Neural Machine Translation (NMT) using Transformer model for English-Indonesian language pair in the spoken language domains, namely colloquial language and speech language. We demonstrate that the domain-adaptation method we use does not only improve the model performance, but is also able to generate translation in more formal language. The most notable formality style transferred is the use of honorifics.

There are still many open research directions for EN-ID and ID-EN NMT systems. In this work, we mostly use the default value of hyperparameters for our Transformer model. An empirical study to explore different set of hyperparameters can be an interesting future work with a goal to build the state-of-the-art model for both language directions. The work can be also followed by model comparison with the previous state-of-the-art RNN-based NMT systems. Besides investigating segments of the translations that may trigger the formality, it is also interesting to conduct further analysis on the style transfer learned by the domain adaptation method in our EN-ID system, not restricted to the formality level.

Acknowledgments

We thank Martin Popel for thoughtful discussions, Ondřej Bojar for the help in the initial work, the mentor and anonymous reviewers for helpful comments. We gratefully acknowledge the support of European Masters Program in Language and Communication Technologies (LCT) and Erasmus Mundus scholarship.

References

Cosmas Krisna Adiputra and Yuki Arase. 2017. Performance of Japanese-to-Indonesian Machine Translation on Different Models. *23rd Annual Meeting of the Speech Processing Society of Japan (NLP2017)*, pages 757–760.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. *ICLR 2015*, pages 1–15.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Confer-*

ence on Machine Translation, pages 272–307, Belgium, Brussels. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Inventory of transcribed and translated talks.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Andi Hermanto, Teguh Bharata Adji, and Noor Akhmad Setiawan. 2015. Recurrent neural network language model for english-indonesian machine translation: Experimental study. In *2015 International Conference on Science in Information Technology (ICSITech)*. IEEE.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *CoRR*, abs/1706.03872.

Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. 2018. A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation. *CoRR*, abs/1806.06957.

Septina Dian Larasati. 2012. Towards an Indonesian-English SMT System : A Case Study of an Under-Studied and Under-Resourced Language , Indonesian. pages 123–129.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *International Workshop on Spoken Language Translation*, Da Nang, Vietnam.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Martin Popel and Ondrej Bojar. 2018. Training tips for the transformer model. *CoRR*, abs/1804.00247.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. *CoRR*, abs/1804.04235.

Herry Sujaini, Kuspriyanto Kuspriyanto, Arry Akhmad Arman, and Ayu Purwarianti. 2014. A novel part-of-speech set developing method for statistical machine translation. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 12(3):581.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

Gongbo Tang, Matthias Müller, Annette Rios, and Rico Sennrich. 2018. Why self-attention? A targeted evaluation of neural machine translation architectures. *CoRR*, abs/1808.08946.

Hai-Long Trieu, Duc-Vu Tran, and Le-Minh Nguyen. 2017. Investigating phrase-based and neural-based machine translation on low-resource settings. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 384–391. The National University (Phillippines).

Brian Tubay and Marta R. Costa-jussà. 2018. Neural machine translation with the transformer and multi-source romance languages for the biomedical WMT 2018 task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 667–670.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, François Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Lukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Evi Yulianti, Indra Budi, Achmad Nizar Hidayanto, Hisar Maruli Manurung, and Mirna Adriani. 2011. Developing indonesian-english hybrid machine translation system. In *Proceedings of the 2011 International Conference on Advanced Computer Science and Information Systems*, pages 265–270.