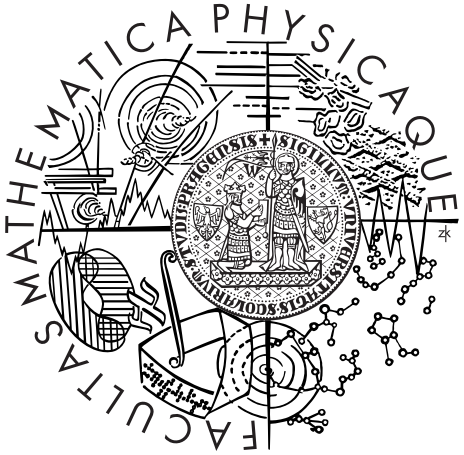
Charles University in Prague
Faculty of Mathematics and Physics

MASTER THESIS



Hoàng Đức Tâm
Duc Tam Hoang

Pivoting Machine Translation for Vietnamese

Institute of Formal and Applied Linguistics

Supervisor of the master thesis: RNDr. Ondřej Bojar, Ph.D.
Study programme: Master of Computer Science
Specialization: Mathematical Linguistics

Prague 2015

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In Prague date July 28, 2015 signature of the author

First and foremost, I want to express my gratitude to my advisor RNDr. Ondřej Bojar, Ph.D, who entrusted me with challenging projects and served as a guiding beacon for over 15 months.

I would like to thank all my friends and colleagues at Institute of Formal and Applied Linguistics for providing me with support. I would especially like to thank Aleš Tamchyna and Vendulka Michlíková, for their invaluable advices in making this thesis.

A special dedication to my family for their unending support throughout my studies.

Název práce: Strojový překlad pro vietnamštinu s pivotním jazykem

Autor: Duc Tam Hoang

Katedra: Ústav Formální a Aplikované Lingvistiky

Vedoucí diplomové práce: RNDr. Ondřej Bojar, Ph.D

Abstrakt: Čeština a vietnamština jsou úředními jazyky České republiky a Vietnamu. Vzhledem k charakteristickým rysům těchto dvou jazyků a nedostatku zdrojů dat je česko-vietnamský strojový překlad velmi náročnou úlohou a překladový nástroj zaměřený speciálně na tento konkrétní jazykový pár nebyl dosud vyvinut. V této práci budujeme statistické překladové systémy pro překlad mezi češtinou a vietnamštinou a zkoumáme možnosti vylepšení kvality překladu pomocí překladu přes pivotní jazyk. Využití pivotního jazyka (jiného přirozeného jazyka) dává možnost zapojit do systému dodatečné jazykové zdroje. Jako pivotní jazyk jsme pro česko-vietnamský překlad vybrali angličtinu a připravili jsme trénovací a testovací korpus pro tyto tři jazykové páry. Pro každý jazykový pár kombinujeme různé zdroje korpusů a podle potřeby zlepšujeme jejich kvalitu pomocí normalizování a filtrování. S metodami překladu přes pivotní jazyk jsme provedli množství experimentů a analyzovali jsme je v realistických podmínkách.

Klíčová slova: statistický strojový překlad, metody překladu přes pivotní jazyk, kaskády systémů, triangulace frázové tabulky

Title: Pivoting Machine Translation for Vietnamese

Author: Duc Tam Hoang

Department: Institute of Formal and Applied Linguistics

Supervisor: RNDr. Ondřej Bojar, Ph.D

Abstract: Czech and Vietnamese are the national languages of the Czech Republic and Vietnam, respectively. The distinctive features and the shortage of resources renders Czech-Vietnamese machine translation into a difficult task, leading to the fact that no effort has been put into developing a translation tool specifically for the language pair. In this thesis, we develop phrase-based statistical machine translation systems for the language pair and investigate the potential to improve the translation quality with pivoting. Pivoting refers to a set of machine translation approaches through which a natural language, called pivoting language, is introduced to solve the problem of data scarcity between source and target languages, one of the most challenging problems of statistical machine translation. Selecting English as the sole pivoting language for Czech-Vietnamese translation, we prepare training and testing corpora for the three language pairs. All possible corpus sources are explored regarding each specific language pair. The next step is to improve quality of the training corpora through normalizing and filtering. Various experiments with pivoting methods are carried out to analyse the performance of pivoting methods in a realistic working condition.

Keywords: Statistical Machine Translation, pivoting methods, system cascades, phrase table triangulation

Table of Contents

1. Introduction

1.1 Motivation of the Thesis

Statistical Machine Translation (SMT) is generally considered the most versatile approach in Machine Translation (MT). Training algorithms of SMT rely on a parallel corpus between the source language and the target language. Therefore, the performance of SMT approaches relies heavily on the quantity and quality of direct bilingual corpora. SMT for under-resourced language pairs is still at a deadlock due to the lack of bilingual corpora. Among thousands of living languages in the world[1], most pairs of languages are resource-poor. Hence, most language pairs cannot benefit from the conventional SMT approaches.

Pivoting method in the past few years has gained serious attention as an alternative method to conduct translation between two languages. The method involves the introduction of one or more natural languages. The added languages lead to the arrival of resources which they share with both the source language and the target language, potentially improving the translation quality. Despite this, the success of most pivoting methods has been reported on the usage of multi-lingual corpora rather than a realistic condition in which the source-pivot resources and the pivot-target resources are unrelated.

In the bag of living languages, Czech and Vietnamese are national languages of the Czech Republic and Vietnam, respectively. They are not under-resourced languages themselves, but the amount of bilingual corpora between the two languages is limited. Despite the fact that there is a large Vietnamese community living in the Czech Republic, recognized as the third largest minority, no effort has been put into developing an MT tool specifically for this language pair. The distinctive features and the lack of resources between the two languages have rendered Czech↔Vietnamese translation into a difficult, yet interesting, problem.

Therefore, in our perspective, it is a high time to investigate the potential SMT methods on an under-resourced language pair, which we have chosen Czech and *Vietnamese*. Combining pivoting MT methods and standard MT methods may pave the way for a good MT system. This can benefit a large group of people as the need for Czech↔Vietnamese translation has been growing rapidly.

1.2 Goals of the Thesis

In a nutshell, the objective of this work is to provide machine translation systems for Czech↔Vietnamese translation. It aims to attain the following goals:

- **Collection of Parallel and Monolingual Corpora**: Corpus is the starting point for every SMT system. For this study, we reuse existing corpora for

[1] Apparently, researchers have not agreed upon the precise number of languages out there

Czech-English and *Vietnamese-English* and we collect *Vietnamese-Czech* bilingual corpora, *Vietnamese-English* bilingual corpora and Vietnamese monolingual corpora. Cleaning the corpora is also an important part of this work.

- **Pivoting methods**: The difference in linguistic features and the lack of resources lead to an anticipation that the direct Czech↔Vietnamese SMT system will not be sufficient. The goal of this thesis is to improve the Czech↔Vietnamese translation with the help of pivoting methods.

In trying to achieve the aforementioned goals, we wish to see how well SMT methods can be used to perform the translation, and whether there is any possibility of utilizing all available methods within a system combination. Any improvement upon the fusion is a step forward in the right direction and becomes the first base for future works.

1.3 Structure of the Thesis

Apart from this introduction, the rest of the thesis is arranged as follows:

- Chapter 2 gives an overview of machine translation. It discusses the background of statistical machine translation in Section 2.1. The chapter further discusses previous studies related to the topic of this thesis in Section 2.2. Besides, a description of Czech and Vietnamese is given in Section 2.3
- Chapter 3 describes the processes of collecting training corpora. It highlights the effort spent on searching for resources. The chapter contains 3 sections, namely 3.1, 3.3 and 3.2, for bilingual corpora and one section, namely 3.4, for monolingual corpora.
- Chapter 4 describes a preparation phase to improve the corpora quality. It contains two sections: 4.1 and 4.2, associated with two cleaning techniques: *normalizing* and *filtering*, respectively.
- Chapter 5 contains the discussion of potential pivoting methods which could be applied in the thesis. It emphasizes detailed scenarios regarding each methodology.
- Chapter 6 contains the description and discussion of the experiments that we conducted. The chapter outlines the settings and the tools in Section 6.1. Afterwards, it presents the detailed set up, results and discussions of each and every methodology in Sections 6.2, 6.3, 6.4, 6.5 and 5.5. In the discussion, we also recall some key concepts which are mentioned in previous Chapters.
- Chapter 7 contains the overall discussion and conclusion, highlighting the contribution which we have made.

2. Background

Machine Translation (MT) is a sub-field of Natural Language Processing (NLP) dedicated to the study of computerized approaches to translating one natural language to another natural language. It has been one of the earliest and most active areas in NLP research since 1700s. A typical MT problem includes transforming a piece of information (text or speech) in one language into a different piece of information (text or speech, respectively) in another language and preserving the original meaning. The two languages are generally called *source language* and *target language*, respectively. MT approaches, which have been proposed for MT problems, can be broadly divided into three main categories: *rule-based* [9] [10] [7], *example-based* [23] and *statistical* [3] [13]. The last two approaches are sometimes called *data-driven* approaches.

Rule-based approaches (sometimes knowledge-based [18]) was introduced for early MT systems. An MT system which is the result of rule-based approaches is generally called *rule-based system*. A classical rule-based system is founded on a set of transformation rules written by human experts. The set of rules must carry a deep linguistic analysis of the source language and the target language. Its quantity and quality, which represents the linguistic knowledge known by the system, plays a critical roles in the system performance. The general translation process of rule-based approaches can be broken down into three steps: *analysis*, *transfer* and *generation*. Firstly, an input in the source language is analysed using linguistic tools. The tools consist of parsers, morphological tools and/or tokenizer depending on how the set of rules is designed. Secondly, the analysed input gets transformed into an intermediate representation by the set of rules. Finally, the intermediate representation gets analysed to create a fluent output. Numerous rule-based approaches have been introduced, which sometimes change the translation process. However, as the scope of this thesis does not lie in the area of rule-based approaches, we provide no further discussion surrounding this specific matter.

Example-based and *statistical* approaches, especially the latter, have gained attention over recent years. They develop a generalized model based on text corpora. The model is derived from actual examples, which are provided by the corpora without significant linguistic analysis. Hence, the MT system performance depends heavily on quality and quantity of the corpora. Notably, this is the reason why these approaches are called *data-driven* approaches. The data driven approaches have an advantage of portability. A system could be re-used for many pairs of languages, provided that training data is sufficient. The difference between *example-based* approaches and *statistical* approaches is the method for handling actual examples. The basic notion of *example-based* approaches is to conduct translation by analogy while *statistical* approaches rely on Bayes rule and statistical theory to maximize probability of the output. Furthermore, *Example-based* approaches prefer to preserve whole sentences while *statistical* approaches breaks everything into phrases or words.

The main objective of this study falls within the statistical approaches to Machine

Translation. Hence, the following section focuses on discussing the background of SMT and its approaches. At the moment, SMT is generally considered the most versatile approach in MT. It has a number of advantages over other translation approaches. The performance of system gets improved by adding further data, which is simple compared to controlling a huge set of hand-crafted rules.

2.1 Background of Statistical Machine Translation

In Statistical Machine Translation(SMT), the problem of translating a sentence from source language F to target language E can be interpreted as a problem of Noisy Channel Model [22]. The symbols E and F come from the first introduction of the model when the translation is from French to English. The noisy channel can be explained as follows:

Assume that all people think in E, but when they express their thoughts, they produce them in F. The task is to find out the original meaning. In other words, to find out the most probable sentence out of all sentences in language E given an input sentence in F.

Therefore, an SMT task could be formalized as follows: Given a sentence f of language F, which need to be translated into sentence e of language E. We represent f and e as segments of words, $f_1^J = f_1...f_j...f_J$ and $e_1^I = e_1...e_i...e_I$. Finding the best translation $\hat{e}_1^I$ for sentence f is equivalent to maximizing the probability of e if we know f:

$$\hat{e}_1^I = argmax(P(e_1^I|f_1^J)) \tag{2.1}$$

The Equation 2.1 is where Bayes Rule comes to play. Any given conditional probability *P(e/f)* can be expressed using Bayes theorem as:

$$P(e_1^I|f_1^J) = \frac{P(f_1^J|e_1^I) \times P(e_1^I)}{P(f_1^J)} \tag{2.2}$$

For one sentence *f*, the maximization of $P(e_1^I|f_1^J)$ in Equation 2.1 is equivalent to the following Equation:

$$\hat{e}_1^I = argmax(P(f_1^J|e_1^I) \times P(e_1^I))) \tag{2.3}$$

In Equation 2.3, $P(f_1^J|e_1^I)$ is called **Translation Model** while $P(e_1^I))$ is called **Language Model**.

Translation Model

For the translation model, a number of approaches have been introduced. The most well-known models are *word-based*, *phrase-based* and *tree-based*.

Word-Based Translation Model

The word-based translation model started from the work on SMT at IBM Candide Project in the late 1980s and early 1990s [13]. The early model for machine translation was based solely on lexical translation, which is the translation of words in isolation. A word was translated by looking it up in a bilingual dictionary. Most words had multiple translations, of which some are more frequent than others. A translation model needs to estimate the lexical translation probability distribution of those translations.

Formally, the problem is to find a function p_f which takes a foreign word w_f and returns a probability for each choice of translation w_e. In SMT, this function is derived by creating a corpus and collecting some statistics over the corpus. A straightforward way is to use the ratio of counts, like in equation 2.4. This method of obtaining a probability distribution from data also satisfies two conditions: probability value belongs to *[0,1]* and summation of probability is *1*. This type of estimation is also called *maximum likelihood estimation* (MLE).

$$p(w_e|w_f) = \frac{count(w_e, w_f)}{count(w_f)} \tag{2.4}$$

The probability distribution lays a foundation to the word-based model of SMT. The conditional probability $p(w_e|w_f)$ is called the translation probability. The model is later processed to translate word-by-word from a sentence into another sentence. This translation assembles an alignment, which is a mapping from the source words to target words. An example of the alignment between the input (source) words and output (target) words can be illustrated as in Table 2.3.

The alignment can be formalized with an alignment function a, which maps input words' positions to output words' position. In Example 2.3, two alignment functions are:

$$a_{VI \to EN} : 1 \to 1, 2 \to 2, 3 \to 3 \tag{2.5}$$
$$a_{EN \to CS} : 1 \to 1, 2 \to 1, 3 \to 2 \tag{2.6}$$

The alignment may be very simple like $a_{VI \to EN}$, where the source word and target counterparts are in exactly the same order. It may be more complicated like $a_{EN \to CS}$, where multiple input words are translated into one output word. The alignment may also be more complicated due to different word order or the number of necessary to express the same concept in different languages. Sometimes, there is no clear equivalent between source words and target words. This problem

is solved by introducing a special *NULL* token, which is aligned to all unknown words.

Overall, the following features affect the alignment model:

- Reordering: Words may be re-ordered during translation.
- One-to-Many Translation: One source word may be translated into multiple target words.
- Dropping Words: Source words may be dropped during translation.
- Inserting Words: Words may be added during translation.

There are a number of options to model the translation probability [19], including popular statistical translation model *IBM-1* to *IBM-5* [2] and Hidden-Markov model. These models are distinguish in translation model but they share the features of single-words lexicon probability.

Phrase-Based Translation Model

While word-based models translate a sentence with words as atomic units, phrase-based models translate phrase as atomic units [13]. One major disadvantage of the word-based model is that contextual information is not taken into account. In many circumstances, the translation depends heavily on surrounding words. This problem is handled solely by the language model rather than the word-based translation model. However, the language model sometimes fails. The phrase-based model offers a way to incorporate the context by learning translation for whole phrases instead of single words. A phrase is simply a sequence of adjacent words, in contrast to the linguistic definition of phrase. It is generally believe that a phrase is more informative than a bag of the same words.

Phrase-based translation model considers a sentence as a sequence of phrases rather than a sequence of words [13]. This brings a number of advantages over the word-based model. First, it solves the problem of non-compositional phrasemes translation. Second, phrases provide a local context in translation. Third, when the data gets larger, the longer phrases can be learned. In an ideal situation, the whole sentence is considered as one phrase and the problem of translation is solved by mapping techniques.

The overall process of translation by Phrase-based model includes segmenting the source input into phrases, translating each phrases into target language and reordering the output phrases.

Language Model

Language model is the probability distribution of a string $S = w_1 w_2 ... w_n = w_1^n$. It reflects how often a string S of words occurs in the given language. There are multiple ways to define a language model, including *word sequences, part-of-speech sequences, tree structures, etc.* In this section, we focus on the *n-gram*

language model over the word sequences [22], which is the model being used in our MT system.

An *n-gram language model* defines the probability distribution by predicting the next word if we know the preceding words [16]. It is the task to estimate the conditional probability: $P(w_n|w_1w_2...w_{n-1})$

Using the chain rule, the probability over a sequence $S = w_1^n$ could be decomposed as follows:

$$P(w_1^n) = P(w_n|w_1w_2...w_{n-1}) \times P(w_{n-1}|w_1w_2...w_{n-2}) \times \ldots \times P(w_2|w_1) \times P(w_1) \tag{2.7}$$

$$= \prod_{i=1}^{n} P(w_i|w_1^{i-1}) \tag{2.8}$$

Using the Markov assumption, the current word is computed only by prior local context, which consists of the last few words. This assumption reduces the problem of computing the whole probability of a word given a long list of preceding words (which is not a trivial task). If we assume that the probability of the next word depends on k preceding words, then the task is more straightforward. In fact, an *n-gram* is a Markov Model with $k = n - 1$. For example, with $k = 1$, the Markov model is called bigram model. It takes only one previous word into account. In this case, the Equation 2.8 is as follows:

$$P(w_1^n) = \prod_{i=1}^{n} P(w_i|w_{i-1}) \tag{2.9}$$

At this point, the remaining task is to compute $P(w_i|w_{i-1})$ for a specific value of i. This is resolved by the Maximum Likelihood Estimation (MLE). This technique estimates the probability by counting the frequencies of the pair (w_{i-1}, w_i), shown in Equation 2.10

$$P(w_i|w_{i-1}) = \frac{count(w_{i-1}, w_i)}{\sum_j count(w_{i-1}, w_j)} \tag{2.10}$$

According to Equation 2.9, the n-gram language model is the product of many fractional conditional probabilities $P(w_i|w_{i-1})$. In case there is no observation of at least one pair (w_{i-1}, w_i), the language model will be given a probability of 0. This is called the problem of unseen data. There are also problems of balance weight between *infrequent n-grams* and *frequent n-grams*. The solution for such problems are called *smoothing techniques*.

