# A STUDY ON MACHINE TRANSLATION FOR LOW-RESOURCE LANGUAGES

By TRIEU, LONG HAI

submitted to
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Written under the direction of
Associate Professor Nguyen Minh Le

September, 2017

# A STUDY ON MACHINE TRANSLATION FOR LOW-RESOURCE LANGUAGES

By TRIEU, LONG HAI (1420211)

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Doctor of Information Science
Graduate Program in Information Science

Written under the direction of
Associate Professor Nguyen Minh Le

and approved by
Associate Professor Nguyen Minh Le
Professor Satoshi Tojo
Professor Hiroyuki Iida
Associate Professor Kiyoaki Shirai
Associate Professor Ittoo Ashwin

July, 2017 (Submitted)

## Acknowledgements

# Abstract

Current state-of-the-art machine translation methods are neural machine translation and statistical machine translation, which based on translated texts (bilingual corpora) to learn translation rules automatically. Nevertheless, large bilingual corpora are unavailable for most languages in the world, called low-resource languages, that cause a bottleneck for machine translation (MT). Therefore, improving MT on low-resource languages becomes one of the essential tasks in MT currently.

In this dissertation, I present my proposed methods to improve MT on low-resource languages by two strategies: building bilingual corpora to enlarge training data for MT systems and exploiting existing bilingual corpora by using pivot methods. For the first strategy, I proposed a method to improve sentence alignment based on word similarity learnt from monolingual data to build bilingual corpora. Then, a multilingual parallel corpus was built using the proposed method to improve MT on several Southeast Asian low-resource languages. Experimental results showed the effectiveness of the proposed alignment method to improve sentence alignment and the contribution of the extracted corpus to improve MT performance. For the second strategy, I proposed two methods based on semantic similarity and using grammatical and morphological knowledge to improve conventional pivot methods, which generate source-target phrase translation using pivot language(s) as the bridge from source-pivot and pivot-target bilingual corpora. I conducted experiments on low-resource language pairs such as the translation from Japanese, Malay, Indonesian, and Filipino to Vietnamese and achieved promising results and improvement. Additionally, a hybrid model was introduced that combines the two strategies to further exploit additional data to improve MT performance. Experiments were conducted on several language pairs: Japanese-Vietnamese, Indonesian-Vietnamese, Malay-Vietnamese, and Turkish-English, and achieved a significant improvement. In addition, I utilized and investigated neural machine translation (NMT), the state-of-the-art method in machine translation that has been proposed currently, for low-resource languages. I compared NMT with phrase-based methods on low-resource settings, and investigated how the low-resource data affects the two methods. The results are useful for further development of NMT on low-resource languages. I conclude with how my work contributes to current MT research especially for low-resource languages and enhances the development of MT on such languages in the future.

**Keywords**: machine translation, phrase-based machine translation, neural-based machine translation, low-resource languages, bilingual corpora, pivot translation, sentence alignment

# Acknowledgements

For three years working on this topic, it is my first long journey that attract me to the academic area. It is also one of the biggest challenges that I have ever dealt with. This work gives me a lot of interesting knowledge and experiences as well as difficulties that require me with the best efforts. At the moment of writing this dissertation as a summary for the PhD journey, it reminds me a lot of support from many people. This work cannot be completed without their support.

First of all, I would like to thank my supervisor, Associate Professor Nguyen Minh Le. Professor Nguyen gives me a lot of comments, advices, discussions in my whole three-year journey from the starting point when I approached this topic without any prior knowledge about machine translation until my last tasks to complete my dissertation and research. Doing PhD is one of the most interesting things in studying, but it is also one of the most challenge things for everyone in the academic career. Thanks to the useful and interesting discussions with professor Nguyen, I have overcome the most difficult periods in doing this research. Not only teach me some first lessons and skills in doing research, professor Nguyen also has interesting and useful discussions that help me a lot in both studying and the life.

I would like to thank the committee: Professor Satoshi Tojo, Professor Hiroyuki Iida, Associate Professor Ittoo Ashwin, Associate Professor Kiyoaki Shirai for their comments. This can be one of the first work in my academic career, that cannot avoid a lot of mistakes and weaknesses. By discussing with the professors in the committee, and receiving their valuable comments, they help me a lot in improving this dissertation.

I also would like to thank my collaborators: Associate Professor Nguyen Phuong Thai for his comments, advices, and experience in sentence alignment and machine translation. I would like to thank Vu Tran, Tin Pham, Viet-Anh Phan for their interesting discussions and collaborations in doing some topics in this research. Thanks so much to Vu Tran, Chien Tran for their technical support.

I would like to thank my colleagues and friends, Truong Nguyen, Huy Nguyen, for their support and encourage. I also would like to give a special thank to professor Jean-Christophe Terrillon Georges for his advices and comments on the writing skills and English manuscripts of my papers, special thank to professor Ho Tu Bao for valuable advices in research. Thanks so much to Danilo S. Carvalho, Tien Nguyen for their comments.

Last but not least, I would like to thank my parents, Thi Trieu, Phuong Hoang, my sister Ly Trieu, and my wife Xuan Dam for their support and encouragement in all time not only in this work but in my life.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Machine Translation

Translation between languages is an important demand of humanity. With the advent of digital computers, it provided a basis for the dream of building machines to translate languages automatically. Almost as soon as electronic computers appeared, people made efforts to build automatic systems for translation, which also opened a new field: *machine translation*. As defined in Hutchins and Somers, 1992 [33], machine translation (MT) is "*computerized systems responsible for the production of translation from one natural language to another, with or without human assistance*".

Machine translation has a long history in its development. Various approaches were explored such as: direct translation (using rules to map input to output), transfer methods (analyzing syntactic and morphological information), and interlingual methods (using representations of abstract meaning). The field attracted a lot of interest from community like: a study of realities of machine translation from US funding agencies in 1966 (ALPAC report), commercial systems from the past (Systran in 1968, Météo in 1976, Logos and METAL in 1980s) to current development by large companies (IBM, Microsoft, Google), and many projects in universities and academic institutes.

Dominated approaches of current machine translation are *statistical machine translation* (SMT) and *neural machine translation* (NMT), which are based on resources of translated texts, a trend of data-driven methods. Previous work cannot succeed with rule-based methods when there are a large number of rules that were so complicated to discover, represent, and transfer between languages. Instead of that, a set of translated texts are used to automatically learn corresponding rules between languages. This trend has shown state-of-the-art results in recent researches as well as applied in the current widely-used MT system, Google.

Translated texts, called *bilingual corpora*, therefore become one of the key factors that affect the translation quality. For more precisely, a *bilingual corpus* (*parallel corpora* or *bilingual corpora* in plural) is a set of sentence pairs of two languages in which two sentences in each pair are the translation of each other. Current MT systems require large bilingual corpora even up to millions of sentence pairs to learn translation rules. There

are many efforts in building large bilingual corpora like Europarl (the bilingual corpus of 21 European languages), English-Arabic, English-Chinese. Building such large bilingual corpora requires many efforts. Therefore, besides bilingual corpora of European languages and some other language pairs, there are few large bilingual corpora for most language pairs in the world. This issue leads to a bottleneck for machine translation in many language pairs that lack large bilingual corpora, called *low-resource languages*. In this work, I define low-resource languages as language pairs that have no or small bilingual corpora (less than one million sentence pairs). Improving MT on low-resource languages becomes an essential task that demands many efforts as well as attracts many interest currently.

## 1.2 MT for Low-Resource Languages

In previous work, solutions have been proposed to deal with the problem of insufficient bilingual corpora. There are two main strategies: *building new bilingual corpora* and *utilizing existed corpora*.

For the first strategy, bilingual corpora can be built manually or automatically. Building large bilingual corpora by human may ensure the quality of corpora; however, it requires a high cost of labor and time. Therefore, automatically building bilingual corpora can be a feasible solution. This task relates to a sub-field: *sentence alignment*, in which sentences that are translation of each other can be extracted automatically [5, 11, 27, 59, 92]. The effectiveness of sentence alignment algorithms affect the quality of the bilingual corpora. In this work, I have improved a problem in sentence alignment namely *out-of-vocabulary*, in which there is insufficient knowledge of bilingual dictionary used for sentence alignment. The proposed method was applied to build a bilingual corpus for several low-resource language pairs and then used to improve MT performance.

For the second strategy, existing bilingual corpora can be utilized to extract translation rules for a language pair called *pivot methods*. Specifically, pivot language(s) are used to connect translation from a source language to a target language if there exist bilingual corpora of source-pivot and pivot-target language pairs [16, 18, 91, 98].

## 1.3 Contributions

There are four main contributions of this dissertation.

First, I have improved a problem in sentence alignment to deal with the out-of-vocabulary problem. In addition, a large multilingual parallel corpus was built to contribute for the development and improving MT on several low-resource language pairs of Southeast Asian: Indonesian, Malay, Filipino, and Vietnamese that there is no prior work on these language pairs.

Second, I propose two methods to improve pivot methods. The first method is to enhance pivot methods by semantic similarity to deal with the problem of lacking information of the conventional triangulation approach. The second method is to improve the conventional triangulation approach by integrating grammatical and morphological knowl-

edge. The effectiveness of the proposed methods were confirmed by various experiments on several language pairs.

Third, I propose a hybrid model that significantly improves MT on low-resource languages by combining the two strategies of building bilingual corpora and exploiting existing bilingual corpora. Experiments were conducted on three different language pairs: Japanese-Vietnamese, Southeast Asian languages, and Turkish-English to evaluate the proposed method.

Fourth, several empirical investigations were conducted on low-resource language pairs using NMT to provide some empirical basis that is useful for further improvement of this method in the future for low-resource languages.

## 1.4 Dissertation Outline

Although MT has shown significant improvement recently, there is still a big issue that requires many efforts in MT: improving MT for low-resource languages because of insufficient training data, one of the key factors in current MT systems. In this thesis, I focus on two main strategies: building bilingual corpora to enlarge training data for MT systems, and exploiting existing bilingual corpora based on pivot methods. I will spend two chapters to describe my proposed methods for the two strategies. Then, one chapter is to present my proposed model that can effectively combine and exploit the two strategies in a hybrid model. Besides the two main strategies, I spend one chapter to present some of my first investigations on utilizing NMT, a successful method recently, on low-resource languages. I start my dissertation by providing necessary background knowledge in Chapter 2 for readers about methods presented in this dissertation. In chapter 3, I describe my proposed methods to improve sentence alignment and a multilingual parallel corpus built from comparable data.[1] Chapter 4 presents my proposed methods in pivot translation that include two main parts: applying semantic similarity; and integrating grammatical and morphological information. In Chapter 5, I present a hybrid model that combines the two strategies. Chapter 6 contains my investigations of utilizing NMT for low-resource languages. Finally, I conclude my work in Chapter 7.

**Building Bilingual Corpora** Chapter 3 is my methods related to the strategy of building bilingual corpora to enlarge the training data for MT, which includes two main parts in this chapter. In the first section, I present my proposed method related to sentence alignment using semantic similarity. Experimental results show the contribution of the proposed method. *This chapter is based on the paper (Trieu et al., 2016 [88]).* The second section is about building a multilingual parallel corpus from Wikipedia that can enhance

[1]In addition, I also have a paper that is based on building a very large monolingual data to train a large language model that significantly improves SMT systems. This system presented in the paper (Trieu et al., 2015 [83]). In the IWSLT 2015 machine translation shared task, the system achieves the state-of-the-art result in human evaluation for English-Vietnamese, and ranked the runner-up for the automatic evaluation.

MT for several low-resource language pairs. *This section is based on the paper (Trieu and Nguyen, 2017 [87]).*

**Pivoting Bilingual Corpora** Chapter 4 introduces my proposed methods related to pivot translation. There are two main sections in this chapter that correlate to two proposed methods in improving the conventional pivot method. The first part presents my proposed method to improve pivot translation using semantic similarity. *This section is based on the paper (Trieu and Nguyen, 2016 [84]).* For the second part, I describe a proposed method that integrates grammatical and morphological to pivot translation. *This section is based on the paper (Trieu and Nguyen, 2017 [85]).*

**A Hybrid Model for Low-Resource Languages** Chapter 5 presents my proposed model that combines the two strategies: building bilingual corpora and exploiting existing bilingual corpora that are described in the previous two chapters. *This section is based on the paper that I submitted to the ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP).* For the second part, I applied this model to Turkish-English that has shown the significant improvement in using the proposed model. *This section is based on the paper (Trieu et al., 2017 [89]).*

**NMT for Low-Resource Languages** Chapter 6 presents my research in utilizing NMT for low-resource languages in various language pairs. This can be a basis for further improvement in the future for low-resource languages. *This chapter is based on the paper (Trieu and Nguyen, 2017 [86]).*

All data, code, and models used in this dissertation are available at `https://github.com/nguyenlab/longtrieu-mt`

# Chapter 2

# Background

In this chapter, I present necessary background knowledge of the main topic and methods in this dissertation, which include: SMT, NMT, pivot methods, and sentence alignment.

## 2.1 Statistical Machine Translation

SMT is a class of approaches in machine translation that build probabilistic models to choose the most probable translation. SMT is based on the Bayes noisy channel model as follows.

Let $F$ be a source-language sentence, and $\hat{E}$ be the best translation of $F$.
$F = f_1, f_2, ..., f_m$
$\hat{E} = e_1, e_2, ..., e_l$
The translation from $F$ to $\hat{E}$ is modeled as follows.

$$\hat{E} = argmax_E P(E|F) = argmax_E \frac{P(F|E)P(E)}{P(F)} = argmax_E P(F|E)P(E) \tag{2.1}$$

There are three components in the models:

- $P(F|E)$ called a *translation model*
- $P(E)$ called a *language model*
- A *decoder*: a component produces the most probable $E$ given $F$

For the translation model $P(F|E)$, the probability that $E$ generates $F$ can be calculated based on two ways: *word-based* (individual words), or *phrase-based* (sequences of words). Phrase-based SMT (Koehn et al., 2003) [44] have showed the state-of-the-art performance in machine translation for many language pairs (Bojar et al., 2013) [4].

### 2.1.1 Phrase-based SMT

Phrase-based SMT uses phrases (a sequence of consecutive words) as atomic units for translation. The source sentence is segmented into a number of phrases. Each phrases is then translated into a target phrase.

Given, $f$: source sentence; $e_{best}$: the best target translation. Then, $e_{best}$ can be computed as follows.

$$e_{best} = argmax_e p(e|f) = argmax_e p(f|e)p_{LM}(e) \tag{2.2}$$

where:

- $p_{LM}(e)$ :the language model
- $p_(f|e)$: the translation model

The translation model $p_(f|e)$ can be decomposed into:

$$p(f_1^{-I}|e_1^{-I}) = \prod_{i=1}^{I} \phi(\overline{f_i}|\overline{e_i}) d(start_i - end_{i-1} - 1) \tag{2.3}$$

where:

- The source sentence $f$ is segmented into $I$ phrases: $\overline{f_i}$
- Each source pharse $f_i$ is translated into a target phrase $e_i$
- $d(start_i - end_{i-1} - 1)$ : *reordering model*; the output phrases can be reordered based on a distance-based reordering model. Let $start_i$ be the first word's position of the source phrase that translates to the $i$th target phrase; $end_i$ be the last word's position of the source phrase; Then, the reordering distance can be calculated as $start_i - end_{i-1} - 1$.

Therefore, the phrase-based SMT model is formed as follows:

$$e_{best} = argmax_e \prod_{i=1}^{I} \phi(\overline{f_i}|\overline{e_i}) d(start_i - end_{i-1} - 1) \prod_{i=1}^{|e|} p_{LM}(e_i|e_1 ... e_{i-1}) \tag{2.4}$$

where: there are three components in the model

- the phrase translation table $\phi(\overline{f_i}|\overline{e_i})$
- the reordering model $d$
- the language model $p_{LM}(e)$

**Tools** For statistical machine translation, several tools have been introduced, which showed the effectiveness and contributed to the development of the field. One of the most well-known system is the phrase-based **Moses** toolkit [43]. Another toolkit based on an n-gram-based statistical machine translation is **Marie** [53]. For integrating syntactic information in statistical machine translation, Li et al., 2009 [47] introduced **Joshua**, an open source decoder for statistical translation models based on synchronous context free grammars. Neubig 2013 [61] presents a system called **Travatar**, a tree-to-string statistical machine translation system. Dyer et al., 2010 introduced **CDEC** [22], a decoder, aligner, and model optimizer for statistical machine translation and other structured prediction models based on (mostly) context-free formalisms. In my work, since I focus on phrase-based machine translation, the powerful and well-known Moses toolkit was utilized in experiments.

One of the core part in phrase-based models is the word alignment. The task can be solved effectively by the system namely GIZA++ [65], an effective training algorithm for alignment models.

### 2.1.2 Language Model

Language model is an essential component in the SMT model. Language model aims to measure how likely it is that a sequence of words can be uttered by a native speaker is the target language. A probabilistic language model $p_{LM}$ should show the correct word order as in the following example:

$$p_{LM}(\text{the car is new}) > p_{LM}\ (\text{new the is car})$$

A method is used in language models called **n-gram** language modeling. In order to predict a word sequence $W = w_1, w_2, ..., w_n$, the model predicts one word at a time.

$$p(w_1, w_2, ..., w_n) = p(w_1)p(w_2|w_1)...p(w_n|w_1, w_2, ..., w_{n-1}) \tag{2.5}$$

The common language models used in machine translation are: **trigram** (the collection of statistics over sequence of three words), or **5-grams**. Some other kinds of n-gram language model include: **unigram** (single word), **bigram** (2-grams or a sequence of two words).

**Tools** For training language models, several effective systems have been proposed such as: KenLM [31], SRILM [78], IRSTLM [24], and BerkeleyLM [38].

### 2.1.3 Metric: BLEU

The **BLEU** metric (**Bi**L**ingual** **E**valuation **U**nderstudy) (Papineni et al., 2002) [66] is one of the most popular automatic evaluation metrics, which are used for evaluation in machine translation currently. The metric is based on matches of larger n-grams with the reference translation.

The BLEU is defined as follows as a model of precision-based metrics.

$$BLEU - n = brevity - penalty\ exp \sum_{i=1}^{n} \lambda_i\ log\ precision_i \tag{2.6}$$

$$brevity - penalty = min(1, \frac{output\_length}{reference\_length}) \tag{2.7}$$

where

- $n$: the maximum order for n-grams to be matched (typically set to 4)
- $prcision_i$: the ratio of correct n-grams of a certain order $n$ in relation to the total number of generated n-grams of that order.
- $\lambda_i$: the weights for the different precisions (typically set to 1)

Therefore, a typically used metric BLEU-4 can be formulated as follows.

$$BLEU - 4 = min(1, \frac{output\_length}{reference\_length}) \prod_{i=1}^{4} precision_i \tag{2.8}$$

For example:
**Output of a system**: I buy a new car this weekend
**Reference**: I buy my car in Sunday
1-gram precision 3/7, 2-gram precision 1/6, 3-gram precision 0/5, 4-gram precision 0/4.

## 2.2 Sentence Alignment

Sentence alignment is an essential task in natural language processing in building bilingual corpora. There are three main methods in sentence alignment: length-based, word-based, and the combination of length-based and word-based.

### 2.2.1 Length-Based Methods

The length-based methods were proposed in [5, 27] based on the number of words or characters in sentence pairs. These methods are fast and effective in some closed language pairs like English-French but obtain low performance in different structure languages like English-Chinese.

### 2.2.2 Word-Based Methods

The word-based methods [11, 36, 51, 55, 97] are based on word correspondences or using a word lexicon. These methods showed better performance than the length-based methods, but they depend on available linguistic resources.

### 2.2.3 Hybrid Methods

In the hybrid methods [59, 92], sentences are first aligned based on a length-based phase; then, the aligned sentences are used to train a word alignment model, which is then used to combine with the length-based phase to extract the final sentence alignment results. As discussed in [59, 74], the hybrid methods are shown to be accurate than the length-based methods due to the utilization of word alignment. In addition, the hybrid methods are faster than the word-based methods and do not depend on the availability of linguistic resources.

Since the advantages of the hybrid methods, I adapted the hybrid methods for the baseline and further develop in my word. I discussed two powerful algorithms in the hybrid methods: *M-align* (the Microsoft sentence aligner [59]) and the *hunalign* [92].

**Microsoft bilingual sentence aligner (Moore, 2002)[59]** In the evaluation of [74], this aligner of the hybrid methods achieved the best performance compared with other sentence alignment approaches.

Let $l_s$ and $l_t$ be the lengths of source and target sentences, respectively. Then, $l_s$ and $l_t$ varies according to Poisson distribution as follows:

$$P(l_t|l_s) = exp^{-l_t r}\frac{(l_s r)^{l_t}}{l_t!} \tag{2.9}$$

Where $r$ is the ratio of the mean length of target sentences to the mean length of source sentences. As shown in [59], the length-based phase based on the Poisson distribution was slightly better than the Gaussian distribution proposed by [5].

$$P(l_v|l_e) = \alpha exp^{-\frac{log(\frac{l_v}{l_e}) - \mu)^2}{2\sigma^2}} \tag{2.10}$$

Where $\mu$ and $\sigma^2$ are the mean and variance of the Gaussian distribution, respectively. The length-based model based on the Poisson distribution was shown to be simpler to estimate than the model based on the Gaussian distribution which has to iteratively estimate the variance $\sigma^2$ using the expectation maximization (EM) algorithm.

Sentence pairs extracted from the length-based phase are then used to train IBM Model 1 [6] to build a bilingual dictionary. The dictionary was then combined with the length-based phase to produce final alignments, which are described as follows:

$$P(s,t) = \frac{P_{1-1}(l_s, l_t)}{(l_s+1)^{l_t}}(\prod_{j=1}^{l_t}\sum_{i=0}^{l_s} tr(t_j|s_i))(\sum_{i=1}^{l_e} f_u(e_i)) \tag{2.11}$$

Where: $tr(t_j|s_i)$ is the probability of the word pair $(t_j|s_i)$ trained by IBM Model 1; $f_u$ is the observed relative unigram frequency of the word in the text in the corresponding language.

**hunalign (Varga et al., 2005) [92]** This algorithm combines the length-based method [27] and a dictionary. When the dictionary is unavailable, a length-based method was used to build a dictionary. This algorithm showed high performance and was applied to build parallel data in several work [77, 82].

## 2.3 Pivot Methods

### 2.3.1 Definition

Given the task of translation from a source language $L_s$ to a target language $L_t$. Let $L_p$ be a third language, and suppose that there exist bilingual corpora of $L_s - L_p$ and $L_p - L_t$. The third language $L_p$ can be used as a bridge for the translation from $L_s$ to $L_t$ using the bilingual corpora although there is no bilingual corpus or only a small bilingual corpus of $L_s - L_t$. This method is called *pivot method*, and $L_p$ is a *pivot language*.

### 2.3.2 Approaches

There are three main approaches in pivot methods: cascade, synthetic, and phrase pivot translation (or triangulation).

**Cascade** In cascade approaches, source sentences are translated to pivot languages using source-pivot bilingual corpora. Then, the translated pivot sentences are translated to target languages based on pivot-target bilingual corpora ([18, 91]).

**Synthetic** In the synthetic approach [18], source-pivot or pivot-target translation models are used to generate a synthetic source-target bilingual corpus. For instance, the pivot side of the source-pivot bilingual corpus is translated into the target language using the pivot-target translation model.

**Triangulation** In triangulation [16, 91, 98], source-pivot and pivot-target bilingual corpora are used to train phrase tables. Then, the source and target phrases are connected via common pivot phrases.

### 2.3.3 Triangulation: The Representative Approach in Pivot Methods

Given $L_s - L_p$, $L_p - L_t$ be bilingual corpora of the source-pivot and pivot-target language pairs. The bilingual corpora are fist used to train two phrase tables. Then, the translation probabilities of source phrases to target phrases are computed based on common pivot phrases by estimating the following feature functions.

**Phrase Translation Probabilities**

$$\phi(\overline{s}_i|\overline{t}_i) = \sum_{\overline{p_i}} \phi(\overline{s}_i|\overline{p_i})\phi(\overline{p}_i|\overline{t}_i) \tag{2.12}$$

where:

- $\overline{s_i}, \overline{p_i}, \overline{t_i}$: the source, pivot, and target phrases
- $\phi(\overline{s}_i|\overline{p_i}), \phi(\overline{p}_i|\overline{t}_i)$: phrase translation probabilities of the source-pivot and pivot-target phrase tables

**Lexical Translation Probability**

$$p_w(\overline{s}|\overline{t}, a) = \prod_{i=1}^{n} \frac{1}{|j|(i,j) \in a|} \sum_{\forall(i,j)\in a} w(s_i|t_j) \tag{2.13}$$

where:

- $(\overline{s}, \overline{t})$: a phrase pair
- $a$: a word alignment
- $i = 1, ..., n$: the source word positions
- $j = 1, ..., m$: the target word positions
- $p_w(\overline{s}|\overline{t}, a)$: the lexical weight

$$w(s|t) = \frac{count(s,t)}{\sum_{s'} count(s',t)} \tag{2.14}$$

where $w(s,t)$: lexical translation probability

**Alignment Induction** The model of alignment induction is illustrated in Figure 2.1.

$$a = (f,e)|\exists p : (f,p) \in a_1 \& (p,e) \in a_2 \tag{2.15}$$

where $f, e$ are source and target words.

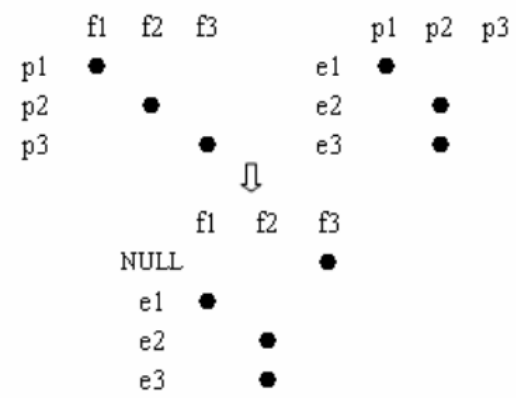



Figure 2.1: **Alignment induction**, Wu and Wang 2007 [98]

### 2.3.4 Previous work

Pivot methods have been applied in some previous work. Schafer and Yarowsky, 2002 [69] used pivot language methods for translation dictionary induction. Wang et al., 2006 [94] used pivot method to improve word alignment. In the work of Callison-Burch et al., 2006 [7], pivot languages were used for paraphrase extraction. Gispert and Marino (2006) [18] used pivot methods for English-Catalan translation by using a Spanish-Catalan SMT system to translate the Spanish side of the English-Spanish parallel corpus into Catalan.

The representative approach in pivot methods, called triangulation, has been proposed in [16, 91, 98].

Pivot translation has been successfully applied in some previous work. [8] applied pivot methods for Arabic-Italian translation via English and showed the effectiveness. In [54], pivot methods were used in translation from Brazilian Portuguese texts to European Portuguese. For a large-scale data set, [41] applied pivot methods on the multilingual Acquis corpus. In recent work, Dabre et al., 2015 [17] utilized a small multilingual corpora for SMT using many pivot languages.

There are several work proposed to improve the triangulation approach. [100] utilized Markov random walks to connect possible translation phrases between source and target languages in order to deal with the problem of lacking information. [23] proposed features to filter low quality phrase pairs extracted by the triangulation. In order to improve phrase translation's scores estimated by the triangulation, [101] proposed a method of pivoting via co-occurence counts of phrase pairs. Miura et al., 2015 [58] proposed a method to solve another problem that the traditional triangulation forgets the pivot information.

In using phrase pivot translation for low-resource languages, Dholakia and Sarkar, 2014 [21] survey previous approaches in pivot translation and applied for several low-resource languages.

## 2.4 Neural Machine Translation

Neural machine translation (NMT) has obtained state-of-the-art performance in machine translation for many languages including Czech-English, German-English, English-Romanian [71]. NMT has been proposed recently as a promising framework for machine translation, which learns sequence-to-sequence mapping based on two recurrent neural networks [14, 79], called encoder-decoder networks. An input sequence is mapped to a continuous vector space as a context vector using a recurrent network called encoder. Meanwhile, the decoder is another recurrent network which generates a target sequence from the context vector. In a basic encoder-decoder network, the dimension of the context vector in the encoder is fixed, which leads to a low performance when translating for long sentences. In order to overcome the problem, [1] proposed a method called attention mechanism, in which the model encodes the most relevant information in an input sentence rather than a whole input sentence into the fixed length context vector. NMT models with the attention mechanism have achieved significantly improvement in many language pairs [28, 34, 50].

