Vietnamese to Chinese Machine Translation via Chinese Character as Pivot*

Hai Zhao[1,2†] **Tianjiao Yin**[3] **Jingyi Zhang**[1,2]
(1) MOE-Microsoft Key Laboratory of Intelligent Computing and Intelligent System
(2) Department of Computer Science and Engineering, Shanghai Jiao Tong University
#800 Dongchuan Road, Shanghai, China, 200240
(3) Facebook, Inc. 1601 Willow Rd. Menlo Park, CA 94025 USA
zhaohai@cs.sjtu.edu.cn, ytj000@gmail.com
zhangjingyizz@gmail.com

Abstract

Using Chinese characters as an intermediate equivalent unit, we decompose machine translation into two stages, semantic translation and grammar translation. This strategy is tentatively applied to machine translation between Vietnamese and Chinese. During the semantic translation, Vietnamese syllables are one-by-one converted into the corresponding Chinese characters. During the grammar translation, the sequences of Chinese characters in Vietnamese grammar order are modified and rearranged to form grammatical Chinese sentence. Compared to the existing single alignment model, the division of two-stage processing is more targeted for research and evaluation of machine translation. The proposed method is evaluated using the standard BLEU score and a new manual evaluation metric, understanding rate. Only based on a small number of dictionaries, the proposed method gives competitive and even better results compared to existing systems.

1 Introduction

The statistical machine translation (SMT) has been well developed from a basis of data-drive idea since the work of (Brown et al., 1993). However, a large amount of parallel corpora are always necessary to build a standard SMT system for a specific language pair, regardless of the possible useful linkages between these two languages. There is existing work that considered using helpful linguistic heuristics to enhance the current SMT (Chu et al., 2012), though their approaches still follow the standard processing pipeline of SMT. For those resource-poor languages, a pivot language will be used as an expedience (Utiyama and Isahara, 2007; Wu and Wang, 2009).

Figure 1: The phrase *Chinese character culture sphere* written in Chinese characters from different regions.

In this work, we focus on machine translation (MT) for language pairs with few parallel corpora but rich linguistic connections. A case study on Vietnamese and Chinese will be done. To exploit the shared linguistic characteristics between the language pair, the common written form, Chinese character, is adopted as a translation bridge. Being the oldest continuously used writing system in the world, Chinese characters are logograms that are still used to write Chinese (汉字/漢字 in Chinese, hànzì in Chinese pinyin) and Japanese (kanji). Such characters were used but are currently less frequently used in Korean (hanja), and were also used in Vietnamese (chữ Hán). All the countries that were historically under Chinese language and culture are unofficially referred to *Chinese character cultural sphere* or Sinosphere. These two terms are often used interchangeably but have different denotations (Matisoff, 1990). A Chinese character writing example of different regions is in Figure 1.

This work was partially supported by the National Natural Science Foundation of China (Grant No.60903119, Grant No.61170114, and Grant No.61272248), and the National Basic Research Program of China (Grant No.2009CB320901 and Grant No.2013CB329401).

†corresponding author

27th Pacific Asia Conference on Language, Information, and Computation pages 250－259

Figure 2: Different scripts for Chinese characters.

There are tens of thousands of Chinese characters, though most of them are minor graphic variants only existing in historical texts as Figure 2. Mastering modern Chinese usually requires knowing 2,000-4,000 characters. Though most words in modern Chinese consist of two or more characters, each Chinese character may correspond to a spoken syllable with a distinct meaning. Being meaning-oriented representation units, Chinese characters are naturally suitable to act as a bridge of semantic representation for translation task. This process will be especially promising as we are working on a language like Vietnamese.

Vietnamese (tiếng Việt) is spoken by about eighty million people. Much of Vietnamese vocabulary has been borrowed from Chinese, and it formerly used a modified Chinese writing system, Chữ Nôm, and given vernacular pronunciation. The Vietnamese alphabet (Quốc Ngữ) in use today is a Latin alphabet with additional diacritics for tones, and certain letters.

In this paper, a novel two-stage approach is proposed for Vietnamese to Chinese MT by adopting Chinese characters as the pivot. Vietnamese syllables will first be converted into Chinese characters according to the meaning equivalence. Then Chinese character sequences in Vietnamese grammar order will be modified and reordered into grammatical Chinese. The proposed approach only requires a small number of linguistic resources, such as bilingual dictionaries and monolingual language model, to work effeciently.

2 Related Work

Only recently have researchers begun to be involved in the domain of Vietnamese language processing. Most work on Vietnamese language processing has to still focus on very basic issues such as corpus building, primary processing tasks, etc.