Smoothing technique decreases the probability of seen events and assigns the leftover probability mass to the unseen events. There are a number of smoothing techniques such as *add 1*, *add less than 1*, *Good Turning estimate*, *linear interpolation*, etc [31].

Another problem is to select the order of the model, i.e the number n. Training lower order language models like unigram, bigram and trigram is criticized as loss of information because it uses limited history. Training higher order language on the other hand reveals the problem of data sparseness and memory demands. Hence, there is a trade off when we choose the order of the language model.

2.2 Related Works

The scope of this thesis contains topics from several sub-fields of statistical machine translation. This section is devoted to describe the works of other people, which either implicitly or explicitly falls in the same category as our thesis. Notably, there are two topics that we cover in the literature review: *Machine Translation for Vietnamese* and *Machine Translation via pivot languages*.

Beside MT, the utility of pivot language has been found among other applications in other natural language processing areas. For example, it plays an important role in cross-lingual information retrieval [33]. However, we do not relate to those usages in our thesis. Furthermore, the term *pivoting method* should be disambiguated from the *pivot* method in information retrieval, which aims to balance the score in the information retrieval language model by integrating document lengths into the document score.

2.2.1 Machine Translation for Vietnamese-Czech

There has been surprisingly little work to date in MT for Vietnamese, especially *Vietnamese-Czech*. The most well-known MT system available for this pair is *Google Translate* [1]. Vietnamese was added into Google Translate at the 11^{st} stage in 2008, the same year as Czech. According to *Wikipedia* [2], Vietnamese and Czech are among groups of language pairs which do not have a direct translation.

Preliminary experiments have shown that available translation services between Vietnamese and Czech are still relatively poor. Their translations involves a wide range of errors, including word choices, word order, valency, et cetera.

2.2.2 Machine Translation via Pivoting Methods

The term *pivot method* is widely used when addressing MT between two natural languages (source language and target language) with the support of other natural

[1] https://translate.google.com/
[2] http://en.wikipedia.org/wiki/Google_Translate

languages, which are called *pivot languages*, *third languages* or *bridge languages*.

Methods

A number of techniques have been developed for pivoting methods. They could be roughly classified into different groups according to the MT internals interacting with the technique.

A trivial solution was to translate the input sentence from source language into the pivot language, then translate the output into the target language. This method was referred as *system cascades*, which name indicated an involvement of more than one MT system. First, a *source-pivot* translation system translated the input from *source* language to *pivot* language. Then, the output was translated by a *pivot-target* translation system. The *system cascades* method was also called *sentence translation*, as opposed to *phrase translation* method [25]. The setting of system cascades could be improved by using n-best translation instead of taking only the best hypothesis. A set of experiments, conducted by Utiyama and colleagues [25], over the translation from *German* to *French* via *English* within the Europarl corpora [12] showed that using *15* hypotheses in the pivot language was generally better than using *1* hypothesis.

The paper [25] also showed the superior of *phrase translation*, compared to *sentence translation*. The method of *phrase translation* merged two phrase tables, *source-pivot* and *pivot-target*, into one *source-target* phrase table. The idea was to join two phrases of source language and target language as the translation of each other if and only if at least a pivot phrase which is paired with both of them existed. A newly created phrase table from *source* language to *target* language was then used to built the new MT system. In 2007, Cohn and Lapata [32] published a method to estimate the feature weights of the new phrase table, which was later used in other works [26] [25]. The paper [32] showed that *triangulation* alone was not as good as the direct translation but a combination of triangulation model and standard model often outperformed the two sub-models. The method of *phrase translation* was widely recognized as *phrase table triangulation*, which name indicated that two phrase tables were triangulated to establish the new phrase table.

Another method to utilize the feature weights of the new phrase table approximated the co-occurrence count of phrase pairs [27]. The co-occurrence counts of the final phrase table were the minimum values of the co-occurrence counts of two phrase tables *source-pivot* and *pivot-target*. Subsequently, the features of pivoting phrase table were computed by the standard phrase extraction method. Their experiment on *Europarl* [12] showed that taking the *minimum* value outperformed other options, including *maximum* and *mean*.

Corpus synthesis was another pivoting method, which creates a synthetic corpus by translating one part of the parallel corpora into another language. If a *pivot-target* translation system was available, one option was to translate the pivot side of a *source-pivot* corpus into the *target* language and create a synthetic *source-target* corpus. The reverse order was also feasible. Translating the pivot side of a *pivot-target* corpus by a *pivot-source* translation system also resulted in a *source-target* corpus.

Česílko [34], a rule-based Czech→Slovak MT system, was utilized to support the translation between English and Slovak [5]. The results showed that the synthetic corpus outperformed both the *direct system* and the *system cascades*.

Bertoldi [1] proposed an approach which obtains n best translated sentences for the new corpus rather than using just the best hypothesis. They reported that the n best translations was expensive but effective against the phrase table approach

Corpus Status

A large number of works on pivoting methods were associated with the potential usage of **multi-parallel corpora** such as Europarl [12]. It was different from another working condition that two **independent corpora**, *source-pivot* corpus and *pivot-target* corpus, were drawn from different sources. Bertoldi [1] put forward a distinction between two problems. If the *source-pivot* corpus and the *pivot-target* were derived from the same set of sentences, or they have a large overlapping chunk, the phrase table manipulation techniques outperformed the system cascades method.

Within the scope of multi-parallel corpora, pivot language could be a factor to prune the conventional method. A potential utility of pivot language was to disambiguate the alignment between source words and target words [15]. The motivation was to use a third language in a multi-lingual corpus to disambiguate the alignment between the source language and the target language. Besides, the third language can be used to filter the less probable phrase pairs from a phrase table [4]. This required two phrase tables: *source-target* direct phrase table and *source-target* triangulation phrase table. A phrase pair in the former phrase table was retained if and only if it appeared in both the direct phrase table and the triangulation phrase table. The pruning technique could be drastic by taking into account only exact phrases in the third language or be tolerant by taking into account the words within the phrases of third language. Finally, the technique was reported to reduce 70% of the phrase table without harming the translation quality.

Pivot Choice

Using more than one pivot language, *ensemble decoding* [20] is introduced to combine multiple pivoting systems. They put forward a strong conclusion over the *Europarl* [12] dataset that the ensemble system outperformed the baseline of direct translation. They drew a conclusion that the more two corpora were overlapping, the higher coverage the triangulation phrase table had. Moreover, the pivot language performed better if it was similar to either the source language or the target language. The remaining problem of this approach was an efficient algorithm to be practical as all of their existing experiments had to be conducted on a small size of corpus and phrase tables.

2.3 Language Features

We have selected Vietnamese and Czech for this study. Before we go into details about the MT task, including corpora, approaches and experiments, this section provides a quick preview of the two languages.

Czech is the official language of the Czech Republic. Besides, native Czech speakers live in the other European countries, especially in Slovakia, and tens of thousands of Czech speakers live in the U.S.A., Canada and Australia. Czech has over 10 million speakers. [8] Along with Slovak, Polish and the High and Low Sorbian, Czech belongs to the western group of Slavic languages. Specifically, Czech belongs to the eastern division of Indo-European languages [3].

Similar to other Slavic languages, Czech has a rich set of inflectional patterns. It has seven cases and four genders (e.g. there are 16 main paradigms for the inflection of nouns) [8]. Czech allows a relatively free word order.

Table 2.1: Czech alphabetical set

a	á	b	c	č	d	d’	e	é	ě	f
g	h	i	í	j	k	l	m	n	ň	o
ó	p	q	r	ř	s	š	t	t’	u	ú
ů	v	w	x	y	ý	z	ž			

Table 2.1 presents the Czech alphabetical set including 41 characters (case insensitive). Czech is written using the Latin alphabet with accents. All Czech letters are included in the Unicode standard. In this table, we exclude character *ch* from the standard Czech alphabetical table as it is the concatenation of two letters.

Vietnamese is the official language of Vietnam. Besides, Vietnamese speakers are found throughout the world (as part of the immigration), especially in SouthEast Asia, North America and Western Europe. Vietnamese is currently recognised as a minority language in the Czech Republic.

Vietnamese is a member of the Austroasiatic language family, a large language family of SouthEast Asia. It has a long history close to Chinese, featuring in a high volume of loan words. The original Vietnamese writing scripts was a modified set of Chinese characters. Modern Vietnamese is written with Latin letters and additional accents in a system developed by the Portuguese missionary Alexander de Rhodes.

For those who are not familiar with Vietnamese language, we provide the basic Vietnamese alphabetical set in Table 2.2.

[3] http://www.czech-language.cz

Table 2.2: Vietnamese alphabetical set

original letters	a	ă	â	b	c	d	đ	e	ê	g	h
	i	k	l	m	n	o	ô	ơ	p	q	r
	s	t	u	ư	v	x	y				
additional letters including tones	à	á	ạ	ả	ã	ằ	ắ	ặ	ẳ	ẵ	ầ
	ấ	ậ	ẩ	ẫ	è	é	ẹ	ẻ	ẽ	ề	ế
	ệ	ể	ễ	í	ì	ị	ỉ	ĩ	ò	ó	ọ
	ỏ	õ	ồ	ố	ộ	ổ	ỗ	ờ	ớ	ợ	ỡ
	ỡ	ù	ú	ụ	ủ	ũ	ừ	ứ	ự	ử	ữ
	ỳ	ý	ỵ	ỷ	ỹ						

The total number of Vietnamese letters is 89 (case insensitive) including 29 original letters from Vietnamese standard alphabet and 60 derived letters which have tones. The different feature of Vietnamese from other languages using Latin script is that Vietnamese uses two layer of diacritics. The first layer of diacritics changes 5 letters of ASCII including 4 vowels *a, e, o, u* and 1 consonant *d* to generate 7 additional letters for the standard alphabet. The second layer of diacritics generally sets tone to the text, attributed to the fact that Vietnamese is a tonal language. This type of accent may not change the pronunciation significantly (especially for European listeners) but it changes the meaning of the word. Two words might have a similar pronunciation but unrelated senses.

Compared to the set of ASCII characters, Vietnamese alphabet omits 4 letters *f, j, w, z*. Technically, these letters are not expected to appear in a Vietnamese text. However, the letters sometimes appear in Vietnamese writing as a substitution of other letters, such as $ph \rightarrow f$ or $qu \rightarrow w$.

Even though sharing the Latin script, Vietnamese and Czech are linguistically different. They represent two extreme cases of *analytic* and *synthetic* languages. The difference in phonetic is not concerned here, as we take the input from written text. While Czech is highly inflectional, Vietnamese is considered non-inflectional. In short, *lemmatization* and *stemming* are of little use in Vietnamese. While Czech allows free word order, Vietnamese has strict rules regarding the matter of word ordering. Most Vietnamese sentences have to follow the pattern of *SVO*.

We provide an example of Vietnamese, Czech and English parallel sentence in Table 2.3

Table 2.3: An example of Vietnamese - English - Czech sentences

Vietnamese	Tôi	là	sinh viên
English	I	am	student
Czech	Jsem		student

Furthermore, while Czech has a standard set of pronouns, Vietnamese pronouns are much more productive. It has a rich set of pronouns, each expresses a specific situation regarding formality, respect and emotional state. Example from

Table 2.3 provides a trivial translation from Vietnamese to Czech, in which the sentence means "I am student". From Czech to Vietnamese, many other variations would be possible and we choose one particular option which suits the situation when the author and the audience are two acquaintances having a formal conversation. Regarding other situations, the Vietnamese pronoun has to change to connote the degree of relationship and kinship.

3. Corpus Collection

Statistical machine translation system demands high quality data, including a parallel corpus and a monolingual corpus, for system training. Firstly, an aligned parallel corpus at sentence level is essential for training the translation model. Two important criteria for parallel corpora are *quantity* and *quality*. A large amount of parallel data is required to produce understandable output. High quality parallel data helps in producing better translation. Besides the parallel bilingual corpus, a large monolingual corpus in the target language is required. Thus monolingual corpus is used to built a language model, which helps to make the output fluent.

Our first goal was to collect as much data as possible for the Czech↔Vietnamese MT. Beside *Vietnamese* and *Czech*, *English* was selected to be the pivot language. Hence, three pairs of languages were taken into account: *Czech-English*, *English-Vietnamese* and *Czech-Vietnamese*. For each of the pairs, different types of corpus sources are possible.

The following sections provide statistics of the data collected for this thesis. We also present the discussion on problems encountered during the searching for available resources.

3.1 Collection of Czech-English Bilingual Corpora

For *Czech-English* pair, our idea was to reuse the large *CzEng* corpus [29]. The up-to-date corpus is the fourth release of CzEng, called *CzEng 1.0*. It contains 15 million parallel sentences, with 233 million English tokens and 206 million Czech tokens.

CzEng 1.0[29] contains data from seven different domains. It is also automatically annotated at surface layer (a-layer) and deep layer (t-layer) of syntactic representation according to *Functional Generative Description*. Seven domains of *CzEng 1.0* consists of *fiction, EU Legislation, Subtitles, Parallel Web Pages, Technical Documentation, News* and *Project Navajo*. They provides a wide range of vocabulary on different domains. As this thesis is done at *UFAL* [1], we gain access to the *CzEng 1.0* corpus, both raw and filtered.

CzEng 1.0 [29] is shuffled at the level of "blocks". Each block is a sequence of no more than 15 consecutive sentences from one source. Each "block" is taken from one of the domains which are indicated in the sentence ID. The original documents thus cannot be reconstructed but some of the information for cross-sentence phenomena is preserved. Individual text "blocks" are combined to numbered files, each file holds about 200 sentence pairs.

CzEng 1.0 is provided in three different formats. The primary format of CzEng is

[1] `https://ufal.mff.cuni.cz/`

Treex XML, which is a successor of the TMT format of TectoMT used in *CzEng 0.9*. Treex XML can be processed using the Treex platform implemented in Perl and available on CPAN. An alternative format is a simple "factored" line-oriented export format. Finally, a plain text format is a simple tab format. It consists of 4 tabs: Sentence pair ID, Filter score, Czech - not tokenized, English - not tokenized. For this thesis, we choose to select the plain text format. The simplest format is selected to be compatible with the format of *English-Vietnamese* and *Czech-Vietnamese* data.

3.2 Collection of Czech-Vietnamese Bilingual Corpora

Given the fact that we are probably the first ones who exploit the task of Czech-Vietnamese translation, we cannot benefit from existing corpora. Our motivation was to simply collect as much data as possible, regardless of the sources. We have tried multiple sources, such as the bilingual newspapers between Vietnamese and Czech, but all of them are very limited.

OPUS - the open parallel corpus is a growing multilingual corpus of translated open source documents [24]. It is a public collection of parallel corpora that can be used and distributed for everyone to prepare experiments on bitexts. The *OPUS* data are provided with a standard encoding format including linguistic annotations. The data of OPUS could be downloaded in multiple formats, such as untokenized data, tokenized data or XML-format data. In our experiments, we choose to use the bilingual data in their untokenized form.

As a growing corpus, OPUS is being developed and data are being added to the pool everyday. After 10 years of development, OPUS now contains data from a large number of resources. However, *Czech-Vietnamese* pair remains to be underdeveloped. The OPUS data for this pair is limited to two groups: *Open Subtitles* and *technical documents*

Open Subtitles data involve three corpora *OpenSubtitles2011*, *OpenSubtitles2012* and *OpenSubtitles2013*. They are collections of documents from Open Subtitle website[2]. They contain the parallel subtitles from movies and television series. As the names indicate, three corpora are associated with three consecutive years from 2011 to 2013. Each corpus consists of a large number of movies. Theoretically, there are three situations regarding the language:

1. A Vietnamese film or TV program which is broadcasted on Czech channel. The Czech subtitles are made for Czech viewers.

2. A Czech film or TV program which is broadcasted on Vietnamese channel. The Vietnamese subtitles are made for Vietnamese viewer.

3. A film or TV program was made in a different language, such as English.

[2] `http://www.opensubtitles.org/`

Its subtitles are translated into both Vietnamese and Czech. The translation may deviate from the original meaning. This situation will double the deviations.

In fact, most of the data that we have come across belong to the third situation. It implies that Vietnamese and Czech have almost no directly translated data.

We decided to merge all three corpora into one single corpus, named *OpenSub*. The corpus contains 1.3 million sentences.

Preliminary experiments have shown that the *OpenSub* corpus is noisy. There are a lot of mistakes in the data which will surely result in poor translation.

Technical documents contain three items: *KDE4*, *GNOME* and *Ubuntu*. They are the manuals and usages for different variants of UNIX operating system and its desktop environment. They are the result of *localizing the operating system* when the system messages, icons and manuals are translated from English into a local language. The corpus thus often contains not just text but also formulas and special expressions such as placeholders.

In total, the group of *technical documents* contains approximately 40 thousand sentence pairs. However, when examining the materials, after carefully checking the meaning compatibility, we have come to a decision to omit the *technical documents* due to the poor quality of the majority of data.

Therefore, among all *OPUS* resources, only *OpenSub* data were selected for the experiment of this thesis.

TAUS (Translation Automation User Society) is a global community for sharing language data. The community was founded in 2007, aiming to provide industry-wide tools for translation. In 2008, the TAUS Data Association was founded with many international companies, aiming to provide a giant multilingual database. Since then, it kept growing with the collaboration of companies and organizations across industry sectors. In principle, the TAUS resources, which are a commercial product, should be freely available for research.

Regarding Czech-Vietnamesepair, the available data consists of 623201 words, which is approximately 31 thousand pairs of sentence. Among which, 84% of TAUS data are software strings and documentation, provided by a support IT company. Hence, the domain of TAUS data is also technical documentation. Another important point worth mentioning is that all TAUS Czech-Vietnamesedata include Matrix TM results. In other words, they were constructed by linking translations memories for different language pairs, a pivoting method for translation memories. As a result, there is no direct data between Vietnameseand Czechin TAUS project.