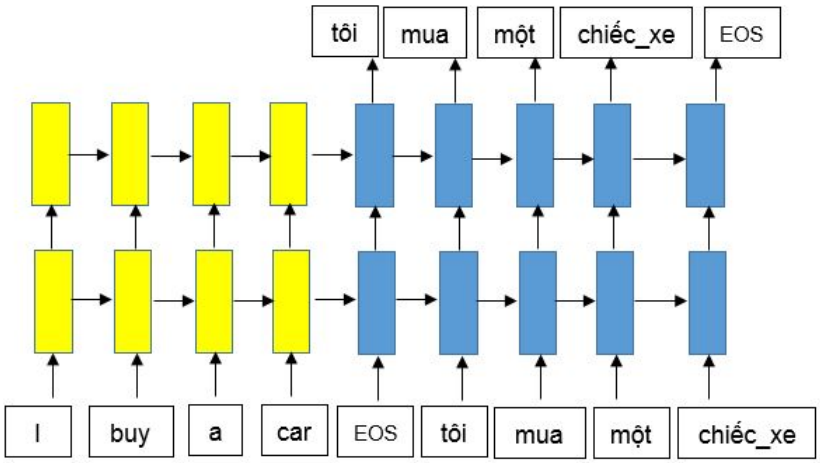



Figure 2.2: **Recurrent architecture in neural machine translation** An example of translation from an English sentence into a Vietnamese sentence; *EOS* marks the end of the sentence

Figure 2.2 illustrates an example of neural machine translation proposed in Sutskever et al., 2014 [79].

Given a source sentence $s = (s_1, ..., s_m)$, and a target sentence $t = (t_1, ..., t_n)$, the goal of a NMT is to model the conditional probability $p(t|s)$. This process bases on the encoder-decoder framework as proposed in [14, 79].

$$logp(t|s) = \sum_{j=1}^{n} logp(t_i|\{t_1, ..., t_{i-1}\}, s, c) \tag{2.16}$$

in which, the source sentence $s$ is represented by the context vector $c$ using the encoder. For each time, a target word is translated based on the context vector using the decoder.

For the decoding, the probability of each target word $t_i$ can be computed as follows.

$$p(t_i|\{t_1, ..., t_{i-1}\}, s, c) = softmax(h_i) \tag{2.17}$$

where $h_i$ is the current target hidden state as in Equation 2.18.

$$h_i = f(h_{i-1}, t_{i-1}, c) \tag{2.18}$$

Finally, for the bilingual corpus $B$, the training objective is computed as in Equation 2.19.

$$I = \sum_{(s,t)\in B} -logp(t|s) \tag{2.19}$$

# Chapter 3

# Building Bilingual Corpora

Bilingual corpora are essential resources for training SMT models. Nevertheless, large bilingual corpora are unavailable for most language pairs. Therefore, building bilingual corpora become an important task to improve SMT models. Sentence alignment, which extracts bilingual sentences from articles, is an essential step in building bilingual corpora. One of the representative methods of sentence alignment is based on the combination of length-based and word correspondences. Sentence pairs are first aligned by the length-based phase based on the correlation of the number of words or characters. The aligned sentence pairs are then used to extract word alignment, which learns a bilingual word dictionary. Finally, the length-based phase is combined with the bilingual word dictionary to generate the alignment output. Nevertheless, when the dictionary does not contain a large vocabulary, it cannot cover a large vocabulary ratio of the input data (*out-of-vocabulary*), which then affects the quality of the final alignment output. I propose a method to deal with the out-of-vocabulary (OOV) problem by using word similarity model extracted from monolingual data sets. The proposed method was then applied to build a bilingual corpus from comparable data to improve SMT for low-resource languages.

In the first section, I propose an approach to deal with the OOV problem in sentence alignment based on word similarity learned from monolingual corpora. Words that were not contained in the bilingual dictionaries were replaced by their similar words from the monolingual corpora. Experiments conducted on English-Vietnamese sentence alignment showed that using word similarity learned from monolingual corpora can reduce the OOV ratio and improve the performance in comparison with some other length-and-word-based sentence alignment methods.

In the second section, the proposed method was applied to build a bilingual corpus from comparable data. The corpus was extracted and processed from Wikipedia automatically. I obtained a multilingual parallel corpus among languages Indonesian, Malay, Filipino, Vietnamese, and English including more than one million parallel sentences of five language pairs. The corpus was evaluated on the task of statistical machine translation, which depends mainly on the availability of parallel data, and obtained promising results. The data sets significantly improved SMT performance and solved the problem of unavailable bilingual data for machine translation.

## 3.1 Dealing with Out-Of-Vocabulary Problem

In sentence alignment methods based on word correspondences, bilingual dictionaries were trained on IBM models can help to produce highly accurate sentence pairs when they contain reliable word pairs with a high percentage of vocabulary coverage. The *out-of-vocabulary (OOV)* problem appears when the bilingual dictionary does not contain word pairs which are necessary to produce a correct alignment of sentences. The higher the OOV ratio, the lower the performance. I propose a method using word similarity learned from monolingual corpora to overcome the OOV problem.

### 3.1.1 Word Similarity Models

Monolingual corpora were used to train two word similarity models separately using a continuous bag-of-word model. In continuous bag-of-words models, words are predicted based on their context, and words that appear in the same context tend to be clustered together as similar words. I adapted a word embedding model proposed by [56] namely *word2vec*, a powerful continuous bag-of-words model to train word similarity. *Word2Vec* computes words' vector based on surrounding words as contexts, and words can be seen as similarity when they appear in the same contexts. The expanded dictionary can help to cover a higher ratio of vocabulary, which reduces the OOV ratio and improves overall performance.

**Algorithm 1:** Word Similarity Using Word Embedding

**Input :** $w_1, w_2$, word2vec
**Output:** $similarity(w_1, w_2)$

```
1  begin
2      most_similar(w1)=word2vec.most_similar(w1)
3      if w2 ∈ most_similar(w1) then
4          similarity(w1, w2)=word2vec.cosine(w1, w2)
5      end
6      else
7          similarity(w1, w2)=0
8      end
9      return similarity(w1, w2)
10 end
```

Algorithm 1 describes computing word similarity using a word embedding model. In order to compute the similarity between the two words $w_1$ and $v_2$, a word embedding model was first trained on a monolingual data. Then, the word similarity was extracted from cosine similarity of the two words (line 4) in the word embedding model.

### 3.1.2 Improving Sentence Alignment Using Word Similarity

There are four phases in the proposed method: length-based phase, training bilingual dictionaries, using word similarity to deal with the OOV problem, and the combination of length-based and word-based methods. The model is illustrated in Figure 3.1.

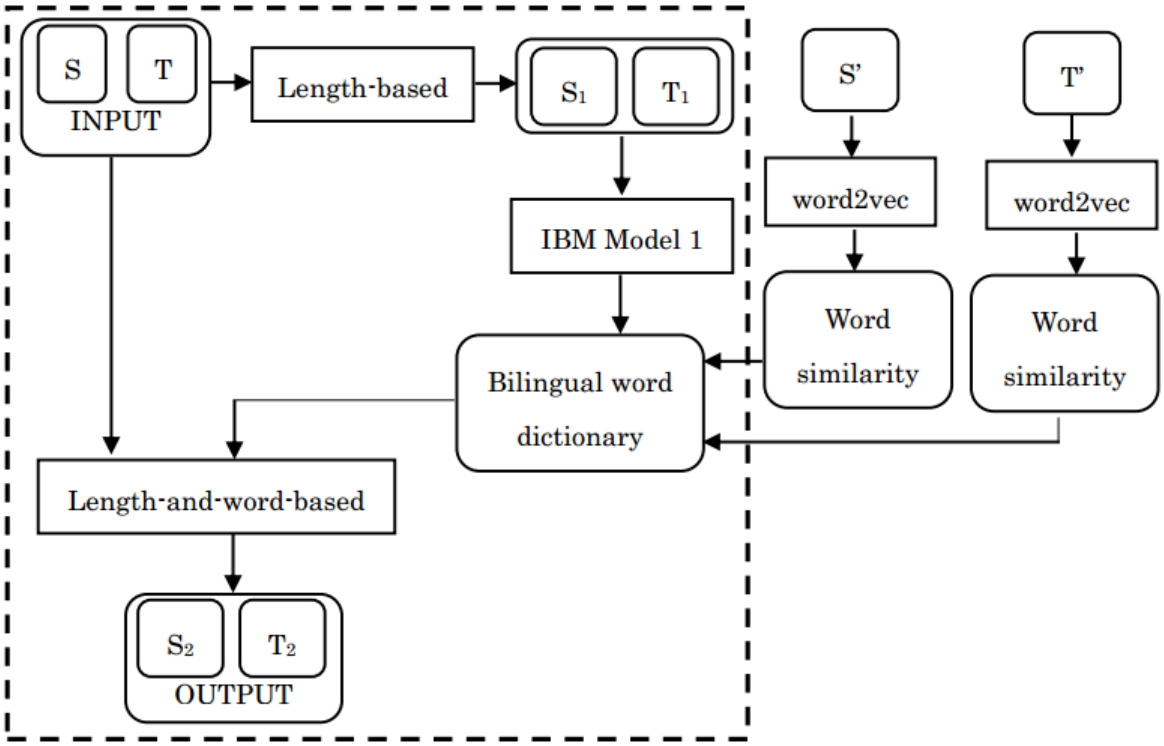



Figure 3.1: **Word similarity for sentence alignment** $S$: the text of source language, $T$: the text of target language; $S_1$, $T_1$: sentences aligned by the length-based phase; $S_2$, $T_2$: sentences aligned by the length-and-word-based phase; $S'$, $T'$: monolingual corpora of the source and target languages, respectively. The components of the length-and-word-based method [59] are bounded by the dashed frame.

As described in Algorithm 2, for each word pair $(s, t)$ in the input word alignment of a language pair $L_s, L_t$, the word alignment can be extended by similar words of $s$ and $t$ in the two word similarity models (lines 3-12). The alignment scores of the new word pairs were computed based on the alignment scores of the input word alignment pairs and the similarity scores (line 5, 9). Finally, the scores were normalized using maximum likelihood to obtain the extended word alignment.

The extended word alignment was then used in the last phase of the baseline sentence alignment algorithm in building bilingual corpora. In the second section, I used the proposed sentence alignment method to build a bilingual corpus from comparable data.

**Algorithm 2:** Extending Word Alignment Using Word Similarity

**Input :** $Wa$, $Ws$, $Wt$
**Output:** $Wm$

**begin**
  $Wm = \emptyset$
  **for** $(s,t) \in Wa$ **do**
    **for** $s' \in Ws$ **do**
      alignment_score$(s',t)$=$similarity(s,s')$*alignment_score$(s,t)$
      $Wm$.add([$(s',t)$,alignment_score$(s',t)$])
    **end**
    **for** $t' \in Wt$ **do**
      alignment_score$(s,t')$=$similarity(t,t')$*alignment_score$(s,t)$
      $Wm$.add([$(s,t')$,alignment_score$(s,t')$])
    **end**
  **end**
  **for** $(s,t) \in Wm$ **do**
    $$normalized\_alignment\_score(s,t) = \frac{alignment\_score(s,t)}{\sum_{t'} alignment\_score(s,t')}$$
  **end**
  return $Wm$
**end**

### 3.1.3 Experiments

**Training Data** I conducted experiments on the sentence alignment task for English-Vietnamese, a low-resource language pair. I evaluated my method on the test set collected from the website.[1] After preprocessing the collected data, I conducted sentence alignment manually to achieve the reference data. The statistics of these data sets are shown in Table 3.1.

In order to produce a more reliable bilingual dictionary, I added an available bilingual corpus to train IBM Model 1, which was collected from the IWSLT2015 workshop.[2] The dataset contains subtitles of TED talks [9]. The IWSLT2015 training data is shown in Table 3.2.

In order to train word similarity models, I used English and Vietnamese monolingual corpora. For English I used the one-billion-words[3] dataset which contains almost 1B words. To build a huge monolingual corpus of Vietnamese, I extracted articles from the web (*www.baomoi.com*)[4]. The data set was then preprocessed to achieve 22 million Vietnamese

[1] http://www.vietnamtourism.com/
[2] https://sites.google.com/site/iwsltevaluation2015/mt-track
[3] http://www.statmt.org/lm-benchmark/
[4] http://www.baomoi.com/

Table 3.1: **English-Vietnamese sentence alignment test data set**

| **Statistics** | **Test Data** |
|---|---|
| Sentences (English) | 1,705 |
| Sentences (Vietnamese) | 1,746 |
| Average length (English) | 22 |
| Average length (Vietnamese) | 22 |
| Vocabulary Size (English) | 6,144 |
| Vocabulary Size (Vietnamese) | 5,547 |
| Reference Set | 837 |

Table 3.2: **IWSLT15 corpus for training word alignment**

| **Statistics** | **iwslt15** |
|---|---|
| Sentences (English) | 129,327 |
| Sentences (Vietnamese) | 129,327 |
| Average length (English) | 19 |
| Average length (Vietnamese) | 18 |
| Vocabulary Size (English) | 46,669 |
| Vocabulary Size (Vietnamese) | 50,667 |

sentences.

**Training Details** The standard preprocessing steps include word tokenization and lowercase. As commonly used for preprocessing data in many tasks like the machine translation competition [5], I utilized the Moses toolkit [6] for preprocessing English. For Vietnamese, since there are a kind of words in Vietnamese called compound words in which a sequence of two or three words can be merged together with a new meaning, I conducted the preprocessing step called word segmentation using the well-known preprocessing tool JVnTextpro [7] for Vietnamese.

For sentence alignment, I implemented the hybrid model (Moore, 2002) [59] using Java. I compared my model with two well-known hybrid methods: **M-align**[8] (Moore, 2002) [59] and **hunalign**[9] (Varga et al., 2005) [92]. For evaluation, I used the commonly used metrics: Precision, Recall, and F-measure [93]. I setup the length-based phase's threshold to 0.99 to extract highest sentence pairs. Then in the length-and-word-based phase, I setup the threshold to 0.9 to ensure a high confidence. The thresholds were set using the same configurations as described in the baseline approach [59].

[5]http://www.statmt.org/wmt17/

[6]http://www.statmt.org/moses/?n=moses.baseline

[7]http://jvntextpro.sourceforge.net/

[8]http://research.microsoft.com/en-us/downloads/aafd5dcf-4dcc-49b2-8a22-f7055113e656/

[9]http://mokk.bme.hu/en/resources/hunalign/

I used the well-known word2vec from gensim python[10] to train two word-similarity models on the monolingual corpora. I set the *CBOW* model with configurations: window size=5, vector size=100, and min count = 10.

Table 3.3: **English-Vietnamese alignment results**; **M-align**: the Microsoft sentence aligner [59], **hunalign**: [92] *Hypothesis*, *Reference*, *Correct*: number of sentence pairs generated by the systems, the reference set, and the correct sentences, respectively

| **Setup** | **M-align** | **hunalign** | **My Method** |
|---|---|---|---|
| Hypothesis | 580 | 1373 | 609 |
| Reference | 837 | 837 | 837 |
| Correct | 412 | 616 | 433 |
| Precision | 71.03% | 44.87% | 71.10% |
| Recall | 49.22% | 73.60% | 51.73% |
| F-measure | 58.15% | 55.75% | 59.89% |

**Results** Experimental results are shown in Table 3.3. Overall, the performance of my model slightly improved the M-align in all scores of precision, recall, and f-measure. My model also gained higher performance than the hunalign. Although the hunalign can achieve the highest recall of 73.60% due to the approach that the hunalign constructs dictionaries, the method produced a number of error results, so this caused the lowest precision.

### 3.1.4 Analysis

**Word Similarity** I describe word similarity models using word embedding with examples. Tables 3.4 and 3.5 show examples of OOV words and their most similar words extracted from the word similarity models. The word similarity models can explore not only helpful similar words in terms of variants in morphology but also words that share the same meaning but different morphemes. There are useful similar words that can have the same meaning as the OOV words like word pairs ("*intends*" and "*aims*") or ("*honours*" and "*awards*"), ("*quát*" (to shout) and "*mắng*" (to scold)), ("*hủy*" (to destroy) and "*phá*" (to demolish)).

**Out-Of-Vocabulary Ratio** A problem of using the IBM Model 1 as in Moore's method was the OOV. When the dictionary cannot cover a high ratio of vocabulary, it decreases the contribution of the word-based phase. The average OOV ratio is shown in Table 3.6. In comparison with M-align, using word similarity in my model reduced the OOV ratio from 7.37% to 4.33% in English and from 7.74% to 6.80% in Vietnamese vocabulary. By using word similarity models I overcame the problem of OOV. The following discussion will show how the word similarity models helped to reduce the OOV ratio.

[10]https://radimrehurek.com/gensim/models/word2vec.html

Table 3.4: **Sample English word similarity**

| OOV Words | Similar Words | OOV Words | Similar Words | OOV Words | Similar Words |
|---|---|---|---|---|---|
| diversifying | diversification | intends | plans | honours | honors |
| diversifying | expanding | intends | aims | honours | prize |
| diversifying | shifting | intends | refuses | honours | prizes |
| diversifying | diversify | intends | prefers | honours | award |
| diversifying | globalizing | intends | seeks | honours | awards |
| | | intends | continues | honours | accolades |

Table 3.5: **Sample Vietnamese word similarity**: the italic words in brackets are corresponding English meaning which were translated by the authors.

| OOV Words | Similar Words | OOV Words | Similar Words |
|---|---|---|---|
| quát (*to shout*) | mắng (*to scold*) | hủy (*to destroy*) | hoại (*to ruin*) |
| quát (*to shout*) | nạt (*to bully*) | hủy (*to destroy*) | dỡ (*to unload*) |
| | | hủy (*to destroy*) | phá (*to demolish*) |

**Sample Alignment** I show an example of how my method deals with the OOV problem in Table 3.7.

The word pairs (*reunification-thống_nhất*) and (*impressively-mạnh_mẽ*) were not covered by the dictionary using IBM Model 1, and this became an example of OOV. Examples of similar word pairs are shown in Table 3.8, and translation word pairs trained by IBM Model 1 are shown in Table 3.9.

Because (*reunification-unification*) was a similar word pair, and the translation word pair (*unification-thống_nhất*) was contained in the dictionary, the new translation word pair (*reunification-thống_nhất*) was then created. Similarly, the new translation word pair (*impressively-mạnh_mẽ*) was created via the similar word pair (*impressively-impressive*) and the translation word pair (*impressive-mạnh_mẽ*). Table 3.10 shows induced translation word pairs. By using word similarity learned from monolingual corpora, a number of OOV words can be replaced by their similar words, which helped to reduce the OOV ratio and improve performance in overall.

## 3.2 Building A Multilingual Parallel Corpus

This section describes applying the proposed method in the first section to build a bilingual corpus from comparable data for low-resource language pairs. For this task, the corpus was built from Wikipedia for Southeast Asian languages: Indonesian, Malay, Filipino, Vietnamese; and the languages paired with English, in which there is no bilingual corpus

Table 3.6: **OOV ratio in sentence alignment**

| Setup | Test | M-align | My Model |
|---|---|---|---|
| #vocab. en | 1,705 | 27,872 | 28,371 |
| #vocab. vi | 1,746 | 25,326 | 25,481 |
| OOV en | NA | 7.37% | 4.33% |
| OOV vi | NA | 7.74% | 6.80% |

Table 3.7: **Sample English-Vietnamese alignment**: the translations to English (italic) were conducted by the authors.

| Language | Sentence |
|---|---|
| English | since the reunification in 1975 , vietnam ' s architecture has been impressively developing . |
| Vietnamese | từ sau ngày đất_nước thống_nhất ( 1975 ) kiến_trúc việt_nam phát_triển khá mạnh_mẽ . |
| (*Meaning*) | *After the country was unified (1975), vietnam's architecture has been developing rather impressively.* |

or only small bilingual corpus available for these language pairs. The corpus was then used to improve SMT.

Wikipedia is a large resource that contains a number of articles in many languages in the world. The freely accessible resource is a kind of comparable data in which many articles are in the same domain in different languages. I can exploit this resource to build bilingual corpora, especially for low-resource language pairs.

Table 3.8: **English word similarity**

| OOV Words | Similar Words | Cosine Similarity |
|---|---|---|
| reunification | independence | 0.71 |
| **reunification** | **unification** | **0.67** |
| reunification | peace | 0.62 |
| impressively | amazingly | 0.74 |
| **impressively** | **impressive** | **0.74** |
| impressively | exquisitely | 0.72 |
| impressively | brilliantly | 0.71 |

Table 3.9: **Sample IBM Model 1** (*Score*: translation probability); the translations to English (italic) were conducted by the authors.

| Score | English | Vietnamese |
|---|---|---|
| 0.597130 | independence | độc_lập (*independent*) |
| 0.051708 | independence | sự_độc_lập (*independence*) |
| **0.130447** | **unification** | **thống_nhất** (*to unify*) |
| 0.130447 | unification | sự_thống_nhất (*unification*) |
| 0.130446 | unification | sự_hợp_nhất (*unify*) |
| 0.551291 | impressive | ấn_tượng (*impression*) |
| **0.002927** | **impressive** | **mạnh_mẽ** (*impressive*) |
| 0.002440 | impressive | kinh_ngạc (*amazed*) |

Table 3.10: **Induced word alignment**; the *(italic)* indicates English meaning

| Score | English | Vietnamese |
|---|---|---|
| 0.215471 | reunification | thống_nhất (*to unify*) |
| 0.369082 | impressively | mạnh_mẽ (*impressive*) |

### 3.2.1 Related Work

Building parallel corpora from webs has been exploited in a long period. One of the first work can be presented in [67]. In order to extract parallel documents from webs, [46] used the similarity of the URL and page content. [90] used matching documents to build parallel data. Meanwhile, [40] used manual involvement for building a multilingual parallel corpus. In the work of [9], a multilingual corpus was built from subtitles of the TED talks website.

For collecting parallel data from Wikipedia, the task has been investigated in some previous work. In [37], parallel sentences are extracted from Wikipedia for the task of multilingual named entity recognition. In [76], parallel corpora are extracted from Wikipedia for English, German, and Spanish. A recent word is proposed in [15], which extract parallel sentences before using an SVM classifier to filter the sentences using some features.

For the Southeast Asian languages, there are few bilingual corpora. A multilingual parallel corpus was built manually in [80]. The corpus is a valuable resource for the languages. Nevertheless, because the corpus is still small with only 20,000 multilingual sentences, and manually building parallel corpora requires many cost of human annotators, automatically extracting large bilingual corpora becomes an essential task for the development of natural language processing for the languages including cross-language tasks like machine translation. In my work, a multilingual parallel corpus of several Southeast Asian languages was built. The corpus was built based on Wikipedia's parallel articles that were collected from the articles' title and inter-language link records. Parallel sentences were extracted based on the powerful sentence alignment algorithm ([59]). The corpus was uti-

lized for improving machine translation on the Southeast Asian low-resource languages, in which there has been no work investigated on this task to our best knowledge.

### 3.2.2 Methods

In order to build a bilingual corpus from Wikipedia, I first extracted parallel titles of Wikipedia articles. Then, pairs of articles were crawled based on the parallel titles. Finally, sentences in the article pairs were aligned to extract parallel sentences. I describe these steps in more detail in this section. The scripts for my methods and extracted bilingual corpus can be accessed at .[11]

**Extracting Parallel Titles** The content of Wikipedia can be obtained from their database dumps.[12] In order to extract parallel titles of Wikipedia articles, I used two resources for each language from the Wikipedia database dumps: the articles' titles and IDs in a particular language (ending with *-page.sql.gz*) and the interlanguage link records (file ends with *-langlinks.sql.gz*).

Table 3.11: **Wikipedia database dumps' resources used to extract parallel titles**; **page (KB)**: the size of the articles' IDs and their titles in the language; **langlinks (KB)**: the size of the interlanguage link records; I collected the two resources for five languages: **en** (English), **id** (Indonesian), **fil** (Filipino), **ms** (Malay), and **vi** (Vietnamese); I used the database that was *updated on 2017-01-20*.

| No. | Data | page (KB) | langlinks (KB) |
|---|---|---|---|
| 1 | en | 1,477,861 | 280,617 |
| 2 | vi | 92,541 | 111,420 |
| 3 | id | 57,921 | 72,117 |
| 4 | ms | 21,791 | 56,173 |
| 5 | fil | 5,907 | 23,446 |

I aims to built a multilingual parallel corpus for several low-resource Southeast Asian languages including Indonesian, Malay, Filipino, and Vietnamese, which there are few bilingual corpora. Furthermore, bilingual corpora of the languages paired with English are also important resources for further research including machine translation. Therefore, I collected the Wikipedia database dumps of the five languages: English, Indonesian, Malay, Filipino, and Vietnamese. Table 3.11 presents the Wikipedia database dumps that I used to extract parallel titles. The English database contains a much larger information in both the articles' titles and the interlanguage link records. Meanwhile, the Filipino database is much smaller, that effects the number of extracted parallel titles as well as final extracted parallel sentences. The extracted parallel titles are presented in Table 3.12.

[11]https://github.com/nguyenlab/Wikipedia-Multilingual-Parallel-Corpus
[12]https://dumps.wikimedia.org/backup-index.html

Table 3.12: **Extracted and processed data from parallel titles**; **Crawled Src Art. (Crawled Trg Art.)**: the number of crawled source (target) articles using the title pairs for each language pair; **Art. Pairs**: the number of parallel articles processed after crawling; **Src Sent. (Trg Sent.)**: the number of source (target) sentences in the article pairs after preprocessing (removing noisy characters, empty lines, sentence splitting, word tokenization).

| No. | Data | Title pairs | Crawled Src Art. | Crawled Trg Art. | Art. Pairs | Src Sent. | Trg Sent. |
|---|---|---|---|---|---|---|---|
| 1 | en-id | 198,629 | 197,220 | 190,954 | 150,759 | 4,646,453 | 990,661 |
| 2 | en-fil | 52,749 | 51,698 | 51,157 | 50,021 | 3,428,599 | 367,276 |
| 3 | en-ms | 204,833 | 201,688 | 199,950 | 160,709 | 2,158,726 | 320,624 |
| 4 | en-vi | 452,415 | 433,124 | 436,488 | 420,919 | 12,130,133 | 3,831,948 |
| 5 | id-fil | 30,313 | 29,961 | 24,946 | 22,760 | 502,457 | 254,216 |
| 6 | id-ms | 98,305 | 88,028 | 89,936 | 68,676 | 452,604 | 403,807 |
| 7 | id-vi | 159,247 | 149,974 | 128,530 | 121,673 | 1,201,848 | 1,878,855 |
| 8 | fil-ms | 25,231 | 21,856 | 25,023 | 21,135 | 202,851 | 243,361 |
| 9 | fil-vi | 36,186 | 30,540 | 35,625 | 28,830 | 267,453 | 723,155 |
| 10 | ms-vi | 133,651 | 118,647 | 116,620 | 105,692 | 560,042 | 1,256,468 |

**Collecting Parallel Articles** After parallel titles of Wikipedia articles were extracted, I collected the article pairs using the parallel titles. I implemented a Java crawler for collecting the articles. The collected data set was then carefully processed in hierarchical steps from articles to sentences, then to word levels. First, noisy characters were removed from the articles. Then, for each article, sentences in paragraphs were splitted so that there is one sentence per line. For each sentence, words were tokenized that separated from punctuations. The sentence and word tokenization steps were conducted using the Moses scripts.[13]

As described in Table 3.12, using the title pairs, I obtained a high ratio of crawled articles. For instance, using 198k title pairs of English-Indonesian, I crawled 197k English articles and 190k Indonesian articles successfully, which there existed the article based on a title. This issue is emphasized because sometimes there is no existed article given a title that will show an error in crawling. For the case of Indonesian-Vietnamese, although there was 159k extracted parallel titles, I obtained 128k Vietnamese articles, which there were more than 30k error or inexistent articles given the set of titles.

**Sentence Alignment** For each parallel article pair, I aligned sentences using the proposed sentence alignment method in the previous section.

After the sentence alignment step, I obtained the parallel data sets which are described in Table 3.13.

[13]https://github.com/moses-smt/mosesdecoder/tree/master/scripts/tokenizer

Table 3.13: **Sentence alignment output**; **Sent. Pairs**: the number of parallel sentences extracted from the **article pairs** after the sentence alignment step

| No. | Data | Article pairs | Sent. Pairs |
|---|---|---|---|
| 1 | en-id | 150,759 | 234,380 |
| 2 | en-fil | 50,021 | 22,758 |
| 3 | en-ms | 160,709 | 198,087 |
| 4 | en-vi | 420,919 | 408,552 |
| 5 | id-fil | 22,760 | 9,952 |
| 6 | id-ms | 68,676 | 83,557 |
| 7 | id-vi | 121,673 | 76,863 |
| 8 | fil-ms | 21,135 | 4,919 |
| 9 | fil-vi | 28,830 | 10,418 |
| 10 | ms-vi | 105,692 | 65,177 |

### 3.2.3 Extracted Corpus

I obtained a multilingual parallel corpus of ten language pairs, which are among Southeast Asian languages and the languages paired with English. In totally, the corpus contains a huge number of parallel sentences up to more than 1.1 million sentence pairs which can be valuable when there is no available bilingual corpora for almost such language pairs. Large bilingual corpora can be extracted such as: English-Vietnamese (408k parallel sentences), Indonesian-English (234k parallel sentences). However, because of the smaller number of the input parallel articles for several language pairs, a much smaller number of parallel sentences were extracted like Indonesian-Filipino (9k) and Filipino-English (22k).