A few studies have been done on Vietnamese related MT, though nearly all MT studies on Vietnamese focus on English as source or target language. As Vietnamese is an under-resourced language, most Vietnamese MT systems adopted rule based methods (Le et al., 2006; Le and Phan, 2009; Le and Phan, 2010).

(Pham et al., 2009) used word-by-word translation incorporated with predefined templates to perform English-Vietnamese translation on weather bulletin texts. The similar strategy was also used in (Hoang et al., 2012) for Vietnamese to Katu language translation on the same domain.

Until very recently, the statistical approach was applied to Vietnamese related MT task. (Nguyen and Shimazu, 2006; Nguyen et al., 2008) used self-defined morphological transformation and syntactic transformation to beforehand solve reordering problem for Vietnamese-English translation. (Thi and Dinh, 2008) introduced a word re-ordering approach that makes use of the syntactic rules extracted from parse tree for English-Vietnamese MT. (Bui et al., 2010) proposed language dependent features to enhance Vietnamese-English SMT. (Nguyen et al., 2012) integrated more knowledge about the topic of the text, part-of-speech and morphology to resolve semantic ambiguity of words during translation. Based on empirical observation, (Nguyen and Dinh, 2012) proposed a group of heuristic patterns to discover the alignment errors. (Bui et al., 2012) proposed a group of rules to split long Vietnamese sentences based on linguistic information to enhance Vietnamese to English MT.

Few studies have been done for MT task between Vietnamese and Chinese as to our best knowledge. For such a low resource language pair, rule based MT systems are too hard to build, and statistical MT systems require too large parallel corpus that is also difficultly acquired. Though Chinese characters have been considered a useful intermediate form for MT, few studies made a full use of them. Instead, most existing approaches focus on the role of Chinese word during translation (Chang et al., 2008; Xu et al., 2008; Dyer et al., 2008; Ma and Way, 2009; Paul et al., 2010; Nguyen et al., 2010). (Chu et al., 2012) exploited shared Chinese characters between Chinese and Japanese to improve the concerned translation performance. The most recent work (Xi et al., 2012)

	Vietnamese	Chinese
Character script	Chữ Nôm	Chinese characters (official now)
Romanized script	Quốc Ngữ (official now)	pinyin

Table 1: Chinese vs. Vietnamese: writing systems

proposed using Chinese character as aligning unit. However, both of the above works are different from ours, in which Chinese characters are used as a pivot for translation task for the first time.

3 Chinese Elements inside Vietnamese

3.1 The Same and The Difference

Most linguists agree that Chinese and Vietnamese belong to two quite different language families. All varieties of modern Chinese are usually categorized as part of the Sino-Tibetan language family. However, opinions are divided on the language family that Vietnamese should belong to, though the most acceptable view is that it is part of the Mon – Khmer branch of the Austroasiatic language family according to the observation that Vietnamese and Khmer share a lot of cognates and basic grammar (Benedict, 1944; Nguyen, 2008).

A writing system comparison between Chinese and Vietnamese is shown in Table 1. An obvious distinction between Vietnamese and Chinese writing is on the role of the Romanized scripts. The Quốc Ngữ is official writing system of Vietnamese today, while pinyin is only an assistant language learning tool for Chinese today.

Both Chinese and Vietnamese, like many languages in East Asia, are analytic (isolating) languages[1]. Neither of them uses morphological marking of case, gender, number or tense. Both languages use word order and function words to convey grammar relationships. As word order or function words are changed, the meaning will be changed accordingly. Moreover, their syntax both conforms to subject-verb-object word order and possesses noun classifier systems.

As each Chinese character in Chinese represents a meaningful unit, a major feature of Vietnamese word-building is that each syllable may be separately used as a meaningful unit. Like Chinese, most Vietnamese words are bi-syllable. Chinese is written without blanks between words and Vietnamese is written with blanks between two syllables instead of words. Thus word segmentation becomes a primary processing for both languages.

Vietnamese is a prop-drop (pronoun-dropping) language, which means that certain classes of pronouns in Vietnamese may be omitted when they are in some sense pragmatically inferable. Chinese also exhibits frequent pro-drop features.

Both Chinese and Vietnamese allow verb serialization. Contrary to subordination in English where one clause is embedded into another, the serial verb construction is a syntactic phenomenon that two verbs are put together in a sequence in which no verb is subordinated to the other.

Different from Chinese on word order, Vietnamese is head-initial, i.e., displaying modified-modifier ordering, but number and noun classifier being before the modified noun. Thus, for example, the *Vietnamese language* in Vietnamese grammar order should not be *Vietnamese language* (Việt Nam tiếng) but *language Vietnamese* (tiếng Việt Nam).