Unfortunately, we came across TAUS data only when the phase of data collection had finished. Hence, this thesis does not benefit from TAUS data.

TED: (Technology, Entertainment, Design) [3] is a global set of conferences run by

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

a private non-profit organization. TED events are held throughout North America and in Europe and Asia, where speakers give their ideas to a number of audiences. The speakers address a wide range of topics within the research and practice of science and culture, often through storytelling. For this study, we were interested in the collection of TED transcripts, which are sometimes available in more than one language. *TED talks* are a special event organized by *TED*. The talks are often translated into multiple languages, which make them a resource of bilingual dataset.

The transcripts are aligned according to the video timing. It could be considered one special type of subtitles. In this case, there is only one narrator throughout the whole video instead of numerous short conversation. This situation somewhat raises the belief that the quality of TED talks is better than movie and television program subtitles. And also the angle of the narrative is closer to written text.

In 2012, *FBK center*[4] published one set of *TED talks* as a parallel corpus in the Web Inventory of Transcribed and Translated Talks(WIT^3). *Vietnamese-Czech* is one pair in the data, yet the size of corpus is small, around 76M sentences.

We crawled all the parallel transcripts until 2015. Our observation shows that the data is not clean. The problem of misalignment is rarer than in *OpenSubtitles* but it is still a serious issue. The original language is *English*. Similar to subtitles, *TED talks* suffer from the deviation in translation from *English* to *Czech* and from *English* to *Vietnamese*.

Table 3.1: Data sizes of Czech↔Vietnamese OPUS and TED data

corpus	**sources**	**sentences**	**src tokens**	**trg tokens**
OpenSub	Open Subtitles 2013	0.6M	3.0M	4.6M
OpenSub	Open Subtitles 2012	0.4M	2.1M	3.1M
OpenSub	Open Subtitles 2011	0.2M	0.9M	1.3M
TED-talk	TED talks 2015	0.17M	1.3M	1.6M
Total	OPUS/TED	1.37M	7.3M	10.6M

Table 3.1 shows the statistics of Czech→Vietnamesecorpora that we have collected. Finally, the final corpus consists of 1.37 million sentence pairs. Based on the natural of the raw data collected from unverified sources, we decided to name it *unreviewed* dataset.

3.3 Collection of English-Vietnamese Bilingual Corpora

Our objective is to collect all *English-Vietnamese* parallel corpora, either professional data or non-professional data. The *English-Vietnamese* data being used in this study belong to two groups:

[4]http://www.fbk.eu/

- *reviewed*: The corpora which were published by previous works.
- *unreviewed*: The newly collected corpora for this study specifically.

The *English-Vietnamese reviewed* data are generally small and fractional. They have been collected from different sources including news and text books.

VLSP corpus

VLSP[5] is a Vietnamese national project *(KC01.01/06-10)*. The VLSP project was to provide official annotated Vietnamese resources for research in the area of natural language processing.

By far VLSP is the most famous dataset in Vietnamese. They not only provide the parallel data but also a *Vietnamese treebank*. Unfortunately, the parallel data is not included in the treebank and there is no available NLP tool to process them. The corpus sources are informatics student books, science books, dictionaries and English books.

From the description, the VLSP corpus consists of 80000 sentences in English and Vietnamese. We extracted data based on the *XML tags* and *id* between English documents and Vietnamese documents. After filtering out all sentences. in which the *id* is missing, the *VLSP* corpus has 93158 sentence pairs, which is more than suggested by the description. After filtering out incompatible *XML tags*, the corpus consists of 80098 sentence pairs. After filtering out identical pairs, 71705 sentence pairs remain. This is the final size of *VLSP* corpus being used in this thesis.

EVB corpus

EVB corpus is a multi-layered *English-Vietnamese* bilingual corpus [17]. The corpus is claimed to have 800000 sentence pairs with 10M Vietnamese/English words, aligned at the sentence level. Besides, 45000 sentence pairs is said to be aligned in word level and have been tagged using other linguistics tags.

However, it turns out that EVB corpus is a commercial product. We contacted the authors and asked for the potential usage in research. Only a small part of data coming from internet news is made free of charge. We use only this free part. After we cleaned the corpus and removed duplicate sentences, the size of EVB corpus that we have is 42844 sentence pairs. For this study, *EVB* corpus is considered small.

[5]http://vlsp.vietlp.org:8080/

UET corpus

The *UET* corpus is an *English-Vietnamese* corpus released by Vietnamese National University. Its source is also internet news with 132636 sentence pairs.

Table 3.2: Data sizes of published English↔Vietnamese corpora

corpus	sentences	English tokens	Vietnamese tokens
VLSP	72K	1.0M	1.0M
EVB	43K	0.8M	0.9M
UET	132K	2.5M	2.5M
Total	247K	4.2M	4.3M

The statistics of *reviewed* data used in this thesis is shown in Table 3.2. The numbers of Vietnamese and English tokens per sentence are approximately the same, around 20 tokens per sentence.

Unreviewed corpora

Similarly to the *Czech-Vietnamese* pair, we also collect data from *OPUS* [24] and *TED*. Because English is a common language, the data available for Vietnamese and English are larger than *Czech-Vietnamese*. The motivation behind this collection is to get as much data as possible. As mentioned above, the clean corpora which were published by other works are relatively small.

The sources of *English-Vietnamese* OPUS data are similar to the sources of *Czech-Vietnamese* data. We treat them in the same way by combining three OpenSubtitles corpora into one single corpus, named *OpenSub*. Finally, we have 1.7 million of parallel subtitle segment pairs from OPUS and 280 thousand parallel transcript segment pairs from TED talks. The statistics of the *English-Vietnamese* corpora from OPUS and TED are shown in Table 3.3.

Table 3.3: Data sizes of English↔Vietnamese OPUS and TED data

corpus	sources	sentences	src tokens	tgt tokens
OpenSub	OpenSubtitles2013	0.9M	5.6M	6.8M
OpenSub	OpenSubtitles2012	0.5M	3.0M	3.7M
OpenSub	OpenSubtitles2011	0.3M	1.8M	2.3M
TED-talk	TED talks 2015	0.28M	2.5M	2.5M
Total	-	1.98M	12.9M	15.3M

3.4 Collection of Monolingual Corpora

This section provides a brief description of the monolingual corpora that we collected for the study. Further details are discussed in the experiment chapter.

On examining the availability of *Czech* and *English* data, we decided to re-use the resources at UFAL[6] for the translation system. The *Czech* side of *CzEng 1.0* is used to built the *Czech* language model. The source of *English* monolingual corpus, which was collected for *WMT13*, is Internet news in 2012.

Table 3.4: Vietnamese monolingual data

Corpus	**Size***(lines)*	**Sources**
MonoNews	0.35M	Internet News
MonoVNTQ	1.46M	Fiction Novels
Total	1.81M	News and Novels

Table 3.4 shows two corpora that we collected for *Vietnamese* language model. Their sources are Internet news (from many Vietnamese newspapers) and fiction novels. Combining two fragmented corpora generates a corpus of 1.8 million sentences. The combined Vietnamese corpus is used throughout the experiments for *Vietnamese* translation.

[6] `https://ufal.mff.cuni.cz/`

4. Data Preparation

As described in the previous chapter, the final data are divided into two groups of corpora: *reviewed* and *unreviewed*. The former, containing *English-Vietnamese* corpora, is used in the experiments without any further preparation. The latter contains larger and noisier corpora, both *Czech-Vietnamese* corpora and *English-Vietnamese* corpora.

The large *unreviewed* data unnecessarily increase the amount of storage and memory. Besides, the lower quality of *unreviewed* corpora can adversely affect the final performance of any SMT method. A noise reduction step is thus important to avoid the degradation.

This chapter discusses two techniques which were adopted to improve the quality of *unreviewed* corpora: normalization and filtering

4.1 Normalizing Corpora

The *unreviewed* corpora are taken from subtitles and transcripts of multi-lingual videos. Most of the data do not have any prescribed writing styles. The noise of *unreviewed* corpora most harmful for SMT can be primarily to man-made errors in subtitles and transcripts, including the haze of spelling errors, the abuse of special characters, etc. Some of the errors are due to bad processing when entering the OPUS or TED collections, e.g. the incorrect language identification.

We are interested in cleaning up the corpora without any deeper knowledge of syntactic analysis. This means that the processing phase has to be carried out on the general format of subtitles and transcripts, with the knowledge of the language.

Regarding subtitles, there are no standard set of rules for either making the subtitles or storing the subtitles. As a product of many independent authors, subtitle formats vary heavily on the authors. We have decided to follow a number of general rules and concepts which are most common across the collection.

Based on the general rules, we have adopted a scheme to normalize the *unreviewed* data as well as to remove the unwanted pairs of sentences. Furthermore, we have to agree on a trade-off point. If the corpora are harshly normalized, we get a corpus with dense but somewhat distorted observations. Otherwise, we have to accept a reasonable amount of noise in the corpus.

Punctuation

Several text patterns can be spotted in subtitles, but only a few are used widely: “three dots”, “bracket” and “parenthesis”. Each of the patterns uses different special characters, which are not part of the sentence translation. However, when the corpora are synced by time frames, the special characters remain in the corpus. Because these patterns do not need to be exactly parallel, they can complicate

the word alignment or lead to noise in the extracted phrases.

Sequence dots (including ending triple dots and starting triple dots): When an expression occurs across time frames, the standard style is to put three dots right after the last character (no space character inserted) of the ongoing frame. In the next time frame, three dots should be used right before the first character of a subtitle (no space character inserted and the first character is in the lowercase form).

The problem is that this segmentation was done for the two languages separately, so an expression spanning several frames may have a different segmentation or may not be segmented at all.

When the data are aligned according to the time, either multiple frames are merged or some frames are dropped. Hence, the sequence of dots can appear at the beginning, at the end or in the middle of a sentence. Our solution was to locate those sequences and substitute them with space characters.

Brackets and parentheses usually contains comments (e.g. explaining the preceding phrase). Such comments are not part of the utterance. Whether or not to put a comment in subtitle is up to translators. Generally, comments appear in only one side of the corpus, so we cannot learn any translation equivalents from them. We decided to filter out all the comments.

Italics on the subtitle text should be used to indicate an off-screen spoken text. It leaves a number of formatting tags in the corpora. We tried to clean the subtitles up by removing the tags and keep the plain text.

There are other tags, but they do not affect quality of the corpora. We left them intact.

Language problems

The large corpora from *OPUS* sometimes encounter serious problem of language recognition. For example, errors in metadata lead to Chinese or Greek appearing in a file labelled as Vietnamese or Czech. We tried to discard these problematic subtitles. However, it is difficult to come up with a perfect solution, we simply did our best.

Besides, language-specific diacritic marks are not displayed correctly if the proper font is not used. The problem happens with *English*, *Czech* and especially *Vietnamese*, a language with two layers of diacritics. A mistaken display of diacritics damages the incoming of a sentence, making the pair of sentences incorrect. We have made another attempt to filter out all sentences which have a font problem.

Text difference

Most subtitle text consists of short conversations. There might be more than one expression (dialogue exchange) in a line. We filter out all pairs of sentences which differ in the number of these segments. We assume that by full stop, exclamation mark, question mark, triple dots and hyphen.

Finally, we discard all pairs of sentences in which two sentences are significantly

different in length.

4.2 Filtering Corpora

As the preprocessing of data acts as a threshold to prune the corpora by removing obvious mistakes, we have decided to apply another filter technique on our direct *Czech-Vietnamese* and *English-Vietnamese* corpora.

We have decided to re-use the *CzEng10 filtering tool* [29] to filter poorly aligned sentence pairs. The tool was originally designed to handle the varying reliability of *CzEng* sources, which demanded an automatic method for recognizing and filtering out poor-quality sentence pairs.

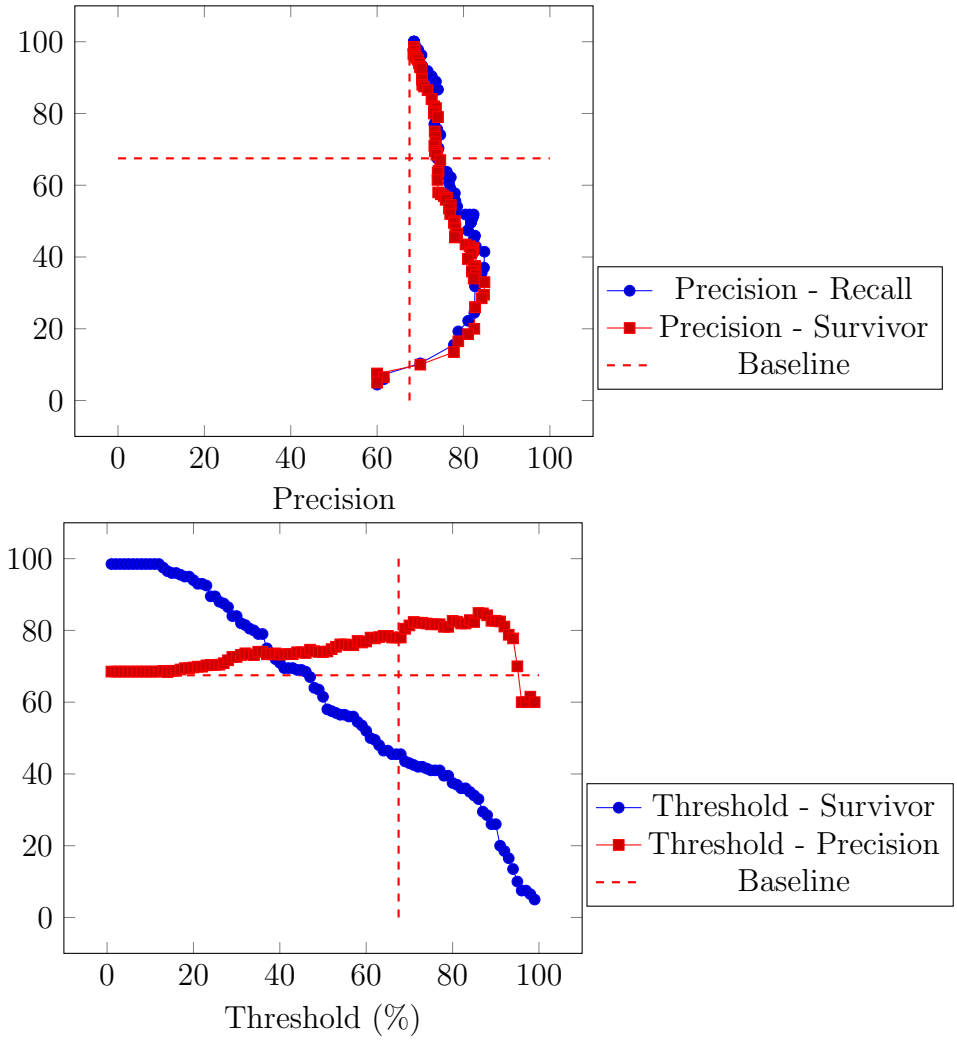


Figure 4.1: Evaluation of Maximum Entropy with Czech-Vietnamese dataset

The *CzEng10 filtering tool* is an update from previous *CzEng* editions, with several new filters and a robust method for their combination. It acts as a classifier which scores every sentence pairs with a set of features. The features are linguistic

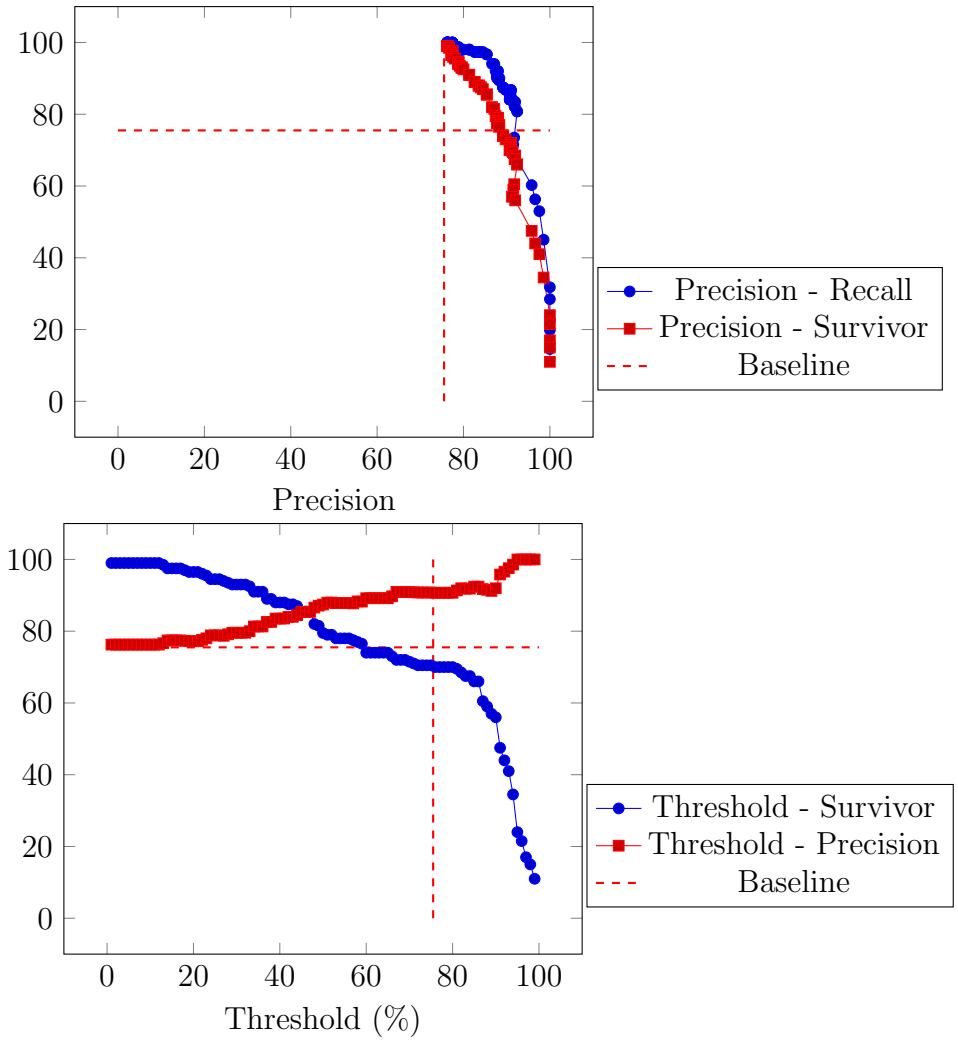


Figure 4.2: Evaluation of Maximum Entropy with English-Vietnamese dataset

or non-linguistic. They are combined to form a single score by training the classifier on manually labelled data. Regarding machine learning method being used, *CzEng* authors have reported that *maximum entropy classifier* outperformed *decision tree classifier* and *naive Bayes classifier*.