Table 3.14: **Extracted Southeast Asian multilingual parallel corpus**

| No. | Data | Sent. Pairs | Src Words | Trg Words | Src Vocab. | Trg Vocab. |
|---|---|---|---|---|---|---|
| 1 | en-id | 234,380 | 4,648,359 | 4,359,976 | 208,920 | 209,859 |
| 2 | en-fil | 22,758 | 447,719 | 399,058 | 42,670 | 44,809 |
| 3 | en-ms | 198,087 | 3,273,943 | 3,221,738 | 156,806 | 148,133 |
| 4 | en-vi | 408,552 | 7,229,963 | 8,373,549 | 274,178 | 222,068 |
| 5 | id-fil | 9,952 | 132,097 | 172,363 | 18,531 | 19,737 |
| 6 | id-ms | 83,557 | 1,464,506 | 1,447,247 | 87,240 | 92,126 |
| 7 | id-vi | 76,863 | 1,014,351 | 1,136,710 | 67,211 | 57,788 |
| 8 | fil-ms | 4,919 | 78,729 | 66,324 | 10,184 | 10,671 |
| 9 | fil-vi | 10,418 | 141,135 | 151,086 | 15,641 | 13,071 |
| 10 | ms-vi | 65,177 | 928,205 | 896,784 | 60,574 | 52,673 |
| | Total | 1,114,663 | – | – | – | – |

Furthermore, I extracted monolingual data sets for the languages: Indonesian, Malay, Filipino, and Vietnamese, which are almost publicly unavailable. The data sets are described in Table 3.15. Large monolingual data sets were obtained such as Indonesian (3.1

Table 3.15: **Monolingual data sets**

| Data set | Sentences | Vocab. | Size (KB) |
|---|---|---|---|
| id | 3,147,570 | 917,861 | 369 |
| fil | 1,034,215 | 252,565 | 113 |
| ms | 1,527,834 | 599,396 | 172 |
| vi | 7,690,426 | 936,137 | 1,033 |

million sentences), Malay (1.5 million sentences), and Vietnamese (up to 7.6 million sentences). I believe that the data sets are also useful for such low-resource languages such as training language models and other tasks.

### 3.2.4 Domain Adaptation

The question now is that how can we utilize the corpus effectively. If there are existing bilingual corpora for the language pairs, which strategies we can use to combine and take advantage the full potential of the corpus. This can be seen as a problem of domain adaptation [45] when the extracted corpus and the existing corpus may come from different domain. Therefore, an effective strategy to combine the two resources is required. I discuss the issue of domain adaptation about the strategies to combine bilingual corpora in this section.

I assume that given a language pair, there exist a bilingual corpus called the *direct corpus*. The corpus extracted from Wikipedia can be used as an additional resource, called the *alignment corpus*. For statistical machine translation (SMT) [44], a bilingual corpus are used to train a phrase table. I used the direct corpus and the alignment corpus to generate two phrase tables called the *direct component* and the *alignment component*. I adapted the linear interpolation [70] to combine the two components. Equation 3.1 describes the combination of the components.

- $d$: the direct component
- $a$: the alignment component

$$p(t|s) = \lambda_d p_d(t|s) + \lambda_a p_a(t|s) \tag{3.1}$$

where $p_d(t|s)$ and $p_a(t|s)$ stand for the translation probability of the direct, alignment, and pivot models, respectively.

The interpolation parameters $\lambda_d$ and $\lambda_a$ in which $\lambda_d + \lambda_a$ were computed by the following strategies.

- *tune*: the parameters were tuned using a development data set which was provided in tuning machine translation models.
- *weights*: the parameters were set based on the ratio of the BLEU scores when using each model separately for decoding on the tuning set.

I evaluated the domain adaptation strategies as well as utilizing the aligned corpus in the experiments section.

### 3.2.5 Experiments on Machine Translation

After the multilingual parallel corpus was built, I evaluate the corpus on machine translation, which aims to improve the machine translation performance using the additional resource. There are two experiments in the evaluation. First, the corpus was used to train translation models, then translate test sets extracted from the Asian Language Treebank corpus [80]. Second, the corpus was used to improve an English-Vietnamese translation system on the shared task IWSLT 2015, which was tested on both phrase-based and neural-based methods.

**SMT on the Asian Language Treebank Parallel Corpus** The parallel corpus extracted from Wikipedia was then used for training SMT models. I aim to exploit the data to improve SMT on low-resource languages.

**Training Data**

I evaluate the corpus on SMT experiments. For development and testing data, I used the ALT corpus (Asian Language Treebank Parallel Corpus) [80], this is a corpus including 20K multilingual sentences of English, Japanese, Indonesian, Filipino, Malay, Vietnamese, and some other Southeast Asia languages. I extracted the development and test sets from the ALT corpus: 2k sentence pairs for development sets, and 2k sentence pairs for test sets.

**Training Details**

I trained SMT models on the parallel corpus using the Moses toolkit [43]. The word alignment was trained using GIZA++ [65] with the configuration *grow-diag-final-and*. A 5-gram language model of the target language was trained using KenLM [31]. For tuning, I used batch MIRA [13]. For evaluation, I used the BLEU scores [66].

**Results**

Table 3.16 describes the experimental results on the development and test sets. It is noticeable that the SMT models trained on the bilingual data aligned from Wikipedia can produce promising results.

For the results on the development sets, I achieved promising results with high BLEU points such as: the Indonesian-Malay pairs (Indonesian-Malay 31.64 BLEU points, Malay-Indonesian 31.56 BLEU points). Similarly, several language pairs also showed high BLEU points such as: English-Vietnamese (30.58 and 23.01 BLEU points), English-Malay (29.85 and 28.87 BLEU points), English-Indonesian (30.56 and 30.14 BLEU points), and Indonesian-Vietnamese (21.85 and 17.41 BLEU points). The language pairs which showed high scores contain a large number of sentences, for instance English-Vietnamese (408k sentence pairs), English-Indonesian (234k sentence pairs), and English-Malay (198k sentence pairs). Nevertheless, since the small number of the extracted corpus on several languages paired with Filipino such as Indonesian-Filipino (9.9k sentence pairs), Malay-Filipino (21.1k sentence pairs), and Vietnamese-Filipino (10.4k sentence pairs), the experimental results

Table 3.16: **Experimental results on the development and test sets (BLEU)**; **Dev (L1-L2)**, **Test (L1-L2)**, **fil-ms**: the translation scores on the development (test) set of the translation from the first language (**L1(fil)**) to the second language (**L2 (ms)**) in the language pair **fil-ms**; **Dev (L2-L1)**, **Test (L2-L1)**, **fil-ms**: the translation on the development (test) set of the inverse translation (from **ms** to **fil**)

| No. | Language Pairs | Dev (L1-L2) | Test (L1-L2) | Dev (L2-L1) | Test (L2-L1) |
|---|---|---|---|---|---|
| 1 | en-id | 30.56 | 28.87 | 30.14 | 29.01 |
| 2 | en-fil | 18.54 | 19.08 | 18.98 | 19.89 |
| 3 | en-ms | 29.85 | 33.23 | 28.87 | 23.82 |
| 4 | en-vi | 30.58 | 34.42 | 23.01 | 22.56 |
| 5 | id-fil | 11.36 | 11.04 | 9.58 | 9.77 |
| 6 | id-ms | 31.64 | 30.21 | 31.56 | 30.11 |
| 7 | id-vi | 21.85 | 22.42 | 17.41 | 17.45 |
| 8 | fil-ms | 7.43 | 8.02 | 8.70 | 9.27 |
| 9 | fil-vi | 5.97 | 6.69 | 6.45 | 7.15 |
| 10 | ms-vi | 15.51 | 18.12 | 11.96 | 13.88 |

showed much lower performance than other language pairs: Indonesian-Filipino (11.36 and 9.58 BLEU points), Malay-Filipino (8.70 and 7.43 BLEU points), and Vietnamese-Filipino (6.45 and 5.97 BLEU points).

Similarly, for the experimental results on the test sets, the language pairs with large bilingual corpora achieved high performance: English-Indonesian (28.87 and 29.01 BLEU points), English-Malay (33.23 and 23.82 BLEU points), English-Vietnamese (34.42 and 22.56 BLEU points). The situation of languages paired Filipino showed the much lower performance: Indonesian-Filipino (11.04 and 9.77 BLEU points), Malay-Filipino (9.27 and 8.02 BLEU points), and Vietnamese-Filipino (7.15 and 6.69 BLEU points).

Figure 3.2 presents experimental results on the development sets (test sets) that vary in several aspects: the translation directions (L1-L2, L2-L1), the corpus's size, and the language pairs. There are several interesting findings from the charts. First, the bigger the corpus's size, the higher the BLEU scores. I sorted the corpus's size increasingly from the left to right. For instance, since the corpora' sizes of language pairs such as Filipino-Malay (4.9k), Indonesian-Filipino (9.9k), and Filipino-Vietnamese (10.4k) are much smaller than that of the language pairs such as Indonesian-Malay (83.5k), English-Indonesian (234k), English-Vietnamese (408k), the BLEU scores also show the correlation of the two language-pair groups: Filipino-Malay, Indonesian-Filipino, Filipino-Vietnamese ($<$10 or $\approx$ 10 BLEU points); Indonesian-Malay, English-Indonesian, English-Vietnamese ($\approx$ 25-30 BLEU points). Second, in the aspect of the translation directions (L1-L2, L2-L1), the scores of the two translations in each language pair are mostly similar to each other in most cases, for instance: English-Indonesian (30.56 and 30.14 BLEU points in the two translation directions on the development set, 28.87 and 29.01 on the test set), Indonesian-Malay (31.64 and 31.56 BLEU points on the development set, 30.21 and 30.11 on the test

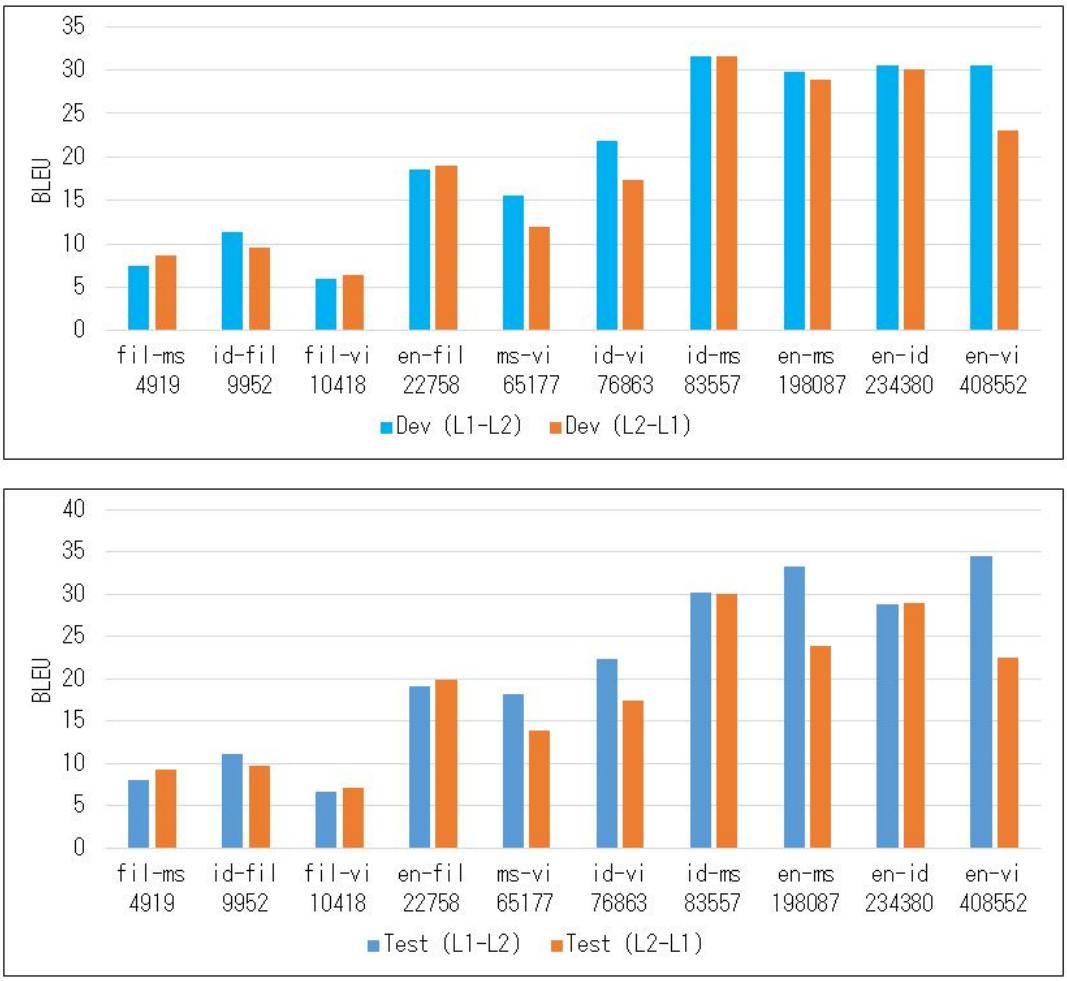



Figure 3.2: **Experimental results on the development and test sets**; the corpus's size is presented for each language pair (**fil-ms 4919**: the Filipino-Malay corpus with 4,919 parallel sentences)

set). Nevertheless, for Vietnamese, the translation scores from a language to Vietnamese are much higher than the translation scores from Vietnamese to that language in most cases, for instance: Malay-Vietnamese (15.51 BLEU point (ms-vi) vs. 11.96 (vi-ms) on the development set, 18.12 (ms-vi) vs. 13.88 (vi-ms) on the test set), Indonesian-Vietnamese (21.85 vs. 17.41 BLEU points on the development set, 22.42 vs. 17.45 BLEU points on the test set), and English-Vietnamese (30.58 vs. 23.01 BLEU points on the development set, 34.42 vs. 22.56 BLEU points on the test set). This problem of the unbalance scores between the two translation directions of a language paired with Vietnamese as well as other language pairs should be further investigated.

**Evaluation on the IWSLT 2015 Machine Translation Shared Task** In this section, I evaluated the extracted corpus on the IWSLT 2015 machine translation shared

task on English-Vietnamese. I aim to evaluate whether the *Wikipedia* corpus can improve some baseline systems on the shared task. In addition, I conducted various experiments of the domain adaptation strategies, statistical machine translation, and neural machine translation using the *Wikipedia* corpus to explore optimal strategies in exploiting the corpus.

**Training Data**

I used the data sets provided by the International Workshop on Spoken Language Translation (IWSLT 2015) machine translation shared task [10], which include three data sets of the training, development, and test sets extracted from subtitles of TED talks.[14] For the training data, the data set called the *constrained* data of 131k parallel sentences. The workshop provided data sets for development and test sets: **tst2012**, **tst2013**, and **tst2015**. In all experiments, I used the **tst2012** for the development set, the **tst2013** and **tst2015** for the test sets.

Table 3.17: **Data sets on the IWSLT 2015 experiments**; **Src Words (Trg Words)**: the number of words in the source (target) side of the corpus; **Src Vocab. (Trg Vocab.)**: the vocabulary size in the source (target) side of the corpus

| **Data** | **Sent.** | **Src Words** | **Trg Words** | **Src Vocab.** | **Trg Vocab.** |
|---|---|---|---|---|---|
| constrained | 131,019 | 2,534,498 | 2,373,965 | 50,118 | 54,565 |
| unconstrained | 456,350 | 8,485,112 | 8,132,913 | 114,161 | 124,846 |
| constrained+Wikipedia | 538,981 | 9,710,389 | 9,017,601 | 288,785 | 345,839 |
| unconstrained+Wikipedia | 864,312 | 15,661,003 | 14,776,549 | 338,424 | 403,581 |
| tst2012 | 1,581 | 28,773 | 27,101 | 3,713 | 3,958 |
| tst2013 | 1,304 | 28,036 | 27,264 | 3,918 | 4,316 |
| tst2015 | 1,080 | 20,844 | 19,951 | 3,175 | 3,528 |

In addition, I used two other data sets for training data: the corpus of National project VLSP (Vietnamese Language and Speech Processing)[15] and the EVBCorpus [62]. The two data sets were merged with the *constrained* data to obtain a large training data set called the *unconstrained* data. The training, development, and test sets are described in Table 3.17.

**Training Details**

I trained translation systems using two methods: SMT and NMT.

**Statistical Machine Translation** In order to train SMT models, I used the well-known Moses toolkit [43]. The GIZA++ [65] was used to train word alignment. For language model, I used KenLM [31] to train 5-gram language models on the target side (Vietnamese) of the training data sets. The parameters were tuned using batch MIRA ([13]). BLEU [66] was used as the metric for evaluation.

[14]https://www.ted.com/talks

[15]http://vlsp.vietlp.org:8080/demo/?page=home

**Neural Machine Translation**

In my work, I based on the model of [71], which are encoder-decoder networks with an attention mechanism [1]. For NMT model, I adopted the attentional encoder-decoder networks combined with byte-pair encoding [71]. In my experiments, I set the word embedding size 500, and hidden layers size of 1024. Sentences are filtered with the maximum length of 50 words. The minibatches size is set to 60. The models were trained with the optimizer Adadelta [99]. The models were validated each 3000 minibatches based on the BLEU scores on development sets. I saved the models for each 6000 minibatches. For decoding, I used beam search with the beam size of 12. I trained NMT models on an Nvidia GRID K520 GPU.

**Results**

**SMT results** Table 3.18 presents experimental results using SMT models. Using the *Wikipedia* corpus, I achieved promising results: 18.40 BLEU point (tst2012), 22.06 (tst2013), and 20.34 (tst2015). When the *Wikipedia* corpus was merged with the constrained data for training data, a significant improvement was achieved especially on the tst2015 (26.69 BLEU point, which improved 1.22 BLEU point from the model using the constrained data). Nevertheless, the domain adaptation strategies show even better performance than the merging setting, in which the *weights* setting model obtained the best performance with +1.74 BLEU point improvement on the tst2015.

Table 3.18: **Experimental results using phrase-based statistical machine translation**; *constrained+Wikipedia*: the constrained data was merged with the *Wikipedia* corpus; *unconstrained*Wikipedia*: interpolation of the two models; *tune, weights*: the two interpolation settings; the **bold** indicates the best results for each setup

| **Model** | **tst2012** | **tst2013** | **tst2015** |
|---|---|---|---|
| Wikipedia | 18.40 | 22.06 | 20.34 |
| constrained | 24.72 | 27.31 | 25.47 |
| constrained+Wikipedia | 24.78 | 27.89 | 26.69 |
| constrained*Wikipedia (tune) | 24.65 | 28.05 | 27.00 |
| constrained*Wikipedia (weights) | **24.95** | **28.51** | **27.21** |
| unconstrained | 34.42 | 27.19 | 25.41 |
| unconstrained+Wikipedia | 33.88 | 27.28 | 26.36 |
| unconstrained*Wikipedia (tune) | 34.44 | 27.55 | 26.68 |
| unconstrained*Wikipedia (weights) | **34.73** | **28.04** | **26.78** |

**NMT results**

The NMT results are described in Table 3.19. From the experimental results, we can observe that the systems obtain the higher scores when the size of training data sets increase (from the Wikipedia, constrained, unconstrained, to the merging in which the unconstrained data was merge with the *Wikipedia* corpus). It is interesting to note that using the *Wikipedia* corpus to enhance the translation systems trained on existed data sets based on NMT achieved the significant improvement up to +4.51 BLEU points on the tst2015. In addition, I compared my systems with the Stanford ([49]), the only system that

used NMT in the IWSLT 2015 for English-Vietnamese translation. My systems outperform the Stanford system with +2.03 on the tst2013 and +0.41 on the tst2015 test sets.

Table 3.19: **Experimental results on neural machine translation (NMT)**; comparison with the Stanford system ([49]) which was the only team using NMT for the shared task.

| **Model** | **tst2012** | **tst2013** | **tst2015** |
|---|---|---|---|
| Stanford ([49]) | – | 26.9 | 26.4 |
| constrained | 20.21 | 23.59 | 17.27 |
| Wikipedia | 15.29 | 18.43 | 17.58 |
| unconstrained | 24.05 | 26.71 | 22.30 |
| unconstrained+Wikipedia | 25.29 (+1.24) | 28.93 (+2.21) | 26.81 (+4.51) |

**SMT vs. NMT** As discussion in the previous results, using the *Wikipedia* corpus to merge with the unconstrained data for the training data shows the significant improvement based on NMT (+4.51 BLEU point on the tst2015, +2.21 BLEU point on the tst2013). Meanwhile, using the same data in training SMT models does not show the improvement. Figure 3.3 illustrates the comparison between the improvement in using the *Wikipedia* corpus to train SMT vs. NMT models.

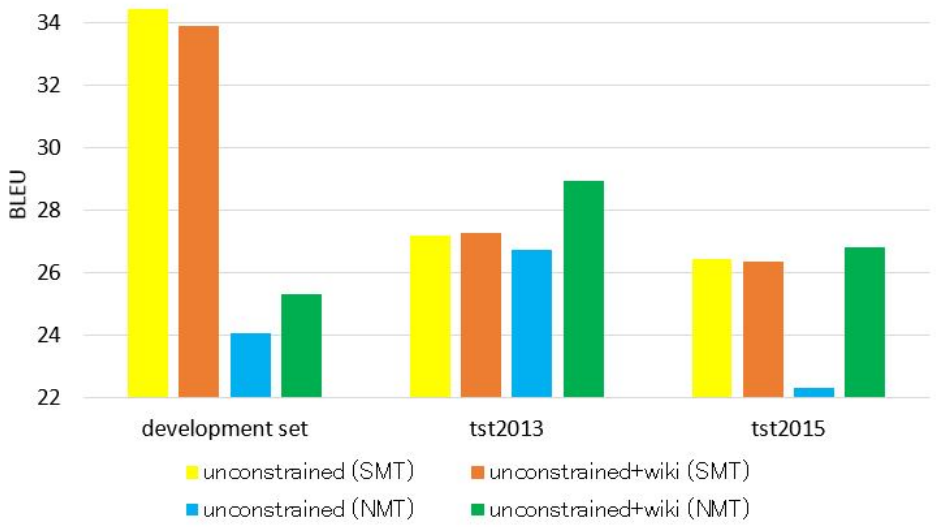



Figure 3.3: **SMT vs NMT in using the Wikipedia corpus**

**IWSLT 2015 systems** I compared my systems with other systems participated in the IWSLT 2015 machine translation shared task: Stanford ([49]), KIT ([30]), PJAIT ([96]), JAIST ([83]). As shown in Table 3.20, by using the *Wikipedia* corpus, my systems outperform the baseline system and four other systems participated in the IWSLT 2015 shared task. The results confirm the contribution of the *Wikipedia* corpus.

Table 3.20: **Comparison with other systems participated in the IWSLT 2015 shared task**; (+) indicates my system outperforms the other system; - several results of the other systems on the **tst2012** and **tst2013** were not reported.

| **System** | **tst2012** | **tst2013** | **tst2015** |
|---|---|---|---|
| PJAIT | - | - | 28.39 |
| JAIST | - | 28.32 (+) | 28.17 |
| KIT | - | - | 26.60 (+) |
| SU | - | 26.9 (+) | 26.4 (+) |
| UNETI | - | - | 22.93 (+) |
| BASELINE | - | - | 27.01 (+) |
| My best system | 24.95 | 28.51 | 27.21 |

## 3.3 Conclusion

In this chapter, I focus on the task of building bilingual corpora to improve SMT on low-resource languages, in which there is no or small bilingual corpora. I proposed a method to deal with the problem of OOV in sentence alignment, an essential step in building bilingual corpora automatically. Then, the method was used to build a multilingual parallel corpus from Wikipedia to improve SMT for several low-resource languages.

In the first section, I propose using word similarity to overcome the problem of OOV in sentence alignment. A word embedding model was trained on monolingual corpora to produce word-similarity models. These models were then combined with the bilingual word dictionary trained on IBM Model 1, which were integrated to length-and-word-based phase in a sentence alignment algorithm. My method can reduce the OOV ratio with similar words learned from monolingual corpora, which leads to an improvement in comparison with some other length-and-word-based methods. Using word similarity trained on monolingual corpora based on a word embedding model reduced the OOV in sentence alignment.

In the second section, I applied the proposed sentence alignment method to build a multilingual parallel corpus of languages: Indonesian, Malay, Filipino, Vietnamese, and English extracted from Wikipedia and processed automatically. The parallel sentences were used in SMT experiments and shown promising results. I released the scripts to extract the data, which can be used to collect parallel data from Wikipedia for other language pairs. The bilingual data set can be used to improve machine translation especially on the low-resource languages in which parallel data are very scarce or unavailable.

I switch to another strategy to improve machine translation on low-resource languages: exploiting existing bilingual corpora, which uses pivot methods to join translations via intermediate languages. I will return back to the methods of sentence alignment in Chapter 5 to introduce a hybrid model for SMT.

# Chapter 4

# Pivoting Bilingual Corpora

One of the main components to build SMT models is the bilingual corpus. For low-resource language pairs which contain small bilingual corpora, there is insufficient training data for SMT models. Previous chapter presents methods in sentence alignment and building bilingual corpora to enlarge the training data for SMT models. Another strategy can be used to improve SMT for low-resource languages is pivot methods, which exploiting existing bilingual corpora via intermediate language(s). Specifically, in order to translate from a source language to a target language, pivot language(s) can be used as a bridge for translations if there exist source-pivot and pivot-target bilingual corpora. *Triangulation*, which extracts source-target phrase pairs via common pivot phrases in the source-pivot and pivot-target phrase tables, is the representative approach in pivot methods. Previous work showed the effectiveness of the triangulation approach in improving SMT. Nevertheless, there are some drawbacks of the conventional triangulation approach: 1) the conventional triangulation approach extracts connections of source phrases to target phrases via common pivot phrases; however, pivot phrases may contain the same meaning even when they are not matched to each and 2) even when pivot phrases are matched to each other, they may contain different meanings when considering to their contexts or grammatical information like part-of-speech.

For the first problem in which pivot phrases may contain the same meaning even when they are not matched to each other, I propose a method to *improve the conventional triangulation approach based on semantic similarity between pivot phrases*. Semantic similarity models of pivot phrases were learnt using several well-known approaches: WordNet, Word2Vec, longest common subsequence, and cosine similarity. In addition to extract connections of source and target phrases based on common pivot phrases, the connections can be enhanced by similar pivot phrases. I conducted experiments on several low-resource language pairs such as Japanese-Vietnamese, Indonesian-Vietnamese, Malay-Vietnamese, and Filipino-Vietnamese. The experiments showed that the semantic similarity models extracted informative connections that improved the conventional triangulation approach.

For the second problem in which common pivot phrases may contain different meanings when considering to their grammatical information like part-of-speech; and a question is that whether additional information like part-of-speech of pivot phrases can help to improve the triangulation approach. I propose a method that *integrating grammati-*

*cal and morphological information of pivot phrases.* Instead of using the surface form of pivot phrases, part-of-speech (POS) information and lemma forms of pivot phrases were integrated to extract connections of source and target phrases via pivot phrases. Experiments were conducted on language pairs of Indonesian-Vietnamese, Malay-Vietnamese, and Filipino-Vietnamese and achieved the improvement of 0.5 BLEU points. Statistical significance tests and several analyses of using different metrics, Spearman rank correlation, and Wilcoxon signed rank tests were carefully conducted to verify the improvement. This shows the effectiveness of integrating grammatical and morphological information in pivot translation.

## 4.1 Semantic Similarity for Pivot Translation

First, I describe semantic similarity models used in my methods. Then, I present methods of applying semantic similarity to improve the conventional triangulation approach.

### 4.1.1 Semantic Similarity Models

For string level similarity, I employ two well-known techniques, namely cosine similarity and longest common subsequence. Cosine similarity [68] is an effective method and commonly used to determine the similarity between two objects [73], [63]. For the well-known longest common subsequence, it has a variety of applications [48] like measuring the similarity between two input strings [3], information retrieval [29]. For word-level similarity, I rely on WordNet-based measure and well as the word embeddings.

**Cosine Similarity** Given two string $s_1$ and $s_2$, the similarity between $s_1$ and $s_2$ can be computed using cosine similarity which is the cosine of the angle between these two vectors representation of $s_1$ and $s_2$.

$$cosine(s_1, s_2) = \frac{v_1 * v_2}{|v_1| * |v_2|} \tag{4.1}$$

where $v_1$ and $v_2$ denote the two vectors representing the two string $s_1$ and $s_2$, respectively.

**Longest Common Subsequence** The similarity of two strings $s_1$ and $s_2$ based on longest common subsequence is defined as follows.

$$d(s_1, s_2) = 1 - \frac{f(s_1, s_2)}{M(s_1, s_2)} \tag{4.2}$$

where $f(s_1, s_2)$ is the length of the longest common subsequence(s) of $s_1$ and $s_2$, and $M(s_1, s_2)$ is the length of the longest string of $s_1$ and $s_2$.

**WordNet** WordNet is a valuable linguistic resource built by annotators that contains relationship between words including synonyms, sets of words that share the same meaning. In this work, I extracted synonyms from WordNet [57] to build word similarity for English.