3.2 Sino-Vietnamese

As a result of close ties with China for more than 2,000 years, quite a few of the Vietnamese lexical elements have Chinese roots. The elements in the Vietnamese derived from Chinese is called Sino-Vietnamese (Hán Việt; 漢越), which accounts for about 30-60% of the Vietnamese vocabulary (LUO, 2011). This vocabulary was originally written with Chinese characters, but like all written Vietnamese, is now written with the Quốc Ngữ, the Latin-based Vietnamese alphabet. Sino-Vietnamese words have a status similar to that of Latin-based words in English: they are used more in formal occasion than in everyday life. Most monosyllabic Sina-Vietnamese are used for word-building morphemes, though a few of them may be directly adopted as words as well.

A lot of Sino-Vietnamese words, such as those in Table 2, have the exactly same meaning as modern Chinese. Some Sino-Vietnamese words (Table 3) are written in the same Chinese characters but represent different meaning from their Chinese counterparts. Some Sino-Vietnamese words (Table 4)are entirely invented by the Vietnamese, which can be directly written in Chines characters

[1] A few linguists strictly define that an isolating language as a type of language with a low morpheme-per-word ratio is a closely related concept of the analytic language, but still different from the latter. In this paper, we do not strictly distinguish these two concepts.

Vietnamese words	Chinese characters	Chinese pinyin	meaning
lịch sử	歷史	lì shǐ	history
định nghĩa	定義	dìng yì	definition
phong phú	豐富	fēng fù	fruitful
thời sự	時事	shí shì	current events

Table 2: A list of Sina-Vietnamese words

Vietnamese in Latin	Vietnamese in chữ Hán	Chinese word	Chinese meaning
linh mục	靈**牧**	**牧**师	clergyman
lí thuyết	**理**說	**理**论	theory
bệnh cảm	病**感**	**感**冒	flu
khẩu trang	**口**裝	**口**罩	mask

Table 4: A list of Sina-Vietnamese words with similar writing and same meaning

but not used in Chinese or no longer used in modern Chinese. Interestingly, though not exactly the same in writing, there is always one character that is shared by both languages for words in Table 4.

Writing Sino-Vietnamese words with Quốc Ngữ may cause some confusions due to the large amount of homophones in Chinese and Sino-Vietnamese. For example, both '明'(*bright*) and '冥'(*dark*) are read or written as *minh* with Quốc Ngữ, thus only using Chinese character can one distinguish the two contradictory meanings of the word "*minh*".

4 Chinese Characters but not Chu Nom

4.1 Why it is Chinese Characters

A Chinese character is regarded as a unity of form (writing), sound (phonetics) and meaning (semantics). An illustrative example of Chinese character is given in Figure 3, which demonstrates a character with written form '福', sound 'fú'(pinyin) and meaning 'good luck'. However, it is not balanced for the three primary factors of Chinese character. The core functionality of Chinese character is being a meaning unit. In fact, Chinese character is neither a good carrier of pronunciation nor stable at written forms: different Chinese variants

Figure 3: Chinese character is a trinity.

and other languages such as Japanese and Korean usually borrow Chinese Characters semantically rather than phonetically, and the Chinese character scripts also continuously evolved in the past 3,000 years as shown in Figure 2.

Meanwhile, the meaning that Chinese character was initially invented to express seldom changes over time. Chinese characters used in the similar way for different languages also share the same or similar meaning, which is especially obvious for Chinese characters borrowed by Japanese (kanji). For example, although the character '山' in Figure 2, has more than 8 different writing scripts, and may be pronounced as *shan* in Mandarin Chinese, either *yama* or *san* in modern Japanese, it is always referred to the meaning 'mountain' in both languages.

In addition, Chinese character writing system is usually more accurate than alphabetic writing systems on expressive ability. In fact, Chinese, Vietnamese and Korean are the victims of a large amount of homophones in their vocabulary. However, modern Korean and Vietnamese that have adopted alphabetic writing systems are more easily plagued by this problem than Chinese as the latter may respite the difficulty by assigning different Chinese characters to respectively record different meanings of the same pronunciation.

4.2 Why it is not Chu Nom

Chữ Nôm is a system of modified and invented characters modeled loosely on Chinese characters, which, unlike the system of Chinese character (chữ Hán), allows for the expression of purely Vietnamese words to any extent.