The tool was designed in *Treex* which allows to include various features. The following features are used in our settings:

- Indication of the source and target sentences' string identity
- Length of the sentences
- The appearance of *source* words on the *target* side vice versa
- The remains of meta-information in the pairs, such as HTML tags or file paths
- Score of symmetrized automatic word alignments obtained by GIZA++

We estimate the precision of the filtering on two independent manually annotated sets of sentence pairs, associated with the *Czech-Vietnamese unreviewed* corpus and the *English-Vietnamese unreviewed* corpus. Each set consists of 1000 sentence pairs.

Figure 4.1 and Figure 4.2 show statistics of the corpora at various thresholds of the score. Two figures shows a significant difference between two corpora. *Czech-Vietnamese* has proven to be more difficult to handle than *English-Vietnamese*. For each language pair, we plot four correlations including the score (*threshold*), the number of remaining pairs (*survivor*), the precision and the recall of the classifier.

The monotonic function between *threshold* and *survivor* is expected. The higher the threshold is, the smaller remaining corpus size we get. Similarly, we would expet a steady increase in precision as we increase the threshold. This is confirmed on the *English-Vietnamese* dataset. However, we reveal that the *threshold* and *precision* do not correlate in *Czech-Vietnamese* dataset. The *precision* peaks at *threshold 0.88* then drops significantly afterwards. It is due to a number of *incorrect* pairs which receive very high scores.

The statistics above indicate different behaviours between two language pairs. This is confirmed by the correlations: *precision-recall* and *precision-survivor*. In the experiment with *English-Vietnamese* dataset, the lowest precision is associated with the highest recall 100%. This is the case when all of the data are kept by specifying the threshold to *0*. The precision increases and the recall decreases when the threshold is raised. In the experiment with *Czech-Vietnamese* dataset, the highest precision is just *82%* instead of *100%* and the recall is approximately *30%*.

The final goal of the filtering process is to maximize the precision (keeping most the correct pairs, which express the same meaning from both sentences). At the same time, the size of the corpora is also an important feature. If we set the *threshold* too low, we filter less sentence pairs and keep more pairs. The result would still be noisy but its size is considerable. On the contrary, if we set the *threshold* high, we filter more sentences pairs and keep less pairs. We receive a small good corpora.

Let's take *Czech-Vietnamese* corpus for example. If we maximize the *Precision*, we have to set the *filter* threshold to *0.88* to receive *85%* correct pairs. However, the size of corpus is only *30%* of the original size. Meanwhile, if we set the threshold to only *0.4*. The corpus size is 75% and the precision is *75%*. To set the best balance, we would need to evaluate the final translation quality with corpora of varying sizes.

Nevertheless, the filtering tool does not guarantee a precise result. The performance on the real dataset may differ from the performance on the test set.

5. Pivoting Methods

In many circumstances (including ours), there are not sufficient data for a direct MT system from source language F to target language E. Over the recent years, pivoting methods have become a candidate to alleviate this problem. The task of pivoting methods in SMT is essentially the usage of another language P to improve the translation from F to E.

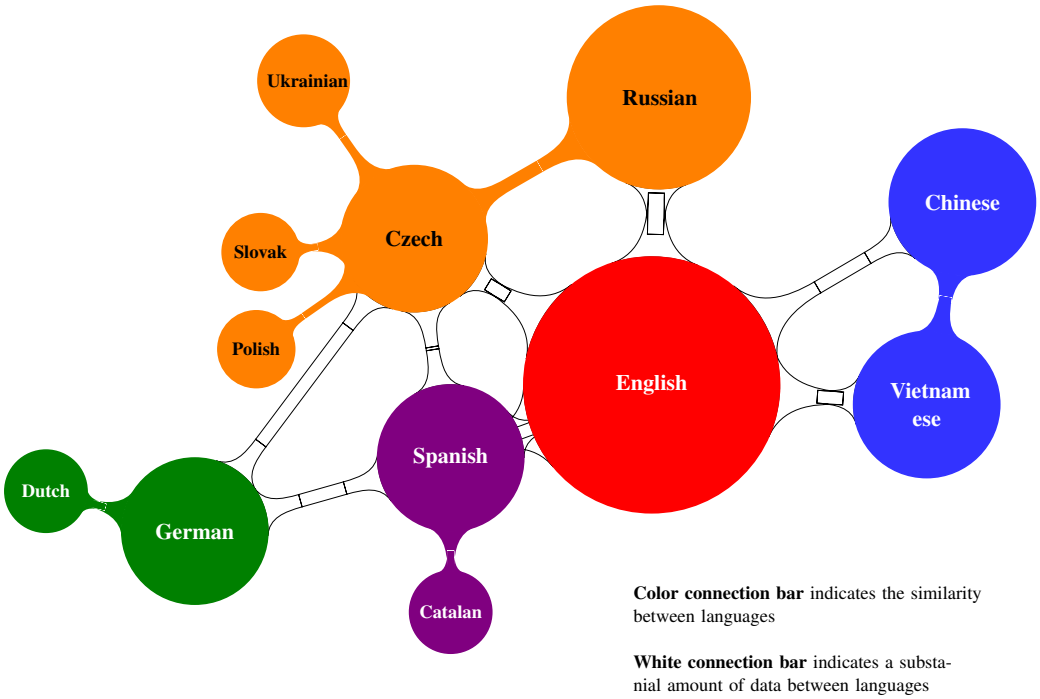


Figure 5.1: Languages for pivoting methods between *Czech* and *Vietnamese*

Talking into account that there are numerous languages besides E and F, including both natural languages and artificial languages, choosing an appropriate third language (sometimes called pivot language or bridge language) is the first important task. In pivoting methods, the ideal would be to use more than one *third language*. As discussed in the previous chapter, for a certain pair of source and target languages, only a few languages could be chosen as the *pivot language*, which fulfil the condition of supporting the translation from F to E. Two of important features, which the pivot language should satisfy, are as follows:

- *Popularity*: The third language should have a better dataset or available MT system with both the two languages. Generally, prominent languages, such as English, fulfil this role.
- *Similarity*: The resemblance between the third language and either the source language or the target language plays an important role. In this

case a superior language in the same family tree is often picked as the pivot language. In this case, superior has the sense of richer corpora or higher popularity. For example, *Czech* for *Slovak* and *Spanish* for *Catalan*

Within the scope of our thesis, *Vietnamese* and *Czech* serve as the source and target languages (in both directions). A significant drawback affecting the quality of MT between Vietnamese and Czech is the linguistic difference between the two languages. Among all languages, *Chinese* is another of the most promising languages and it is similar to Vietnamese. Taking another route, *Czech* shares linguistic features with other similar Slavic languages, among which *Russian* has the most speakers. However, neither *Chinese* nor *Russian* has sufficient resources to act as the bridge language between Vietnamese and Czech.

After carefully examining the potential of all languages, we decided to select Englishas the sole pivot language. Figure 5.1 shows the candidates for the role of pivot language. English is the only language, either natural or artificial, which provides sufficient corpora to perform the translation. An important point to understand is that English is a Germanic language, and thus it is not closely linguistically related to neither Czech nor *Vietnamese*. This can potentially have a harmful effect on the system performance.

Another important point is that the goal of pivoting methods is different from the standard problem of multi-source translation. In the context of SMT, it is possible to transform the pivoting problem into the multi-source problem by translating the input into various pivot hypotheses in different languages.

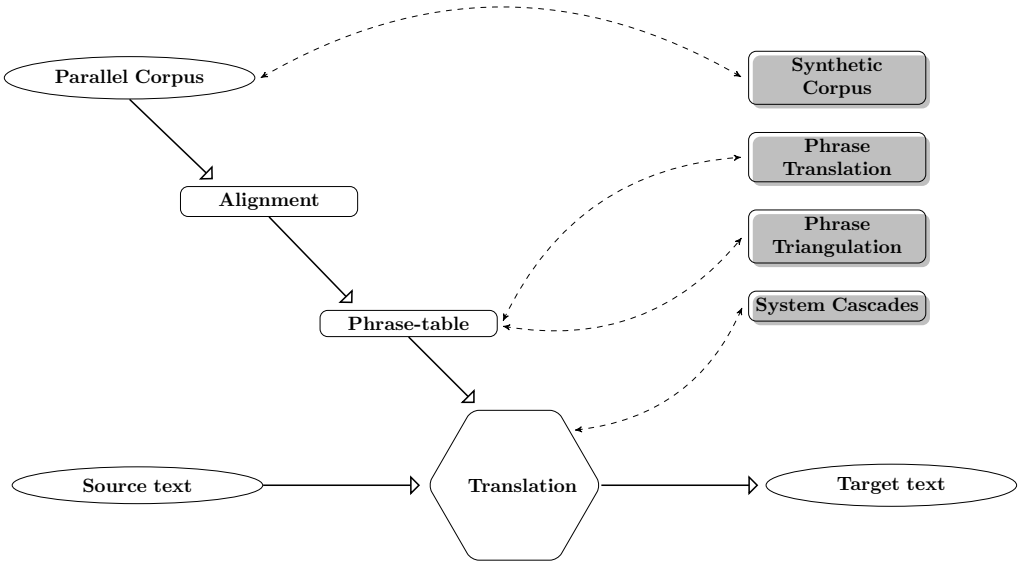


Figure 5.2: A schematic overview of the SMT process and the interaction with various pivoting methods

Pivoting methods can be applied in various steps of what we call "the process of SMT". Each method relies on the availability and usability of the pivot language. Figure 5.2 shows an example of how pivoting methods interact with the SMT process.

5.1 Synthetic Corpus

As mentioned above, the *synthetic corpus* method involves translating the pivot side of one corpus (either source-pivot corpus or pivot-target corpus) to obtain a source-target corpus with one side synthetic. The obtained corpus is then used to build the MT system between source language and target language.

Within the scope of this thesis, the specific task is to translate the Englishside of a corpus to convert it into a Czech-Vietnamese synthetic corpus. There are two scenarios proposed which achieve this task:

1. Translating the Englishside of our English-Vietnamese corpora into Czech. This scenario requires an English→Czech MT system. The output of this scenario is a Czech-Vietnamese corpus with the Czech side synthetic.

2. Translating the Englishside of an Czech-English corpus, which, in this thesis, is *CzEng 1.0*. This scenario requires an English→Vietnamese MT system. The output of this scenario is again a Czech-Vietnamese corpus now with the Vietnamese side synthetic.

A major limitation affecting the quality of the synthetic corpora is the performance of the English→Vietnamese and English→CzechMT systems, which are not prominent due to the difference in language and the small dataset. Another drawback is the effort spent on translating a whole corpus, especially when the corpus size is large. For example, the size of *CzEng 1.0* corpus makes it quite expensive to follow the second scenario, despite the fact that it possibly works better than the first scenario.

Within the work of this thesis, *synthetic corpus* was one of the last approaches that we tried. Following the first scenario, we obtain a synthetic Czech→Vietnamese corpus. After carefully analysing the result of this experiment, we concluded that the *synthetic corpus* does not attain any significant improvement over other methods. Hence, we omit the synthetic corpus approach from the text altogether and we focus on other approaches rather than the second scenario.

Another option to bear in mind is the possibility of generating more than one hypothesis during the translation of corpus. Even though this option is deemed memory consuming, it is a possibility for future work.

5.2 Phrase Table Translation

The *phrase table translation* approach involves translating the pivot side of a phrase table into either source language or target language. Within this thesis, there are also two scenarios for this approach, associated with translating the English-Vietnamesephrase table and translating the Czech-English phrase table respectively.

In the technical view, the *phrase table translation* approach is very similar to the *synthetic corpus* approach. However, phrase table obtained from the translation of another phrase table is different from the phrase table obtained from the synthetic corpus. In the context that phrases are the atomic units of phrase-based MT system, translating the phrase table preserves the structure (i.e. phrase seqmentation) of the old phrase table while *synthetic corpus* approach creates a new phrase table.

The task of translating a phrase table poses a different challenges compared to the task of translating a corpus. The input is a phrase, which is generally shorter than a sentence. Thus, the role of language model is limited during translation. It is also common that the input phrase is a phrase pair in the MT model which is used to translate the phrase table. It is likely that the whole phrase is chosen over fragmented phrases.

5.3 Phrase Table Triangulation

The *phrase table triangulation* approach, along with the *phrase-translation* approach, manipulates the phrase table in SMT process. In this approach, triangulation is done by merging two phrase tables instead of translating the pivot language to either source language or target language. This methods requires two phrase tables: *source-pivot* phrase table T_{SP} and *pivot-target* phrase table T_{PT}.

Each phrase table consists of four features functions, one alignment mapping and three occurrence counts.

For example, two phrase tables T_{VE} (Vietnamese→English phrase table) and T_{EC} (English→Czech phrase table) provide us with following values, in which v, e and c are phrases in Vietnamese, English and Czech respectively.

- 4 phrase translation probabilities for both direction $\phi(v|e)$, $\phi(e|c)$ and $\phi(c|e)$, $\phi(e|v)$
- 4 phrase lexical probabilities for both direction $p_w(v|e)$, $p_w(e|c)$ and $p_w(c|e)$, $p_w(e|v)$
- 2 alignment mapping tables $align_{v\to e}$ and $align_{e\to c}$
- 3 occurrence counts from T_{VE}: c_v, c_e^1, c_{ve}
- 3 occurrence counts from T_{EC}: c_c, c_e^2, c_{ec}

The problem of merging T_{SP} and T_{PT} can be broadly divided into three tasks.

The first task is to construct source and target pairs of phrases. In general, phrase table triangulation method connects s and t whenever there exists a pivot phrase p such that $s - p$ is listed in T_{SP} and $p - t$ is listed in T_{PT}.

The second task is to estimate word alignment for linked phrases. Given the two alignment mappings a_{sp} and a_{pt} of component phrase tables, we have to

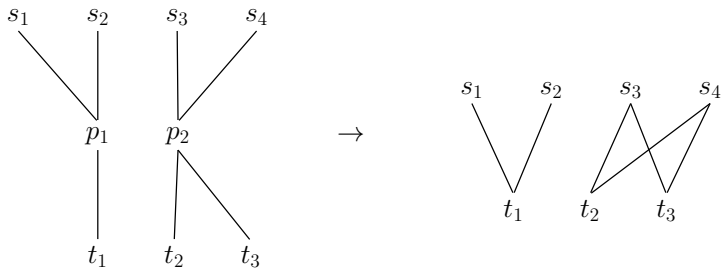


Figure 5.3: Constructing source-target alignment

construct the source-target alignment a. The task is arguably solved by tracking the alignments from each source word $s_i \in s$ over any pivot word $p_k i \in p$ to each target word $t_j \in t$ as illustrated in Figure 5.3. Formally:

$$(s_i, t_j) \in a \Leftrightarrow \exists p_j : (s_i, p_k) \in a_{sp} \;\&\; (p_k, t_j) \in a_{pt} \tag{5.1}$$

The third task is to estimate the four features function of new phrase table T_{PT}, phrase translation probabilities for both directions, $\phi(s|t)$ and $\phi(t|s)$ and lexical translation probabilities for both direction, $p_w(t|s)$, $p_w(s|t)$. This task gives rise to two different approaches: using posterior probabilities and using co-occurrence counts.

Estimate probabilities using the probabilities available

The problem of triangulation is viewed as a generative probabilistic process on two sets of phrase pairs, *s-t* and *p-t*. The conditional distribution $p(s|t)$ needs to be estimated, in which the arguments denote the source and target phrase. If we assume that a pivot phrase from a different language is independent from the two languages, we can find the conditional over source-target phrase pair by marginalising out the pivot phrase

$$\begin{aligned} p(s|t) &= \sum_p p(s|p,t) \times p(p|t) \\ &\approx \sum_p p(s|p) \times p(p|t) \end{aligned} \tag{5.2}$$

Equation 5.2 imposes a simpler conditional probability over the source and target phrase. In a linguistics sense, the similarity of sense between the source phrase and the target phrase depends on how many similar senses in the third language they share.

$$\begin{aligned} p(s|t) &\approx \sum_p p(s|p) \times p(p|t) \\ &\approx \max_p p(s|p) \times p(p|t) \end{aligned} \tag{5.3}$$

In Equation 5.3, the conditional probability is simplified further, in which only the most prominent sense in the third language is concerned.

Generally, this method requires all target phrases in the source-pivot phrase table to be found in the source side of the pivot-target phrase table. While this can be quite probable in multi-parallel corpora, it poses a problem when the two corpora are drawn from different sources. This obviously applies to our corpora.

From two Equations 5.2 and 5.3, there are two scenarios for phrase triangulation method. The first scenario takes into account all the middle phrases which are paired with both source phrase s and target phrase t. The second scenario considers only the intermediate phrase which yields the highest probability among all phrases which are paired with sources phrase and target phrase. We call them the summing and maximizing scenario for short.

Summing scenario

In the *summing scenario*, the translation probability ϕ and the lexical probability p_w of the new phrase table are computed as follows:

$$\begin{aligned}
\phi(s|t) &= \sum_{\overline{p}} \phi(s|\overline{p}) \times \phi(\overline{p}|t) \\
\phi(t|s) &= \sum_{\overline{p}} \phi(\overline{p}|s) \times \phi(t|\overline{p}) \\
p_w(s|t) &= \sum_{\overline{p}} p_w(s|\overline{p}) \times p_w(\overline{p}|t) \\
p_w(t|s) &= \sum_{\overline{p}} p_w(\overline{p}|s) \times p_w(t|\overline{p})
\end{aligned} \tag{5.4}$$

Maximizing scenario

In the *mazimizing scenario*, the translation probability ϕ and the lexical probability p_w of the new phrase table are computed as follows:

$$\begin{aligned}
\phi(s|t) &= \max_{\overline{p}} \phi(s|\overline{p}) \times \phi(\overline{p}|t) \\
\phi(t|s) &= \max_{\overline{p}} \phi(\overline{p}|s) \times \phi(t|\overline{p}) \\
p_w(s|t) &= \max_{\overline{p}} p_w(s|\overline{p}) \times p_w(\overline{p}|t) \\
p_w(t|s) &= \max_{\overline{p}} p_w(\overline{p}|s) \times p_w(t|\overline{p})
\end{aligned} \tag{5.5}$$

Estimate probabilities by the co-occurrence counts

Phrase probabilities, as established above, are based on the probabilities from the two source phrase tables. The following method uses the underlying co-occurrence counts to estimate indirect probabilities. The occurrence counts of phrase pair (s, t) are constructed from the occurrence counts of two source phrase pairs: (s, p) and (p, t).

