The English-Vietnamese Machine Translation System for IWSLT 2015

[1,2] Viet Tran Hong, [2,3] Huyen Vu Thuong, [2,4] Trung Le Tien, [2,5]Luan Nghia Pham, [2]Vinh Nguyen Van

[1]University Of Economic And Technical Industries, Hanoi, Vietnam
[2]University of Engineering and Technology-Vietnam National University, Hanoi, Vietnam
[3]ThuyLoi University, Hanoi, Vietnam
[4]Open University, Hanoi, Vietnam
[5]Haiphong University, Haiphong, Vietnam
thviet@uneti.edu.vn,huyenvt@tlu.edu.vn,trunglt@hou.edu.vn,nghialuan@gmail.com,vinhnv@vnu.edu.vn

1. Introduction

In this paper we have described our system for IWSLT2015 machine translation. Focusing primarily on the English-Vietnamese and Vietnamese-English translation direction. Our additions for Moses phrase-based SMT and Phrasal SMT include two language model with monolingual training set for English and Vietnamese.

We submitted two systems to IWSLT 2015 evaluations for English to Vietnamese Machine Translation and Vietnamese to English Machine Translation. Our systems is including sub-systems: 6 based on Phrasal toolkit [Green et al.2014] and 6 others base Moses toolkit [Koehn et al.2007b]. The systems conducted with IWSLT 2015 data using with extension language model using monolingual training data.

2. Data and Pre-Processing

We perform to pre-processing data from IWSLT 2015 for dev, test, train dataset. We convert from formatted xml data to have parallel data. These data are tokenizer for both Vietnamese and English. With Vietnamese data we use VnTokenizer [Phuong-Le Hong2008]. Filter the corrupt characters and the larger sentence of length 300. With English data, we also use tokenizer for segmentation. After that, we conducted experiment for IWSLT 2015 data.

3. Monolingual Data

We expand the language model using Monolingual Data. For English-Vietnam translation, we used data with the crawl from electronic newspaper in Vietnam. We install the tool library used crawler Jsoup to collect 1GB of data and used for training. With Vietnamese-English translation, we use one part of the data WMT2015 collect 1GB of data and used for training.

4. Brief description of the baseline Phrase-based SMT

Phrase-based SMT, as described by [Koehn et al.2003] translates a source sentence into a target sentence by decomposing the source sentence into a sequence of source phrases, which can be any contiguous sequences of words (or tokens treated as words) in the source sentence. For each source phrase, a target phrase translation is selected, and the target phrases are arranged in some order to produce the target sentence. A set of possible translation candidates created in this way is scored according to a weighted linear combination of feature values, and the highest scoring translation candidate is selected as the translation of the source sentence.

Moses [Koehn et al.2007b] is a statistical machine translation system that allows automatically train translation models for any language pair. When we have a trained model, an efficient search algorithm quickly finds the highest probability translation among the exponential number of choices.

Beside Moses, nowadays, Phrasal [Green et al.2014] is also a toolkit for phrase-based SMT. It is a state-of-the-art statistical phrase-based machine translation system, written in Java. At its core, it provides much the same functionality as the core of Moses.

5. Experiment

We present our experiments to translate from English to Vietnamese in a statistical machine translation system. We compare Phrasal and Moses by evaluation with IWSLT 2015 data. We evaluated our approach on English-Vietnamese machine translation tasks, and show that it can significantly outperform the baseline phrase-based SMT system by extended Language model.

The performances of the statistical machine translation systems in our experiments are evaluated by the BLEU scores [Papineni and Zhu2002].

745 sentences in IWSLT15.TED.dev2010 as our dev set on which we tuned the feature weights, and report results on the 1046 sentences of the IWSLT15.TED.tst2015 test set.

Corpus	Sentence pairs	Training Set	Development Set	Test Set
General	123957	122132	745	1080
			English	Vietnamese
Training	Sentences		122132	
	Average Length		15.93	15.58
	Word		1946397	1903504
	Vocabulary		40568	28414
Development	Sentences		745	
	Average Length		16.61	15.97
	Word		12397	11921
	Vocabulary		2230	1986
Test	Sentences		1046	
	Average Length		16.25	16.13
	Word		17023	16889
	Vocabulary		2701	2759

Table 1: The Summary statistical of data sets: English-Vietnamese

In order to extract the translation grammar necessary for our model, we used the provided Europarl and News Commentary parallel training data. The lowercased and tokenized training data was then filtered for length and aligned using the GIZA++ [Och and Ney2003] implementation of IBM Model 4 to obtain one-to-many alignments in both directions and symmetrized by combining both into a single alignment using the grow-diag-final-and method and Berkeley Aligner [DeNero and Klein2007]. We constructed a 4-gram language model using the SRI language modeling toolkit [Stolcke2002] and KenLM [Heafield2011] from the provided English monolingual training data and Vietnamese monolingual training data from crawler web data. Since the beginnings and ends of sentences often display unique characteristics that are not easily captured within the context of the model, and have previously been demonstrated to significantly improve performance, we explicitly annotate beginning and end of sentence markers as part of our translation process. We used the 745 sentences in IWSLT15.TED.dev2010 as our dev set on which we tuned the feature weights, and report results on the 1046 sentences of the IWSLT15.TED.tst2015 test set. (122131 train.tags + 125531 train + 1GB mono data)

6. Evaluation

We conducted some experiments the following:

- Using the state of art Phrase-based SMT Moses:
  - with SMT Moses Decoder [Koehn et al.2007a] and SRILM. We trained a 4 gram language model using interpolate and kndiscount smoothing with 1GB Vietnamese monolingual data for English-Vietnamese translate direction and 1GB English monolingual data for Vietnamese-English translate direction.
  - Before extracting phrase table, we use GIZA++ to build word alignment with grow-diag-final-and algorithm. Besides using pre-processing, we also used default reordering model in Moses Decoder: using word-based extraction (wbe), splitting type of reordering orientation to three class (monotone, swap and discontinuous msd), combining backward and forward direction (bidirectional) and modeling base on both source and target language (fe).
- with SMT Phrasal:
  - We also trained with 1GB Vietnamese monolingual data for English - Vietnamese translate direction and 1GB English monolingual data for Vietnamese-English translate direction a 4 gram language model with 1GB.
  - Before extracting phrase table, we use berkeley aligner to build word alignment with grow-diag-final-and algorithm. Besides using pre-processing, we also used default reordering model in Phrasal.

6.1. English-to-Vietnamese Translation

We conducted 6 experiments: 3 base on Phrasal and 3 base on Moses. Using 4 gram for building language model with monolingual following:

- Using train.tags.en-vi.vi as monolingual data for building language model.
- Combine train.tags.en-vi.vi and train.vi as monolingual data for buiding language model.
- Combine train.tags.en-vi.vi and train.vi and 1GB crawler web data from news site in Vietnam as monolingual data for building language model.

PHRASAL					
No	System	Experiments	BLEU	Description	N-GRAM
1	En-Vn	RUN01	22.16	Baseline System using monoligual data from training set	4
2		RUN02	22.59	Baseline System using monoligual data from 1GB monoligual web crawler data	
3		RUN03	22.90	Baseline System using monoligual data from training set combine training from IWSLT 2015 and 1GB monoligual from web crawler data	
MOSES					
No	System	Experiments	BLEU	Description	N-GRAM
1	En-Vn	RUN01	22.70	Baseline System using monoligual data from training set	4
2		RUN02	22.93	Baseline System using monoligual data from 1GB monoligual web crawler data	
3		RUN03	23.15	Baseline System using monoligual data from training set combine training from IWSLT 2015 and 1GB monoligual from web crawler data	

Figure 1: The experiment our systems for English to Vietnamese translation direction

Figure 1 described results our experiments for English to Vietnamese translation direct. Highest BLEU score is 23.15 for English-Vietnamese translation system with the IWSLT 2015 data.

6.2. Vietnamese-to-English Translation

We conducted 6 experiments: 3 base on Phrasal and 3 base on Moses. Using 4 gram for building language model with monolingual following:

- Using train.tags.vi-en.en as monolingual data for building language model.
- Combine train.tags.vi-en.en and train.en as monolingual data for building language model.
- Combine train.tags.en-vi.en and train.en and 1GB English data from WMT2015 as monolingual data for building language model.

PHRASAL					
No	System	Experiments	BLEU	Description	N-GRAM
1	Vn-En	RUN01	17.37	Baseline System using monoligual data from training set	4
2		RUN02	17.95	Baseline System using monoligual data from 1GB monoligual web crawler data	
3		RUN03	20.18	Baseline System using monoligual data from training set combine training from IWSLT 2015 and 1GB monoligual from web crawler data	
MOSES					
No	System	Experiments	BLEU	Description	N-GRAM
1	Vn-En	RUN01	17.19	Baseline System using monoligual data from training set	4
2		RUN02	17.56	Baseline System using monoligual data from 1GB monoligual web crawler data	
3		RUN03	19.72	Baseline System using monoligual data from training set combine training from IWSLT 2015 and 1GB monoligual from web crawler data	

Figure 2: The experiment our systems for Vietnamese to English translation direction

Figure 2 described results our experiments for Vietnamese to English translation direction. Highest BLEU score is 20.18 for Vietnamese-English translation system with IWSLT 2015 data.

7. Conclusions

In this paper, we has described an an empirical study for English-Vietnamese Statistical Machine Translation. We attempted to tackle the problem of training SMT on parallel data. The extend of the monolingual training set to build language model for training SMT could lead results be more stable and better enough. We evaluated our approach on English-Vietnamese machine translation tasks with Moses toolkit and Phrasal toolkit (state-of-the-art phrase-based and hierarchical statistical MT systems). The experiment results showed that our approach achieved statistically improvements in BLEU scores .

8. Acknowledgements

This work described in this paper has been partially funded by Hanoi National University (QG.15.23 project)

9. References

[DeNero and Klein2007] John DeNero and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 17–24, Prague, Czech Republic, June. Association for Computational Linguistics.

[Green et al.2014] Spence Green, Daniel Cer, and Christopher D. Manning. 2014. Phrasal: A toolkit for new directions in statistical machine translation. In *In Proceddings of the Ninth Workshop on Statistical Machine Translation*.

[Heafield2011] Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

[Koehn et al.2003] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pages 127–133. Edmonton, Canada.

[Koehn et al.2007a] 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. 2007a. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL, Demonstration Session*.

[Koehn et al.2007b] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007b. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and*

demonstration sessions, pages 177–180. Association for Computational Linguistics.

[Och and Ney2003] Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

[Papineni and Zhu2002] Salim Roukos-Todd Ward Papineni, Kishore and WeiJing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *ACL*.

[Phuong-Le Hong2008] Azim Roussanaly Vinh-Ho Tuong Phuong-Le Hong, Huyen-Nguyen Thi Minh. 2008. A hybrid approach to word segmentation of vietnamese texts. In *In Proceedings of the 2nd International Conference on Language and Automata Theory and Applications*, Springer, LNCS 5196.

[Stolcke2002] Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, volume 29, pages 901–904.