The character set for Chữ Nôm is extensive, up to 20,000, arbitrary in composition and inconsistent in pronunciation [2]. The Chữ Nôm characters can be divided into two groups: those borrowed from Chinese and those invented especially for Vietnamese. The characters borrowed from Chinese are used to represent either Chinese loan words or native Vietnamese words. For the former case, the character may have more than one pronunciation. For the latter case, the character may be only used phonetically, regardless of the original standard meaning of it in Chinese. For example, the Chinese character '沒' (méi, means *none*

[2]Online resources on Chữ Nôm can be found at the following links, http://nomfoundation.org/ and http://www.chunom.org/.

Vietnamese Meaning	Vietnamese words	Chinese characters	Chinese pinyin	Chinese meaning
method	phương tiện	方便	fāng biàn	convenience
office building	văn phòng	文房	wén fáng	study room
rich	phong lưu	風流	fēng liú	romantic
full-grown	phương phi	芳菲	fāng fēi	flowers and plants

Table 3: A list of Sina-Vietnamese words with same writing but different meaning

Figure 4: The 20 most frequent Nom characters in which red bold ones are not used in Chinese.

in Chinese) is used to represent the Vietnamese word một (means *one* in Vietnamese).

Figure 4 shows top 20 most frequent Nom characters. As we are finding a pivot written form for Vietnamese to Chinese MT, Chữ Nôm looks like a good candidate. However, three reasons make Chữ Nôm unable to fulfill the task. First, too many characters in Chữ Nôm belong to the Vietnamese-only type, which can be neither recognized by modern Chinese nor naturally mapped to commonly used Chinese characters. Second, Chinese characters that are still popularly used in modern Chinese may be only phonetically borrowed by Chữ Nôm. Third, Chữ Nôm has never been standardized, which may lead to multiple writing choices for the same Vietnamese syllable.

5 The Proposed Approach

Chinese character is a powerful representation as an ideographic writing system, for text written with Chinese characters, even if grammatically incorrect, it is understandable and even readable for people who know Chinese characters but speak different languages of Sinosphere. Vietnamese as an analytical language, its individual syllable has similar ideographic property. Vietnamese is perhaps more suitable to adopt an ideographic writing system like Chinese characters. Therefore, we first attempt to find a proper Chinese character to record each syllable of Vietnamese text in accordance with its contextual meaning. In this way, we will have a Vietnamese text written with Chinese characters. Then, with additional processing, the Chinese character sequences in the Vietnamese grammar are converted into grammatical Chinese sentences. The proposed approach is divided into two stages as the following.

Stage 1: Syllable-to-Character Conversion

To find a matching Chinese character for a Vietnamese syllable, bilingual dictionaries are necessary to provide possible character candidates. However, we still need more heuristics to determine which character should be chosen according to the context in which the Vietnamese syllable is located. As multisyllable Vietnamese word usually has a unique Chinese equivalent, we propose to first perform word segmentation over the Vietnamese sentence and then convert the segmented words into Chinese. Relying on a pre-specified dictionary, the maximum matching algorithm as shown in Algorithm 1 that is traditionally used for Chinese word segmentation will be applied to Vietnamese word segmentation [3].

Two bilingual dictionaries are used for Vietnamese word segmentation and Chinese character conversion.

The first dictionary is a Sino-Vietnamese vocabulary. Sino-Vietnamese vocabulary will play a core role during the conversion. For all known Sino-Vietnamese words, we can simply determine their Chinese character equivalents without ambiguities. Vietnamese and Chinese actually share the same words on the Sino-Vietnamese vocabulary, and the only difference is behind the written forms, either Vietnamese Quốc Ngữ or Chinese characters. In this work, we use all bisyllable Sino-Vietnamese vocabulary from (LUO, 2011), which includes 10,900 Vietnamese-Chinese word pairs. This dictionary will be referred to D_s hereafter.

For the part beyond the Sino-Vietnamese words in Vietnamese text, there is no such a dictio-

[3] This algorithm is more precisely referred to as the forward maximal matching algorithm.

Algorithm 1 Maximal matching algorithm

INPUT1 Dictionary $D = \{w_i|\}$, and $maxlen$ is the maximal word length inside D.
INPUT2 Syllable sequence $c_0c_1..c_{n-1}$
Let i=0
while i<n **do**
  let m=$min(maxlen, n-i)$
  while m>1 **do**
    if $c_ic_{i+1}..c_{i+m-1}$ is a word in D **then**
      $i = i+m$, and set a segmentation mark before c_{i+m}.
      break
    else
      $m = m-1$
    end if
  end while
  if m==1 **then**
    Set a segmentation mark before c_{i+1}
  end if
  $i = i+m$
end while

nary that meet our requirements, we have to seek help from online resources for a quick but inaccurate solution. We let a crawler collect Vietnamese texts[4] from the Internet and then feed the Google translator[5] with the text, so that we obtain a loose parallel corpus between Vietnamese and Chinese. After each Vietnamese monosyllable and each Chinese character segmented as a word, we may obtain an aligned phrase table by using bidirectional GIZA++ alignment (Och and Ney, 2003). We perform two steps of pruning on the phrase table. First, only those aligned phrases that have the same numbers for both Vietnamese syllables and Chinese characters will be kept. Second, If a Vietnamese phrase is mapped to multiple Chinese phrases, then only the one with the highest aligning probability will be conserved. Regarding both Vietnamese syllables and Chinese characters in the phrase table as words, we finally build the second bilingual dictionary D_g with 6.8 million word pairs.