Formally, given the co-occurrence counts $c(s,p)$ and $c(p,t)$, the task is to find a function f which estimates the co-occurrence count $c(s,t)$ of the new phrase table in Equation 5.6.

$$c(s,t) = \sum_{p} f(c(s,p), c(p,t)) \tag{5.6}$$

Once the co-occurrence count $c(s,t)$ is estimated for all phrases s and t, $c(s)$ and $c(t)$ of the final phrase table can be computed by summing up the count $c(s,t)$, for every t and s respectively.

$$\begin{aligned} c(s) &= \sum_{t} c(s,t) \\ c(t) &= \sum_{s} c(s,t) \end{aligned} \tag{5.7}$$

Then, the phrase tables features are computed as maximum-likelihood estimate using those values. The co-occurrence count $c(s,p)$ gives the number of times that two phrases s and p appear together in the source-pivot corpus. Similarly, the co-occurrence count $c(p,t)$ gives the times that two phrases p and t appear together in the pivot-target corpus. From the co-occurrence counts $c(s,p)$ and $c(p,t)$, the co-occurrence count $c(s,t)$ is zero if the pivot phrases are not idential.

We experiment with four options to compute $c(s,t)$ as follows:

$$\begin{aligned} c_{min}(s,t) &= \sum_{p} min(c(s,p), c(p,t)) \\ c_{max}(s,t) &= \sum_{p} max(c(s,p), c(p,t)) \\ c_{ar}(s,t) &= \sum_{p} \frac{c(s,p) + c(p,t)}{2} \\ c_{ge}(s,t) &= \sum_{p} \sqrt{c(s,p) \times c(p,t)} \end{aligned} \tag{5.8}$$

Because the source-pivot and pivot-target parallel corpus significantly differ in size, the *minimum* function is likely to take the counts of the smaller corpus (Vietnamese-English). The *maximum* function is likely to take the counts of the larger corpus (English-Czech). The *arithmetic mean* and *geometric mean* relatively moderate this problem but they still favour the larger value.

From four options for function *f*, there are four scenarios to re-compute the co-occurrence counts. When the co-occurrence count $c(s,t)$ is calculated, the phrase and lexical probabilities are estimated as Equation 5.9:

$$
\begin{aligned}
p_{\cdot}(s|t) &= \frac{c_{\cdot}(s,t)}{\sum_s c_{\cdot}(s,t)} \\
p_w(s|t,a) &= \prod_{i=1}^{n} \frac{1}{|j|(i,j) \in a|} \sum_{(i,j)\in a} w(s_i|t_j)
\end{aligned}
\tag{5.9}
$$

For the lexical probabilities, the alignment a between the source word s_i and the target word t_j is established in the second task of triangulation method. The lexical translation probability w between source word s_i and target word t_j must be computed beforehand as follows:

$$
w(s_i|t_j) = \frac{c(s_i,t_j)}{\sum_{s'} c(s',t_j)} \tag{5.10}
$$

While phrase-triangulation method sounds appealing, it suffers from a few problems, especially when the corpora come from different sources. First, the triangulation encounters twice the problem that a standard source-target phrase table encounters. The errors when two phrases are paired incorrectly will compound. This leads to noisier estimates than in the source-target phrase table. Secondly, the noise from phrase table leads to a problem that the phrase triangulation method omits a larger quantity of rare phrases while it increases the quantity of common phrases. Thirdly, alignment errors lead to the situation that one common phrase in the pivot language is paired with numerous phrases. The common phrase then acts as a bridge for many different phrases from source language to target language. This not only produces incorrect phrase pairs but also leads to a huge size of the triangulated phrase table. Finally, as in other methods, the assumption of independence in conditional probability might result in a loss of information.

Other publications often compare their proposed approach to the baseline, which is often either *direct translation* or *system cascades*. Our thesis intended to unite them all in a realistic experiment.

5.4 System Cascades

The *system cascades* method is one of the most straightforward ones. However, it is often the most stable approach to translate one language to another language without direct parallel data. A brief introduction of the method is mentioned in Chapter 1.

While other pivoting methods are conducted during the training process, the *system cascades* method is done during the translation process. The problem of finding the best sentence $\hat{(e)}$ for a foreign input sentence f could be defined as: maximizing the translation score from source sentence f to a pivot sentence p, then from p to target sentence e. However, investigating all possible pivot

sentences p is too expensive. In the context of SMT process, the pivot sentence p is taken from the output of the first translation from source language to pivot language.

$$\hat{e} \approx \arg\max_{e,p_i} p_{smt}(p_i|f) \times p_{smt}(e|p_i) \tag{5.11}$$

Equation 5.11, p_i, in which $i \in [1, n]$, is a pivot hypothesis of the first MT system which translates source language into pivot language. The hypothesis then becomes the input sentence of the second MT system which translates pivot language into target language.

If $n = 1$, which means that the first MT system provides 1-best translation, the scenario is simple enough to be done manually. In this case, the source input sentence is translated twice by two consecutive MT systems respectively. When n increases, more pivot hypotheses are taken into consideration. On the one hand, this allows the translation process to be optimized. On the other hand, this raises the problem of memory requirements. In this study, a comparison between the performance of $n = 1$ and $n > 1$, is reported as part of the experiments with *system cascades* methods.

5.5 Phrase Table Interpolation

Often, it is possible to find at least some directly parallel data, so pivoting is not the only option for translating the source to the target.

As discussed in Chapter 3 and 4, the Czech-Vietnamese corpora were collected and cleaned up. Therefore, we adopt a method of combination by interpolating the standard phrase table with phrase tables obtained from pivoting methods. The interpolation is expected to gain improvement if the strengths of one method will offset the weaknesses of the other approach.

To get the best from both the direct and the pivoted translation, we would like to combine the two systems. *System combination* is the task of combining multiple systems to produce an output better than its component systems. It is a challenging MT task, which various approaches have been proposed over the past few years [21] [35] [36] [28]. Within the scope of this thesis, we limit our exploration of combination methods to just one: *phrase table interpolation.*

Phrase table interpolation method merges multiple phrase tables into one phrase table, hoping that the obtained phrase table containes all potential phrase pairs with reasonable feature scores. The new feature scores of a phrase pair are estimated by interpolating feature scores of the phrase pair in all component phrase tables. Formally:

$$\begin{aligned}
\phi(s|t) &= \sum_{i=0}^{n} \lambda_i \times \phi_i(s|t) \\
\phi(t|s) &= \sum_{i=0}^{n} \lambda_i \times \phi_i(t|s) \\
p_w(s|t) &= \sum_{i=0}^{n} \lambda_i \times p_w^i(s|t) \\
p_w(t|s) &= \sum_{i=0}^{n} \lambda_i \times p_w^i(t|s)
\end{aligned} \tag{5.12}$$

Equations 5.12 show how the translation probability ϕ and the lexical probability p_w are estimated. The index i specifies one of the input phrase tables, each of which is given the interpolation coefficient λ_i on the condition that $\sum_{i=0}^{n} \lambda_i = 1$.

For this study, we have chosen the uniform distribution of λ_i to build a simple phrase table from other phrase tables. The uniform distribution is generally considered as a robust option. The investigation on different distributions as well as other system combination approaches can be a potential direction for future work.

6. Experiments and Results

This chapter describes experiments performed on the dataset that we have collected (see Chapter 3 for detailed description). It is divided into sections which contain the implementation plan which helps quantify and interpret the results and the experimental results themselves. The experiments are named, presented and discussed with regards to the translation quality, equivalent effort, improvements and degradations. The discussion also shows the comparison between different settings to highlight the contribution of each method.

We carried out various sets of experiments, each set focuses on a specific issue. An experimental set consists of a number of systems and SMT entities, which are given constant names. The four main experimental sets reported in this chapter are as follows: *baseline experiments*, *experiments with corpus filtering* which emphasize the improvement gained during preprocessing phase, *experiments with pivoting methods* which focus on multiple MT methods between Vietnamese and Czech without the direct parallel corpus, *experiments with system combination* which describe all possible methods and combine them in the the final system.

6.1 Experimental Setup

To facilitate correlation analysis among different systems, a consistent setting is used throughout all the experiments. All experimental systems are built in the same environment. They are trained with the same language models. They are optimized on the same development corpus and tested by the same test corpus. Finally, they are also evaluated by the same metric.

In this section, we describe the general setting which are shared by all experiments. For viewers who are familiar with SMT, the setting also provides a glimpse into how well the experimental systems will perform.

Translation System

For mass production of experiments, we have chosen the simple conventional phrase-based model instead of the state-of-the-art approaches, which demand extensive effort and knowledge. In this section, we provide the detail of the toolkit used to built the phrase-based model as well as to perform translation.

The statistical phrase-based machine translation system, Moses [14], is employed in this work to conduct experiments for Czech↔Vietnamese, Czech↔English and English↔Vietnamese translation. **Moses** is a statistical phrase-based machine translation system, which is "a complete out-of-the-box translation system for academic research. It consists of all the components needed to pre-process data, train the language models and the translation models".

We use *eman*[1], an experiment manager, to manage the translation process of *Moses*. *Eman* is an SMT tool for managing large numbers of computational experiments [30]. It facilitates both productivity and search for the best configuration and parameters of *Moses*.

Eman manages the translation process by breaking the whole pipeline into atomic tasks called *steps*. A *step* may correspond to a process in translation, such as training a language model, running tuning or translating a test set. A *step* may also correspond to a toolkit being used in the translation, such as preparing the Moses environment or downloading the latest version of *Treex*. The individual steps depend on each other. In other words, steps are combined to form experiments.

Each *step* has a type, such as *tm* (translation model) or *lm* (language model). A step is created by executing a corresponding *command*, which is generated by a corresponding *seed*.

Test Set Preparation

The details of data preparation for training are discussed in Chapters 3 and 4. This subsection focuses on the creation of test data.

General procedure of data preparation is to divide the whole dataset into training set, development set and test set. However, this method may lead to a few major disadvantages:

- Once the test set is created by dividing the whole dataset into three subsets, it is a local test set. It is not comparable to the performance of other systems.
- The test set, development set and training set have the same source. The test set is an *in-domain* test set. We aim to focus on a realistic condition, in which the training model and the test set are independent.
- Given the fact that we do not possess a multi-lingual dataset, three languages pairs demand three different test sets. The results will not be comparable between different language pairs.

Therefore, we decide to create a multi-lingual testing corpus, which is independent upon the training data. A *golden test set* is created by translating the test set of *WMT13 translation task*[2] into Vietnamese. The *WMT13 test set* is a multi-sources corpora. It contains data in 6 languages: *English, Czech, French, Spanish, German* and *Russian*. The source of *WMT13 test set* is internet news.

The translation of WMT13 test set are conducted through two stages, involving human translators. All translators are Vietnamese students who are studying in an English speaking University in Europe.

[1] `https://ufal.mff.cuni.cz/eman/`

[2] `http://www.statmt.org/wmt13/translation-task.html`

Firstly, we gather a group of 9 translators to translate the *English* side of WMT13 test set into *Vietnamese*. The data are divided into 52 pieces of text according to the sources of news. Each translator, working separately, is given a number of topics to translate. Out of 9 translators, 2 people decide to manually translate the test set and 7 people choose to post-edit the Vietnamese translation [3] of the test set.

Secondly, we gather another group of 6 people to post-edit the output of the first translation. This step aims to improve fluency of the Vietnamese output.

After two stages, a Vietnamese corpus with 3000 sentences is added to the *WMT13 test set*, creating a *golden corpus*, namely *newstest2013*. For this study, the *newstest2013* test contains three languages, *English*, *Czech* and *Vietnamese*. The *golden corpus* is divided into the development set and the test set, each contains 1500 sentences in every language.

Evaluation Measures

Evaluation is one of the most difficult tasks in MT. Along with the manual evaluation, we have selected the *BLEU* [37] to be the evaluation metric for this thesis. BLEU measures the similarity between candidate and reference by comparing the n-gram of two sentences. The n-gram matches for *BLEU* are independent to position. Based on the modified number of n-gram precision, the *BLEU* metric ranges from 0 to 1. However, general usage scales the score to 0 and 100. If the candidate is identical to the reference, the score is 100. If there are no similarities between the candidate and the reference, the score is 0.

Types of Experiments

In this study, we carry out a number of experiments for Czech→Vietnamese translation and Vietnamese→Czech translation.

We start with the baseline experiments followed by the experiments to observe the effect of a variety of improvement techniques that are applied to get the better translation quality. The main categories of experiments in this study are the following:

- Baseline Experiments
- Experiments with the corpus filter tool
- Experiments with various pivoting methods
- Experiments with a simple method for phrase table combination

[3] by Google Translate

By default, experiments are performed using the parallel corpora and monolingual corpora for training set. Development data set and test set come from the *golden set*, namely *newstest2013*, unless stated otherwise.

Language Models

While other sections primarily focus on the manipulation of bilingual corpora to create the translation model, this section is devoted to the language model. Even though the language model is not the main concern of this work, it contributes to the final result. Statistics of monolingual data provide an insight into the reason behind system performance.

As mentioned in *Chapter 2*, three monolingual corpora are collected for this study. They are used to built the language models in *English*, *Czech* and *Vietnamese*. The language models are consistently used in all experiments.

Table 6.1: Analysis of corpora used for language model

Language	**Corpus**	**Sentences**	**Overlap**	**OOV**	**OOV Unique**
English	mononews2012	14.7M	17	0.462%	2.266%
Czech	CzEng 1.0	14.8M	31	1.259%	3.435%
Vietnamese	mononews-vi monoVNTQ-vi	1.8M	4	1.570%	8.508%

Table 6.1 shows basic statistics between the monolingual corpora and the corpus *newstest2013*. In the table, *OOV* (Out-Of-Vocabulary) is the ratio of words which appear in the test corpus and do not appear in the training corpus. The *Overlap* column shows the number of identical sentences between the test corpus and the training corpus. Based on the statistic of the table, the overlapping ratio is considered low, 0.57% in English, 1.07% in Czech and 0.13% in Vietnamese. This implies the difference between the language model and the test/development data. An independent language model avoids a unbalanced model which could achieve high result with one test set and perform poorly with others.

The monolingual corpora for Czech and English languages are substantial, approximately 14.8 million sentences each. In contrast, the Vietnamese monolingual corpus is small, around 1.8 million sentences. This results in more out-of-vocabulary items in Vietnamese language model than OOV items in English language model and Czech language model. As for the translation models, the statistics is discussed in other sections. The column OOV Unique represents type-level OOV: the percentage of test set vocabulary not available in the vocabulary of the training corpora.

6.2 Baseline Systems

Our baseline setup is a plain phrase-based translation model derived from the direct bilingual corpora. The main focus of this thesis is the translation between

Vietnamese and Czech (both Vietnamese→Czech and Czech→Vietnamese). After carefully analysing, English is chosen to be the pivot language. Therefore, the set of baseline systems consists of 6 MT systems among three languages (Vietnamese ↔ Czech, Czech ↔ English and Vietnamese ↔ English). Even though only two baselines act as the starting point of the experiments, the comparison between all of the baselines illustrates the translation quality of machine translation regarding those languages.

Two sets of baseline are prepared. Each set serves a different purpose. For clarity, we assign each constructed system a unique identifier in the form $\mathbf{S}X_1X_2X_3$. X_1 disambiguates the translation pairs, in which Czech↔Vietnamese has $X_1 = 0$, English↔Vietnamese has $X_1 = 1$ and Czech↔English has $X_1 = 2$. X_2 indicates the group of systems, including *baseline 1*, *baseline 2* and further methods. X_3 is the identification number to distinguish systems of the same group.