**Algorithm 3:** Word Similarity Using WordNet

**Input :** $w_1, w_2$, wordnet
**Output:** $similarity(w_1, w_2)$

1 **begin**
2 $synsets_1$=wordnet.synsets($w_1$)
3 $synsets_2$=wordnet.synsets($w_2$)
4 $share\_syns(w_1, w_2) = |synsets_1 \cap synsets_2|$
5 $$similarity(w_1, w_2) = \frac{share\_syns(w_1, w_2)}{\sum_{w_i} share\_syns(w_1, w_i)}$$
6 return $similarity(w_1, w_2)$
7 **end**

I describe computing word similarity using WordNet in Algorithm 3. WordNet provides synonym in a term namely synsets (lines 2-3). I defined the $share\_syns(w_1, w_2)$ (line 4) as the number of shared words of the two synsets, which denotes the strength of similarity between the two words. The strength was then normalized using maximum likelihood to obtain the similarity score of the two words.

**Word embeddings** For word similarity using word embeddings, I used the same method as described in Algorithm 1 in Chapter 3.

### 4.1.2 Semantic Similarity for Triangulation

**Conventional Triangulation** In triangulation [16,91,98], source-pivot and pivot-target bilingual corpora are used to train phrase tables. Then, the source and target phrases are connected via common pivot phrases.

Given a source phrase $s$ and target phrase $t$ of the source-pivot phrase table $T_{SP_s}$ and the pivot-target phrase table $T_{P_tT}$, the phrase translation probability is estimated via common pivot phrases $p$ based on the following feature function.

$$\phi(t|s) = \sum_{p \in (T_{SP_s}) \cap (T_{P_tT})} \phi(p|s)\phi(t|p) \tag{4.3}$$

Previous researches showed the effectiveness of this method when source-target bilingual corpora are unavailable or in limited amounts. Nevertheless, the conventional triangulation does not extract sufficient information because some pivot phrases can contain the same

meaning even when they are not matched to each other. Therefore, I proposed pivoting via similar pivot phrases to extract more informative connections.

**Pivoting via Similar Phrases** The phrases $s$ and $t$ can be connected via similar pivot phrases $p_s$ and $p_t$ as described in Equation 4.4.

$$\phi(t|s) = \sum_{p_s \in P_s, p_t \in P_t} \phi(p_s|s)\phi(t|p_t)\Theta(p_s, p_t) \tag{4.4}$$

Where, $\Theta(p_s, p_t)$ denotes the similarity between $p_s$ and $p_t$ based on the similarity models using the cosine similarity, longest common subsequence, WordNet, and word embeddings described in Section 4.1.1.

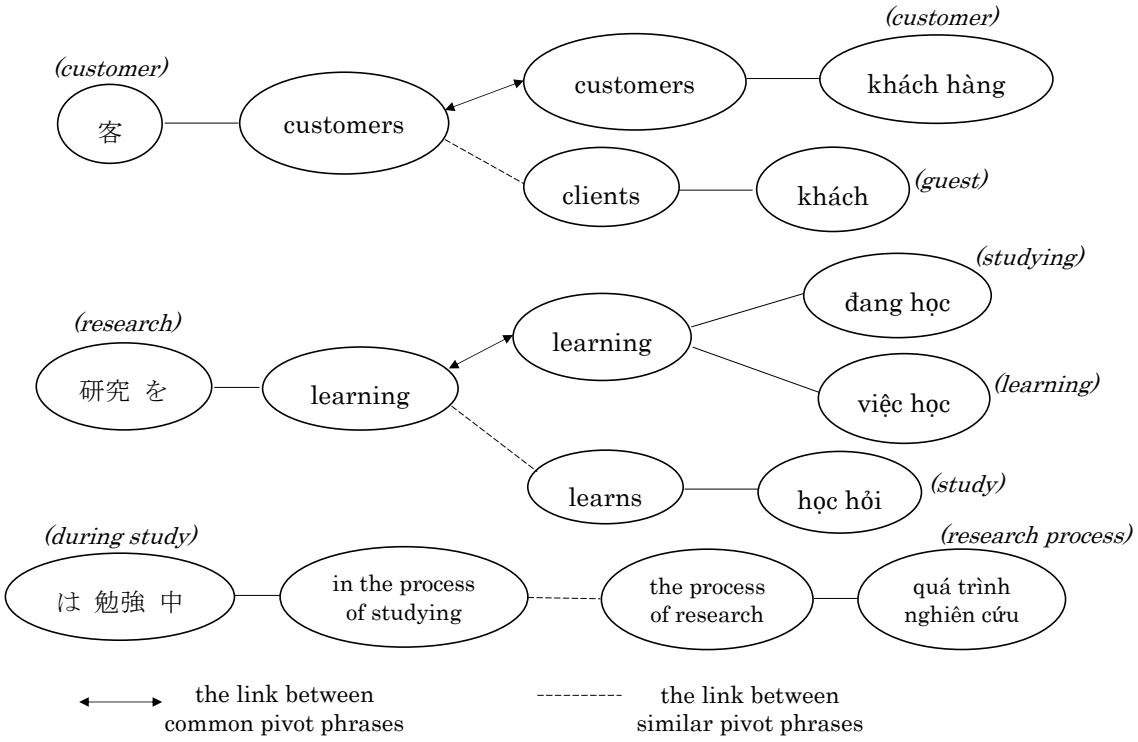



Figure 4.1: **Semantic similarity for pivot translation** Source phrases (Japanese) are connected to target phrases (Vietnamese) via pivot phrases (English). The meanings of Japanese and Vietnamese phrases are translated by the authors in the italic.

I present examples of pivot similarity in Figure 4.1. In the conventional pivot translation,

source and target phrases are connected via common pivot phrases. In my method, some informative connections between source and target phrases can be exploited via similar pivot phrases.

**Pivot Phrase Table** After using the pivot similarity method, I obtain a phrase table of source and target phrases. I then combine the two phrase tables using the conventional pivot method and the similarity pivot method. Since connections were based on the similarity between pivot phrases, there might contain some less reliable phrase pairs. The two phrase tables were combined using the back-off combination,[1] in which only new phrase pairs of the pivot similarity phrase table were added to the common pivot phrase table to create a phrase table for decoding.

For next sections, I will present experiments that applied the proposed models on machine translation for several low-resource languages: Japanese-Vietnamese, Southeast Asian languages (translation from Indonesian, Malay, and Filipino to Vietnamese). The experiments on these languages are novel when there is no prior work of machine translation on those language pairs.

### 4.1.3 Experiments on Japanese-Vietnamese

I conducted experiments on Japanese-Vietnamese, a language pairs that shows various challenges for phrase-based machine translation. First, the languages contain different characteristics in language structures. Second, there is no large bilingual data available, in which about only 100k parallel sentences can be obtained online from several resources like TED talks [9], Bibble [2], and OPUS [82].

**Training Data** I used a small bilingual data extracted from the TED talks [9],[3] and the Bibble data.[4] For pivot translation, I used English as the pivot language, and I used two bilingual corpora of Japanese-English and English-Vietnamese. For Japanese-English, I used the Kyoto corpus [60].[5] The bilingual data of English-Vietnamese includes the VLSP corpus,[6] the English-Vietnamese training data in the IWSLT 2015,[7] and an in-house bilingual corpus used in the system [83] participated in the IWSLT 2015. Table 4.1 describes the bilingual data for Japanese-Vietnamese experiments.

I collected the development and test data sets from several webs such as dongdu.edu.vn, *kaizen.vn*, and *duhoc.daystar.com.vn* including bilingual news and novels, which are described in Table 4.2.

[1] http://www.statmt.org/moses/?n=Advanced.Models#ntoc7
[2] http://homepages.inf.ed.ac.uk/s0787820/bible/
[3] https://wit3.fbk.eu/mt.php?release=2012-02-plain
[4] http://homepages.inf.ed.ac.uk/s0787820/bible/
[5] http://www.phontron.com/kftt/
[6] http://vlsp.hpda.vn:8080/demo/?page=resources
[7] http://workshop2015.iwslt.org/

Table 4.1: **Bilingual corpora for Japanese-Vietnamese pivot translation** (**ja**: Japanese, **en**: English, **vi**: Vietnamese). **Src Words, Trg Words**: number of source, target words; **Src Vocab, Trg Vocab**: the source, target vocabulary; **Src Avg len, Trg Avg len**: the average length of source, target sentences.

| **Languages** | **ja-vi** | **ja-en** | **en-vi** |
|---|---|---|---|
| **Sentence Pairs** | 83,313 | 329,882 | 456,350 |
| **Src Words** | 2,076,083 | 6,085,131 | 8,485,112 |
| **Trg Words** | 2,138,623 | 5,911,486 | 10,843,982 |
| **Src Vocab** | 37,689 | 114,284 | 114,161 |
| **Trg Vocab** | 19,411 | 161,655 | 62,933 |
| **Src Avg len** | 25 | 18 | 19 |
| **Trg Avg len** | 25 | 18 | 24 |

Table 4.2: **Japanese-Vietnamese development and test sets**

| **Data set** | **Dev** | **Test** |
|---|---|---|
| **Sentence Pairs** | 1,200 | 1,266 |
| **Src Words** | 12,955 | 24,332 |
| **Trg Words** | 15,274 | 23,815 |
| **Src Vocab** | 2,798 | 4,052 |
| **Trg Vocab** | 1,686 | 2,487 |
| **Src Avg len** | 11 | 19 |
| **Trg Avg len** | 13 | 19 |

**Training Details** I conducted baseline experiments using the well-known Moses toolkit for phrase-based machine translation [43]. The word alignment was trained using GIZA++ [65], an effective training algorithm for alignment models, with the configuration *grow-diag-final-and*. A 5-gram language model of the target language was trained using KenLM [31]. The KenLM has been shown to be effective in both time and memory costs, which outperformed other well-known packages for language model like: SRILM [78], IRSTLM [24], and BerkeleyLM [38] in terms of speed and memory consumption. Tuning parameters were performed based on the batch MIRA [13], which has been shown to be simple and effective and outperforms other tuning strategies including the traditional MERT approach (the Minimum Error Rate Training [64]). The evaluation was conducted based on the commonly used metrics: BLEU [66].

For pivot translation, I implement the baseline triangulation method [98] using Java. I compare my system with the TmTriangulate[32]. As discussed in [23], the triangulation method can generate a very big phrase table, which contains noisy phrase pairs. In this work, I implement the triangulation method of [98] and created a modification to filter the phrase table. Specifically, for each source phrase, I extract only *n-best* candidate of target phrases.

As described in the method section, in order to compute similarity between pivot

Table 4.3: **Monolingual data sets of Japanese, English, Vietnamese**. **Words**: number of words, **Vocab**: vocabulary, **Avg len**: the average length of sentences

| **Language** | **ja** | **en** | **vi** |
|---|---|---|---|
| **Sentences** | 52,741,702 | 30,301,028 | 16,201,114 |
| **Words** | 775,785,793 | 768,648,884 | 485,087,517 |
| **Vocab** | 4,118,306 | 2,425,337 | 850,650 |
| **Avg len** | 15 | 25 | 30 |

phrases, I used the methods including cosine similarity, longest common subsequence, WordNet, and word embeddings. I trained a word embedding model of English using the English monolingual data (Table 4.3). I used the similarity approaches to score the similarity between English phrases in the two sides of English phrases in the Japanese-English and English-Vietnamese phrase tables.

**Results** Table 4.4 presents the experimental results of pivot translation and the proposed method. I used the linear interpolation to combine the phrase tables of the triangulation baseline with the proposed pivot translation using similarity. Nevertheless, the combination with pivot similarity (5.32 BLEU score) does not improve the pivot baseline model (5.52 BLEU score). This indicates that the pivot similarity phrase table may contain some noisy phrase pairs which hurt the performance. Therefore, I used another technique to combine the phrase tables namely back-off in which the phrase pairs in the pivot similarity table are added to the combined table if they do not exist in the pivot baseline table. As described in Table 4.4, the back-off combination improved the pivot baseline (increased from 5.52 to 5.68 BLEU scores).

Table 4.4: **Japanese-Vietnamese pivot translation results**

| **Model** | **Dev (BLEU)** | **Test (BLEU)** |
|---|---:|---:|
| triangulation | 3.46 | 5.52 |
| triangulation-similarity (interpolation) | 3.0 | 5.32 |
| triangulation-similarity (backoff) | 3.42 | 5.68 |

### 4.1.4 Experiments on Southeast Asian Languages

**Training Data** For training data, I used two resources: TED data [9] and the ALT corpus (Asian Language Treebank Parallel Corpus) [80]. I extracted parallel data for the Southeast Asian language pairs from the TED data. First, I collected the monolingual data of the TED talks from the web[8] of Indonesian, Malay, Filipino, and Vietnamese. Then, I created a multilingual parallel data for Indonesian, Malay, Filipino paired with Vietnamese by using the *talk id* and the *seekvideo id* in the data to extract parallel sentences. For the

[8]https://wit3.fbk.eu/

ALT corpus, this includes 20K multilingual sentences of English, Japanese, Indonesian, Filipino, Malay, Vietnamese, and some other Southeast Asian languages. I divided the ALT corpus for Indonesian, Malay, Filipino, and Vietnamese into three data sets: 16K sentences for training data, 2K sentences for development data, and 2K sentences for test data. The training sets extracted from the ALT corpus were combined with the data aligned from the TED talks to train SMT models as the baseline models. Table 4.5 presents the training, development, and test data sets.

Table 4.5: **Bilingual corpora of Southeast Asian language pairs** (id: Indonesian, ms: Malay, fil: Filipino, vi: Vietnames)

| **Data set** | **Sentences** | **Src Words** | **Trg Words** | **Src Vocab** | **Trg Vocab** | **Src Avg len** | **Trg Avg len** |
|---|---|---|---|---|---|---|---|
| id-vi | | | | | | | |
| Training | 226,239 | 1,932,460 | 2,822,894 | 52,935 | 29,896 | 9 | 12 |
| Dev | 1,982 | 46,518 | 67,630 | 7,847 | 5,075 | 23 | 34 |
| Test | 2,074 | 47,574 | 68,014 | 8,082 | 5,339 | 23 | 33 |
| ms-vi | | | | | | | |
| Training | 33,399 | 504,642 | 731,486 | 29,019 | 18,353 | 15 | 22 |
| Dev | 1,982 | 46,555 | 67,630 | 7,506 | 5,075 | 23 | 34 |
| Test | 2,074 | 48,255 | 68,014 | 7,736 | 5,339 | 23 | 33 |
| fil-vi | | | | | | | |
| Training | 22,875 | 521,681 | 614,650 | 39,955 | 17,458 | 23 | 27 |
| Dev | 1,982 | 57,874 | 67,630 | 9,482 | 5,075 | 29 | 34 |
| Test | 2,073 | 59,496 | 67,934 | 9,594 | 5,335 | 29 | 33 |

The training data of Indonesian-Vietnamese contains 226K parallel sentences; however, the Malay-Vietnamese and Filipino-Vietnamese are in very limited amounts with only 33K sentence pairs (Malay-Vietnamese) and 22K sentence pairs (Filipino-Vietnamese).

For pivot translation experiments, English was used as the pivot language, and I used the same combination of the TED data and the training sets in the ALT corpus for Indonesian, Malay, Filipino, Vietnamese paired with English. These training data sets are described in Table 4.6.

Table 4.6: **Bilingual data for pivot translation of Southeast Asian language pairs** (id: Indonesian, ms: Malay, fil: Filipino, vi: Vietnames, en: English)

| **Data set** | **Sentences** | **Src Words** | **Trg Words** | **Src Vocab** | **Trg Vocab** | **Src Avg len** | **Trg Avg len** |
|---|---|---|---|---|---|---|---|
| id-en | 244,858 | 2,086,659 | 2,413,216 | 55,520 | 57,562 | 9 | 10 |
| ms-en | 31,608 | 502,329 | 559,802 | 29,210 | 32,660 | 16 | 18 |
| fil-en | 21,951 | 523,078 | 469,603 | 40,127 | 30,674 | 24 | 21 |
| vi-en | 377,736 | 4,446,502 | 3,562,696 | 36,661 | 67,325 | 12 | 9 |

**Monolingual Data** For English and Vietnamese monolingual data sets, I employed the monolingual data sets in the Japanese-Vietnamese experiments. For Indonesian, Malay, and Filipino monolingual data sets, I extracted all sentences from Wikipedia, which were used in experiments of building bilingual corpora. I describe the monolingual data in Table 4.7.

Table 4.7: **Monolingual data sets**: id (Indonesian), ms (Malay), fil (Filipino)

| **Language** | **id** | **ms** | **fil** |
|---|---|---|---|
| **Sentences** | 1,478,986 | 596,097 | 682,939 |
| **Words** | 25,525,803 | 10,903,878 | 13,785,021 |
| **Vocab** | 494,688 | 339,906 | 221,637 |
| **Avg len** | 17 | 18 | 20 |

**Training Details** I applied the triangulation baseline and the pivot similarity with the same methods as described in the Japanese-Vietnamese experiments. The *back-off* setting was used to combine the pivot baseline and the pivot similarity phrase tables.

Table 4.8: **Pivot translation results of Southeast Asian language pairs**

| **Model** | **Dev (BLEU)** | **Test (BLEU)** |
|---|---|---|
| direct: id-vi | 29.97 | 30.46 |
| triangulation: id-vi | 24.82 | 33.37 |
| triangulation-similarity: id-vi | 24.96 | 33.50 |
| direct: ms-vi | 30.09 | 32.81 |
| triangulation: ms-vi | 25.43 | 35.01 |
| triangulation-similarity: ms-vi | 25.66 | 35.12 |
| direct: fil-vi | 22.10 | 24.29 |
| triangulation: fil-vi | 18.23 | 25.74 |
| triangulation-similarity: fil-vi | 18.42 | 25.87 |

**Results** Experiments of pivot translation as described in Table 4.8 showed some interesting results.

First, the triangulation in all language pairs significantly improved the direct models including Indonesian-Vietnamese (30.46 BLEU score to 33.37 BLEU score), Malay-Vietnamese (32.81 BLEU score to 35.01 BLEU score), Filipino-Vietnamese (24.29 BLEU score to 25.74 BLEU score).

The second results were the improvement of the similarity pivot for the pivot baseline: Indonesian-Vietnamese (33.37 to 33.50 BLEU), Malay-Vietnamese (35.01 to 35.12 BLEU), Filipino-Vietnamese (25.74 to 25.87 BLEU). The results confirmed the contribution of the proposed similarity pivot method although the improvement was still slight.

## 4.2 Grammatical and Morphological Knowledge for Pivot Translation

The conventional pivot method, which connect source to target phrases via common pivot phrases, lacks some potential connections when pivoting via the surface form of pivot phrases. In this section, I improve the pivot translation method by integrating grammatical and morphological information to connect pivot phrases instead of using only the surface form. Experiments were conducted on several Southeast Asian low-resource language pairs: Indonesian-Vietnamese, Malay-Vietnamese, and Filipino-Vietnamese.

### 4.2.1 Grammatical and Morphological Knowledge

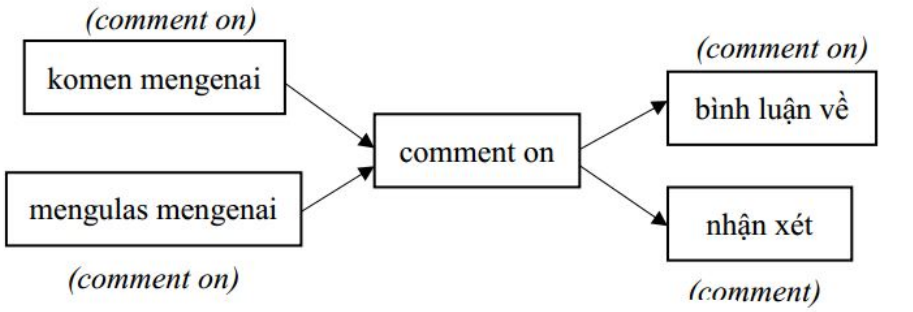
In order to integrate grammatical and morphological information to pivot translation, I trained factored models [42] on the source-pivot and pivot target bilingual corpora. First, the part-of-speech and lemma forms of words in the pivot sides of the bilingual corpora were generate to enrich information for pivoting via pivot phrases. Since English was used for the pivot language, analyzing part-of-speech (POS tagging) and lemma forms were conducted using the Stanford CoreNLP [52]. Then, the bilingual corpora in which words of the pivot sides in the bilingual corpora contain part-of-speech tags and lemma forms were used to train source-pivot and pivot-target phrase tables. After that, source-target phrase pairs were extracted based on common pivot phrases using Equation 4.3.

**Part-Of-Speech Tags in Pivot Translation** The connections via common pivot phrases can be enriched by integrating grammatical information. As shown in Table 4.9, when adding grammatical information (part-of-speech tags), the pivot phrase *commented on* was divided into two cases: *commented|VBD on|IN* and *commented|VBN on|IN*. Due to adding the part-of-speech information, a new connection to the target phrase *"đã bình luận về" (English meaning: commented on)* was employed instead of only one connection to the target phrase as in the surface model.

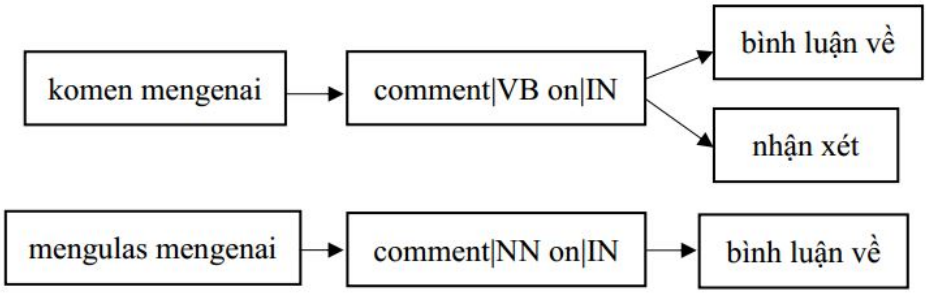
Table 4.9: **Examples of grammatical information for pivot translation** (POS tags: VBD (Verb, past tense), VBN (Verb, past participle), NN (Noun, singular or mass), VB (Verb, base form), IN (Preposition)); *Italic words*: English meaning.

| **Model** | **Source (Malay)** | **Pivot (English)** | **Target (Vietnamese)** |
|---|---|---|---|
| surface | mengulas mengenai *(comment on)* | commented on | bình luận về *(comment on)* |
| POS | mengulas mengenai *(comment on)* | commented\|VBD on\|IN | bình luận về *(comment on)* |
| | mengulas mengenai *(comment on)* | commented\|VBN on\|IN | đã bình luận về *(commented on)* |

Integrating grammatical information also helps to filter connections via pivot phrases.




a) Pivot translation via the surface form




b) Integrating POS factor to pivot translation

Figure 4.2: **Pivoting using syntactic information**

Figure 4.2 describes an example of using part-of-speech information (POS factor) in pivot translation that help to filter connections. For the case of pivoting via surface form of pivot phrases (Figure 4.2a), the two Malay phrases *komen mengenai* and *mengulas mengenai* were both connected to the pivot phrase *comment on*. Nevertheless, when integrating POS factor (Figure 4.2b), the two source phrases were connected to two different pivot phrases (*comment|VB on|IN* and *comment|NN on|IN*). The separated connections help to classify the connections in more detailed, that refine the translation probabilities.

**Lemma Forms in Pivot Translation** As shown in [42], using lemma forms can improve the problem of sparse training data in SMT. I propose connecting phrases via the lemma form instead of the surface form of pivot phrases.

Figure 4.3 describes an example of using lemma forms in pivot translation. Because of the sparse data problem, the Malay word *golongan* cannot be connected to any Vietnamese target word. However, when using the lemma form (Figure 4.3b), because the pivot words *class* and *classes* share the same lemma form (*class*), a new informative connection was generated for the source word *golongan*.

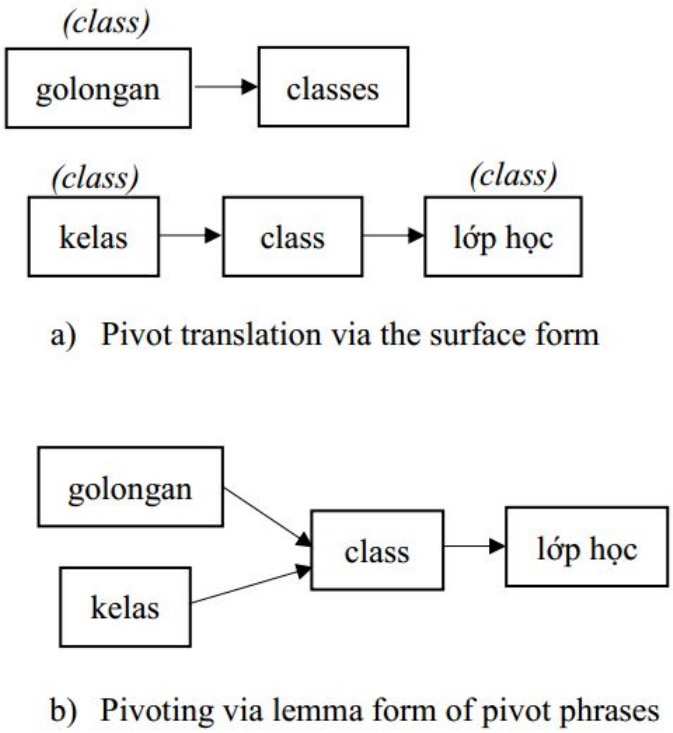



Figure 4.3: **Pivoting using morphological information**; *the italic words* indicate the English meaning

### 4.2.2 Combining Features to Pivot Translation

Because part-of-speech information and lemma forms of pivot phrases were added to improve the baseline pivot method (trained on the surface form of pivot phrases), I introduce a model that combines three following components:

- *base*: the triangulated phrase table based on the surface form of pivot phrases
- *pos*: the triangulated phrase table based on factored training using part-of-speech tags of pivot phrases
- *lem*: the triangulated phrase table based on factored training using lemma forms of pivot phrases

A combined phrase table was generated using linear interpolation [70], in which the probability of a target phrase $t$ given a source phrase $s$ can be computed by Equation 4.5.

$$p(t|s) = \lambda_{base} p_{base}(t|s) + \lambda_{pos} p_{pos}(t|s) + \lambda_{lem} p_{lem}(t|s) \quad (4.5)$$

where $p_{base}(t|s)$, $p_{pos}(t|s)$, and $p_{lem}(t|s)$ stand for the translation probability of the baseline, POS tags, and lemma models, respectively.

The interpolation parameters $\lambda_{base}$, $\lambda_{pos}$, and $\lambda_{lem}$ in which $\lambda_{base} + \lambda_{pos} + \lambda_{lem} = 1$ were computed by the following strategies.

- *tune*: the parameters were tuned using a development data set.
- *weights*: the parameters were set based on the ratio of the BLEU scores when using each model separately for decoding the tuning set.

### 4.2.3 Experiments

**Training Data** I conducted experiments of translation from Indonesian, Malay, and Filipino to Vietnamese, which are Southeast Asian low-resource language pairs. I used English for the pivot language, which is one of the most common language in the world.

Table 4.10: **Southeast Asian bilingual corpora for training factored models**

| Data | Sentences |
|---|---|
| Malay-English | 31,608 |
| Indonesian-English | 244,858 |
| Filipino-English | 21,951 |
| English-Vietnamese | 377,736 |

For training data, I used two resources: TED data [9] and the ALT corpus (Asian Language Treebank Parallel Corpus) [80]. A multilingual parallel data set was collected from the TED talks[9] for training data of Indonesian, Malay, Filipino, English, and Vietnamese. The ALT corpus includes 20K multilingual sentences of English, Japanese, Indonesian, Filipino, Malay, Vietnamese, and some other Southeast Asian languages. I divided the ALT corpus into three sets: training (16K), development (2K) and test (2K). The training part (16K) was combined with the TED data set for training translation models. For monolingual data of Vietnamese to train a language model, I collected articles from the website *http://www.baomoi.com/*, which contains 16M sentences. Table 4.10 presents the training data sets.

**Training Details** Experiments were conducted using the well-known SMT toolkit, Moses [43]. The word alignment was trained using GIZA++ [65] with the configuration *grow-diag-final-and*. A 5-gram language model of the target language was trained using KenLM [31]. For tuning, I used the batch MIRA [13]. BLEU scores [66] were used as the metric for evaluation.

For pivot translation, I implemented the triangulation approach of Wu and Wang, 2007 [98] using Java. As reported in [23], one of the issue of the triangulation approach is the very big size of the induced pivot phrase table. Therefore, I filtered the induced phrase table by a *n-best* strategy, in which $n$ best candidates of target phrases were selected for each source phrase. I chose *n=10* in my experiments to ensure the reliability of phrase pairs. I also compared my system with the TmTriangulate [32], a python implementation of the triangulation method [98].