Given a Vietnamese sentence, we apply the maximal matching algorithm twice to accomplish the word segmentation. When a word is segmented according to the Vietnamese part of bilingual dictionary, it will be automatically converted into Chinese characters according to the corresponding Chinese part of the dictionary. The Sino-Vietnamese dictionary D_s is first adopted. If there are still undetermined parts in the sentence after the first round of segmentation and conversion, then the dictionary D_g will be used.

Stage 2: Restating and Reordering

As Vietnamese uses a different modifier-modified order, which is the most difference from Chinese, its text, even though written in Chinese characters, cannot be fully understood by one who only knows Chinese. Therefore, we introduce this stage of processing to polish the Chinese character sequences in Vietnamese word order. Note that it is entirely a monolingual processing task.

The first difficulty that we should consider is that not all Vietnamese words are the exactly same as their Chinese counterparts. To alleviate this difficulty, we tentatively replace a Vietnamese word written in Chinese character by another related word. A Chinese synonym dictionary[6] with 77,000 items is therefore used to enumerate all these possible related words.

To determine each best related word and reorder the character sequence into Chinese word order, we use language model trained on Chinese text following equation (1).

$$\{w_0^*w_1^*..w_n^*\} = \underset{\forall\omega(w_i)\text{and}\{w_0'w_1'..w_n'\}}{\text{argmax}} \prod_{i=1}^{n}(P(\omega(w_i)'|\omega(w_{i-m+1})'\omega(w_{i-m})'..\omega(w_{i-1})')), \quad (1)$$

where $\omega(w_i)$ represent a related word of w_i and $\{w_0'w_1'..w_n'\}$ is a permutation of $\{w_0w_1..w_n\}$. To prevent from generating too many reordering possibilities, the distance between the original position of each word and its new location is limited to less than 4 words. The above output sequence can be decoded through a Viterbi style algorithm.

6 Experiments

We manually collect 2,046 sentence pairs as test set to evaluate the proposed approach. We report the MT performance using the original BLEU metric (Papineni et al., 2002). A trigram Chinese language model is trained on the text with segmentation that is extracted from the People' Daily[7]

[4]The Vietnamese corpus has 77 M bytes, 0.86 million Vietnamese sentences and 13 million Vietnamese monosyllables.

[5]http:// translate.google.com/?hl=en#vi/zh-CN/

[6]The Word Forest of Synonyms: http://www.ir-lab.org/

[7]It is the most popular newspaper in China.

Systems	Ours/Stage 1	Ours/Stage 1+2	Google
BLEU	14.5	18.6	20.3

Table 5: BLEU scores for the proposed system.

Systems	Ours/Stage 1	Ours/Stage 1+2	Google
/wo ref.	65.4	67.1	62.3
/w ref	62.3	63.6	60.7

Table 6: Understanding rates without any Vietnamese knowledge.

from 1993 to 1997. To segment the Chinese text, the maximal matching segmentation algorithm with Chinese side words of the above two bilingual dictionaries are used.

The results with Google translation comparison are given in Table 5. With limited support linguistic resources, the proposed approach gives a very competitive result as the Google translator does.

Although it goes without saying, actually all the state-of-the-art MT systems are far from the requirement of being serious publishing or any official usages. Most of the current MT outputs are used, tacitly in fact, for a rough understanding of texts written in other languages that readers do not know at all. We will evaluate the results of the proposed approach from this sense. A group of human evaluation experiments are done based on the following scoring rules. For each translated sentence, a human evaluator will determine if the rough meaning of the sentence is understandable, and the sentence will be given score (1) 1.0 if the sentence can be fully understood; (2) 0.5 if the sentence seems understandable but not so certain; (3) 0.0 if the meaning of sentence cannot be captured at all. We define understanding rate for the given test set, $U_r = \frac{\sum_{i=1}^{N} \alpha_i}{N}$, where N is number of sentences in the test set and α_i is the evaluation score given to the i-th sentence.