- **Baseline 1**: Translation model is drawn from the original data (without pre-processing). They present the quality of plain phrase-based SMT system and the quality of parallel corpora available. They are the first systems in our experiments, acting as a starting point for all other systems. The set of *baseline 1* contains six systems:

  1. S001 - Direct translation from Czech to Vietnamese
  2. S002 - Direct translation from Vietnamese to Czech
  3. S101 - Direct translation from English to Vietnamese
  4. S102 - Direct translation from Vietnamese to English
  5. S201 - Direct translation from Czech to English
  6. S202 - Direct translation from English to Czech

- **Baseline 2**: Translation model was built after the data was pre-processed (including normalizing and filtering). Normalizing phase corrects sentences and removes bad sentence pairs based on the features of the corpus (subtitles). Filtering phase handles the quality of corpora by scoring the sentence pairs. The goal of *baseline 2* is to present the result of phrase-based SMT system with the best quality of parallel corpora. These systems are also in the experiments with other methods. Even though the final result would be higher if we used a state-of-the-art MT system in the pivoting approaches, the usage of our own baseline shows how much we gain by utilizing the data at hand. The set of *baseline 2* contains six systems:

  1. S011 - Direct translation from Czech to Vietnamese
  2. S012 - Direct translation from Vietnamese to Czech
  3. S111 - Direct translation from English to Vietnamese
  4. S112 - Direct translation from Vietnamese to English
  5. S211 - Direct translation from Czech to English
  6. S212 - Direct translation from English to Czech

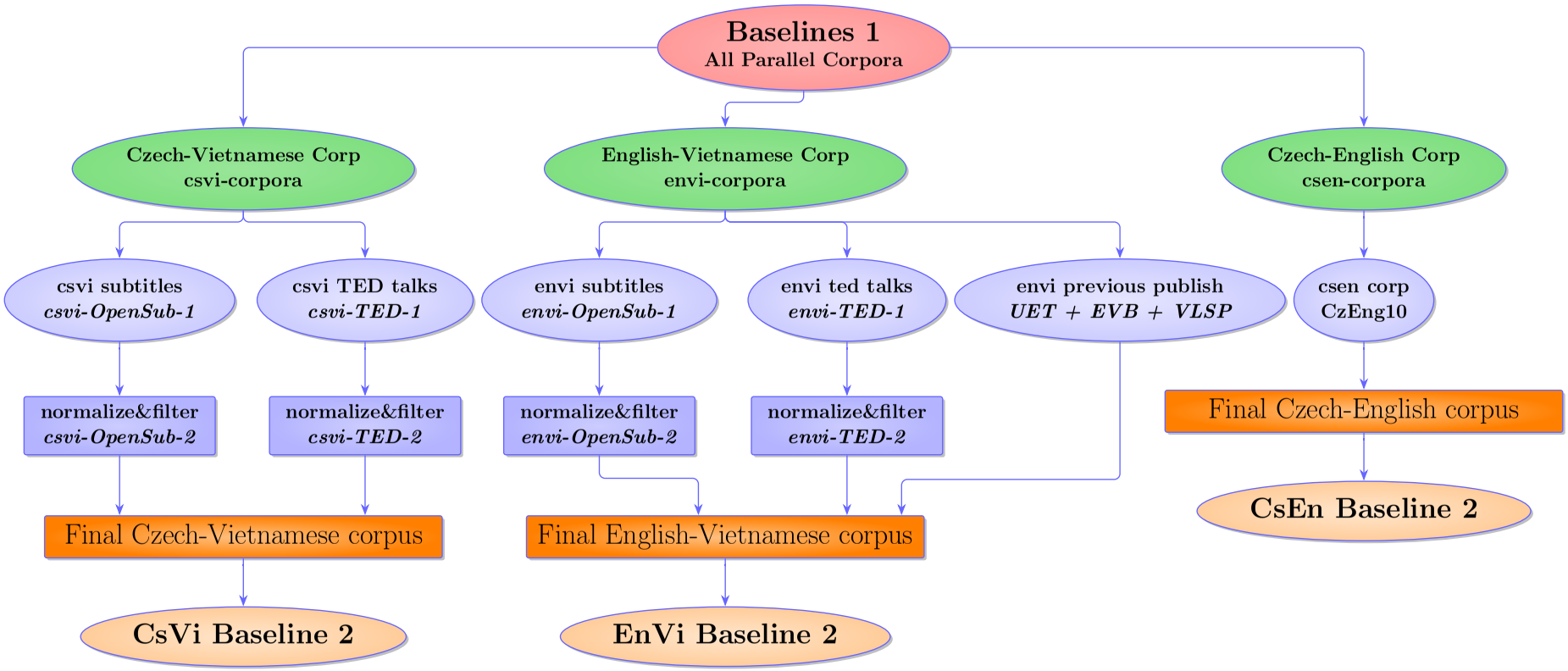


Figure 6.1: The corpora for two sets of baseline system

Figure 6.1 shows the overall process from original corpora to the baselines. Each pair of languages has its own datasets, which are processed differently.

Translation between Czech and English

Czech and English are the most familiar SMT pair at UFAL [4]. Instead of collecting fragmented data, this work is solely based on *CzEng 1.0* corpus [29]. We decided not to do any further pre-processing phase. Therefore, *baseline 1* system and *baseline 2* system between these two languages are the same.

Although *CzEng 1.0* is a large corpus, 15 million sentence pairs, the overlapping ratio between the *CzEng 1.0* and the *newstest2013* corpus are relatively small. Table 6.2 shows the statistics between the two corpora.

Table 6.2: Word analysis of Czech↔English corpora

Language	Line count	Word count	Overlap	OOV	OOV unique
English	14.8M	235.67M	34	0.690%	3.136%
Czech	14.8M	205.17M	31	1.259%	3.435%

Finally, the result of Czech-English baselines are as follows:

Table 6.3: BLEU scores of Czech↔English baselines

Source	Target	System	BLEU
Czech	English	S201, S211	23.23
English	Czech	S202, S212	15.26

Table 6.3 shows the BLEU scores of the baseline between *Czech* and *English*. It is considered a difficult language pair for SMT. The difficulty is attributed by the rich inflectional morphology and free word order of Czech. Further analysis on the matter could be seen in other work [29]. When translating into Czech, analysis shows that the word lemma is generally correct but its morphological properties such as the case are often incorrect.

Translation between Vietnamese and English

As mentioned in Chapter 3, we collect both *reviewed* existing corpora and *unreviewed* corpora. On the one hand, the *unreviewed* corpora are normalized and filtered. On the other hand, the *reviewed* corpora are leaved intact. In order to confirm the English↔Vietnamese evaluation, a further analysis of the bilingual corpora are carried out:

Table 6.4: Word analysis of English↔Vietnamese corpora

Language	Line count	Word count	Overlap	OOV	OOV unique
English	1.35M	12.8M	14	1.598%	7.782%
Vietnamese	1.35M	12.2M	3	1.720%	9.630%

[4] https://ufal.mff.cuni.cz/

Table 6.4 shows a basic analysis of the final English-Vietnamese corpora. There is an apparent similarity in the statistics between Vietnamese and English. Two languages have an approximately equal number of words per sentence, so does the out-of-vocabulary ratio. Partially, this is the result of similar linguistic features. Both English and Vietnamese are strict in word order. While English s a small number of inflectional morphemes, Vietnamese does not inflect words at all. Therefore, it is predictable that the translation between English and Vietnamese is easier than for other pairs, such as Czech-English.

Table 6.5: BLEU scores of English↔Vietnamese regarding fragmented data

Source	**Target**	**Corpus**	**Status**	**Corpus Size**	**BLEU**
Vietnamese	English	reviewed	intact	0.24M	33.32
Vietnamese	English	unreviewed	unnormalized	2.04M	30.49
Vietnamese	English	unreviewed	normalized	1.77M	30.98
English	Vietnamese	reviewed	intact	0.24M	32.85
English	Vietnamese	unreviewed	unnormalized	2.04M	29.28
English	Vietnamese	unreviewed	normalized	1.77M	29.96

Table 6.5 shows that the BLEU scores in translation between *English* and *Vietnamese* are relatively high. There are two main reasons. First, the characteristics of two languages share a number of common features. Both of them have a small number of word forms. Both of them rely heavily on the word order. Second, we obtained the test set by post-editing a machine translation of WMT13 news test, which may result in a favourable setting for phrase-based translation.

Even though the size of *unreviewed* corpora is much larger than the size of *reviewed corpora*, the BLEU score is considerably lower. It is because of the fact that *unreviewed* data is noisy, and even the normalization phase could only mitigate the problem.

Table 6.6: BLEU scores of English↔Vietnamesebaselines

Source	**Target**	**System**	**Baselines**	**BLEU**
Vietnamese	English	S101	Baseline 1	33.33
Vietnamese	English	S111	Baseline 2	33.88
English	Vietnamese	S201	Baseline 1	32.84
English	Vietnamese	S211	Baseline 2	34.45

Table 6.6 shows the BLEU scores of the baselines after we merge two sets of corpora together. English→Vietnamese translation gains more than Vietnamese→English translation. The improvement in BLEU score is 1.61 BLEU point and 0.54 BLEU point respectively. When the two corpora are merged, it provides a greater set of lexical choices. In the phrase table, it is more common for one English word to have various Vietnamese counterparts than one Vietnamese word to have various English counterparts. On average, the ratio is approximately 1.49 according to our observation. This is the main reason why the English→Vietnamese MT system benefits significantly from adding data to the English↔Vietnamese corpus.

Translation between Czech and Vietnamese

To the best of our knowledge, our project is the very first research which take the Czech-Vietnamese pair into account. The lack of data and the difficulty posed by difference between two languages have become probably the major reason why it did not attract sufficient interest so far.

The final goal of this work is to improve translation between Czech and Vietnamese. As discussed above, all Czech↔Vietnamese data are obtained as the *unreviewed* ones. Compared to English↔Vietnamese counterpart, the corpora are not only smaller in size but also noisier.

Table 6.7: BLEU scores of Czech ↔ Vietnamese baselines

System	Source	Target	Status	Corpus Size	BLEU
S002	Vietnamese	Czech	unnormalized	1.33M	5.34
N/A	Vietnamese	Czech	normalized	1.15M	7.32
S012	Vietnamese	Czech	normed&filtered	1.09M	7.62
S001	Czech	Vietnamese	unnormalized	1.33M	8.78
N/A	Czech	Vietnamese	normalized	1.15M	10.59
S011	Czech	Vietnamese	normed&filtered	1.09M	10.57

Table 6.7 shows the result of Czech↔Vietnamese baselines. Overall, there is a significant improvement after the normalizing stage but almost no improvement from filtering stage. The promising result gains from normalizing stage comes from the noise elimination. A large chunk of simply bad sentence pairs is removed. It helps the alignment to pick up the correct alignment between two languages. Moreover, after a sentence is normalized, clauses (phrases) are handled better than unnormalized data when there are non-alphabetical characters appearing between words, e.g. unnecessary punctuations. The process of filtering data will be discussed further in the following section.

Table 6.7 also shows a poor performance of Czech↔Vietnamese translation. The best translation into Czech achieves just *7.62* BLEU points and the translation into Vietnamese gets *10.57* BLEU points. The errors of translation spread across all types. The poor performance of Czech↔Vietnamese translation is also caused by the difference of two languages. This problem is similar to Czech↔English translation, in which a baseline from a corpus of *15M* sentences achieves just approximately 15 BLEU points.

6.3 Experiments with Corpus Filter

In the previous section, we reported improvements made by the first data preparation phase, called *normalizing*, regarding the performance of *baseline systems*. This section further focuses on the importance of corpus quality by reporting the experiments with the second data preparation phase, *filtering*.

As described in Chapter 4, after normalizing the *unreviewed* corpora, we decided to filter out poorly word-aligned sentence pairs with *the filtering tool from CzEng 1.0*. Section 4.2 shows an analysis of the tool on a manual annotated dataset taken randomly from Czech-Vietnamese and English-Vietnamese *unreviewed* corpora. Here, various experiments performed on the whole *unreviewed* dataset are reported. They highlight the real effect of the filtering tool on the performance of SMT systems.

Detailed description of the filtering phase is shown in Chapter 4. This section briefly discusses the usage of *the filtering tool from CzEng 1.0* in a practical point of view. Two corpora are taken as the input of the filtering tool, namely Czech-Vietnamese *unreviewed* corpus and English-Vietnamese*unreviewed* corpus. Regarding the *English-Vietnamese* corpora, we decided to filter the *unreviewed* corpus while leaving the *reviewed* corpus intact.

The filtering tool from CzEng 1.0 offered a machine learning system to score every *source-target* pairs based on a combination of features. The tool is designed to work in *Treex*. It supports a wide ranges of features on different layers of *Prague Dependency Treebank*, notably *w*-layer, *m*-layer, *a*-layer and *t*-layer. Unfortunately, the available resources of *Vietnamese* supports only the lower layers. Thus, the final feature combination in our experiments contains following features:

- Alignment cumulation
- Alignment score
- Different number of tokens
- Identical sentences
- Letter count
- Long sentence
- Long word
- Reordering quantity
- Repeated character
- Special characters ratio
- Suspicious character

Machine Learning classifier based on the annotated data, see Section 4.2, allows the possibility of setting different thresholds. In the evaluation process, the classifier scores every sentence pair of the dataset. The sentence pair is kept only if the score is greater or equal than the threshold. Section 4.2 shows that the correlation between the score and the real quality of word-aligned sentence pair is not trivial. It also shows a difference in the behaviour of two corpora, Czech-Vietnamese and *English-Vietnamese*.

For each threshold, we obtain a new smaller corpus from the remaining sentence pairs. An SMT system based on the new corpus is constructed and evaluated in a consistent setting. Figure 6.2 and figure 6.3 show the BLEU scores for increasing thresholds in Czech→Vietnamese corpus and English→Vietnamese corpus respectively.

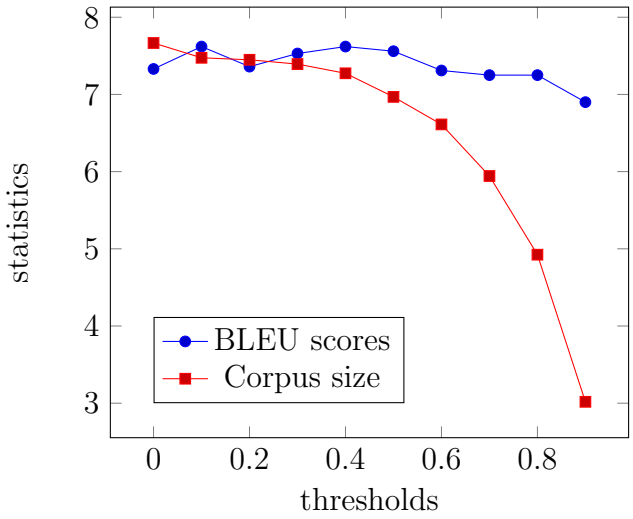


Figure 6.2: Performance of corpus regarding filter thresholds on Czech→Vietnamese translation

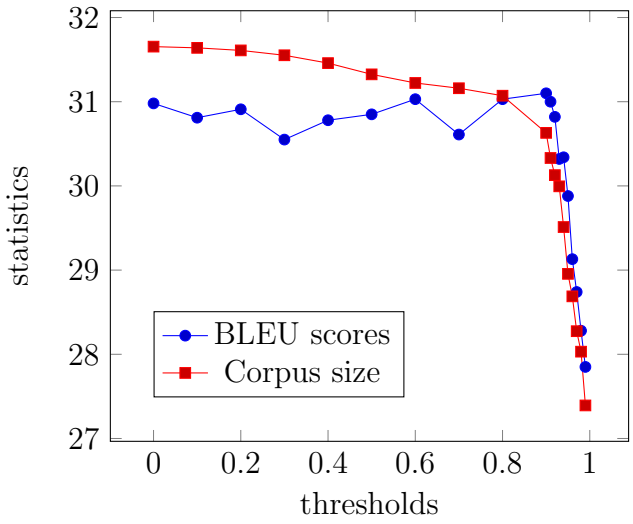


Figure 6.3: Performance of corpus regarding filter thresholds on English→Vietnamese translation

Two Figures 6.2 and 6.3 show different behaviours of English→Vietnamese translation and Czech→Vietnamese translation when the threshold increase from 0 to 1. An important point to understand is that the size of corpus always gets smaller (to the point of 0 remaining sentence pair) when the threshold increases. A small corpus size generally leads to a poor performance of SMT, thus affecting the BLEU score.

The two line graphs of *corpus size* show two different patterns regarding the two corpora. The size of Czech→Vietnamese corpus gradually drops as the threshold increases. It is the direct result of an even distribution of sentence quality scores. On the other hand, the English→Vietnamese corpus size slightly dwindles until the threshold is 0.9 and then it plunges. It is because the majority of English→Vietnamese sentence pairs get high score. The point $threshold = 0.9$ is also the point when the English→Vietnamese translation acquires highest result, making this point a perfect threshold for filtering tool. On the contrary, the BLEU score of Czech→English translation fluctuates on a wide range of thresholds. The problem, discussed on Section 4.2, is a trade off between the quantity and quality of Czech→Vietnamese corpus. It shows that the threshold chosen based on the precision and recall of the annotated data, 0.4, is proved to be a good pick. It preserves a 82% of Czech→Vietnamese corpus while achieves the highest BLEU score.

To put it briefly, experiments conducted on the corpus filtering tool have shown promising results. In this light, the final corpora are established with *filtering*.

6.4 Experiments with Pivoting Methods

This section discusses the experiments using pivoting methods (see Chapter 5 for detailed description). It focuses on implementing and discussing results of Czech↔Vietnamese translation via *English*. The rest of the section continues with a brief description of the data available, then detailed scenarios of pivoting methods.

As discussed in previous chapters, we have decided to use only one pivot language, *English*, for Czech↔Vietnamese translation. Hence, the task for pivoting method was to build a Czech→Vietnamese MT system and a Vietnamese→Czech MT system based on our *English-Vietnamese* and *English-Czech* corpora.

Firstly, the final set of English-Vietnamese corpora of corpora was as follows:

1. VLSP corpus
2. EVB corpus
3. UET corpus
4. A normalized and pruned corpus of Open Subtitles
5. A normalized and pruned corpus of Ted Talks

The first three corpora were merged to form the English-Vietnamese*reviewed* corpus. The last two corpora were merged to form the English-Vietnamese*unreviewed* corpus after normalizing and filtering. On preparing resources for the pivoting methods, which required English↔VietnameseMT systems, two SMT baselines were built based on the combination of *reviewed* and *unreviewed* corpora. They

were *S111* and *S112*, which were derived from two phrase tables *PT111* and *PT112* respectively.

Secondly, the final Czech-English corpus was *CzEng 1.0*.

Finally, the MT problem could be formulated as follows: "Given the Czech-English corpus, namely *CzEng 1.0*, and the English-Vietnamesecorpora, namely *reviewed* and *unreviewed*, the task is to deliver the translation between Czech and Vietnamese in both directions". The problem could be resolved using pivoting methods, each of which leads to a separate system. For this thesis, we do not include the experiments of *synthetic corpus* method.