[9]https://wit3.fbk.eu/

**Results** Experimental results are presented in (Tables 4.11-4.13). There are several findings from the experimental results. First, using additional knowledge of part-of-speech and lemma forms improved the baseline triangulation trained on the surface form of pivot phrases. Second, combining the baseline model (surface form) with the pos (part-of-speech) and lemma models by using the interpolation method improved the baseline model.

Table 4.11: **Results of using POS and lemma forms** (BLEU); **baseline-pos-lemma**: the interpolation model of the baseline (pivot via surface form), pos (pivot via factored models of part-of-speech), and lemma (pivot via factored models of lemma). **weights**, **tune**: the interpolation settings presented in Section 4.2.2.

| **Model** | **Dev** | **Test** |
|---|---|---|
| TmTriangulate | 25.06 | 35.09 |
| baseline | 25.78 | **35.86** |
| pos | 25.82 | 35.99 (+0.13) |
| lemma | 24.93 | 35.62 |
| baseline-pos (weights) | 25.89 | 36.04 (+0.18) |
| baseline-pos (tune) | 25.89 | 36.08 (+0.22) |
| baseline-lemma (weights) | 25.76 | 36.20 (+0.34) |
| baseline-lemma (tune) | 25.79 | 36.12 (+0.26) |
| pos-lemma (weights) | 25.72 | 36.11 (+0.25) |
| pos-lemma (tune) | 25.83 | 36.19 (+0.33) |
| baseline-pos-lemma (weights) | 25.81 | **36.38 (+0.52)** |
| baseline-pos-lemma (tune) | 25.89 | 36.25 (+0.39) |

Table 4.12: **Indonesian-Vietnamese experimental results (BLEU)**

| **Model** | **Dev** | **Test** |
|---|---|---|
| TmTriangulate | 24.60 | 32.83 |
| baseline | 25.51 | **33.83** |
| pos | 25.54 | 33.87 (+0.04) |
| lemma | 24.68 | 32.89 (+0.06) |
| baseline-pos (weights) | 25.65 | 33.91 (+0.08) |
| baseline-pos (tune) | 25.62 | 33.87 (+0.04) |
| baseline-lemma (weights) | 25.38 | 34.07 (+0.24) |
| baseline-lemma (tune) | 25.48 | **34.18 (+0.35)** |
| pos-lemma (weights) | 25.38 | 33.87 (+0.04) |
| pos-lemma (tune) | 25.53 | 34.01 (+0.18) |
| baseline-pos-lemma (weights) | 25.51 | 34.05 (+0.22) |
| baseline-pos-lemma (tune) | 25.64 | 33.94 (+0.11) |

Specifically, combining the baseline model with the pos and lemma models improved from 0.1 to 0.5 BLEU scores for the Malay-Vietnamese experiments (Tables 4.11). For

Table 4.13: **Filipino-Vietnamese experimental results (BLEU)**

| **Model** | **Dev** | **Test** |
|---|---|---|
| TmTriangulate | 17.51 | 25.29 |
| baseline | 18.07 | **26.02** |
| pos | 18.22 | 25.95 |
| lemma | 17.33 | 25.38 |
| baseline-pos (weights) | 18.16 | **26.16 (+0.14)** |
| baseline-pos (tune) | 18.17 | 25.98 |
| baseline-lemma (weights) | 17.96 | 25.90 |
| baseline-lemma (tune) | 18.09 | 25.89 |
| pos-lemma (weights) | 18.00 | 25.93 |
| pos-lemma (tune) | 18.12 | 25.96 |
| baseline-pos-lemma (weights) | 18.16 | 26.00 |
| baseline-pos-lemma (tune) | 18.19 | 26.01 |

Indonesian-Vietnamese, the highest improvement comes from the combination of the baseline and lemma models (+0.35 BLEU score). Unexpectedly, for the Filipino-Vietnamese experiments, the combination of all models: baseline, pos, and lemma does not show any improvement. Meanwhile, the baseline-pos combination model slightly improved the baseline model (+0.14 BLEU).

For the comparison of my system with the TmTriangulate, my system with the filtering strategy showed much better performance.

Table 4.14: **Input factored phrase tables** (**Src-Pvt**, **Pvt-Trg**: source-pivot, pivot-target phrase pairs, **Common**: common pivot phrases)

| **Model** | **Src-Pvt** | **Pvt-Trg** | **Common** |
|---|---|---|---|
| **Malay-Vietnamese** | | | |
| baseline | 83,914 | 395,983 | 33,573 |
| pos | 89,424 | 420,790 | 36,404 |
| lemma | 86,359 | 405,051 | 29,585 |
| **Indonesian-Vietnamese** | | | |
| baseline | 239,172 | 395,983 | 63,596 |
| pos | 250,286 | 420,790 | 68,404 |
| lemma | 240,426 | 405,051 | 54,808 |
| **Filipino-Vietnamese** | | | |
| baseline | 79,671 | 395,637 | 26,026 |
| pos | 82,151 | 420,441 | 27,517 |
| lemma | 80,265 | 404,712 | 23,172 |

### 4.2.4 Analysis

In this subsection, I analyze the results on various aspects: the effect of adding grammatical and morphological information in improving pivot translation; out-of-vocabulary ratio, conducting statistical significance tests, conducting Spearman rank correlation on several metrics, conducting Wilcoxon signed rank test to verify the improvement of the proposed method.

Table 4.15: **Extracted phrase pairs by triangulation** (**Pairs**, **Src**, **Trg**: source-target phrase pairs, source phrases, target phrases, respectively)

| Model | Pairs | Src | Trg |
|---|---|---|---|
| **Malay-Vietnamese** | | | |
| baseline | 94,776 | 16,936 | 24,868 |
| pos | 93,972 | 16,939 | 24,858 |
| lemma | 101,904 | 17,404 | 26,072 |
| baseline-pos | 112,529 | 17,587 | 26,475 |
| baseline-lemma | 125,903 | 17,942 | 27,559 |
| pos-lemma | 116,529 | 17,406 | 26,553 |
| baseline-pos-lemma | 131,983 | 17,942 | 27,742 |
| **Indonesian-Vietnamese** | | | |
| baseline | 128,206 | 20,492 | 27,511 |
| pos | 126756 | 20449 | 27474 |
| lemma | 134,594 | 20,931 | 28,562 |
| baseline-pos | 144,375 | 20,760 | 28,865 |
| baseline-lemma | 161,985 | 21,148 | 29,917 |
| pos-lemma | 155,648 | 20,933 | 29,103 |
| baseline-pos-lemma | 168,905 | 21,148 | 30,119 |
| **Filipino-Vietnamese** | | | |
| baseline | 133,003 | 22,115 | 23,258 |
| pos | 131,015 | 22,046 | 23,216 |
| lemma | 143,429 | 22,635 | 24,317 |
| baseline-pos | 150,599 | 22,419 | 24,557 |
| baseline-lemma | 173,227 | 22,901 | 25,661 |
| pos-lemma | 165,930 | 22,646 | 24,896 |
| baseline-pos-lemma | 180,941 | 22,907 | 25,862 |

**1. The Effect of Factored Models** I evaluated the effect of factored models to pivot translation. The factored models enrich linguistic knowledge of part-of-speech and lemma. When using factored models for training source-pivot and pivot-target phrase tables, the numbers of input phrase pairs were increased, which lead to increase the number of common pivot phrases and the extracted phrase pairs in the induced phrase table output. Table 4.14 illustrates the statistics of input phrase tables using factored models.

Because of adding knowledge of part-of-speech and lemma, the number of phrase pairs in the combined phrase tables was significantly increased, which cover larger ratio of vocabulary. This leads to an improvement on the baseline model. As presented in Table 4.15, the numbers of phrase pairs in the baseline models were significantly increased when adding part-of-speech and lemma information: Malay-Vietnamese (+37,207 phrase pairs, or 39.26%), Indonesian-Vietnamese (+40,699 phrase pairs, or 31.75%), and Filipino-Vietnamese (+47,938 phrase pairs, or 36.04%).

**2. Out-Of-Vocabulary Ratio** I analyzed the out-of-vocabulary (OOV) ratio when using models for decoding on test sets. Using part-of-speech and lemma information enriched linguistic knowledge for the baseline model trained on only the surface form of pivot phrases. As described in Table 4.16, the OOV ratios of the baseline models were reduced such as Malay-Vietnamese (-1.65%), Indonesian-Vietnamese (-0.77%), and Filipino-Vietnamese (-0.79%).

Table 4.16: **Out-Of-Vocabulary ratio**

| **Model** | **OOV phrases** | **OOV ratio (%)** |
|---|---|---|
| baseline (ms-vi) | 1,510 | 20.42 |
| pos | 1,481 | 20.05 |
| lemma | 1,432 | 19.35 |
| baseline-pos-lemma (weights) | 1,390 | 18.77 |
| baseline-pos-lemma (tune) | 1,387 | 18.77 |
| baseline (id-vi) | 1,180 | 15.40 |
| pos | 1,186 | 15.49 |
| lemma | 1,136 | 14.82 |
| baseline-pos-lemma (weights) | 1,124 | 14.63 |
| baseline-pos-lemma (tune) | 1,159 | 15.12 |
| baseline (fil-vi) | 2,628 | 30.24 |
| pos | 2,643 | 30.23 |
| lemma | 2,609 | 29.74 |
| baseline-pos-lemma (weights) | 2,567 | 29.45 |
| baseline-pos-lemma (tune) | 2,605 | 29.46 |

**3. More Rubust BLEU** Since the improvement of the proposed method is still limited in some cases, I conducted statistical significance tests to verify the improvement. The significance tests were based on a method called bootstrap resampling in machine translation [39].

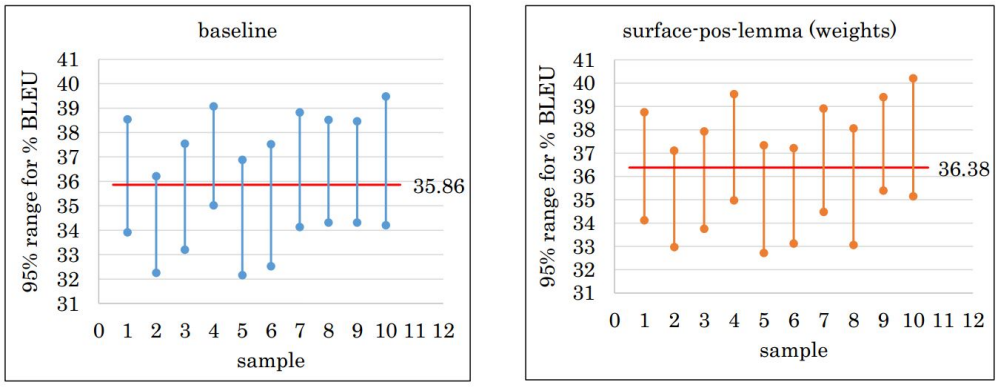
*Broad sample.* First of all, a set of $n$ test sets of $m$ sentences was extracted from an original test set using a technique called *broad sample*. For instance, given an original test set of 30,000 sentences, this original test set is divided into 10 test sets, in which each test set contains 3000 sentences. The sample $i^{th}$ includes the sentences $i, i + 10, ..., i +$

29990. This is to avoid the issue that translation quality of neighbouring sentences will be highly correlated with one another. This is because factors that influence translation quality tend to be clustered, e.g. the language style, topic etc will usually remain constant over neighbouring sentences. Therefore, estimates of the translation quality over these sentences will be inaccurate and biased. *Broad sampling* avoids this issue by creating test sets containing sentences from different parts of the corpus.

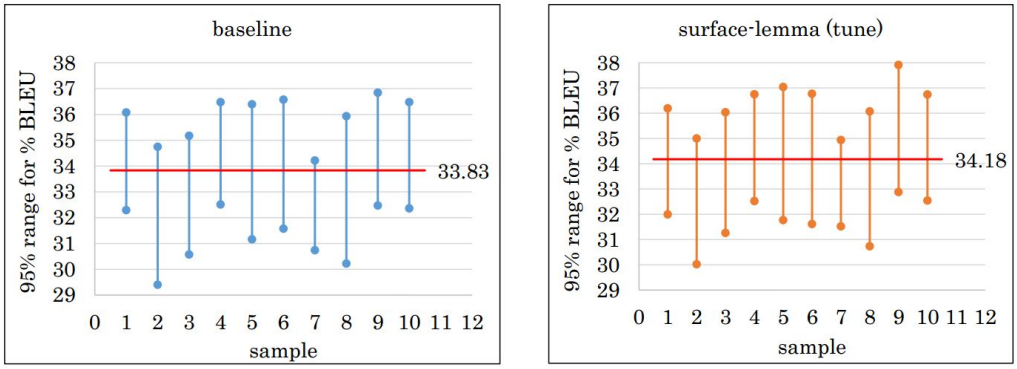
*Bootstrap resampling.* Given the BLEU scores reported in Tables 4.11-4.13 as the true BLEU scores, we want to compute with a confidence or find intervals for the true BLEU score. Specifically, given $m$ as the true BLEU score, we find the interval $[m - d, m + d]$ with a confidence $q$ (or $p - level = 1 - q$) (typically, $q = 0.95, (p - level = 0.5)$, a 95% confidence interval).

I used the original test sets of 2,000 sentences to create 10 test sets with size of 200 sentences using the broad sample technique. Then, for each of 10 test sets, a large number of 100 "virtual" test sets was generated by randomly drawing with replacement. For each of the "virtual" tets sets, I calculated the BLEU score using the baseline and proposed systems; then, drop top 3 and bottom 2, leaving us with 95% of BLEU scores in an interval $[a, b]$. By the law of large numbers, then the score $[a, b]$ approaches the 95% confidence interval. That is, we can say that the true BLEU scores lies in the range $[a, b]$ with a probability of 0.95.

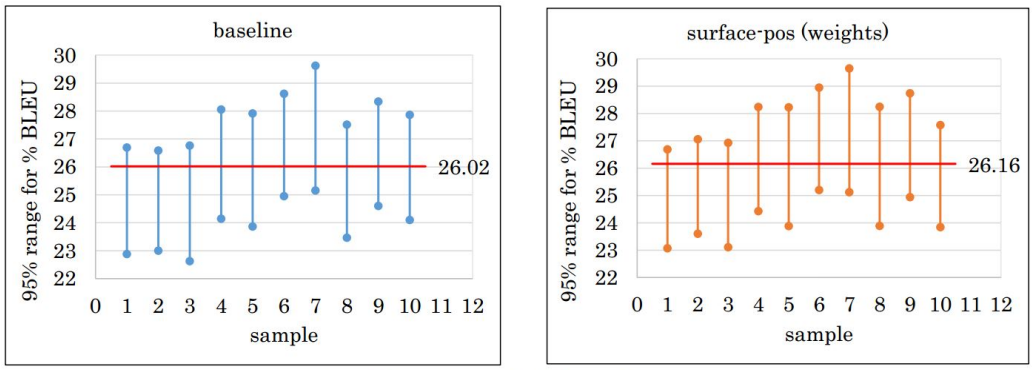
The confidence intervals are illustrated in Figure 4.4.




a) Confidence intervals Malay-Vietnamese




b) Confidence intervals Indonesian-Vietnamese




c) Confidence intervals Filipino-Vietnamese

Figure 4.4: **Confidence intervals**; bootstrapped 95% confidence intervals of true BLEU on 10 broad samples of 200 sentences: assuming the 2,000 sentence BLEU as true score (the red line), no mistake was made

*Paired bootstrap resampling.* For each sample extracted from the *broad sample* step, a large number of "virtual" test sets of size $n$ sentences was randomly drawn with replacement. For each "virtual" test set, two translation systems that we want to compare were used to translate the test set and compare the performance. If one system outperforms the other system 95% of the time, we draw the conclusion that the system is better with 95% statistical significance.

*Results.* For generating the broad samples and random test sets, I used the same test sets with the size of 2000 sentences as described in Section 4.2.3 for the original test set. Two sample sizes of 200 and 400 sentences were used to extract 10 and 5 test sets respectively using the broad sample technique. Then, for each test set, a large number of "virtual" 100 test sets were randomly drawn with replacement.

Table 4.17: **Results of statistical significance tests**

| **System Comparison** | **BLEU** | **Size 200** | **Size 400** |
|---|---|---|---|
| ms-vi: baseline-pos-lemma (weights) better than baseline | 0.52% | 40% | 60% |
| id-vi: baseline-lemma (tune) better than baseline | 0.35% | 40% | 80% |
| fil-vi: baseline-pos (weights) better than baseline | 0.14% | 20% | 0% |

Experimental results are presented in Table 4.17. For Indonesian-Vietnamese, although the improvement on the original test set was still limited (0.35%), for 80% of samples (size 400) and 40% of samples (size 200) we draw the conclusion the proposed system is better than the baseline system with at least 95% statistical significance. For Filipino-Vietnamese, for 20% of samples (2/10 samples size 200) we draw the conclusion the proposed system is better than the baseline system with 95% statistical significance. For all cases, no wrong conclusion was drawn, in which the proposed system does not improve the baseline system.

**4. Spearman Rank Correlation** .

*Evaluation on different metrics.* I used different metrics to evaluate the performance of the systems rather than BLEU to have a better evaluation on many aspects. I used two other well-known metrics in machine translation: TER [75] and METEOR [20]. Experimental results are presented in Table 4.18.

*Spearman rank correlation.* In this analysis, I would like to examine the correlation between the performance of the baseline and the combined systems on different metrics. First, for each language pair, each of the systems in Table 4.18 was ranked by metrics, which created three ranked list in BLEU, TER, and METEOR (presented in Table 4.19). Then, I computed the Spearman rank correlation between pairs of metrics. The aim with Spearman ranked correlation is to show that there is a high degree of overlap between 2 ranked-lists.

As presented in Table 4.20, the high correlation of BLEU-TER and BLEU-METEOR in Indonesian-Vietnamese and Malay-Vietnamese showed that the combined system improved the baseline system with the high correlation not only in BLEU but in the other metrics. Nevertheless, for Filipino-Vietnamese, it showed a lower correlation when the combined model outperformed the baseline in only one setting (baseline-pos (weights)) according to BLEU; however, by the combined model outperformed the baseline in all settings on METEOR. Regardless of the metrics used, the performance of the systems remains consistent. Especially, the baseline-pos-lemma (weights) and baseline-pos-lemma (tune) models are ranked first and second, which shows that my method works well regardless of the metric.

**5. Wilcoxon Signed-Rank Test** In order to measure statistical significance of the performance of the proposed systems compared to the baseline, I used Wilcoxon signed ranks test [95], a robust non-parametric test for statistical comparisons of classifiers [19], and does not make any normality assumption about the data.

Let $H_0$ be the null hypothesis, that the proposed systems do not improve the baseline (i.e. the baseline and the proposed systems achieve the same performance), and $H_a$ be the alternative hypothesis, that the proposed systems outperform the baseline.

Given the original test sets (2,000 sentences), a set of $N$ test sets was generated using the broad sample technique (in my experiments, I used $N = 5$ and $N = 10$ with the size of 400 and 200 sentences, respectively). For the proposed systems, I chose the system that achieved the best performance for each language pair: *baseline-post-lemma (weights)* for Malay-Vietnamese, *baseline-lemma (tune)* for Indonesian-Vietnamese, and *baseline-pos (weights)* for Filipino-Vietnamese.

Let $R^+$ be the sum of ranks for the data set on which the proposed translation system outperformed the baseline system, and $R^-$ the sum of ranks for the opposite; let $d_i$ be the difference between the performance on the $i - th$ data set of the proposed method and the baseline. The statistics $z$ is distributed approximately normally.

Table 4.18: **Experimental results on different metrics: BLEU, TER, METEOR**

| Model | Dev (BLEU) | BLEU | TER | METEOR |
|---|---|---|---|---|
| baseline (ms-vi) | 25.78 | 35.86 | 0.4849 | 0.3298 |
| pos | 25.82 | 35.99 (+0.13) | 0.4823 | 0.3299 |
| lemma | 24.93 | 35.62 | 0.4891 | 0.3273 |
| baseline-pos (weights) | **25.89** | 36.04 (+0.18) | 0.4822 | 0.3316 |
| baseline-pos (tune) | **25.89** | 36.08 (+0.22) | 0.4822 | 0.3313 |
| baseline-lemma (weights) | 25.76 | 36.20 (+0.34) | 0.4826 | 0.3312 |
| baseline-lemma (tune) | 25.79 | 36.12 (+0.26) | 0.4837 | 0.3317 |
| pos-lemma (weights) | 25.72 | 36.11 (+0.25) | 0.4839 | 0.3304 |
| pos-lemma (tune) | 25.83 | 36.19 (+0.33) | 0.4821 | 0.3311 |
| baseline-pos-lemma (weights) | 25.81 | **36.38 (+0.52)** | **0.4805** | 0.3323 |
| baseline-pos-lemma (tune) | **25.89** | 36.25 (+0.39) | 0.4812 | **0.3324** |
| baseline (id-vi) | 25.51 | 33.83 | 0.5107 | 0.3209 |
| pos | 25.54 | 33.87 (+0.04) | 0.5101 | 0.3214 |
| lemma | 24.68 | 32.89 | 0.5215 | 0.3178 |
| baseline-pos (weights) | **25.65** | 33.91 (+0.08) | 0.5114 | 0.3219 |
| baseline-pos (tune) | 25.62 | 33.87 (+0.04) | 0.5107 | 0.3216 |
| baseline-lemma (weights) | 25.38 | 34.07 (+0.24) | 0.5093 | 0.3220 |
| baseline-lemma (tune) | 25.48 | **34.18 (+0.35)** | **0.5082** | 0.3220 |
| pos-lemma (weights) | 25.38 | 33.87 (+0.04) | 0.5106 | 0.3215 |
| pos-lemma (tune) | 25.53 | 34.01 (+0.18) | 0.5092 | 0.3217 |
| baseline-pos-lemma (weights) | 25.51 | 34.05 (+0.22) | 0.5094 | **0.3223** |
| baseline-pos-lemma (tune) | 25.64 | 33.94 (+0.11) | 0.5102 | 0.3217 |
| baseline (fil-vi) | 18.07 | 26.02 | 0.5948 | 0.2811 |
| pos | **18.22** | 25.95 | 0.5938 | 0.2818 |
| lemma | 17.33 | 25.38 | 0.6017 | 0.2776 |
| baseline-pos (weights) | 18.16 | **26.16 (+0.14)** | **0.5917** | 0.2824 |
| baseline-pos (tune) | 18.17 | 25.98 | 0.5959 | 0.2825 |
| baseline-lemma (weights) | 17.96 | 25.90 | 0.5954 | 0.2814 |
| baseline-lemma (tune) | 18.09 | 25.89 | 0.5979 | 0.2822 |
| pos-lemma (weights) | 18.00 | 25.93 | 0.5936 | 0.2815 |
| pos-lemma (tune) | 18.12 | 25.96 | 0.5952 | 0.2822 |
| baseline-pos-lemma (weights) | 18.16 | 26.00 | 0.5953 | 0.2827 |
| baseline-pos-lemma (tune) | 18.19 | 26.01 | 0.5949 | **0.2829** |

Table 4.19: **Ranks on different metrics**

| No. | Model | BLEU | TER | METEOR |
|---|---|---|---|---|
| | **Malay-Vietnamese** | | | |
| 1 | baseline | 10 | 10 | 10 |
| 2 | pos | 9 | 6 | 9 |
| 3 | lemma | 11 | 11 | 11 |
| 4 | baseline-pos (weights) | 8 | 4.5 | 4 |
| 5 | baseline-pos (tune) | 7 | 4.5 | 5 |
| 6 | baseline-lemma (weights) | 3 | 7 | 6 |
| 7 | baseline-lemma (tune) | 5 | 8 | 3 |
| 8 | pos-lemma (weights) | 6 | 9 | 8 |
| 9 | pos-lemma (tune) | 4 | 3 | 7 |
| 10 | baseline-pos-lemma (weights) | 1 | 1 | 2 |
| 11 | baseline-pos-lemma (tune) | 2 | 2 | 1 |
| | **Indonesian-Vietnamese** | | | |
| 1 | baseline | 10 | 8.5 | 10 |
| 2 | pos | 8 | 5 | 9 |
| 3 | lemma | 11 | 11 | 11 |
| 4 | baseline-pos (weights) | 6 | 10 | 4 |
| 5 | baseline-pos (tune) | 8 | 8.5 | 7 |
| 6 | baseline-lemma (weights) | 2 | 3 | 2.5 |
| 7 | baseline-lemma (tune) | 1 | 1 | 2.5 |
| 8 | pos-lemma (weights) | 8 | 7 | 8 |
| 9 | pos-lemma (tune) | 4 | 2 | 5.5 |
| 10 | baseline-pos-lemma (weights) | 3 | 4 | 1 |
| 11 | baseline-pos-lemma (tune) | 5 | 6 | 5.5 |
| | **Filipino-Vietnamese** | | | |
| 1 | baseline | 2 | 4 | 10 |
| 2 | pos | 7 | 3 | 7 |
| 3 | lemma | 11 | 11 | 11 |
| 4 | baseline-pos (weights) | 1 | 1 | 4 |
| 5 | baseline-pos (tune) | 5 | 9 | 3 |
| 6 | baseline-lemma (weights) | 9 | 8 | 9 |
| 7 | baseline-lemma (tune) | 10 | 0 | 5.5 |
| 8 | pos-lemma (weights) | 8 | 2 | 8 |
| 9 | pos-lemma (tune) | 6 | 6 | 5.5 |
| 10 | baseline-pos-lemma (weights) | 4 | 7 | 2 |
| 11 | baseline-pos-lemma (tune) | 3 | 5 | 1 |

Table 4.20: **Spearman rank correlation between metrics**

| Language | BLEU-TER | BLEU-METEOR | TER-METEOR |
|---|---|---|---|
| **ms-vi** | 0.7153 | 0.7818 | 0.7517 |
| **id-vi** | 0.8368 | 0.9309 | 0.6270 |
| **fil-vi** | 0.6091 | 0.5194 | 0.1093 |

$$R^+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i) \tag{4.6}$$

$$R^- = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i) \tag{4.7}$$

$T = min(R^+, R^-)$

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}} \tag{4.8}$$

With $\alpha$=0.05, the null-hypothesis can be rejected if $z$ is smaller than -1.96.

Tables 4.21-4.26 present the results on Wilcoxon signed rank tests. For all cases, the null hypothesis is rejected, which confirms the improvement of the proposed systems. Especially, the null hypothesis is rejected for Filipino-Vietnamese although the improvement is small on the true BLEU scores (+0.14, Table 4.13). We can conclude that the difference between the baseline and proposed systems is statistically significant.