The first group of results with U_r are given in Table 6. There are two types of results in the table, the first human evaluator gives score without allowing to read the translation reference, while the second is allowed to read the reference to verify and modify his score after he already gives a score.

The second group of results are given on the condition that human evaluators are taught about grammar difference between Vietnamese and Chinese that Vietnamese is a head-initial language. The results are given in Table 7. The second group of human evaluation results are slightly better than the first group. With limited linguistic knowledge, human evaluator can finish the necessary grammar conversion by himself for better understanding on the translated text.

Systems	Ours/Stage 1	Ours/Stage 1+2	Google
/wo ref.	68.5	72.9	69.3
/w ref	66.1	71.6	67.7

Table 7: Understanding rates with limited Vietnamese knowledge.

Systems	Sentences
Source	Du khách Tây Ban Nha thưởng thứ trà tại Trâm Anh quán.
Target	西班牙游客在簪缨馆品茶。
Stage 1	游客 西班牙 赏识 茶 在 簪缨 店 .
Stage 2	西班牙 游客 赏识 茶 在 簪缨 店 .
Google	西班牙游客享受茶在英国的前哨基地。

Table 8: Vietnamese-to-Chinese translation.

Overall, the proposed approach gives satisfactory results on Vietnamese to Chinese translation with quite limited linguistic input. Our system gives competitive results as the existing system in terms of BLEU, and outperforms the latter according to the newly introduced evaluation metric. These comparisons show that though the translation output of our system is not up to its best on word transformation and ordering (that is mostly concerned by the BLEU score, and mostly determined by grammar translation stage.), but it possesses better understandability, which is mostly determined by semantic translation stage.

7 Error Analysis

Rough manual inspection shows that both work stages introduce factors that lead to poor translation. For Stage 1, most errors occur because the target Chinese characters are incorrectly given by the dictionary D_g at the very beginning. For those Vietnamese syllables that have multiple conversion options in writing as Chinese characters, it is surprising that we find few examples on such types of character selection errors. This observation suggests that a direct refine work on D_g may be hopeful to give significant performance improvement. Furthermore, it is also useful to enrich the current Sino-Vietnamese dictionary D_s as up to

now it only includes bisyllable words.

For Stage 2, it looks like that word order adjusting does not work well, though one can see a BLEU score increasing after the processing of Stage 2. In fact, U_r scores in Table 6 and 7 demonstrate that word order only has a marginal effect over the understanding (or guessing) the translated sentence. Later, according to word order difference between Chinese and Vietnamese, we may especially adjust the order of Vietnamese words with specific part-of-speech so that the translation results can be further improved.

Table 8 shows an actual translation output by our system. For a detailed English explanation, please refer to the appendix.

8 Semantic Translation vs. Grammar Translation

Using Chinese character as an intermediate form, a two-stage MT approach has been proposed. We loosely refer the first stage as semantic translation, and the second as grammar translation. The semantic translation is called because syllable to character conversion is based on semantic equivalence rather than anything else, such as phonetics or written forms. The grammar translation includes two monolingual subtasks, word restating and reordering, to let the expression more fluent.

Standard SMT integrates semantic and grammar translation into one word/phrase alignment model, which partially make researchers working on MT lose focus. We say that the proposed two-stage MT processing strategy allows translation research more focused. For example, our experiments demonstrate a higher understanding rate but lower BLEU score for the same MT outputs. If we loosely regard that understanding rate measures the semantic aspect of MT performance and BLEU measures the grammar factor, then we now have a chance to see that a poor grammar translation is an obstacle on the way to let MT outputs become really useful on a formal occasion.

9 Other Language Pairs

Now we consider if the proposed approach can be extended to other language pairs. Though it is a two-stage translation, language specific properties are actually concerned only at Stage 1, and Stage 2 may work on any other target language in principle. Thus we may only focus on Stage 1.

A simple maximal matching algorithm with bilingual dictionaries can be adopted to perform semantic translation as Stage 1 because two essential language facts, (1) Vietnamese and Chinese share a very large vocabulary, and (2) They both belong to analytic (isolating) languages, which means that there is nearly a full correspondence between a single word/character/syllable and a single aspect of meaning. Motivated by this observation, it is possible to extend this work to other languages in Sinosphere, such as Korean and Japanese. The following gives reasons why both languages meet the above conditions.

Let us first consider the vocabulary. The exact proportion of Sino-Korean vocabulary is still a matter of debate. (Sohn, 2001) stated that it is between 50 – 60%. For Sino-Japanese, it usually has an estimation of 40-50%.