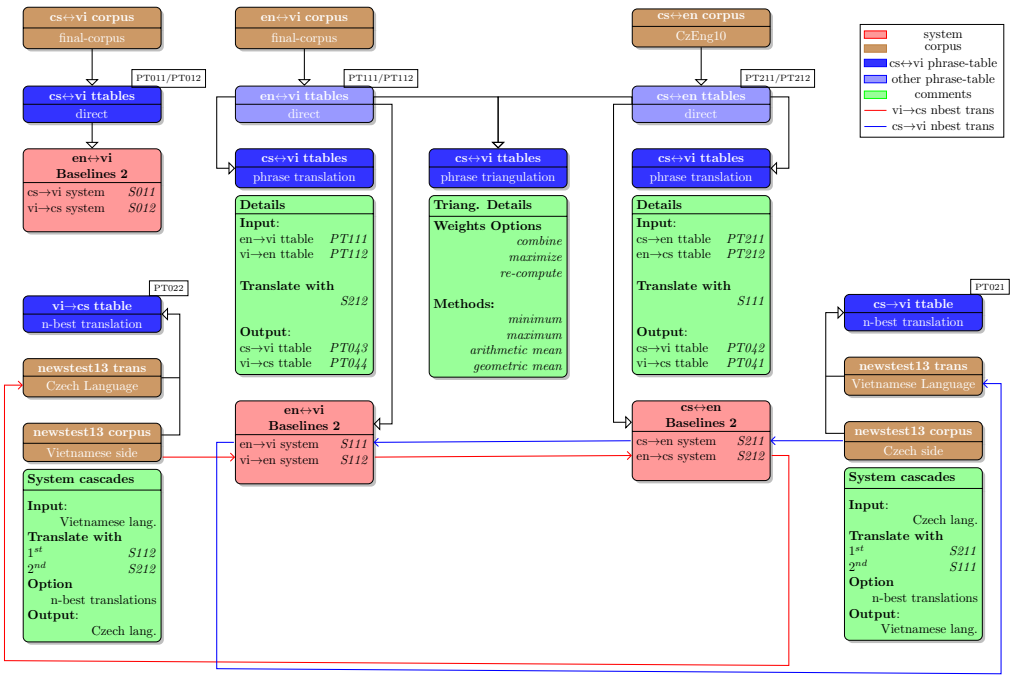


Figure 6.4: An overview of our MT systems and phrase tables

Figure 6.4 shows the overall framework which produces multiple *Czech-Vietnamese* phrase tables in our experiments. Each phrase table represents a specific scenario of either direct translation or pivoted translation.

System 1: System Cascades

According to the *system cascades* approach, the input in source language was first translated into pivot language, then the pivot hypothesis was translated into target language. In our experiment, we tried out different numbers of n. Hence, this method is also refered to as *n-best translation* method.

The theory behind this approach is to maximize the score of $P(f|e)$ among the possible pivot sentences p_i in which i ranges from *1* to n. If $i = 1$, it is the simple

scenario that one sentence is translated twice by two MT systems.

$$P(f|e) = argmax_{i=0}^{n}(P(f|p^i) \times P(p^i|e)) \tag{6.1}$$

From an experimental point of view, this approach involves two completed machine translation systems. The first system translates a sentence from the source language (i.e. Vietnamese or Czech) to n best sentences in pivot language (i.e. English). The second system translates those n sentences into n outputs in target language (i.e. Czech or Vietnamese). Then, $n \times n$ target sentences are ranked. Their scores are the combination of the source-pivot score and pivot-target score.

Table 6.8 shows the result of translating between *Vietnamese* and *Czech* via *English*. The Czech→Vietnamese translation direction is done through S211 and S111. The Vietnamese→Czech translation direction is done through S212 and S112.

Table 6.8: System cascades of Czech↔Vietnamese translation via English

N	1	2	5	10	20	30	50	75	100
cs→vi	9.05	9.19	9.33	9.50	9.70	9.70	9.80	**9.82**	9.82
vi→cs	13.35	13.51	13.65	13.71	13.77	**13.83**	13.73	13.75	13.79

A glance at Table 6.8 shows that BLEU scores of n-best translation approach increases when n increases. It is reasonable that the larger n outperforms $n = 1$. More hypotheses in the pivot language increase the chance of finding a good final translation. However, the improvement between n-best translation and *1*-best translation is only marginal.

Compared to Czech↔Vietnamese baselines (S011 and S012), *system cascades* performed significantly better. The result was somewhat beyond our expectation, given the fact that *system cascades* was a straightforward approach.

Table 6.9: N-gram analysis of Vietnamese→English translation

N	1	2	3	4	5	6
Vietnamese	10487	5114	2409	615	90	13
English	10627	4671	2316	814	220	68

Table 6.9 shows the distribution of phrase lengths used in the translation from Vietnamese to English. The average length of Vietnamese phrases is 1.652 while the average length of English phrases is 1.693. The analysis shows that the number of long phrases which got translated from Vietnamese to English was relatively small. Instead, various short English phrases are concatenated to construct a translated sentence.

For example, a Vietnamese phrase "số lượng người thiệt mạng do sét đánh" was partitioned into three pieces "số lượng", "người thiệt mạng" and "do sét đánh". These pieces were translated into "the number of", "people killed" and "by lightning" respectively. The translated piece were merged together to form an English

phrase "the number of people killed by lightning". Overall, the translation process was highly accurate.

The promising performance of *system cascades* relied on the fact that the method uses complete translation steps. This led to flexible phrase handling compared to other pivoting methods, which generate artificial phrase tables. During the translation process, pivoting sentences were broken into phrases separately for each of the two phrase tables. Only a small portion of phrases remained intact during the process. In most of the cases, the segmentation into phrases was different for the pivot-target translation and for the source-pivot translation.

For the English→Czech translation, the input is the obtained hypotheses from Vietnamese→English translation. There is no guarantee that the English hypotheses are grammatically correct. We are interested in the number of phrases which are translated from Vietnamese to English and remained as one phrase when they are translated into Czech.

Table 6.10: N-gram analysis of English→Czech translation

N	1	2	3	4	5	6	7
English	12705	5163	1814	578	155	27	7
Czech	15253	3726	1098	280	71	17	4
Preserving Segmentation Phrases	6461	1524	500	151	27	3	0

Let us take an example of *system cascades* with the Vietnamese→Czech translation. The last row of Table 6.10 shows the number of English phrases surviving two translations. Surviving phrases are the phrases composed by Vietnamese→English translation and preserved when the English hypothesis is decomposed during English→Czech translation. It is shown that the ratio of surviving phrases drops rapidly when n increases from 1 to 7. If unigram $n = 1$ is not taken into account, the percentage is about 25%.

Compared to other pivoting methods, system cascades bring both advantages and disadvantages. On one hand, the method is robust in handling phrases. The translation process adapts to the different segmentations needed by the two phrase tables rather than totally relying on the segmentation of the first phrase table. On the other hand, it creates noises during the translation process. The sentence meaning can deviate greatly after translating twice by two different systems.

Arguably, the quality of the first MT system (source-pivot translation) plays a crucial role in the final performance. If the first MT system produces a low quality output, the second MT cannot deliver a good translation. This problem is affected by the choice of the pivot language.

Phrase Table Translation Method

The *phrase table translation method* involves translating one side (pivot language) of a phrase table into source language or target language. Given two phrase tables: *source-pivot* phrase table and *pivot-target* phrase table obtained from the

corresponding corpora, there are two options: translating the *pivot* side of the former phrase table into *target* language by the system derived from the later phrase table or translating the *pivot* side of the later phrase table into *source* language by the system derived from the former phrase table.

The output of this approach is Czech↔Vietnamese synthetic phrase table and Czech↔VietnameseMT system, which can be drawn from two sources:

- Translation of English↔Vietnamese phrase table
    1. Obtain Czech→Vietnamese phrase table (*PT043*) by translating the source side of English→Vietnamese phrase table (*PT111*) with *S212*. Final result is a Czech→Vietnamese MT system (*S043*)
    2. Obtain Vietnamese→Czech phrase table (*PT044*) by translating the target side of Vietnamese→English phrase table (*PT112*) with *S212*. Final result is a Vietnamese→Czech MT system (*S044*)
- Translation of Czech↔English phrase table
    1. Obtain Czech→Vietnamese phrase table (*PT042*) by translating the target side of Czech→English phrase table (*PT211*) with *S111*. Final result is a Czech→Vietnamese MT system (*S042*)
    2. Obtain Vietnamese→Czech phrase table (*PT041*) by translating the source side of English→Czech phrase table (*PT212*) with *S111*. Final result is a Vietnamese→Czech MT system (*S041*)

For this method, it is necessary to have both English language model and Vietnamese language model. This additional information might play an important role in the translation. Overall, this method generates two systems for each translation direction of Czech↔Vietnamese. They are named as *System 2*. The difference when translating English↔Vietnamese phrase table and Czech↔English phrase table is the system used to translate it. For English↔Vietnamese phrase table, an English→Czech system is required while an English→Vietnamese system is required for Czech↔English phrase table. The performance of SMT systems decides the quality of the Czech↔Vietnamese phrase table. Previous analysis has shown that English→Vietnamese systems generally performs significantly better than English→Czech systems.

Table 6.11: BLEU score of phrase translation for Czech↔Vietnamese

System	**Direction**	**Number of phrases**	**OOV**	**BLEU**
S041	Vietnamese→Czech	20044	2594	**8.40**
S044	Vietnamese→Czech	17530	596	7.34
S042	Czech→Vietnamese	20044	2345	**12.09**
S043	Czech→Vietnamese	17530	654	9.67

Table 6.11 shows the statistics of two methods with regards to the translation between Vietnamese and Czech. The artificial phrase table translated from

English→Vietnamese phrase table encounters more out-of-vocabulary phrases than the artificial phrase table derived from Czech→English phrase table even though PT041 is larger than PT044. This is because the source side of PT044 is natural language while the source side of PT041 is made-up language. When it comes to PT042 and PT043, the issue is different. PT042 not only has larger size but is also translated by better MT systems. It results in a larger gap of BLEU scores between the two systems.

Overall, the artificial phrase tables obtained from *CzEng 1.0* phrase tables acquire more promising results than the artificial phrase tables obtained from English↔Vietnamese phrase tables. It shows that the source phrase tables and the translation systems have greater impact on the final performance, compared to the translated side of phrase tables.

The result has shown that *phrase table translation approach* surpasses the baseline of direct translation but falls behind the *system cascades*. Between two pivoting methods, *phrase table translation approach* is bound to the phrases in the phrase table while *system cascades* handle phrase boundaries in a more flexible way. Especially when the overall quality of phrase table is low, n-best translation has proved to be superior in handling the translation.

Phrase Table Triangulation Method

As discussed in Chapter 5, phrase table triangulation method generates an artificial *source-target* phrase table by directly joining two phrase tables, *source-pivot* and *pivot-target*. No other resource such as a parallel corpus is needed. Given the input from two phrase tables, there are two options for computing phrase translation probabilities and lexical translation probabilities of a *source-target* phrase table: We can either manipulate the original phrase co-occurrence counts or we can combine the probability estimates.

Pivoting the Co-occurrence Count of Phrase Pairs

From the phrase count in the two phrase tables, we establish the phrase count of *source-target* phrase table. Then the phrase translation probability and lexical probability are computed based on the new counts and word alignment. When approximating the co-occurrence counts of the *source-target* phrase pairs, we can combine the $s - p$ and $p - t$ counts using *minimum*, *maximum*, *arithmetic mean* and *geometric mean*.

The mean values are expected to perform well when the source-pivot and pivot-target parallel corpus differ greatly in size. In that case, *minimum* function would always take the value from the small corpus while *maximum* function would take the value from the large corpus. In our case, the *Vietnamese-English* corpus has the size of 10% the size of *English-Czech* corpus. The difference is significant but not extreme.

Pivoting the Features Probability of Phrase Pairs

The probabilistic formulation of phrase translation distribution in *source-target*

phrase table is estimated by multiplying phrase translation probabilities from the *source-pivot* phrase table and the *pivot-target* phrase table. As discussed above, one *source* phrase might be paired with one *target* phrase via one or more *pivot* phrase. The desired conditional over the *source-target* pairs could be the *maximum* or the *summation* of all the triples *source-pivot-target*.

Two options to estimate the distribution of *source-target* pairs could be explained as: (max) choosing the most prominent translation of an expression or (sum) considering all the translations of an expression.

Empirical Evaluation of Phrase Table Triangulation Options

To sum up, we proposed 6 options to pivot two phrase tables into one final phrase table. All 6 options will result in the same set of *source-target* pairs but they differ in the scores. In general, the approach seems to generate a lot of noisy *source-target* pairs caused by common phrases. For example, one common phrase p in *third* language is paired with n phrases in *source* language. At the same time, it is paired with m phrases in *target* language. After triangulating, the final phrase table will consist of $n \times m$ phrases, most of which are inaccurate. The more common p is, the higher the values of m and n are.

We wrote a program performing *phrase table triangulation* for all 6 options. Using the script, six Czech-Vietnamesephrase tables were generated for each translation direction.

Table 6.12: Six options of Phrase triangulation for Vietnamese→Czech direction

Direction	**Method**	**Option**	**BLEU**
Vietnamese→Czech	Co-occurrence Count	minimum	7.24
Vietnamese→Czech	Co-occurrence Count	maximum	6.38
Vietnamese→Czech	Co-occurrence Count	arithmetic-mean	6.25
Vietnamese→Czech	Co-occurrence Count	geometric-mean	7.05
Vietnamese→Czech	Probabilities	sumarization	**7.44**
Vietnamese→Czech	Probabilities	maximization	7.21

Table 6.12 shows the results of all 6 options when translating from Vietnamese to Czech. To our surprise, all 6 options achieved lower BLEU scores than other pivoting methods. The primary reason was the high level of noise created by triangulation. The source of the problem lies in the fact that phrases are combined without considering any context or different meanings of the expressions. Besides, re-computing co-occurrence count appeared to be less effective than re-computing the probabilities directly. The primary reason was the difference between two phrase tables. The Czech-English phrase table was much larger than the English-Vietnamese phrase table. Hence, the gap in co-occurrence counts were large. When co-occurrence counts were estimated, the noisy phrase pairs became a legitimate pair with the counts mostly taken from the small phrase table. They were treated equally when the new co-occurrence is estimated. Hence, the noisy pairs acquired probabilities as high as the valid pairs. This situation worsened when the new co-occurrence counts were computed based on *maximum* or *arithmetic-mean* because of the high number of occurrence count of common phrases in *third* language.

Another observation shows that computation of the new probability favours *summation* over *maximization*. It is reasonable that the final probability of a *source-target* pairs should be computed over all middle-phrases rather than just one phrase. One unit (word or phrase) may have more than one translation in other language.

Compared to other methods, phrase triangulation appears to be less effective even though it covers a wider set of senses for every word. A further analysis on this matter will be discussed in section of *combination method*. However, it is worth pointing out that the phrase table of phrase triangulation method provides many good translation regardless of its poor final result.

Table 6.13: Six options of Phrase triangulation for Czech→Vietnamese direction

Direction	**Method**	**Option**	**BLEU**
Czech→Vietnamese	Co-occurrence Count	minimum	9.86
Czech→Vietnamese	Co-occurrence Count	maximum	7.64
Czech→Vietnamese	Co-occurrence Count	arithmetic-mean	6.95
Czech→Vietnamese	Co-occurrence Count	geometric-mean	9.24
Czech→Vietnamese	Probabilities	sumarization	**10.28**
Czech→Vietnamese	Probabilities	maximization	9.64

Table 6.13 shows the result of six options when translating from Czech to Vietnamese. Overall, the result provides a similar picture as the *Vietnamese→Czech* direction. Computing the new features by the features of two component phrase tables again outperforms the method which recomputes co-occurrence counts. In conclusion, phrase table triangulation method is shown to be less effective than *system cascades* regardless of the extensive effort. In the final combination, we include only the best triangulation option, marked as *System 3*.

6.5 Experiments with Phrase Table Interpolation

In previous sections, multiple methods including both direct approach and pivoting approaches were described. This section reports an experiment combining obtained phrase tables to build an interpolated phrase table in both directions. The system which was built based on interpolated phrase table can be seen as a combination of all the preceding systems.

For Vietnamese→Czech translation, a set of Vietnamese→Czech phrase tables are constructed as follows:

- One phrase table from the Vietnamese-Czech corpus, which are normalized and filtered.
- Two phrase tables from the phrase table translation approach. They are corresponding to the translations of Czech-English phrase table and English-Vietnamesephrase table.

- One phrase table from the phrase table triangulation approach.
- One phrase table obtained from the output of system cascades.

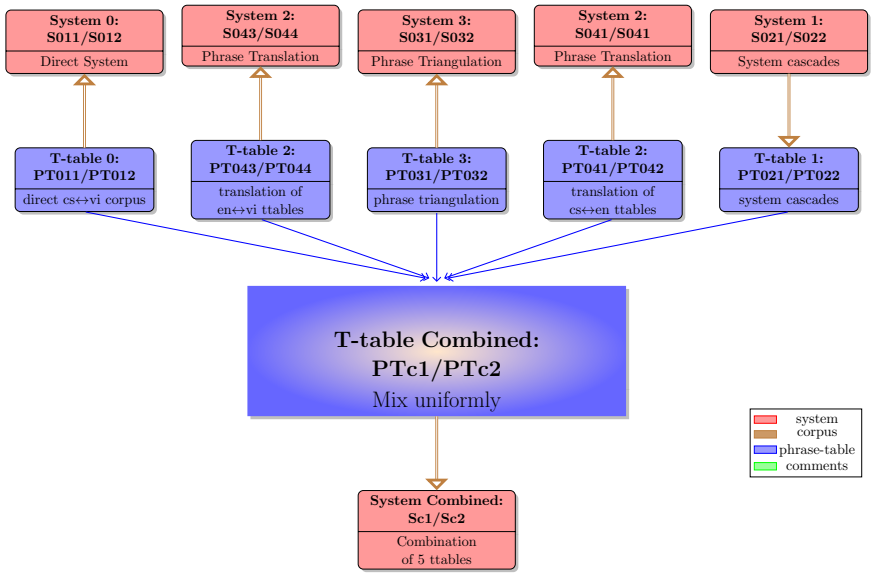


Figure 6.5: Overview of the combined system and its components

Figure 6.5 shows a basic combination of multiple systems. All phrase tables were combined uniformly to create the final phrase table. The newly created phrase table was then used to build the final system.

Table 6.14: System combination for Vietnamese→Czech direction

Description	**Ttable Size**	**Phrase Contrib.**	**BLEU**
Direct Translation	8.7M	1137	7.62
Trans. of en→cs ttable	53.21M	3802	8.4
Trans. of vi→en ttable	19.2M	2812	7.24
Phrase Triangulation	61.5M	1255	7.44
System Cascades	0.08M	5306	9.82
Combination	95M	8957	**10.12**

Table 6.14 shows size and scores of the final combination. The column *Phrase Contribution* shows the number of phrases selected by the translation, regarding each component phrase table. Among 5 source phrase tables, the synthetic phrase table obtained from the n-best translation proved to be the most effective. Its contribution contained long phrases while other phrase tables contributed mostly short phrases.