Table 4.21: **Wilcoxon on Malay-Vietnamese (BLEU)**

| **no.** | **model** | **baseline** | **proposed** | **difference** | **abs** | **rank** |
|---|---|---|---|---|---|---|
| 1 | set-1 | 35.55 | 35.90 | +0.35 | 0.35 | 1 |
| 2 | set-2 | 35.26 | 35.85 | +0.59 | 0.59 | 3 |
| 3 | set-3 | 35.40 | 35.80 | +0.4 | 0.4 | 2 |
| 4 | set-4 | 36.70 | 37.30 | +0.6 | 0.6 | 4 |
| 5 | set-5 | 35.76 | 36.41 | +0.65 | 0.65 | 5 |
| | $R^+$=15, $R^-$=0, $T$=0, $N$=5, $z$= -2.0226 | (rejected) | | | | |

Table 4.22: **Wilcoxon on Indonesian-Vietnamese (BLEU)**

| **no.** | **model** | **baseline** | **proposed** | **difference** | **abs** | **rank** |
|---|---|---|---|---|---|---|
| 1 | set-1 | 33.97 | 34.05 | +0.08 | 0.08 | 1 |
| 2 | set-2 | 32.14 | 32.74 | +0.6 | 0.6 | 5 |
| 3 | set-3 | 33.21 | 33.58 | +0.37 | 0.37 | 2 |
| 4 | set-4 | 34.48 | 34.90 | +0.42 | 0.42 | 3 |
| 5 | set-5 | 34.03 | 34.49 | +0.46 | 0.46 | 4 |
| | $R^+$=15, $R^-$=0, $T$=0, $N$=5, $z$= -2.0226 | (rejected) | | | | |

Table 4.23: **Wilcoxon on Filipino-Vietnamese (BLEU)**

| **no.** | **model** | **baseline** | **proposed** | **difference** | **abs** | **rank** |
|---|---|---|---|---|---|---|
| 1 | set-1 | 25.68 | 25.82 | +0.14 | 0.14 | 2 |
| 2 | set-2 | 26.16 | 26.42 | +0.26 | 0.26 | 4 |
| 3 | set-3 | 25.28 | 25.62 | +0.34 | 0.34 | 5 |
| 4 | set-4 | 26.34 | 26.52 | +0.18 | 0.18 | 3 |
| 5 | set-5 | 26.07 | 26.11 | +0.04 | 0.04 | 1 |
| | $R^+$=15, $R^-$=0, $T$=0, $N$=5, $z$= -2.0226 | (rejected) | | | | |

Table 4.24: **Wilcoxon on Malay-Vietnamese (BLEU)**

| **no.** | **model** | **baseline** | **proposed** | **difference** | **abs** | **rank** |
|---|---|---|---|---|---|---|
| 1 | set-1 | 36.11 | 36.44 | +0.33 | 0.33 | 3.5 |
| 2 | set-2 | 34.27 | 35.16 | +0.89 | 0.89 | 9 |
| 3 | set-3 | 35.30 | 35.87 | +0.57 | 0.57 | 7 |
| 4 | set-4 | 36.96 | 37.17 | +0.21 | 0.21 | 1 |
| 5 | set-5 | 34.38 | 34.92 | +0.54 | 0.54 | 6 |
| 6 | set-6 | 34.98 | 35.36 | +0.38 | 0.38 | 5 |
| 7 | set-7 | 36.21 | 36.54 | +0.33 | 0.33 | 3.5 |
| 8 | set-8 | 35.50 | 35.73 | +0.23 | 0.23 | 2 |
| 9 | set-9 | 36.39 | 37.43 | +1.04 | 1.04 | 10 |
| 10 | set-10 | 37.09 | 37.85 | +0.76 | 0.76 | 8 |
| | $R^+$=55, $R^-$=0, $T$=0, $N$=10, $z$= -2.80306 | (rejected) | | | | |

Table 4.25: **Wilcoxon on Indonesian-Vietnamese (BLEU)**

| **no.** | **model** | **baseline** | **proposed** | **difference** | **abs** | **rank** |
|---|---|---|---|---|---|---|
| 1 | set-1 | 34.0 | 33.97 | -0.03 | 0.03 | 1 |
| 2 | set-2 | 31.91 | 32.42 | +0.51 | 0.51 | 7 |
| 3 | set-3 | 33.15 | 33.69 | +0.54 | 0.54 | 8 |
| 4 | set-4 | 34.49 | 34.62 | +0.13 | 0.13 | 2 |
| 5 | set-5 | 33.7 | 34.19 | +0.49 | 0.49 | 6 |
| 6 | set-6 | 33.94 | 34.13 | +0.19 | 0.19 | 4 |
| 7 | set-7 | 32.38 | 33.07 | +0.69 | 0.69 | 9 |
| 8 | set-8 | 33.27 | 33.46 | +0.19 | 0.19 | 3 |
| 9 | set-9 | 34.47 | 35.18 | +0.71 | 0.71 | 10 |
| 10 | set-10 | 34.35 | 34.77 | +0.42 | 0.42 | 5 |
| | $R^+$=54, $R^-$=1, $T$=1, $N$=10, $z$= -2.70113 | (rejected) | | | | |

Table 4.26: **Wilcoxon on Filipino-Vietnamese (BLEU)**

| **no.** | **model** | **baseline** | **proposed** | **difference** | **abs** | **rank** |
|---|---|---|---|---|---|---|
| 1 | set-1 | 24.55 | 24.54 | -0.01 | 0.01 | 1 |
| 2 | set-2 | 24.71 | 25.37 | +0.66 | 0.66 | 10 |
| 3 | set-3 | 24.86 | 25.07 | +0.21 | 0.21 | 5 |
| 4 | set-4 | 25.95 | 26.04 | +0.09 | 0.09 | 2 |
| 5 | set-5 | 26.04 | 26.33 | +0.29 | 0.29 | 7.5 |
| 6 | set-6 | 26.82 | 27.11 | +0.29 | 0.29 | 7.5 |
| 7 | set-7 | 27.56 | 27.42 | -0.14 | 0.14 | 3 |
| 8 | set-8 | 25.7 | 26.16 | +0.46 | 0.46 | 9 |
| 9 | set-9 | 26.75 | 27.03 | +0.28 | 0.28 | 6 |
| 10 | set-10 | 26.09 | 25.89 | -0.2 | 0.2 | 4 |
| | $R^+$=47, $R^-$=8, $T$=8, $N$=10, $z$= -1.98762 | (rejected) | | | | |

**6. Sample Translations** I describe examples of using part-of-speech and lemma information in improving pivot translation in Tables 4.27-4.29. Some phrases that were not translated by the baseline model can be translated by the proposed models using part-of-speech and lemma information.

Table 4.27: **Sample translations: improving pivot translation by using POS and lemma factors** ; **baseline, lemma, pos**: the translation generated by the baseline, lemma, pos models, respectively; the *italic phrases* indicate the phrases that were not translated by the baseline model; the **bold phrases** indicate the translation by the pos and lemma models.

| **Setup** | **Example** |
|---|---|
| **input** | FDA berkata ia sedang mengkaji semula keputusan itu , yang mempunyai tiga puluh hari untuk mematuhinya . |
| **baseline** | FDA cho biết họ đang xem xét lại quyết định này , những người đã ba mươi ngày để *mematuhinya* . |
| **lemma** | FDA cho biết nó đang xem xét lại các quyết định , với ba mươi ngày để **tuân theo** . |
| **reference** | FDA cho biết đang xem xét lại phán quyết này , và có ba mươi ngày để tuân theo . |
| **meaning** | the FDA says it is reviewing the ruling , which it has thirty days to comply with . |
| **input** | beliau dijangka hadir di mahkamah juvana minggu depan . |
| **baseline** | ông ta dự kiến sẽ có mặt tại tòa án *juvana* vào tuần tới . |
| **pos** | ông ta dự kiến sẽ có mặt tại tòa án **thiếu niên** vào tuần tới . |
| **reference** | cậu bé dự kiến sẽ xuất hiện tại tòa án vị thành niên vào tuần tới . |
| **meaning** | he is expected to appear in juvenile court next week . |

Table 4.28: **Sample translation: Indonesian-Vietnamese**

| **Setup** | **Example** |
|---|---|
| **input** | menurut estimasi serikat , antara tahun 2006 dan 2007 hampir 250 penambang tewas dalam kecelakaan . |
| **baseline** | theo *estimasi* công đoàn , giữa năm 2006 và năm 2007 gần 250 thợ mỏ đã thiệt mạng trong vụ tai nạn . |
| **lemma** | theo **ước tính** của công đoàn , giữa năm 2006 và năm 2007 gần 250 thợ mỏ đã thiệt mạng trong vụ tai nạn . |
| **reference** | theo ước tính của công đoàn , giữa năm 2006 và 2007 gần 250 thợ mỏ đã thiệt mạng trong các vụ tai nạn . |
| **meaning** | according to union estimations , between 2006 and 2007 nearly 250 miners died in accidents . |

Table 4.29: **Sample translation: Filipino-Vietnamese**

| **Setup** | **Example** |
|---|---|
| **input** | ang linya sa pagitan ng York at Leeds ay isinara ng ilang oras , na nagpaantala sa ibang serbisyo . |
| **baseline** | tuyến đường giữa của York và Leeds đã bị đóng cửa trong nhiều giờ , *nagpaantala* khác của dịch vụ . |
| **pos** | theo đường giữa của York và Leeds đã bị đóng cửa trong nhiều giờ , **trì hoãn** của các dịch vụ khác . |
| **reference** | tuyến đường giữa York và Leeds bị chặn trong nhiều giờ , làm đình trệ nhiều dịch vụ khác . |
| **meaning** | the line between York and Leeds was closed for several hours , delaying other services . |

## 4.3 Pivot Languages

In previous sections, only one language is used for pivot, and English is typically used because of this common language. In this section, I investigate several other configurations related to pivot languages. First, I investigate whether other languages can be used effectively for pivot rather than English. Second, I experiment a technique called rectangulation, which translates a source language to a target language via two pivot languages rather than one pivot language. Specifically, a pivot language (*pivot1*) was used as pivot to translate from the source language to another pivot language (*pivot2*); then, the *pivot2* was used to bridge the translation from the source to the target language.

### 4.3.1 Using Other Languages for Pivot

**Setup** I conducted experiments for translation from Indonesian, Malay, and Filipino to Vietnamese using the Asian Language Treebank corpus. The corpus, which contains 20,084 multilingual sentences, was divided into three sets: 18k sentences for training, 1k sentences for tuning, and 1k sentences for test sets. I used the other pivot languages for the triangulation rather than English. For instance, Filipino and Malay were used for pivot languages to translate from Indonesian to Vietnamese, etc.

**Results** Experimental results are presented in Table 4.30. Using other languages for pivot still obtained reasonable results compared with using English, and even higher performance in some cases. For instance, when Malay was used for pivot language to translate from Indonesian to Vietnamese, a better performance was achieved compared with using English for pivot (24.44 vs. 23.53 BLEU points). For the other cases, using English for pivot obtained the highest performance (18.54 BLEU for Filipino-Vietnamese and 26.33 BLEU for Malay-Vietnamese). Finally, I combined three languages for pivot translation to investigate the performance. The combination setting obtained much improvement rather than using one pivot language.

Table 4.30: **Using other languages for pivot**; For each translation, two other languages were used for pivot rather than English only; the three pivot languages were combined using interpolation in the previous section (the last lines), similar to Equation 4.5.

| **Pivot (id-vi)** | **Dev** | **Test** | **Pivot (fil-vi)** | **Dev** | **Test** | **Pivot (ms-vi)** | **Dev** | **Test** |
|---|---|---|---|---|---|---|---|---|
| en | 19.46 | 23.53 | en | 13.12 | **18.54** | en | 21.56 | **26.33** |
| fil | 17.91 | 20.73 | id | **13.42** | 18.31 | fil | 20.18 | 23.01 |
| ms | **19.92** | **24.44** | ms | 13.30 | 18.35 | id | **22.31** | 25.14 |
| | | | | | | | | |
| en-fil-ms | 20.38 | 24.89 | en-id-ms | 14.15 | 19.21 | en-fil-id | 22.76 | 27.30 |

### 4.3.2 Rectangulation for Phrase Pivot Translation

I conducted experiments based on a method called rectangulation, using two languages for pivot. Specifically, the translation from a source language to a target language can be bridged via two pivot languages. First, one pivot language (*pivot1*) was used to translate from the source language to the second pivot language (*pivot2*). Then, the *pivot2* was used as pivot to translate from the source language to the target language. This can be applicable when there is no pivot language in which bilingual corpora between the source and the target paired with the pivot language exist. Therefore, we need to use one more intermediate language for pivot.

**Results** Using the same data sets as described in Section 4.3.1, experimental results are showed in Table 4.31. In most cases, using rectangulation obtained the lower performance than single pivot language like English, Malay or Indonesian. Nevertheless, rectangulation achieved higher performance than single Filipino as in Indonesian-Vietnamese (22.24 vs. 20.73 BLEU points), Malay-Vietnamese (24.61 vs. 23.01 BLEU points). Additionally, rectangulation achieved quite promising results, which are not much lower than using single pivot. This indicates that when there is no single pivot language, rectangulation can be considered as a solution.

Table 4.31: **Using rectangulation for phrase pivot translation**; A pair of pivot languages was used, for instance Malay and English *(ms-en)* were used for pivot to translate from Indonesian to Vietnamese; the rectangulation was compared with using a single pivot language.

| **Pivot (id-vi)** | **Dev** | **Test** | **Pivot (fil-vi)** | **Dev** | **Test** | **Pivot (ms-vi)** | **Dev** | **Test** |
|---|---|---|---|---|---|---|---|---|
| en | 19.46 | 23.53 | en | 13.12 | **18.54** | en | 21.56 | **26.33** |
| fil | 17.91 | 20.73 | id | **13.42** | 18.31 | fil | 20.18 | 23.01 |
| ms | **19.92** | **24.44** | ms | 13.30 | 18.35 | id | **22.31** | 25.14 |
| ms-en | 18.98 | 22.24 | id-en | 12.79 | 17.68 | id-en | 21.11 | 24.61 |

## 4.4 Conclusion

In this chapter, I focus on pivot methods that take advantage existing bilingual corpora to improve SMT for low-resource languages. I proposed two methods to solve several problems of the triangulation approach using semantic similarity and integrating grammatical and morphological information for pivot translation.

In the first section of this chapter, I present a method based on semantic similarity between pivot phrases in phrase pivot translation. Conventional phrase pivot translation is based solely on common pivot phrases, which still lacks informative source-target phrase

pairs. I used some strategies for similarity between pivot phrases: string similarity measures for phrases containing more than one word; WordNet and word embeddings for word similarity. Experiments show that using these methods can extract more informative phrases for pivot translation, which improves pivot translation. Nevertheless, since the phrase pairs extracted by the traditional triangulation method were a large portion (about 90% in my empirical study), the additional phrase pairs extracted by the similarity models showed a small improvement.

In the second section, I propose a method to improve pivot translation using grammatical and morphology information. Part-of-speech tags and lemma forms were added for the triangulation method instead of using only the surface form of pivot phrases. Experiments were conducted on several Southeast Asian low-resource language pairs: Indonesian-Vietnamese, Malay-Vietnamese, Filipino-Vietnamese. Experimental results showed that integrating part-of-speech and lemma information improved the triangulation method trained on the surface form of pivot phrases by 0.52 BLEU score. This indicates the effectiveness of integrating grammatical and morphological information in pivot translation.

Additionally, I conducted several experiments related to choosing pivot languages: using other languages for pivot rather than English only as commonly used; rectangulation, a technique that uses two pivot languages together. I reported the results, which can be useful for research on phrase pivot translation.

This chapter is a strategy of utilizing existing bilingual corpora while Chapter 3 is a strategy of building bilingual corpora to enlarge training data for SMT models. The question is how can we apply both the two strategies effectively and whether these strategies can be combined to exploit the potential of the strategies efficiently. I will move to Chapter 5 to investigate and answer this question.

# Chapter 5

# Combining Additional Resources to Enhance SMT for Low-Resource Languages

The previous two chapters present methods of the two strategies: building bilingual corpora to enlarge the training data for SMT models (Chapter 3), and exploiting (or pivoting) existing bilingual corpora (Chapter 4). A necessary demand is that whether we can exploit both the strategies in a model that is able to further improve SMT performance. I combined components of building bilingual corpora and pivoting bilingual corpora as additional resources to enhance SMT on low-resource languages. Specifically, sentence alignment is applied to build a bilingual corpus that was then used to train a phrase table, called *alignment component*. When using pivot methods, a phrase table of a source-target language pair, called *pivot component*, can be generated from source-pivot and pivot-target bilingual corpora. For the source-target language pair, a direct bilingual corpus may exist that can be exploited to train a phrase table, called *direct component*. The components were combined to enhance SMT. I adopted linear interpolation [70] for combining the components.

Experiments were conducted on three different low-resource language pairs: Japanese-Vietnamese; Southeast Asian language pairs of Indonesian, Malay, Filipino, and Vietnamese; and the European language pair: Turkish-English. The combined models achieved the significant improvement from 2-3.0 BLEU points, which show the effectiveness of the model in improving SMT on low-resource languages.

## 5.1 Enhancing Low-Resource SMT by Combining Additional Resources

One of the main problems of SMT for low-resource languages is the unavailability of large bilingual data. In this work, I aim to further enhance SMT for low-resource language pairs by combining the additional resources as the following components.

**Direct Component** For a language pair, some small bilingual corpora may exist. In that case, I train a phrase table using the existed bilingual data, called *direct component*.

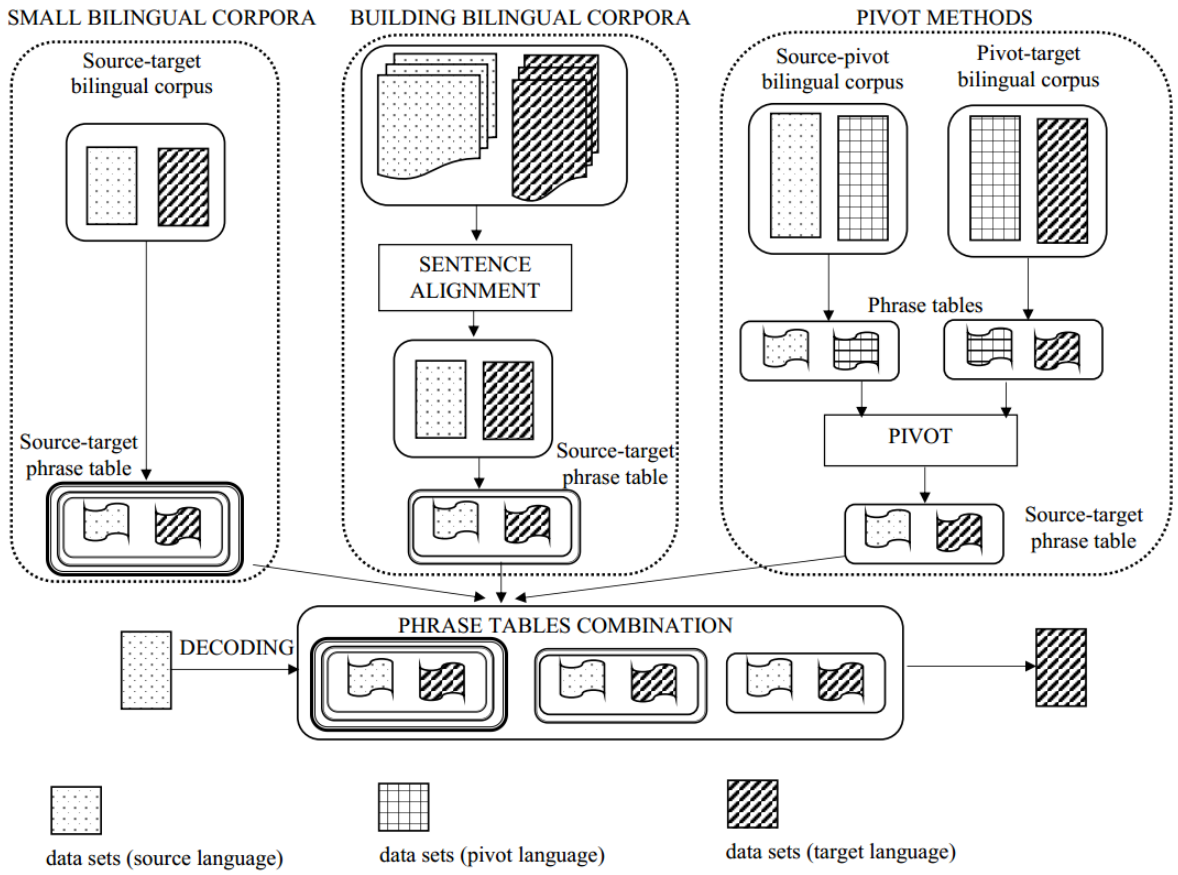



Figure 5.1: **A combined model for SMT on low-resource languages**

**Alignment Component** Since one of the main problems in SMT for low-resource languages is that large bilingual corpora are unavailable. Therefore, we need to enlarge the bilingual data from comparable data. The bilingual corpora were then used to train a phrase table, called *alignment component*.

**Pivot Component** When source-pivot and pivot-target bilingual corpora exist, I build a phrase table based on phrase pivot translation, called *pivot component*.

**Combining Components** The three components were combined to generate a phrase table for decoding. I adapted the linear interpolation [70] for combining phrase tables. Equation 5.1 describes the combination of the components.

- $d$: the direct component

- $a$: the alignment component
- $tr$: the pivot component

$$p(t|s) = \lambda_d p_d(t|s) + \lambda_a p_a(t|s) + \lambda_{tr} p_{tr}(t|s) \quad (5.1)$$

where $p_d(t|s)$, $p_a(t|s)$, and $p_{tr}(t|s)$ stand for the translation probability of the direct, alignment, and pivot models, respectively.

The interpolation parameters $\lambda_d$, $\lambda_a$, and $\lambda_{tr}$ in which $\lambda_d + \lambda_a + \lambda_{tr} = 1$ were computed by the following strategies.

- *tune*: the parameters were tuned using a development data set.
- *weights*: the parameters were set based on the ratio of the BLEU scores when using each model separately for decoding the development data set.

Figure 5.1 describes the combined model.

## 5.2 Experiments on Japanese-Vietnamese

### 5.2.1 Training Data

For training data, I used the same training data in the Japanese-Vietnamese pivot experiments as described in previous chapter (Section 4.1.3). Specifically, the Japanese-Vietnamese corpus includes 83k bilingual sentences of the TED talks and the Bibble corpora. The Japanese-English corpus contains 329k bilingual sentences from the Kyoto corpus, and the English-Vietnamese corpus includes 456k sentence pairs from the VLSP corpus and the IWSLT2015 corpus.

For training language models and word embeddings, the monolingual data sets were used: 52M sentences of Japanese extracted from Wikipedia, 30M sentences of English from the 1 Billion Word Language Model Benchmark.[1], and 16M Vietnamese sentences extracted from the website *baomoi.com*[2].

The development and test sets were extracted from several webs such as dongdu.edu.vn, *kaizen.vn*, and *duhoc.daystar.com.vn* including bilingual news and novels: 1200 bilingual sentences for the development set and 1,266 bilingual sentences for the test set.

### 5.2.2 Training Details

I conducted baseline experiments using the Moses toolkit [43]. The word alignment was trained using GIZA++ [65] with the configuration *grow-diag-final-and*. A 5-gram language model of the target language was trained using KenLM [31]. For tuning, I used the batch MIRA [13]. For evaluation, I used the BLEU scores [66].

[1]http://www.statmt.org/lm-benchmark/
[2]http://www.baomoi.com/

### 5.2.3 Main Results

Using the baseline setting to train on the Japanese-Vietnamese bilingual data (*direct model*), the baseline system obtained 6.18 BLEU score on the test set, and 8.26 BLEU score on the development set (Table 5.1).

Table 5.1: **Japanese-Vietnamese results on the direct model** - trained on the small Japanese-Vietnamese bilingual corpus (83k bilingual sentences)

| **System** | **Development** (BLEU) | **Test** (BLEU) |
|---|---|---|
| direct | 8.26 | 6.18 |

The scores indicate a low performance which requires investigations in both aspects: the challenges in terms of the languages (different structures between Japanese and Vietnamese), and the limited amounts of the training data. For the challenges of the languages' structures, it requires further researches. In this work, I focus on improving the problem of limited bilingual data by using methods of pivot translation and building bilingual corpora from comparable data.

I combined the two components: building bilingual data and the pivot translation with the baseline model to create a framework for improving SMT. Using the linear interpolation to combine phrase tables, I describe the experimental results in Table 5.2.

Table 5.2: **Japanese-Vietnamese results on the combined models** - *direct*, *pivot*, *wiki*: the direct, pivot, alignment components; *direct-pivot*: the combination of the direct and pivot models; *tuning*, *weights*: the combination settings

| **System** | **Dev (BLEU)** | **Test (BLEU)** |
|---|---|---|
| direct | 8.26 | 6.18 |
| direct-wiki (tuning) | 8.59 | 6.65 |
| direct-wiki (weights) | 8.57 | 6.69 |
| direct-pivot (tuning) | 8.82 | 7.73 |
| **direct-pivot (weights)** | 8.41 | **8.69** |
| combined model (tuning) | 8.78 | 7.86 |
| combined model (weights) | 8.31 | 8.66 (+2.48) |

The experimental results showed that using the *weights* setting in the linear interpolation produced better performance than the *tuning* setting. The aigned corpus from Wikipedia improved the direct model with +0.47 to +0.51 BLEU scores. For pivot translation, this improved significantly with +2.51 BLEU score. my framework significantly improved the direct model trained on small bilingual corpora (+2.48 BLEU). This showed the contribution of the framework in improving the SMT on low-resource data. Nevertheless, this experimental result also indicates a problem in combining the three components in which the combination with the aligned data from Wikipedia slightly decreased the performance of the combination: direct and pivot only. This can be affected by some

aspects: the overlap in vocabulary between the pivot and the aligned models; the noisy sentences in the aligned model.

**Improvement on the baseline system with low BLEU points** Although my model achieved a significant improvement on Japanese-Vietnamese with +2.5 BLEU point, a question can be whether the improvement on such baseline system with low performance is reliable. Working on machine translation for the language pair like Japanese-Vietnamese faces many challenges. First, the languages differ in language structures: Subject-Verb-Object in Vietnamese versus Subject-Object-Verb in Japanese. Second, several information can be hidden in Japanese such as pronoun. Fortunately, statistical methods in machine translation can discover some translation rules automatically even on such challenges as long as we provide a large parallel data for training. Nevertheless, such large parallel data also does not exist between Japanese and Vietnamese. Therefore, although showing promising results in translation for many language pairs, statistical methods also failed with such low-resource languages as Japanese-Vietnamese. The challenges are doubled because the task for machine translation now has to deal with not only the differences in linguistics but also the limited training data.

In order to verify whether the proposed model actually improve the baseline or not, I conducted a set of tests called statistical significance tests as discussed in previous chapter. First, I choose a big test set of Japanese-Vietnamese from the Asian Lanugage Treebank corpus [80] including 20,084 parallel sentences. The baseline and proposed systems were used to test on the big test set. Then, I generated a larger number of test sets by using the broad sample and randomly drawn test sets methods as discussed in previous chapter. A size of 400 sentences was used for each test set. Experimental results are presented in Table 5.3 and Table 5.4.

Table 5.3: **Results of Japanese-Vietnamese on the big test set**

| **System** | **BLEU** |
|---|---|
| direct | 9.82 |
| combined model (weights) | 10.28 |
| combined model (tune) | 10.19 |

Table 5.4: **Results of statistical significance tests on Japanese-Vietnamese**

| **System Comparison** | **BLEU** | **Size 400** |
|---|---|---|
| combined model (weights) better than direct | 0.46% | 33% |
| combined model (weights) better than direct | 0.37% | 33% |

The experimental results reveal that the proposed systems outperform the baseline system on even a very big test set. Furthermore, the statistical significance test sets also

confirm that the proposed systems actually outperform the baseline system. We draw 33% correct conclusion that the proposed systems outperform the baseline system, and there are no wrong conclusion.

## 5.3 Experiments on Southeast Asian Languages

I conducted experiments using the combined model on other languages including translations from Indonesian, Malay, and Filipino to Vietnamese, which are low-resource Southeast Asian language pairs that have not been yet investigated according to my best knowledge.

### 5.3.1 Training Data

I used the same training data of the Southeast Asian pivot translation as described in the previous chapter (Section 4.1.4). The training data was extracted from the TED corpus [9] and the ALT corpus [80]: 226k parallel sentences for Indonesian-Vietnamese, 33k parallel sentences for Malay-Vietnamese, and 22k parallel sentences for Filipino-Vietnamese. For each language pair, the development set includes 2k bilingual sentences, and the test set includes 2,074 bilingual sentences, which were extracted from the ALT corpus.

English was used for the pivot language. Bilingual corpora for training pivot models were extracted from the TED corpus and the ALT corpus: Indonesian-English (244k parallel sentences), Malay-English (31k parallel sentences), Filipino-English (21k parallel sentences), and English-Vietnamese (377k parallel sentences).

Monolingual data sets of Indonesian, Malay, and Filipino were extracted from Wikipedia: Indonesian (1.4M sentences), Malay (596k sentences), and Filipino (682k sentences). For Vietnamese, the same monolingual data as described in the previous chapter was used that is 16M sentences extracted from the *baomoi.com*, a website of articles in Vietnamese.

### 5.3.2 Training Details

I used the same training setup as in the Japanese-Vietnamese setting.

### 5.3.3 Main Results

Table 5.5 presents the baseline experimental results. Unlike the baseline results of Japanese-Vietnamese, the baseline models on the Southeast Asian languages showed the much higher performance with 30.36 BLEU score (Indonesian-Vietnamese), 32.81 BLEU score (Malay-Vietnamese), and 24.29 BLEU score (Filipino-Vietnamese) although the training data sets are limited especially in Malay-Vietnamese and Filipino-Vietnamese.

I finally combined the alignment model and the proposed pivot translation with the direct model for the final objective that aims to improve SMT for low-resource languages. The components were combined using the linear interpolation based on the *weights* setting

Table 5.5: **Southeast Asian results on the direct models (BLEU)** - id: Indonesian, ms: Malay, fil: Filipino, vi: Vietnamese

| System | Development | Test |
|---|---|---|
| id-vi | 29.97 | 30.46 |
| ms-vi | 30.09 | 32.81 |
| fil-vi | 22.10 | 24.29 |

(the ratio of BLEU scores on the test set). Table 5.6 presents the experimental results of the framework.

Table 5.6: **Southeast Asian results on the combined model**

| Model | Dev (BLEU) | Test (BLEU) |
|---|---|---|
| direct: id-vi | 29.97 | 30.46 |
| direct-wiki: id-vi | 30.02 | 30.48 |
| **direct-pivot: id-vi** | 29.85 | **33.60** |
| my framework: id-vi | 30.04 | 33.46 (+3.0) |
| direct: ms-vi | 30.09 | 32.81 |
| direct-wiki: ms-vi | 30.09 | 32.69 |
| direct-pivot: ms-vi | 29.81 | 35.73 |
| **my framework: ms-vi** | 29.97 | **35.85 (+3.04)** |
| direct: fil-vi | 22.10 | 24.29 |
| direct-wiki: fil-vi | 21.98 | 24.34 |
| direct-pivot: fil-vi | 21.84 | 26.67 |
| **my framework: fil-vi** | 21.83 | **26.69 (+2.40)** |

For experimental results described in Table 5.6, although the aligned models have shown promising results without the direct bilingual data, the models do not show significant improvement when combined with the direct models which produced quite high performance. Meanwhile, pivot translation models obtained much improvement when combined with the baseline models. Finally, in my framework, I obtained a significant improvement for the low-resource setting (the direct model with limited bilingual data). Specifically, my framework obtained +2.4 to +3.04 BLEU scores improvement, which were significantly contributed by the pivot translation, and a small contribution from the aligned corpora. In overall, when bilingual data sets are unavailable or in limited amounts, which are popular for most languages, the proposed framework can be a solution to help improving SMT on such low-resource languages.