Both Korean and Japanese belong to agglutinative languages, which seems that the above second condition is not met. However, using Chinese character based writing traditionally or currently, a stable correspondence between meaning and writing for both languages can be generally found [8]. From the writing perspective, both Korean and Japanese are *quite isolating*.

10 Conclusions and Future Work

This paper presents a two-stage conversion method for the MT task between resource-scarce language pairs that both belong to the isolating language type, such as Vietnamese and Chinese, and other languages in Sinosphere that demonstrate observable isolating language characteristics.

Chinese character as the heart of the evolution of languages in Sinosphere is selected as an intermediate equivalent form during translation. In detail, Chinese character sequences subject to source language grammar play a pivot role. Compared to existing translation system, the proposed method, with a small number of linguistic resources, gives competitive or even better results in terms of standard BLEU score or a newly introduced human evaluation metric, understanding rate.

It is worth noting that we have only made a very preliminary attempt with respect to the proposed approach. For example, during semantic translation, in addition to bilingual dictionary, we do

[8] For example, though the meaning *moutain* is pronounced as *yama* or *san* in Japanese, it can be always written as the same Chinese character 山.

not use any other context information to effectively determine the target Chinese characters. During grammar translation, we do not use any language-specific features to improve the target language generation. Exploring all the potentials, it is expected to receive even better results.

References

Paul K. Benedict. 1944. Thai, Kadai and Indonesian: A new alignment in southeastern Asia. *American Anthropologist*, 44(2):576–601.

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

Thanh Hung Bui, Le Minh Nguyen, and Akira Shimazu. 2010. Using rich linguistic and contextual information for tree-based statistical machine translation. In *2011 International Conference on Asian Language Processing*, pages 189–192, Harbin, China, December.

Thanh Hung Bui, Le Minh Nguyen, and Akira Shimazu. 2012. Sentence splitting for Vietnamese-English machine translation. In *2012 Fourth International Conference on Knowledge and Systems Engineering*, pages 156–160, Danang, Vietnam, August.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232, Columbus, Ohio, USA, June.

Chenhui Chu, Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. 2012. Exploiting shared Chinese characters in Chinese word segmentation optimization for Chinese-Japanese machine translation. In *Proceedings of the 16th EAMT Conference*, pages 35–42, Trento, Italy, May.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing word lattice translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1012–1020, Columbus, OH, USA.

Thi My Le Hoang, Thi Bong Phan, and Huy Khanh Phan. 2012. Building a machine translation system in a restrict context from Ka-Tu Language into Vietnamese. In *2012 Fourth International Conference on Knowledge and Systems Engineering*, pages 167–172, Danang, Vietnam, August.

Manh Hai Le and Thi Tuoi Phan. 2009. Three algorithms for word-to-phrase machine translation. In *2009 International Conference on Asian Language Processing*, pages 328–331, Singapore, December.

Manh Hai Le and Thi Tuoi Phan. 2010. Lexical gap in English-Vietnamese machine translation: What to do? In *2010 International Conference on Asian Language Processing*, pages 265–269, Harbin, China, December.

Manh Hai Le, Chanh Thanh Nguyen, Chi Hieu Nguyen, and Thi Tuoi Phan. 2006. Dictionaries for English-Vietnamese machine translation. In *The 21st International Conference on the Computer Processing of Oriental Languages*, pages 363–369, Singapore, December.

Wenqing LUO. 2011. *A Study on Bi-syllable Sino-Vietnamese Words in Vietnamese: A Comparison with Chinese (in Vietnamese)*. World Publishing Corporation, Guangzhou, China.

Yanjun Ma and Andy Way. 2009. Bilingually motivated domain-adapted word segmentation for statistical machine translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 549–557, Athens, Greece, April. Association for Computational Linguistics.

James A. Matisoff. 1990. On megalocomparison. *Language*, 66(1):106–120.

Giang Thanh Nguyen and Dien Dinh. 2012. Improving English-Vietnamese word alignment using translation model. In *2012 IEEE International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RVIF)*, Ho Chi Minh City, Vietnam, February.

Thai Phuong Nguyen and Akira Shimazu. 2006. Improving phrase-based statistical machine translation with morphosyntactic transformation. *Machine Translation*, 20:147–166.

Vinh Van Nguyen, Thai Phuong Nguyen, Akira Shimazu, and Minh Le Nguyen. 2008. A reordering model for phrase-based machine translation. In *GoTAL - 6th International Conference on Natural Language Processing*, pages 476–487, Gothenburg, Sweden, August.

ThuyLinh Nguyen, Stephan Vogel, and Noah A. Smith. 2010. Nonparametric word segmentation for machine translation. In *Proceedings of COLING-2010*, pages 815–823, Beijing, China, August.