On the other hand, the phrase table triangulation method appears to be less productive. It produces the largest phrase table, yet the number of phrases selected from this phrase table is the second smallest one.

Let us take an example to illustrate why phrase triangulation method generates much more data than other methods, yet it is less effective. The word "horká" is a common Czech one-word phrase. It had 4 translations in the direct phrase table PT, in which 1 translation is correct for the given phrase. The artificial phrase table translated from Czech-English phrase table provides 1 correct translation. The artificial phrase table translated from English↔Vietnamese does not have any translation, and so does the synthetic phrase table of system cascades. The phrase triangulation method generates 106 translations, almost uniform in weight. Another problem of phrase triangulation is that the translations deviate significantly from the word sense of "horká". For rare words, our observation showed that phrase table triangulation methods actually provided better quality of phrase pairs. However, the proportion of rare phrases is also relatively small. Hence, it is indeed best to combine this method with other methods.

The smallest contribution is made by the direct phrase table between Vietnamese and Czech. It reflected the quality and the small size of the direct corpus, even when the corpus had been cleaned and filtered.

It is a positive result that the final combination achieved a relative high score compared to other systems. The statistics shows that most of the phrases come from the system cascades method, which has also been the most versatile method. Very similar observations also hold for Czech→Vietnamese translation.

Table 6.15: System combination for Czech→Vietnamese direction

Description	**Ttable Size**	**Phrase Contrib.**	**BLEU**
Direct Translation	8.7M	1025	10.59
Trans. of cs→en ttable	54.1M	3762	12.09
Trans. of en→vi ttable	19.2M	2950	9.67
Phrase Triangulation	61.5M	1212	9.86
System Cascades	0.08M	5212	**13.83**
Combination	95M	8642	**13.80**

Table 6.15 shows that the performance of combination system is similar to the performance of system cascades. Other phrase tables provide a wider set of lexical choices for translation but the phrases from system cascades were generally selected. The primary reason is that the synthetic phrase table of system cascades is obtained from the test set itself.

6.6 Discussion and Future Work

In previous chapters, we described our experiments to build Czech↔Vietnamese SMT systems and to improve their quality by pivoting methods. This chapter analyses all the steps which have been done. It provides the assessment of the outcome, primary reasons as well as potential directions for future work.

First of all, our decision to build the baseline based on simple phrase-based statistical machine translation enables the experiments to be comparable, yet hurts the performance of SMT systems among Czech, English and Vietnamese. The usage of third party services, such as a state-of-the-art MT system, would definitely shore up the translation quality.

Regarding the translation quality, the very low BLEU scores of Czech→Vietnamese translation indicates an ill performance of SMT between the two languages. After carefully examining the outputs, we confirm that the constructed systems suffer from severe errors, notably lexical choice, word order and grammatical error. There are a few major reasons. Firstly, the linguistic difference between a functional language and an analytic language results in a highly challenging language pair. Secondly, our experimental setting relies heavily on statistics. It does not exploit the features of natural language. Finally, the majority of current corpora comes from video subtitles (including transcripts), which lead to a degradation in the corpus quality. Moreover, the development and test set, however, is derived from *WMT 13 test set* (news test), which partially turns the translation issue into a cross-domain translation problem.

As described in Chapter 3, some of the data sources, which are freely available for research, have been left out during the making of this thesis. Carefully exploiting the new sources would not only increase the size of dataset but also improve the diversity of corpus domains.

For this work, we decided to focus on the pivoting methods rather than to exploit the language features of *Czech* and *Vietnamese*. During the phase of data preparation using *CzEng 1.0* filtering tool, this work relied on a combination of a set of very basic language features. The experiments acquired promising results, highlighting the effectiveness of the filtering tool. However, if we implemented some linguistically more informed features, the performance would likely increase. This involves a more thorough research of natural language processing tools for Vietnamese and Czech. The language features could also be applied on the translation process to enhance the quality of translation.

Although the pivoting experiments are not fully in line with our expectations, the results are positive. It is worth mentioning that our initial goal, beating the baseline by pivoting method, is accomplished. Most of pivoting SMT systems, except the ones based on *phrase table triangulation* method, achieve better results than the first baseline systems.

The fact that phrase table triangulation method acquires lower scores than other pivoting methods is an unexpected result. The phrase table triangulation method demands greater effort and seems promising at the first sight. The phrase table

obtained is significantly larger than other phrase tables. It provides more translations for every phrase. However, analysis shows that the distribution acquired from phrase table triangulation method is greatly unbalanced between rare phrases and frequent phrases. The majority of the phrase table are short and common phrase pairs, in which many of those are mismatched. This is a key point which distinguishes the pivoting translation, which is addressed by this thesis, with the multi-source translation. The independence between *source-pivot* corpora and *pivot-target* corpora results in a large amount of noise created by triangulating two phrase tables. This problem gives rise to a future work, which involves pruning the triangulated phrase table or using the method only for rare wordss.

Of all approaches to the *phrase table triangulation*, our experiments show that pivoting the co-occurrence counts, which seems more principled, is less effective than the conventional method which computes the new probabilities based on the source probabilities. There are two main reasons. The first reason is the size difference of the two phrase tables. The estimate of co-occurrence is bias one phrase table, regarding the scenarios. The second reason is the exposure to noise. It leads to an uniform distribution of the newly created phrase pairs. Hence, the gap between the correct and incorrect phrase pairs are relatively small.

Among all pivoting methods, *system cascades* obtained the best results. During the translation process, sentences are broken into phrases and then re-joint before the second stage system gets them. The flexible process delivers a higher quality output than the rigid synthetic phrase tables. Our experiments also show that the improvement of n-best translation over *1*-best translation is marginal.

Finally, the system combination based on mixing phrase table achieves promising results, which are slightly better than the component systems. This is a positive result that we are proceeding in the right direction. One potential direction for future work would be to convert the text input into a word lattice input, which allows multi-lingual input.

7. Conclusion

The thesis describes and experiments with various SMT methods designed for Vietnamese. The methods, including pivoting methods and standard methods, are implemented and evaluated, to compare their performance in a stable setting. The system performance is evaluated based on a *golden test* of multilingual corpus, manually translated from the *WMT13 test set*.

We decided to experiment with the phrase-based SMT approach completely based on the resources that we have collected, rather than using third party services. The thesis describes a consistent work in which all pivoting methods are implemented, evaluated and compared with each other. This is one contribution of our work.

The work starts with the collection of bilingual corpora and monolingual corpora for three languages, namely Czech, English and Vietnamese. Regarding three languages, their resources are collected in specific ways, such as reusing clean corpora published by other works, gathering the noisy data from *OPUS* and crawling the raw data from *TED* talks. Afterwards, we further clean up the set of noisy corpora, called *unreviewed* corpora, by the data description and the filtering tool of *CzEng 1.0*. Based on the remaining data, we continue with the plain phrase-based SMT to prepare multiple baseline systems among the three languages. A comparison between two baseline systems, before and after the data are cleaned, highlights the importance of the data preparation phase.

We continue this thesis with various pivoting methods: *system cascades*, *phrase table translation* and *phrase table triangulation*. Each method is analysed, implemented and evaluated in, again, several different settings. Firstly, the *system cascades* includes the *1*-best translation and n-best translation. Secondly, the *phrase table translation* includes two scenarios to translate the English side of English↔Vietnamese phrase table into Czech or to translate the English side of Czech↔English phrase table into Vietnamese. Lastly, the *phrase table triangulation*, the most discussed method, contains two scenarios to re-compute the phrase table features, either based on old features from two source phrase tables or by re-computing co-occurrence counts of phrase pairs. Regarding the last method, each of its scenario again contains several options to construct the new phrase table features. The experiments indicate that *system cascades* with n-best translation is the most robust and effective method

Various systems, built by different methods, give rise to a further option to combine them into one final system. We adopt a simple method for system combination, by merging the phrase tables associated with the proposed methods. For systems which do not contain a standard phrase table, a synthetic phrase table is obtained based on the development and test set. The combined phrase table provides the decoder with more possible paths to optimize. Finally, the combined systems performed better than their sub-systems.

In this study, we only scratch the surface of translation issues among under-resourced language pairs, focusing on a specific pair: Czech and *Vietnamese*. We have built the baseline for Czech↔Vietnamese translation for both direc-

tions: Czech→Vietnamese and Vietnamese→Czech. The work consists of various steps, such as collecting fragmented corpora, cleaning the corpora, implementing SMT systems and evaluating the performance. This is one of the main contribution that we made. Besides, regarding English↔Vietnamese language pair, another contribution is our effort to unite non-commercial corpora for two final English↔Vietnamese baseline systems, again for both directions.

In a nutshell, we have described our experiments in Czech↔Vietnamese translation using pivoting methods of phrase-based machine translation. In the future, we would like to further improve the quality of Czech↔Vietnamese translation with the aforementioned methods to alleviate the weak points of each pivoting method. It is worth pointing out that Czech↔Vietnamese translation is difficult, yet intriguing. A good MT system between the two languages has the great potential for growth. To sum up, we have made the following contribution:

- Preparation of training and testing corpora for translation between English, Czech and Vietnamese.
- Experiments with corpus normalization and filtering.
- A wide range of experiments with pivoting methods.
- An experiment with a simple combination of all the examined approaches.

Bibliography

[1] Bertoldi, Nicola, Barbaiani, Madalina, Federico, Marcello and Cattoni, Roldano. *Phrase-based statistical machine translation with pivot languages*. International Workshop on Spoken Language Translation. Honolulu, Hawaii, USA, October 20-21, 2008. 143–149.

[2] Brown Peter F., Pietra Vincent J. Della, Pietra Stephen A. Della , and Mercer, Robert L.. *The Mathematics of Statistical Machine Translation: Parameter Estimation*, Computational Linguistics. MIT Press. Cambridge, MA, USA. 1993.

[3] Chai Chunguang, Du Jinhua, Wei Wei, Zhou Keyan, Liu Peng, He Yanqing, and Zong, Chengqing. *NLPR translation system for IWSLT 2006 evaluation campaign*, International Workshop on Spoken Language Translation, IWSLT 2006, Keihanna Science City, Kyoto, Japan, November 27-28, 2006. 91–94.

[4] Chen, Yu, Eisele, Andreas and Kay, Martin. *Improving Statistical Machine Translation Efficiency by Triangulation*. Proceedings of the International Conference on Language Resources and Evaluation Marrakech, Morocco, 2008.

[5] Galusčáková, Petra and Bojar, Ondřej. *Improving SMT by Using Parallel Data of a Closely Related Language*. Human Language Technologies - The Baltic Perspective - Proceedings of the Fifth International Conference Baltic HLT 2012, Tartu, Estonia, 4-5 October 2012. 58–65.

[6] A. de Gispert and J. B. Mario. *Catalan-english statistical machine translation without parallel corpus: bridging through spanish*. in Proc. of 5th International Conference on Language Resources and Evaluation (LREC), Genoa, Italy, 2006

[7] Habash, Nizar and Dorr, Bonnie J. *Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation*, Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users. London, UK, 2002. 84–93.

[8] Hajič, Jan; Panevová, Jarmila; Hajičová, Eva; Sgall, Petr; Pajas, Petr; Štěpánek, Jan; Havelka, Jiří; Mikulová, Marie; Žabokrtský, Zdeněk; Ševčíková-Razímová, Magda ; and Urešová, Zdeňka. *Prague Dependency Treebank 2.0 (PDT 2.0)*, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, 2006,

[9] Hutchins, W.J. and Somers, H.L. *An Introduction to Machine Translation*, Academic Press. London.

[10] Knight Kevin, Chander Ishwar, Haines Matthew, Hatzivassiloglou Vasileios, Hovy Eduard H., Iida Masayo, Luk Steve K., Okumura

Akitoshi, Whitney Richard and Yamada, Kenji. *Integrating Knowledge Bases and Statistics in MT*. Proceedings of the Conference of the Association for Machine Translation in the Americas. 1994.

[11] Koehn, Philipp, Och, Franz J and Marcu, Daniel. *Statistical Phrase-Based Translation*, In HLT-NAACL: Human Language Technology Conference of the North American. Chapter of the Association for Computational Linguistics, 127-133. 2003.

[12] Koehn, Philipp. *Europarl: A Parallel Corpus for Statistical Machine Translation*, Machine Translation Summit X. Phuket, Thailand, 2005. 79–86.

[13] Koehn, Philipp. *Statistical Machine Translation*. Cambridge University Press. New York, NY, USA. 2010.

[14] Koehn, Philipp; Hoang, Hieu; Birch, Alexandra; Callison-Burch, Chris; Federico, Marcello; Bertoldi, Nicola; Cowan, Brooke; Shen, Wade; Moran, Christine; Zens, Richard; Dyer, Chris; Bojar, Ondřej; Constantin, Alexandra; and Herbst, Evan. *Moses: Open Source Toolkit for Statistical Machine Translation*, Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Association for Computational Linguistics. Stroudsburg, PA, USA, 2007. 177–180.

[15] Kumar, Shankar, Och, Franz and Macherey, Wolfgang. *Improving Word Alignment with Bridge Languages*. Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 209 N. Eighth Street, East Stroudsburg, PA, USA, 2007.

[16] Manning, Christopher D and Schütze, Hinrich. *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA, USA, 1999. 0-262-13360-1.

[17] Ngo Quoc Hung, Winiwarter Werner, and Wloka, Bartholomäus. *EVBCorpus - A Multi-Layer English-Vietnamese Bilingual Corpus for Studying Tasks in Comparative Linguistics*, Proceedings of the 11th Workshop on Asian Language Resources. AFNLP. Singapore, 2013.

[18] Nirenburg, Sergei. *Knowledge-based machine translation*, Machine Translation. 1989. 5–24.

[19] Och Franz Josef, and Ney, Hermann. *A Comparison of Alignment Models for Statistical Machine Translation*, Proceedings of the 18th Conference on Computational Linguistics - Volume 2. COLING '00. Saarbrücken, Germany. 2000.

[20] Razmara, Majid and Sarkar, Anoop. *Ensemble Triangulation for Statistical Machine Translation*. Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013. 252–260.

[21] Rosti Antti-Veikko I., Ayan Necip Fazil Ayan, Xiang Bing, Matsoukas Spyridon, Schwartz Richard M. and Dorr Bonnie J. *Combining Outputs from Multiple Machine Translation Systems*, Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Rochester, New York, USA, 228–235, 2007.

[22] Shannon, Claude. E. *A Mathematical Theory of Communication.* The Bell System Technical Journal. vol 27. 1948. 379–423, 623–656.

[23] Somers, Harold. *Example-based Machine Translation.* Machine Translation. vol 14. 1999. 113–157.

[24] Tiedemann, Jörg. *Parallel Data, Tools and Interfaces in OPUS*, Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012 2214–2218.

[25] Utiyama, Masao and Isahara, Hitoshi. *A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation.* Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Rochester, New York, USA, 2007. 484–491.

[26] Wu, Hua and Wang, Haifeng. *Pivot Language Approach for Phrase-Based Statistical Machine Translation*, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic, 2007.

[27] Zhu Xiaoning, He Zhongjun, Wu Hua, Zhu Conghui, Wang Haifeng and Zhao, Tiejun. *Improving Pivot-Based Statistical Machine Translation by Pivoting the Co-occurrence Count of Phrase Pairs*, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, 2014. 1665–1675.

[28] Barrault Loic. *Open Source Machine Translation System Combination.* Prague Bulletin of Mathematical Linguistics - Special Issue on Open Source Machine Translation Tools, Charles University. 2010.

[29] Bojar Ondřej, Žabokrtský Zdeněk, Dušek Ondřej, Galuščáková Petra, Majliš Martin, Mareček David, Maršík Jiří, Novák Michal, Popel Martin, and Tamchyna, Aleš. *The Joy of Parallelism with CzEng 1.0*, Proceedings of LREC2012. European Language Resources Association. Istanbul, Turkey, 2012.

[30] Bojar Ondřej and Tamchyna Aleš. *The Design of Eman, an Experiment Manager.* Prague Bull. Math. Linguistics. 39–58. 2013.

[31] Chen Stanley F. and Joshua Goodman. *An Empirical Study of Smoothing Techniques for Language Modeling.* Proceedings of the 34th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics. Santa Cruz, California. 1996.

[32] Cohn, Trevor and Lapata, Mirella. *Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora.* Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, 2007.

[33] Gollins Tim and Sanderson Mark. *Improving cross language retrieval with triangulated translation.* In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01). New York, NY, USA, 90-95.. 2001.

[34] Hajič Jan, Kuboň Vladislav, and Homola Petr. *Česílko is a tool enabling the fast and efficient translation from one source language into many target languages, which are mutually related.*. Attribution-NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0). 2012.

[35] Heafield Kenneth and Lavie Alon. *Combining Machine Translation Output with Open Source, The Carnegie Mellon Multi-Engine Machine Translation Scheme. Machine Translation Tools* Prague Bulletin of Mathematical Linguistics - Special Issue on Open Source Machine Translation Tools, Charles University. 2010.

[36] Matusov Evgeny, Leusch Gregor, Banchs Rafael E., Bertoldi Nicola, Dechelotte Daniel, Federico Marcello, Kolss Muntsin, Lee Young-Suk, Marino Jose B., Paulik Matthias, Roukos Salim, Schwenk Holger and Ney Hermann. *System Combination for Machine Translation of Spoken and Written Language.* IEEE Transactions on Audio, Speech and Language Processing. vol 16. 1222-1237. 2008.

[37] Papineni Kishore, Roukos Salim, Ward Todd, and Zhu Wei-Jing. *BLEU: A Method for Automatic Evaluation of Machine Translation.* Association for Computational Linguistics. Philadelphia, Pennsylvania. 311–318. 2002.

List of Tables

List of Figures

Common Abbreviations and Terms

Symbols	Meaning
NLP	Natural Language Processing
SMT	Statistical Machine Translation
MT	Machine Translation
unreviewed	Data collected from OPUS and TED
reviewed	Data published by other works
ttable	Phrase table
F	Source language
E	Target language
f	Sentence in the source language
e	Sentence in the target language
s	Phrase in the source language
t	Phrase in the target language
p	Phrase in the pivot language