## 5.4 Experiments on Turkish-English

In this section, I present the experiments on Turkish-English using the proposed model, and the results that I submitted to the shared task of machine translation (WMT17).[3]

**System Combination** I exploited two resources to enhance machine translation for the low-resource setting: a bilingual corpus extracted from Wikipedia, and bilingual corpora of Turkish and English paired with the six pivot languages. My goal now is to utilize the resource most effectively. I introduce a system combination which includes the following components. First, I trained a phrase table based on the Wikipedia bilingual corpus, called *align* component. Second, using the phrase pivot translation, I obtained pivoted phrase table, called the *pivot* components. Additionally, I trained a phrase table using the Turkish-English training data, called *baseline* component. The components were combined to generate a phrase table for decoding. I adapted the linear interpolation [70] for combining phrase tables. Equation 5.2 describes the combination of the components.

$$\begin{aligned} p(t|s) =& \lambda_d p_d(t|s) + \lambda_a p_a(t|s) \\ &+ \lambda_{p_1} p_1(t|s) + \lambda_{p_2} p_2(t|s) + \lambda_{p_3} p_3(t|s) \\ &+ \lambda_{p_4} p_4(t|s) + \lambda_{p_5} p_5(t|s) + \lambda_{p_6} p_6(t|s) \end{aligned} \tag{5.2}$$

Where $p_d(t|s)$, $p_a(t|s)$, and $p_{tr}(t|s)$ stand for the translation probability of the *baseline*, *align*, and the *pivot* components, respectively. $p_i(t|s), i = 1..6$ stand for the translation probability of the six pivoted phrase tables.

The interpolation parameters $\lambda_d$, $\lambda_a$, and $\lambda_{p_i} (i = 1..6)$ in which $\lambda_d + \lambda_a + \lambda_{p_i} = 1$ were tuned using the development set (*newsdev2016*) provided by the shared task.

### 5.4.1 Training Data

For training data, I used the data provided by the shared task: 207k Turkish-English bilingual sentences that was extracted from the SETIMES2 corpus [81]. The corpus was preprocessed: word tokenization, truecase, and clean that keep sentences of 1-80 words using the Moses scripts.[4]

For development and test sets, I used the *newsdev2016* (1,001 bilingual sentences) for tuning parameters, the *newstest2016* (3,000 bilingual sentences) and *newstest2017* (3,007 bilingual sentences) for evaluation. These data sets are provided by the shared task WMT17.

For monolingual data, I used the monolingual data sets provided by the shared task: 40M sentences of Turkish and 40M sentences of English.

I used bilingual data sets of the SETIMES2 corpus [81][5], the same resource of the Turkish-English training data in the shared task, for training phrase pivot translation. I used six pivot languages to bridge the translation between Turkish and English: Bulgarian,

[3] http://www.statmt.org/wmt17/translation-task.html
[4] https://github.com/moses-smt/mosesdecoder/tree/master/scripts/tokenizer
[5] http://opus.lingfil.uu.se/SETIMES2.php

Bosnian, Greek, Macedonian, Romanian, and Albanian. The bilingual corpora for the pivot translation are presented in Table 5.7.

Table 5.7: **Bilingual corpora for Turkish-English pivot translation** (the number of parallel sentences); *tr*: Turkish, *en*: English; *tr-pivot*: the bilingual corpus of Turkish and the pivot language

| No. | Pivot language | tr-pivot | pivot-en |
|---|---|---|---|
| 1 | bg (Bulgarian) | 206k | 213k |
| 2 | bs (Bosnian) | 133k | 138k |
| 3 | el (Greek) | 206k | 226k |
| 4 | mk (Macedonian) | 202k | 207k |
| 5 | ro (Romanian) | 205k | 212k |
| 6 | sq (Albanian) | 206k | 227k |

### 5.4.2 Training Details

I conducted baseline experiments using the Moses toolkit [43]. The word alignment was trained using GIZA++ [65] with the configuration *grow-diag-final-and*. A 5-gram language model of the target language was trained using KenLM [31]. For tuning, I used batch MIRA [13]. For evaluation, I used the BLEU scores [66].

### 5.4.3 Results

Experimental results are presented in Table 5.8 and Table 5.9. The results of my system can be accessible at the Shared Task on Machine Translation (WMT17). [6]

Table 5.8: **Experimental results on the Turkish-English** (BLEU); **baseline (align)**: the system trained on the baseline (the aligned Wikipedia) bilingual corpus; **pivot (bs), pivot (6)**: the phrase pivot translation system using one pivot language (bs: Bosnian) or using all of the 6 pivot languages; **baseline-pivot(6)-align**: the system combination of the baseline, align, and 6 pivot components.

| Model | newsdev2016 | newstest2016 | newstest2017 |
|---|---|---|---|
| baseline | 12.28 | 12.3 | 12.0 |
| align | 7.67 | 8.1 | 7.9 |
| pivot (bs) | 7.47 | 11.0 | 7.6 |
| baseline-align | 13.35 | 12.9 (+0.6) | 12.7 (+0.7) |
| baseline-pivot(bs) | 12.39 | 13.1 (+0.8) | 12.4 (+0.4) |
| baseline-pivot(bs)-align | 13.02 | 13.0 (+0.7) | 12.7 (+0.4) |
| baseline-pivot(6)-align | 14.04 | **13.7 (+1.4)** | **13.1 (+1.1)** |

[6] http://matrix.statmt.org/

Table 5.9: **Experimental results on the English-Turkish translation** (BLEU).

| **Model** | **newsdev2016** | **newstest2016** | **newstest2017** |
|---|---|---|---|
| baseline | 8.66 | 9.3 | 9.9 |
| align | 5.96 | 6.3 | 6.6 |
| pivot (bs) | 6.01 | 8.2 | 6.3 |
| baseline-align | 8.87 | 9.3 | 10.0 (+0.1) |
| baseline-pivot | 9.01 | 9.6 (+0.3) | 9.7 |
| baseline-pivot(bs)-align | 8.98 | 9.6 (+0.3) | 9.9 |
| baseline-pivot(6)-align | 10.11 | **9.7 (+0.4)** | **10.4 (+0.5)** |

**Building A Parallel Corpus** A bilingual corpus of Turkish-English was built from Wikipedia. Table 5.10 presents the input data and the results of building the corpus. After extracting parallel titles and collecting 2M articles in Turkish and 3M articles in English, I obtained 48k bilingual sentences. The corpus was cleaned to keep sentences of 1-80 words.

Table 5.10: **Building a bilingual corpus of Turkish-English from Wikipedia.** (the number of parallel sentences)

| | **Turkish** | **English** |
|---|---|---|
| Input articles | 188,235 | 192,512 |
| Input sentences | 2,030,931 | 3,023,324 |
| Bilingual articles | 184,154 | 184,154 |
| Aligned articles | 22,100 | 22,100 |
| Aligned sentences | 48,554 | 48,554 |

Although the aligned Wikipedia corpus contains a small number of parallel sentences (48k) compared with the direct training data (207k), the phrase-based models trained on the Wikipedia corpus showed a quite promising result: 7.9 BLEU point on the Turkish-English and 6.6 BLEU point on the English-Turkish. When the baseline model was combined with the align model, I achieved a significant improvement: +0.6 and +0.7 BLEU points on the Turkish-English of the *newstest2016* and *newstest2017*, respectively. The results showed the effectiveness of the extracted corpus to enhance machine translation on the low-resource setting. Nevertheless, the task becomes more challenge on the English-Turkish. Although the Wikipedia corpus showed the contribution on the Turkish-English translation, there was no improvement on the English-Turkish translation when I achieved only +0.1 BLEU point on the *newstest2017*.

**Phrase Pivot Translation** For the phrase pivot translation models, using one pivot language (bs: Bosnian) showed the competitive performance on the newstest2016 of the Turkish-English: 11.0 BLEU point vs. 12.3 BLEU point (baseline), or 8.2 BLEU point vs. 9.3 BLEU point (baseline) on the English-Turkish.

When the pivot model (using one pivot language of Bosnian) was combine with the baseline model, I achieved the improvement on both translation directions: +0.8 BLEU point on the Turkish-English, and +0.3 BLEU point on the English-Turkish of the *newstest2016*. For the newstest2017, I achived the improvement only on the Turkish-English (+0.4 BLEU point).

The results confirmed the contribution of the phrase pivot translation. Nevertheless, there was no improvement on some cases. Therefore, I seek to the combination of all components: the baseline, align, and pivot components (from one pivot language to six pivot languages).

**System Combination** I would like to exploit the components most effectively to improve machine translation on the low-resource setting. The baseline, align, and pivot components were combined in a model. When using one pivot language (Bosnian), I achieved the improvement in most cases: +0.7 and +0.4 BLEU points on the *newstest2016* and *newstest2017* of the Turkish-English. For the English-Turkish, I achieved the improvement of +0.3 BLEU point on the *newstest2016*; however, there was no improvement on the *newstest2017*, in which the pivot model did not showed the contribution.

Interestingly, using six pivot languages showed the significant improvement in all settings. For the Turkish-English, I achieved +1.4 and +1.1 BLEU points on the *newstest2016* and *newstest2017*, respectively. For the English-Turkish, the combined system showed +0.4 BLEU point (newstest2016) and +0.5 BLEU point (newstest2017).

I submitted my systems using the system combination of the baseline, align, and six pivot languages in the phrase pivot translation.

## 5.5 Analysis

In this section, I discuss several aspects of the combined model that improves the SMT performance for low-resource languages.

### 5.5.1 Exploiting Informative Vocabulary

Table 5.11: **Out of vocabulary ratio on the combined model (%)**

| **Model** | **ja-vi** | **id-vi** | **ms-vi** | **fil-vi** |
|---|---|---|---|---|
| direct | 20.68 | 12.97 | 16.66 | 24.89 |
| wiki | 48.05 | 28.84 | 37.62 | 61.72 |
| direct-wiki | 18.95 | 12.56 | 15.41 | 24.29 |
| pivot | 41.51 | 13.61 | 18.05 | 26.64 |
| pivot-similarity | 39.46 | 12.56 | 17.23 | 24.73 |
| direct-pivot | 14.46 | 11.52 | 15.05 | 22.11 |
| direct-pivot-wiki | 13.47 | 11.02 | 14.10 | 21.67 |

I present the out of vocabulary ratio in Table 5.11. There are some interesting observations from the ratio. First, using the interpolation of direct models and pivot and wiki models reduced the out of vocabulary ratio, which leads to the improvement in all language pairs. Second, the performance was effected by not only the OOV ratio. For instance, although the OOV ratio in the direct model of Japanese-Vietnamese (20.68) is lower than that of Filipino-Vietnamese (24.89), the BLEU score of the Japanese-Vietnamese (6.18) is much lower than that of Filipino-Vietnamese (24.29). This problem can be caused by other reasons like reordering or the differences in language structures.

By using the combined model that combines the two strategies: building bilingual corpora (alignment component), and taking advantage existing bilingual corpora (pivot component), a set of informative vocabulary can be exploited that overcomes the problem of sparse data in which the direct model contains only a small vocabulary size. This is one of the main factors to improve SMT on low-resource languages using the proposed model.

### 5.5.2 Sample Translations

I present examples of the OOV problem in Table 5.12. Pivot and wiki models generate correct translations that can improve the OOV problem in the baseline models.

Table 5.12: **Sample translations (Japanese-Vietnamese)**. **input** (input sentence of ja (Japanese)); **reference** (reference sentence of vi (Vietnamese)); **Meaning**: the English meaning translated by the authors in *the italic*; **direct**: trained on the direct corpus; **direct-pivot**: the combined model of the direct and pivot components; the underline: phrases were not translated by the direct model (OOV). The **bold words**: correct translations generated by the proposed models that showed the improvement

| **System** | **Examples** |
|---|---|
| **ja (input)** | むかし むかし 、 ある 山寺 に 、 和尚 さん と とん ちの きく 小僧 さん が い まし た 。 |
| **vi (reference)** | ngày xửa ngày xưa , ở ngôi chùa trên núi nọ , có vị trụ trì và chú tiểu lanh trí . |
| **Meaning** | *Once upon a time, in a temple on the mountain, there was a monk and a quickwitted novice.* |
| **direct** | vì , lúc xưa , 山寺 和尚 小僧 bởi do chạy trốn ! ừ phải , và ngợi khen mà ngợi khen . |
| **direct-pivot** | lâu rồi , đã có một số **núi** được cho là **tu sĩ** , do đó , người Nhật Bản ) nở . |
| **direct-pivot-wiki** | lâu rồi , đã có một số **núi** được cho là **tu sĩ** , do đó , người Nhật Bản ) nở hoa . |
| **ja (input)** | 「 よせ よせ 、 大阪 まで は とても 遠く て 、 たいへん だ ぞ 。 ケロ 」 |
| **vi (reference)** | - Này này , Đến Osaka thì rất là xa và vất vả lắm đấy . ộp ộp |
| **Meaning** | "*Alright, it is very far to Osaka, it is very hard*" |
| **direct** | 大阪 , kéo " rất xa , cho rằng , " ケロ ! . |
| **direct-pivot** | " cho **đến Osaka** , kéo , rất , rất , rất xa ケロ . " |
| **direct-pivot-wiki** | " cho **đến Osaka** , kéo , rất , rất xa ケロ là của bạn . |

Table 5.13: **Sample translations** - on the combined model (Indonesian-Vietnamese, Malay-Vietnamese)

| System | Examples |
|---|---|
| **id (input)** | peluncuran yang sukses akan membuat Korea Selatan bisa menjadi pemain di dalam bisnis komersial luar angkasa yang nilai industrinya berkisar USD250 miliar . |
| **vi (reference)** | một vụ phóng thành công sẽ có thể giúp Hàn Quốc trở thành một thành viên trong các thương vụ phóng không gian thương mại , một ngành công nghiệp trị giá khoảng US $ 250 tỷ . |
| **Meaning** | *A successful launch will enable Korea to become a member of the commercial space launch , an industry worth about US $ 250 billion* |
| **direct** | phóng thành công sẽ tạo ra Hàn Quốc có thể trở thành một vận động viên trong kinh doanh thương mại không gian mà giá trị industrinya xoay USD250 tỷ . |
| **direct-pivot** | phóng thành công sẽ tạo ra Hàn Quốc có thể trở thành cầu thủ trong kinh doanh thương mại không gian giá trị **công nghiệp** kéo dài USD250 tỷ . |
| **direct-pivot-wiki** | phóng thành công sẽ tạo ra Hàn Quốc có thể trở thành cầu thủ trong kinh doanh thương mại không gian giá trị **công nghiệp** kéo dài USD250 tỷ . |
| **ms (input)** | kenyataan jawatankuasa itu mendapat sokongan daripada pencinta alam sekitar dan juga dari beberapa industri sektor tenaga . |
| **vi (reference)** | tuyên bố của Ủy ban đã thu hút được sự ủng hộ từ các nhà hoạt động môi trường cũng như từ một số ngành công nghiệp về lĩnh vực năng lượng . |
| **Meaning** | *The Commission's statement has garnered support from environmentalists as well as from some industries in the field* |
| **direct** | tuyên bố của ủy ban này đã nhận được sự ủng hộ từ pencinta môi trường và cũng từ một số ngành công nghiệp của khu vực năng lượng . |
| **direct-pivot** | tuyên bố của ủy ban này đã nhận được sự ủng hộ từ **các nhà môi trường** cũng như từ một số ngành công nghiệp khu vực năng lượng . |
| **direct-pivot-wiki** | tuyên bố của ủy ban này đã nhận được sự ủng hộ từ **các nhà môi trường** và cũng từ một số ngành công nghiệp khu vực năng lượng . |

Table 5.14: **Sample translations** - using the combined model (Filipino-Vietnamese))

| System | Examples |
|---|---|
| **fil (input)** | ang balat ng lalaki ay nakadikit na sa tela ng silya , ayon pa sa mga opisyal ng batas , isa sa kanila ay kinailangan pang itapon ang kanyang uniporme pagkatapos pumunta sa nasabing tahanan . |
| **vi (reference)** | da của người đàn ông đã dính chặt với vải của ghế tựa , các nhân viên thực thi pháp luật cho biết , một trong số đó đã phải vứt bỏ đồng phục của mình khi trở về nhà . |
| **Meaning** | *The man's skin was attached to the fabric of the chair , one of them had to throw away his uniform when he returned home , said law enforcement officers* |
| **direct** | da của người đàn ông này còn nối với dệt may của chiếc ghế , cho biết các quan chức của pháp luật , một trong số đó đã được vứt bỏ uniporme của mình sau khi đi trong ngôi nhà này . |
| **direct-pivot** | da của những người đàn ông đã gắn liền với với dệt may của chiếc ghế , cũng theo các quan chức của luật pháp , một người trong số họ đã vứt bỏ **đồng phục** của mình sau khi tham dự trong các ngôi nhà . |
| **direct-pivot-wiki** | da của những người đàn ông đã gắn liền với với dệt may của chiếc ghế , cũng theo các quan chức của luật pháp , một trong số đó là được vứt bỏ **đồng phục** của mình sau khi tham dự trong các ngôi nhà . |

## 5.6 Conclusion

In this chapter, I seek to the problem of whether the two strategies of building and pivoting bilingual corpora can be combined effectively to exploit the additional data and further improve SMT performance. I present a combined model that combines several components for SMT on low-resource languages: the *pivot component* (for the strategy of exploiting existing bilingual corpora), the *alignment component* (trained on the corpus built from comparable data for the strategy of building corpus to enlarge training data for SMT models), and the *direct component* (trained on any available existed direct corpus).

Three language pairs were used in experiments to evaluate the proposed model: Japanese-Vietnamese; Southeast Asian languages (Indonesian, Malay, Filipino, Vietnamese); and Turkish-English. Various settings of the combined model were conducted among components and the interpolation settings: the *tuning* that was based on a given tuning set, and the *weights* that was based on the ratio of BLEU scores when decoding the tuning set. Experimental results showed a significant improvement when applying the proposed model in which +2.0 to +3.0 BLEU points improvement were achieved although there exist small direct bilingual corpora on the low-resource language pairs. This confirms the

effectiveness and contribution of the proposed model in improving SMT for low-resource languages.

The three chapters 3, 4, and 5 present methods related to two strategies of building bilingual corpora and exploiting existing bilingual corpora that were applied and evaluated on SMT, statistical methods. Another kind of methods that has effectively and successfully applied in several rich language pairs recently: neural machine translation (NMT). The problem is whether NMT can be applied successfully to low-resource languages. The next chapter presents investigations of utilizing NMT on low-resource languages.

# Chapter 6

# Neural Machine Translation for Low-Resource Languages

Previous chapters present methods of the two strategies in improving SMT for low-resource languages: building bilingual corpora and exploiting existing bilingual corpora that were applied and evaluated on statistical methods (SMT). Recent work of neural machine translation (NMT) has shown the effectiveness on rich language pairs, which include several upto tens million of bilingual sentences like English-German, English-French. Neural-based and phrase-based methods have shown the effectiveness and promising results in the development of current machine translation. The two methods are compared on some European languages, which show the advantages of the neural machine translation. Nevertheless, there are few work of comparing the two methods on low-resource languages, which there are only small bilingual corpora. The problem of unavailable large bilingual corpora causes a bottleneck for machine translation for such language pairs. In this chapter, I present a comparison of the phrase-based and neural-based machine translation methods on several Asian language pairs: Japanese-English, Indonesian-Vietnamese, and English-Vietnamese. Additionally, I extracted a bilingual corpus from Wikipedia to enhance machine translation performance. Experimental results showed that when using the extracted corpus to enlarge the training data, neural machine translation models achieved the higher improvement and outperformed the phrase-based models. This work can be useful as a basis for further development of machine translation on the low-resource languages. Additionally, a recent work in transfer learning for low-resource neural machine translation achieved promising results and showed potentials for my work in further developing this method. I discuss this transfer learning method and several strategies that can be developed in further research.

## 6.1 Neural Machine Translation

In my work, I based on the model of [71], which are encoder-decoder networks with an attention mechanism [1]. First, I briefly discuss some background information. Then, I discuss building synthetic training data in pivot translation using NMT.

In addition to some background knowledge of neural-based machine translation as presented in Section 2.4, I discuss two essential background: attention mechanism and dealing with rare and unknown words in neural-based machine translation, which were used in my experiments.

### 6.1.1 Attention Mechanism

As shown in [1], the translation performance decreases when translating long sentences. Instead of encoding entire the input sentence into the context vector, the most relevant information of the input sentence is encoded into the single, fixed-length vector. The representation $c$ for the source sentences is set as follows.

$$c = [\bar{h_1}, ..., \bar{h_m}] \tag{6.1}$$

There are two stages in the function $f$ in Equation 2.18: attention context and extended recurrent neural network (RNN). In the attention context, an alignment vector $a_i$ is learned by comparing the previous hidden $h_{i-1}$ with individual source hidden states in the context vector $c$; then the model derives a weighted average ($c_i$) of the source hidden states based on the alignment vector $a_i$. For the second stage, extended RNN, the RNN unit is expanded for the context vector $c_i$ in addition to the previous hidden state $h_{i-1}$ and the current input $t_{i-1}$ to compute the next hidden state $h_i$.

### 6.1.2 Byte-pair Encoding

In order to overcome the problem of out-of-vocabulary, [72] proposed a method for open-vocabulary translation by encoding rare and unknown words as sequences of subword units. This is because various word classes can be translated by smaller units like compositional translation for compounds, phonological and morphological transformations for cognates and loanwords. In order to do that, words are segmented using byte-pair encoding that originally devised as a compression algorithm [26].

## 6.2 Phrase-based versus Neural-based Machine Translation on Low-Resource Languages

SMT and NMT models have shown successfully in language pairs in which large bilingual corpora are available such as English-German, English-French, Chinese-English, and English-Arabic. There are some work that evaluated the phrase-based versus neural-based methods such as the comparison of the two methods on English-German [2], the comparison on 30 translation directions on the United Nations Parallel Corpus [35]. Nevertheless, for low-resource settings like Asian language pairs which contain only small bilingual corpora, there are few work of the comparison of the two methods on such language pairs. Additionally, the problem of unavailable large bilingual corpora causes a bottleneck for machine translation on such languages.

In this work, I compared the SMT and NMT methods on several low-resource language pairs. The standard phrase-based SMT is the well-known Moses toolkit [43]. For NMT models, I utilized the state-of-the-art model [71] in the WMT 2016,[1] which used encoder-decoder networks with attention mechanism and open-vocabulary translation. Experiments were conducted on Asian language pairs: Japanese-English, Indonesian-Vietnamese, and English-Vietnamese with only small bilingual corpora. Furthermore, in order to overcome the problem of unavailable large bilingual corpora, I extracted a bilingual corpus from Wikipedia to enhance machine translation on both SMT and NMT models. Moreover, I aim to evaluate the effects of enlarging training data to the two different machine translation methods and to the overall performance. Experimental results showed meaningful findings in the comparison of the two machine translation methods on the low-resource settings. This work can be useful as a basis for further development of NMT as well as machine translation in general on the low-resource languages. The scripts, corpora, and trained models used in this research can be found at the repository.[2]

### 6.2.1 Setup

I conducted experiments on Asian language pairs: Japanese-English, Indonesian- Vietnamese, and English-Vietnamese using the two machine translation methods: SMT and NMT. Additionally, I extracted a bilingual corpus from Wikipedia to enhance the machine translation on both of the two methods.

For SMT models, I used the Moses toolkit [43]. The word alignment was trained using GIZA++ [65] with the configuration *grow-diag-final-and*. A 5-gram language model of the target language was trained using KenLM [31]. For tuning, I used the batch MIRA [13]. For evaluation, I used the BLEU scores [66].

For NMT models, I adapted the attentional encoder-decoder networks combined with byte-pair encoding [71]. In our experiments, I set the word embedding size 500, and hidden layers size of 1024. Sentences are filtered with the maximum length of 50 words. The minibatches size is set to 60. The models were trained with the optimizer Adadelta [99]. The models were validated each 3000 minibatches based on the BLEU scores on development sets. I saved the models for each 6000 minibatches. For decoding, I used beam search with the beam size of 12. I trained NMT models on an Nvidia GRID K520 GPU.

### 6.2.2 SMT vs. NMT on Low-Resource Settings

**Experiments on Japanese-English** I conducted experiments on Japanese-English using the Kyoto bilingual corpora [60]. The training data includes 329,882 parallel sentences. For the development and the test data, there are 1,235 parallel sentences in the development set and 1,160 parallel sentences in the test set (see Table 6.1 for the data sets).

[1] http://www.statmt.org/wmt16/

[2] https://github.com/spt41bk/MT-LowRec

Table 6.1: **Bilingual data set of Japanese-English** of the training set (Train), development set (Dev), and test set (Test); **Words**: the number of tokens split by spaces; **Vocabulary**: the number of distinct words

| | **Train** | **Dev** | **Test** |
|---|---|---|---|
| Sentences | 329,882 | 1,235 | 1,160 |
| Japanese Words | 6,085,131 | 34,403 | 28,501 |
| English Words | 5,911,486 | 30,822 | 26,734 |
| Japanese Vocabulary | 114,284 | 4,909 | 4,574 |
| English Vocabulary | 161,655 | 5,470 | 4,912 |

Experimental results of Japanese-English translation are showed in Table 6.2. The NMT model obtained 11.91 BLEU point on the development set. For the test set, the model achieved 14.91 BLEU point after training 20 epochs. Meanwhile, the SMT model obtained the higher performance: +1.18 BLEU point on the development set, and +2.86 BLEU point on the test set. The experimental results indicated that for a small bilingual corpus (329k parallel sentences of the Japanese-English Kyoto corpus), the SMT model showed the higher performance than the NMT model.

Table 6.2: **Experimental results in Japanese-English translation** (BLEU)

| **Model** | **Dev** | **Test** |
|---|---|---|
| SMT | 13.09 | 17.75 |
| NMT | 11.91 | 14.91 |

**Experiments on Indonesian-Vietnamese** I conducted experiments on the Indonesian-Vietnamese language pairs, which has yet investigated on machine translation to our best knowledge. For training data, I used two resources: TED data [9] and the ALT corpus (Asian Language Treebank Parallel Corpus) [80]. I extracted Indonesian-Vietnamese parallel sentences from the TED data. For the ALT corpus, I dived the Indonesian-Vietnamese bilingual corpus into three parts: 16,000 sentences for training, 1,000 sentences for the development set, and 1,084 sentences for the test set. I combined the Indonesian-Vietnamese TED data with the training set extracted from the ALT corpus to create 226,239 training sentence pairs. The data sets are described in Table 6.3.

I showed the experimental results of the Indonesian-Vietnamese translations in Table 6.4. The NMT model achieved 14.48 BLEU point on the development set and 14.98 BLEU point on the test set after training 22 epochs. Meanwhile, the SMT model obtained the much higher performance: 27.37 BLEU point on the development set and 30.17 BLEU point on the test set.

Table 6.3: **Bilingual data sets of Indonesian-Vietnamese translations**

| | **Train** | **Dev** | **Test** |
|---|---|---|---|
| Sentences | 226,239 | 1,000 | 1,084 |
| Indonesian Words | 1,932,460 | 22,736 | 25,423 |
| Vietnamese Words | 2,822,894 | 32,891 | 36,026 |
| Indonesian Vocabulary | 52,935 | 4,974 | 5,425 |
| Vietnamese Vocabulary | 29,896 | 3,517 | 3,751 |

Table 6.4: **Experimental results on Indonesian-Vietnamese translation** (BLEU)

| **Model** | **Dev** | **Test** |
|---|---|---|
| SMT | 27.37 | 30.17 |
| NMT | 14.48 | 14.98 |

**Experiments on English-Vietnamese** I conducted experiments on English- Vietnamese using the data sets of the IWSLT 2015 machine translation shared task [10]. The *constrained* training data contained 130k parallel sentences from the TED talks.[3] I used the *tst2012* for the development set, *tst2013* and *tst2015* for the test sets.

In addition, I used two other data sets to enlarge the training data from the two resources: the corpus of National project VLSP (Vietnamese Language and Speech Processing)[4] and the EVBCorpus [62]. The two data sets were merged with the *constrained* data to create a large training data called *unconstrained* data. This aims to investigate how the large training data affects the SMT and NMT models.