Quy Nguyen, An Nguyen, and Dien Dinh. 2012. An approach to word sense disambiguation in English-Vietnamese-English statistical machine translation. In *2012 IEEE International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RVIF)*, Ho Chi Minh City, Vietnam, February.

Thien Giap Nguyen. 2008. *A Brief History on Vietnamese Study (in Vietnamese)*. Vietnam Education Press, Hanoi, Vietnam.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael Paul, Andrew Finch, and Eiichiro Sumita. 2010. Integration of multiple bilingually-learned segmentation schemes into statistical machine translation. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 400–408, Uppsala, Sweden, July. Association for Computational Linguistics.

Son Bao Pham, Giang Binh Tran, Dang Duc Pham, Kien Chi Phung, and Kien Trung Nguyen. 2009. An information extraction approach to English-Vietnamese weather bulletins machine translation. In *2009 First Asian Conference on Intelligent Information and Database Systems*, pages 161–166, Dong hoi, Quang binh, Vietnam, April.

Ho-Min Sohn. 2001. *The Korean Language*. Cambridge University Press, Cambridge, UK.

Hong-Nhung Nguyen Thi and Dien Dinh. 2008. A syntactic-based word re-ordering for English-Vietnamese statistical machine translation system. In *The Tenth Pacific Rim International Conference on Artificial Intelligence (PRICAI-08)*, pages 809–818, Hanoi, Vietnam, December.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Proceedings of NAACL HLT 2007*, pages 484–491, Rochester, NY, USA, April.

Hua Wu and Haifeng Wang. 2009. Revisiting pivot language approach for machine translation. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 154–162, Singapore, August.

Ning Xi, Guangchao Tang, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2012. Enhancing statistical machine translation with character alignment. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 285–290, Jeju, Republic of Korea, July.

Jia Xu, Jianfeng Gao, Kristina Toutanova, and Hermann Ney. 2008. Bayesian semi-supervised chinese word segmentation for statistical machine translation. In *Proceedings of COLING-2008*, pages 1017–1024, Manchester, UK, August.

APPDENDIX

We now give a detailed English explanation to the translation example output by our system. A word-by-word translation is shown in Table 9. The meaning of source Vietnamese is that '*Spanish tourists enjoy tea at Tram Anh teahouse.*' We analyze two problematic words in our translation. The first word *thưởng thứ* is Sino-Vietnamese, its exact Chinese form is right '赏识'. Unfortunately, these two characters in Chinese as a word means '*appreciate*' instead of '*enjoy*' as its Vietnamese counterpart. Furthermore, Stage 2 also fails to rectify this meaning-drift word due to the limitation of our synonymous dictionary. However, if we only concern about the first character '赏' of the word '赏识', it will be acceptable for Chinese readers, as two basic senses of '赏' are '*enjoy*' and '*award*', though '赏茶' is not a usual expression in Chinese for saying '*enjoy tea*'. The second inexact conversion about '*quán*' comes from building the second bilingual dictionary D_g. As in the aligned phrase table, '*quán* is translated onto '店'(diàn in Chinese pinyin, means '*shop*' or '*building/facilities for business purpose*') with a higher probability than the expected exact one, ' 馆'(guǎn, means *building (group) for specific purpose*). Though the character '店' is not the expected translation, it is rough in line with the original meaning of source phrase and acceptable for most Chinese readers. Generally, most Vietnamese names are supposed to have standard forms written in Chinese characters. Using a pivot language, Vietnamese names are hard to exactly be translated into Chinese. For our example, in addition to the unique mismatched character, the named entity *Trâm Anh quán* has been exactly translated, which can be hardly done by Google translator, as we surmise, using English as a pivot language. In fact, the meaning of the Google translator output is '*Spanish tourists enjoy tea at the British outpost*'.

English	Vietnamese	Chinese	Stage 1
tourist	du khách	游客	√
Spain	Tây Ban Nha	西班牙	√
enjoy	thưởng thứ	赏识	?
tea	trà	茶	√
at	tại	在	√
Tram Anh	Trâm Anh	簪缨	√
house,inn,etc	quán	店	?

Table 9: Vietnamese-to-Chinese translation: word by word conversion.

Overall, by speculating that ' 赏识茶' is a typo of ' 赏茶', a Chinese reader can easily guess the true meaning of the source Vietnamese, '西班牙游客在簪缨馆赏茶(*Spanish tourists enjoy tea at Tram Anh shop*)' from the translation given by our system, '西班牙游客赏识茶在簪缨店(*Spanish tourists appreciate tea at Tram Anh shop*)'.