Table 6.5: **Experimental results English-Vietnamese translations** (BLEU); **constrained (SMT)**: the model trained on the constrained data using SMT; **unconstrained (NMT)**: the model trained on the unconstrained data using NMT

| **System** | **tst2012** | **tst2013** | **tst2015** |
|---|---|---|---|
| constrained (SMT) | 23.80 | 26.54 | 24.42 |
| constrained (NMT) | 20.21 | 23.59 | 17.27 |
| unconstrained(SMT) | 34.42 | 27.19 | 25.41 |
| unconstrained(NMT) | 24.05 | 26.71 | 22.30 |

Experimental results of English-Vietnamese are presented in Table 6.5. In overall, the SMT model obtained the higher performance than the NMT model (26.54 vs. 23.59 BLEU points on the *tst2013* using the *constrained* data, 25.41 vs. 22.30 BLEU points on the *tst2015* using the *unconstrained* data). Another point is the effect of enlarging the training

[3]https://www.ted.com/talks

[4]http://vlsp.vietlp.org:8080/demo/?page=home

data using the *unconstrained* data set. Enlarging the training data (increasing from 130k to 456k parallel sentences) improved both SMT and NMT models. Specifically, the SMT model achived +0.65 BLEU point on the *tst2013* and +0.99 BLEU point on the *tst2015*. The interesting point is that the NMT model showed the higher improvement than the SMT model when using the *unconstrained* data: +3.12 BLEU point on the *tst2013* and +5.03 BLEU point on the *tst2015*.

### 6.2.3 Improving SMT and NMT Using Comparable Data

**Building An English-Vietnamese Bilingual Corpus from Wikipedia** As presented in Chapter 3, I used the Wikpedia database dumps to extract parallel titles, which were updated on *2017-01-20*. After collecting, processing, and aligning sentences in parallel articles, I obtained 408,552 parallel sentences for English-Vietnamese. The extracted corpus are available at the repository of this work.

Table 6.6: **Experimental results of English-Vietnamese using the corpus extracted from Wikipedia** (BLEU); **Wikipedia (NMT)**: the model trained on the extracted corpus from Wikipedia using NMT models; **unconstr+Wikipedia**: the unconstrained data was merged with the *Wikipedia* corpus for the training data; **Stanford**: the Stanford system [49] participated in the IWSLT 2015 shared task, which is the only team using NMT on the English-Vietnamese translation

| System | tst2012 | tst2013 | tst2015 |
|---|---|---|---|
| Wikipedia (SMT) | 18.40 | 22.06 | 20.34 |
| Wikipedia (NMT) | 15.29 | 18.43 | 17.58 |
| unconstrained(SMT) | 34.42 | 27.19 | 25.41 |
| unconstrained(NMT) | 24.05 | 26.71 | 22.30 |
| unconstrained+Wikipedia(SMT) | 33.88 | 27.28 | 26.36 |
| unconstrained+Wikipedia(NMT) | 25.29 | 28.93 | 26.81 |
| Stanford | – | 26.9 | 26.4 |

**Improving SMT and NMT models** I evaluated the extracted bilingual corpus in improving SMT and NMT models. Experimental results are shown in Table 6.6. There are several interesting findings from this experiment. First, although using only the *Wikipedia* corpus to train SMT and NMT models, I obtained promising results: 20.34 BLEU point using SMT and 17.58 BLEU point using NMT on the *tst2015*. Second, when the *Wikipedia* corpus was merged with the *unconstrained* for the training data, both SMT and NMT models achieved the improvement. For the SMT model, the improvement was +0.09 BLEU point on the *tst2013* and +0.95 BLEU point on the *tst2015*. Meanwhile, the NMT model showed the higher improvement with +2.22 BLEU point on the *tst2013* and up to +4.51 BLEU point on the *tst2015*. The next interesting point is that when using the large training data (more than 800k parallel sentences of merging 456k sentences the *unconstrained* with

408k sentences of the *Wikipedia* corpus), the NMT model outperformed the SMT model: 28.93 BLEU point vs. 27.28 BLEU point on the *tst2013*, 26.81 BLEU point vs. 26.36 BLEU point on the *tst2015*.

Additionally, I compared our NMT system which trained on the merged *unconstrained* and *Wikipedia* corpus with the Stanford system [49], the only team using NMT on English-Vietnamese participated in the IWSLT 2015 shared task. Using the *Wikipedia* corpus to enlarge the training data showed the effectiveness when our system achieved the better performance than the Stanford system: 28.93 vs. 26.9 BLEU points on the *tst2013*, and 26.81 vs. 26.4 BLEU points on the *tst2015*. This also showed the effectiveness of the NMT model when enlarging the training data.

## 6.3 A Discussion on Transfer Learning for Low- Resource Neural Machine Translation

Recent development in machine translation shows the promising results and effectiveness of neural-based machine translation in some language pairs with large bilingual data. The problem of limited parallel data in low-resource languages causes a bottleneck for successfully applying neural-based machine translation on such languages. A few work consider designing strategies to apply neural machine translation on low-resource languages. Firat et al., 2016 proposed a method for zero-resource translation [25] and obtained some promising results; however, the method was only tested on a small number of training data of English, French, and Spanish, not on actual low-resource language pairs. Cheng et al., 2017 [12] introduce a joint training algorithm for pivot-based neural machine translation and achieved improvements on several European language pairs of Spanish, English, and French with very large bilingual corpora.

A recent research that can be closedly related to my work is the transfer learning for low-resource neural machine translation introduced in Zoph et al., 2016 [102]. The paper presents a transfer learning method that first used a rich-resource language pairs of French-English to train a neural-based machine translation model called a parent model; then, the parameters of the parent model were used to initialize and constrain for training a neural machine translation model on Uzbek-English, a low-resource language pair. The model achieved a significant improvement. I discuss in this section this transfer learning method and the application as well as the potential of further development based on this method to my work on neural machine translation for low-resource languages.

There are several advantages of the transfer method proposed in [102] that can be suitable, applicable, as well as further improved in my work. First, the method achieved a significant improvement on the actual low-resource language pair. Second, the method is simple but effective that can be easily to apply and develop. Third, one of the characteristic of this method is that the target language of the parent model needs to be the same with the target language of the child model. It opens a potential for my work when I focus on low-resource language pairs, and I can take advantage bilingual corpora of the low-resource language paired with a rich-resource language to train a parent model.

From the discussion on the potential of applying and further extending the transfer learning method for low-resource neural machine translation, I discuss several directions that can be developed in further research. First, instead of using a language pair to train the parent model, I consider utilize a set of language pairs that contain the target language to train a set of parent models, and then join those models to initialize for the child model. This is because bilingual corpora on a set of language pairs for training parent models can be exist, and we can take advantage those resources. Second, the transfer method of [102] focused mainly on transfer the vocabulary of the target language. I consider about transferring not only the target but also the source language. In order to do that, we can used two bilingual corpora of the source and the target language in the child model paired with rich-resource languages to train two parent models. Then, we transfer the vocabulary and parameters from the parent models to the child model with the source and the target sides separately. A joint strategy between the two parent models with the single child model is required to produce an effective transfer result. These strategies can be conducted in further development for my work in future research.

## 6.4 Conclusion

In this chapter, I present some first investigations of utilizing NMT on low-resource language pairs. Recent methods of phrase-based and neural-based have showed the promising directions in the development of machine translation. Neural machine translation models have been applied successfully on several language pairs with large bilingual corpora available. The phrase-based and neural-based methods are also compared and evaluated on some European language pairs. Nevertheless, there is still a bottleneck in SMT and NMT on low-resource language pairs when large bilingual corpora are unavailable. In this work, I conducted a comparison of SMT and NMT methods on several Asian language pairs which contain small bilingual corpora: Japanese-English, Indonesian-Vietnamese, and English-Vietnamese. In addition, a bilingual corpus was extracted from Wikipedia to enhance the machine translation performance and investigate the effects of the extracted corpus on the two machine translation methods. Experimental results showed meaningful findings. For a small bilingual corpus, SMT models showed the better performance than NMT models. Nevertheless, when enlarging the training data with the extracted corpus, both SMT and NMT models were improved, in which NMT models showed the higher improvement and outperformed the SMT models. This work can be useful for further improvement for machine translation on the low-resource languages. Additionally, I discuss a promising method of using transfer learning for low-resource neural machine translation, which is suitable for my current work. Several strategies are discussed for further development using the transfer learning for neural-based machine translation on low-resource languages.

# Chapter 7

# Conclusion

In this dissertation, my goal is to improve machine translation for low-resource languages, in which there are no or small bilingual corpora. Machine translation has a long history in development, and the dominated methods currently in MT are statistical MT and neural MT based on translated texts (bilingual corpora), a trend of data-driven methods to learn translation rules automatically. Although recent methods in MT have shown promising results, and some MT systems can generate increasingly good translation quality, one of the issues in current MT is that there is insufficient training data for most languages in the world exception for several rich languages like English, German, French, Chinese. Improving MT on low-resource languages therefore becomes an essential task currently. I have focused on two main directions: building bilingual corpora to enlarge traing data for SMT models, and exploiting existing bilingual corpora using pivot methods. Another method that utilizes NMT for low-resource languages is also investigated. Chapter 1 - *Introduction* briefly describes the whole story of this dissertation starting from the development process of MT to current methods and locate the problem that requires further investigations and contribution of researchers: *improving MT for low-resource languages*. I list and describe my findings and contributions to solve the problem that I completed for three years working in this topic. The outline of this dissertation is also described to help readers easily capture the structure and information flow presented in this dissertation. In Chapter 2 - *Background*, I provide readers necessary knowledge that help to understand methods as well as terminologies presented in this dissertation. It also aims to provide a brief survey related to my methods to help readers capture more knowledge about the topic.

Chapter 3 - *Building Bilingual Corpora* presents my methods in building bilingual corpora to enlarge training data for SMT models. There are two parts in this chapter: 1) improving sentence alignment by using word similarity learnt from monolingual corpora to deal with the out-of-vocabulary problem and 2) building a multilingual parallel corpus from comparable data. In the first part, word similarities were extracted from monolingual data using word embedding models. The word similarity models were used to enhance informative vocabulary for word alignment, a phase in sentence alignment. This helps to cover more informative vocabulary that reduces OOV ratio and improve sentence alignment. Experimental results on English-Vietnamese showed the contribution of

the proposed method. For the second part, the proposed method was used in building a multilingual parallel corpus among several Southeast Asian languages: Indonesian, Malay, Filipino, and Vietnamese, and between these languages paired with English. A corpus of 900k parallel sentences were extracted from Wikipedia. Experimental results on MT using the extracted corpus present promising results and improvement for the low-resource language pairs.

Chapter 4 - *Pivoting Bilingual Corpora* presents methods in another strategies: exploiting existing bilingual corpora based on pivot methods. Triangulation, the representative approach in pivot methods shows effectiveness in SMT when direct bilingual corpora are unavailable. However, there are several problems of the triangulation that may lack information, which are based on common pivot phrases to connect source phrases to target phrases in source-pivot and pivot-target phrase tables. I propose two methods to overcome the problems. First, semantic similarity was used to connect pivot phrases. The similarity models were based on several approaches such as cosine similarity, longest common subsequence, WordNet, and word embeddings. Experimental results on Japanese-Vietnamese and Southeast Asian language pairs showed the contribution of the proposed method although the method can improve slightly. For the second method, grammatical and morphological information were used to provide more knowledge for pivot connections. Experiments were conducted on Indonesian-Vietnamese, Malay-Vietnamese, and Filipino-Vietnamese that show a significant improvement by 0.5 BLEU points. This indicates the effectiveness of integrating grammatical and morphological information in pivot translation.

Chapter 5 - *A Hybrid Model for SMT on Low-Resource Languages* present my proposed model that combines the two components: the alignment component that was trained from the bilingual data created by the alignment methods described in Chapter 3, the pivot component that was generated by pivot translation. The two components can be combined with the direct component that was trained on any available direct bilingual corpus. I adopted linear interpolation for combining components using two settings: *weights* and *tuning* in which the weights mean the interpolation parameters computed by the BLEU ratio of the components on a test set while the tuning mean the interpolation parameters tuned by using a tuning set. Experiments were conducted on three low-resource language pairs: Japanese-Vietnamese, Southeast Asian languages (Indonesian, Malay, Filipino, Vietnamese), and Turkish-English. Experimental results confirm the effectiveness and contribution of the proposed model when a significant improvement was achieved with +2.0 to +3.0 BLEU points even when there are only small direct bilingual corpora. The hybrid model contributes a solution to improve SMT on low-resource languages.

Chapter 6 - *Neural Machine Translation for Low-Resource Languages* describes my investigations on utilizing NMT for low-resource languages. Although NMT has been successfully applied in several rich languages, there are few work of NMT on low-resource languages. In this chapter, NMT was utilized for low-resource languages such as Japanese-English, Indonesian-Vietnamese, Czech-Vietnamese, English-Vietnamese. A pivot-based method was also conducted on Czech-Vietnamese translation using NMT, in which a pseudo Czech-Vietnamese bilingual corpus was synthesized using NMT models trained

on Czech-English and English-Vietnamese bilingual corpora. The work on this chapter provides empirical investigations of NMT for low-resource languages, which can be used for further improvement.

# Bibliography


[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[2] Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*, 2016.

[3] Lasse Bergroth, Harri Hakonen, and Timo Raita. A survey of longest common subsequence algorithms. In *String Processing and Information Retrieval, 2000. SPIRE 2000. Proceedings. Seventh International Symposium on*, pages 39–48. IEEE, 2000.

[4] Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44. Association for Computational Linguistics, August 2013.

[5] Peter F Brown, Jennifer C Lai, and Robert L Mercer. Aligning sentences in parallel corpora. In *Proceedings of ACL*, pages 169–176. Association for Computational Linguistics, 1991.

[6] Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[7] Chris Callison-Burch, Philipp Koehn, and Miles Osborne. Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 17–24. Association for Computational Linguistics, 2006.

[8] Mauro Cettolo, Nicola Bertoldi, and Marcello Federico. Bootstrapping Arabic-Italian SMT through comparable texts and pivot translation. In *Proceedings of EAMT*, 2011.

[9] Mauro Cettolo, Christian Girardi, and Marcello Federico. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, 2012.

[10] Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. The iwslt 2015 evaluation campaign. *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, 2015.

[11] Stanley F Chen. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of ACL*, pages 9–16. Association for Computational Linguistics, 1993.

[12] Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. Neural machine translation with pivot languages. *arXiv preprint arXiv:1611.04928*, 2016.

[13] Colin Cherry and George Foster. Batch tuning strategies for statistical machine translation. In *Proceedings of HLT/NAACL*, pages 427–436. Association for Computational Linguistics, 2012.

[14] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[15] Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. Integrated parallel sentence and fragment extraction from comparable corpora: A case study on chinese–japanese wikipedia. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 15(2):10:1–10:22, December 2015.

[16] Trevor Cohn and Mirella Lapata. Machine translation by triangulation: making effective use of multi-parallel corpora. In *Proceedings of ACL*, pages 728–735. Association for Computational Linguistics, June 2007.

[17] Raj Dabre, Fabien Cromieres, Sadao Kurohashi, and Pushpak Bhattacharyya. Leveraging small multilingual corpora for smt using many pivot languages. In *Proceedings of HLT/NAACL*, pages 1192–1202. Association for Computational Linguistics, 2015.

[18] Adrià De Gispert and Jose B Marino. Catalan-english statistical machine translation without parallel corpus: bridging through spanish. In *Proceedings of LREC*, pages 65–68. Citeseer, 2006.

[19] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[20] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014.

[21] Rohit Dholakia and Anoop Sarkar. Pivot-based triangulation for low-resource languages. In *Proc. AMTA*, pages 315–328, 2014.

[22] Chris Dyer, Jonathan Weese, Hendra Setiawan, Adam Lopez, Ferhan Ture, Vladimir Eidelman, Juri Ganitkevitch, Phil Blunsom, and Philip Resnik. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations*, pages 7–12. Association for Computational Linguistics, 2010.

[23] Ahmed El Kholy, Nizar Habash, Gregor Leusch, Evgeny Matusov, and Hassan Sawaf. Language independent connectivity strength features for phrase pivot statistical machine translation. In *Proceedings of ACL*, pages 412–418. Association for Computational Linguistics, 2013.

[24] Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. Irstlm: an open source toolkit for handling large scale language models. In *Interspeech*, pages 1618–1621, 2008.

[25] Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T Yarman Vural, and Kyunghyun Cho. Zero-resource translation with multi-lingual neural machine translation. *arXiv preprint arXiv:1606.04164*, 2016.

[26] Philip Gage. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38, 1994.

[27] William A Gale and Kenneth W Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102, 1993.

[28] Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. On using monolingual corpora in neural machine translation. In *CoRR 2015*, 2015.

[29] AnYuan Guo and Hava T Siegelmann. Time-warped longest common subsequence algorithm for music retrieval. In *ISMIR*, 2004.

[30] Thanh-Le Ha, Teresa Herrmann, Jan Niehues, Mohammed Mediani, Eunah Cho, Yuqi Zhang, Isabel Slawik, and Alex Waibel. The kit translation systems for iwslt 2013. In *Proceedings of the International Workshop on Spoken Language Translation*, 2013.

[31] Kenneth Heafield. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics, 2011.

[32] Duc Tam Hoang and Ondřej Bojar. Tmtriangulate: A tool for phrase table triangulation. *The Prague Bulletin of Mathematical Linguistics*, 104(1):75–86, 2015.

[33] William John Hutchins and Harold L Somers. *An introduction to machine translation*, volume 362. Academic Press London, 1992.

[34] Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. Montreal neural machine translation systems for wmt'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, pages 134–140, 2015.

[35] Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. Is neural machine translation ready for deployment? a case study on 30 translation directions. *arXiv preprint arXiv:1610.01108*, 2016.

[36] Martin Kay and Martin Röscheisen. Text-translation alignment. *Computational Linguistics*, 19(1):121–142, 1993.

[37] Sungchul Kim, Kristina Toutanova, and Hwanjo Yu. Multilingual named entity recognition using parallel data and metadata from wikipedia. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 694–702. Association for Computational Linguistics, 2012.

[38] Adam Pauls Dan Klein. Faster and smaller n-gram language models. *Proceeding HLT*, 11.

[39] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395, 2004.

[40] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, Phuket, Thailand, September 2005.

[41] Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 462 machine translation systems for europe. In *Proceedings of the MT Summit XII*. International Association for Machine Translation, 2009.

[42] Philipp Koehn and Hieu Hoang. Factored translation models. In *EMNLP-CoNLL*, pages 868–876, 2007.

[43] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*, pages 177–180. Association for Computational Linguistics, 2007.

[44] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of HLT/NAACL*, pages 48–54. Association for Computational Linguistics, 2003.

[45] Philipp Koehn and Josh Schroeder. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the second workshop on statistical machine translation*, pages 224–227. Association for Computational Linguistics, 2007.

[46] Bo Li and Juan Liu. Mining Chinese-English parallel corpora from the web. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP)*, 2008.

[47] Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren NG Thornton, Jonathan Weese, and Omar F Zaidan. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139. Association for Computational Linguistics, 2009.

[48] George S Lueker. Improved bounds on the average length of longest common subsequences. *Journal of the ACM (JACM)*, 56(3):17, 2009.

[49] Minh-Thang Luong and Christopher D Manning. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, 2015.

[50] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421, 2015.

[51] Xiaoyi Ma. Champollion: A robust parallel text sentence aligner. In *Proceedings of LREC*, pages 489–492, 2006.

[52] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60, 2014.

[53] José B Marino, Rafael E Banchs, Josep M Crego, Adrià de Gispert, Patrik Lambert, José AR Fonollosa, and Marta R Costa-Jussà. N-gram-based machine translation. *Computational Linguistics*, 32(4):527–549, 2006.

[54] Luis Marujo, Nuno Grazina, Tiago Luis, Wang Ling, Luisa Coheur, and Isabel Trancoso. BP2EP - adaptation of Brazilian Portuguese texts to European Portuguese. In *Proceedings of EAMT*, pages 129–136, 2011.

[55] I Dan Melamed. A geometric approach to mapping bitext correspondence. In *Proceedings EMNLP*. Association for Computational Linguistics, 1996.

[56] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[57] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[58] Akiva Miura, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Improving pivot translation by remembering the pivot. In *ACL (2)*, pages 573–577, 2015.

[59] Robert C Moore. *Fast and accurate sentence alignment of bilingual corpora*. Springer, 2002.

[60] Graham Neubig. The Kyoto free translation task. http://www.phontron.com/kftt, 2011.

[61] Graham Neubig. Travatar: A forest-to-string machine translation engine based on tree transducers. In *ACL (Conference System Demonstrations)*, pages 91–96, 2013.

[62] Quoc Hung Ngo, Werner Winiwarter, and Bartholomäus Wloka. Evbcorpus-a multi-layer english-vietnamese bilingual corpus for studying tasks in comparative linguistics. In *Proceedings of the 11th Workshop on Asian Language Resources (11th ALR within the IJCNLP2013)*, pages 1–9, 2013.

[63] Hieu Nguyen and Li Bai. Cosine similarity metric learning for face verification. *Computer Vision–ACCV 2010*, pages 709–720, 2011.

[64] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics, 2003.

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

[66] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318. Association for Computational Linguistics, 2002.

[67] Philip Resnik. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics (ACL)*, 1999.

[68] Gerard Salton. Automatic text analysis. *Science*, 168(3929):335–343, 1970.

[69] Charles Schafer and David Yarowsky. Inducing translation lexicons via diverse similarity measures and bridge languages. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–7. Association for Computational Linguistics, 2002.

[70] Rico Sennrich. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of EAMT*, pages 539–549, 2012.

[71] Rico Sennrich, Barry Haddow, and Alexandra Birch. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation (WMT)*, 2016.

[72] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

[73] Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18(3):491–504, 2014.

[74] Anil Kumar Singh and Samar Husain. Comparison, selection and use of sentence alignment algorithms for new language pairs. In *Proceedings of the ACL Workshop on Building and using Parallel texts*, pages 99–106. Association for Computational Linguistics, 2005.

[75] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, 2006.

[76] Dan Ştefănescu and Radu Ion. Parallel-wiki: A collection of parallel sentences extracted from wikipedia. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2013)*, pages 24–30, 2013.

[77] Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*, 2006.

[78] Andreas Stolcke et al. Srilm-an extensible language modeling toolkit. In *Interspeech*, volume 2002, page 2002, 2002.

[79] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems (NIPS)*, pages 3104–3112, 2014.

[80] Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. Introducing the asian language treebank (alt). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 1574–1578, 2016.

[81] Jörg Tiedemann. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria, 2009.

[82] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *LREC*, volume 2012, pages 2214–2218, 2012.

[83] Hai-Long Trieu, Thanh-Quyen Dang, Phuong-Thai Nguyen, and Le-Minh Nguyen. The jaist-uet-miti machine translation systems for iwslt 2015. In *Proceedings of The 12th International Workshop on Spoken Language Translation (IWSLT)*, 2015.

[84] Hai-Long Trieu and Le-Minh Nguyen. Applying semantic similarity to phrase pivot translation. In *Proceedings of The 28th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2016.

[85] Hai-Long Trieu and Le-Minh Nguyen. Enhancing pivot translation using grammatical and morphological information. In *Proceedings of The 15th International Conference of the Pacific Association for Computational Linguistics (PACLING)*, 2017.

[86] Hai-Long Trieu and Le-Minh Nguyen. Investigating phrase-based and neural-based machine translation on low-resource settings. In *The 31st Pacific Asia Conference on Language, Information and Computation*, 2017.

[87] Hai-Long Trieu and Le-Minh Nguyen. A multilingual parallel corpus for improving machine translation on southeast asian languages. In *Proceedings of The 16th Machine Translation Summit (MTSummit XVI)*, 2017.

[88] Hai-Long Trieu, Le-Minh Nguyen, and Phuong-Thai Nguyen. Dealing with out-of-vocabulary problem in sentence alignment using word similarity. In *Proceedings of The 30th Pacific Asia Conference on Language, Information and Computation (PACLIC 30)*, 2016.

[89] Hai-Long Trieu, Trung-Tin Pham, and Le-Minh Nguyen. The jaist machine translation systems for wmt 17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 405–409, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[90] Masao Utiyama and Hitoshi Isahara. Reliable measures for aligning Japanese-English news articles and sentences. In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 72–79, 2003.

[91] Masao Utiyama and Hitoshi Isahara. A comparison of pivot methods for phrase-based statistical machine translation. In *Proceedings of HLT/NAACL*, pages 484–491. Association for Computational Linguistics, April 2007.

[92] Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. Parallel corpora for medium density languages. *Amsterdam studies in the theory and history of linguistic science series 4*, 292:247, 2007.

[93] Jean Véronis and Philippe Langlais. Evaluation of parallel text alignment systems. In *Parallel text processing*, pages 369–388. Springer, 2000.

[94] Haifeng Wang, Hua Wu, and Zhanyi Liu. Word alignment for languages with scarce resources using bilingual corpora of other language pairs. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 874–881. Association for Computational Linguistics, 2006.

[95] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[96] Krzysztof Wołk and Krzysztof Marasek. Pjait systems for the iwslt 2015 evaluation campaign enhanced by comparable corpora. In *Proceedings of the International Workshop on Spoken Language Translation*, 2015.

[97] Dekai Wu. Aligning a parallel english-chinese corpus statistically with lexical criteria. In *Proceedings ACL*, pages 80–87. Association for Computational Linguistics, 1994.

[98] Hua Wu and Haifeng Wang. Pivot language approach for phrase-based statistical machine translation. In *Proceedings of ACL*, pages 856–863. Association for Computational Linguistics, June 2007.

[99] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *CoRR*, 2012.

[100] Xiaoning Zhu, Zhongjun He, Hua Wu, Haifeng Wang, Conghui Zhu, and Tiejun Zhao. Improving pivot-based statistical machine translation using random walk. In *Proceedings of EMNLP*, pages 524–534. Association for Computational Linguistics, October 2013.

[101] Xiaoning Zhu, Zhongjun He, Hua Wu, Conghui Zhu, Haifeng Wang, and Tiejun Zhao. Improving pivot-based statistical machine translation by pivoting the co-occurrence count of phrase pairs. In *Proceedings of EMNLP*, pages 1665–1675. Association for Computational Linguistics, 2014.

[102] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*, 2016.

# Publications

## JOURNALS

[1] Long Hai Trieu, Vu Duc Tran, Ashwin Ittoo, Minh Le Nguyen, **Leveraging Additional Resources for Improving Machine Translation on Asian Low-Resource Languages**, ACM Transactions on Asian and Low-Resource Language Information Processing **(revised)**

[2] Long Hai Trieu, Thai Phuong Nguyen, Minh Le Nguyen, **A New Feature to Improve Moore's Sentence Alignment Method**, VNU Journal of Science: Computer Science and Communication Engineering 31, no. 1, 2015

## INTERNATIONAL CONFERENCES

[1] Long Hai Trieu, Minh Le Nguyen, **Investigating Phrase-Based and Neural-Based Machine Translation on Low-Resource Settings**, in The 31st Pacific Asia Conference on Language, Information and Computation, 2017

[2] Long Hai Trieu, Minh Le Nguyen, **A Multilingual Parallel Corpus for Improving Machine Translation on Southeast Asian Languages**, in Proceedings of the Machine Translation Summit XVI, 2017

[3] Long Hai Trieu, Tin Trung Pham, Minh Le Nguyen, **The JAIST Machine Translation Systems for WMT 17**, in Proceedings of Second Conference on Machine Translation (WMT17), 2017

[4] Long Hai Trieu, Minh Le Nguyen, **Enhancing Pivot Translation Using Grammatical and Morphological Information**, the 2017 Conference of the Pacific Association for Computational Linguistics, 2017

[5] Long Hai Trieu, Minh Le Nguyen, **Applying Semantic Similarity to Phrase Pivot Translation**, in Proceedings of the 28th IEEE International Conference on Tools with Artificial Intelligence, 2016

[6] Long Hai Trieu, Thai Phuong Nguyen, Minh Le Nguyen, **Dealing with Out-Of-Vocabulary Problem in Sentence Alignment Using Word Similarity**, in Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation, 2016

[7] Long Hai Trieu, Quyen Thanh Dang, Thai Phuong Nguyen, Minh Le Nguyen, **The JAIST-UET-MITI Machine Translation Systems for IWSLT 2015**, in Proceedings of the 12th International Workshop on Spoken Language Translation, 2015

**INTERNATIONAL CONFERENCES (NOT RELATED TO THE DISSERTATION)**

[1] Vu Duc Tran, Anh Viet Phan, Long Hai Trieu, **An Approach for Retrieving Legal Texts**, in Proceedings of the Ninth International Workshop on Juris-informatics (JURISIN 2015)

[2] Son Truong Nguyen, Anh Viet Phan, Huy Thanh Nguyen, Long Hai Trieu, Phuong Ngoc Chau, Tin Trung Pham, Minh Le Nguyen, **Legal Information Extraction/Entailment Using SVM-Ranking and Tree-based Convolutional Neural Network**, in Proceedings of the Tenth International Workshop on Juris-informatics (JURISIN 2016)

[3] Long Hai Trieu, Hiroyuki Iida, Nhien Bao Hoang Pham, Minh Le Nguyen, **Towards Developing Dialogue Systems with Entertaining Conversations**, in Proceedings of the 9th International Conference on Agents and Artificial Intelligence (ICAART 2017)