Comparing Different Criteria for Vietnamese Word Segmentation

Quy T. Nguyen[1] *Ngan L.T. Nguyen*[2] *Yusuke Miyao*[2]
(1) University of Informatics, Hochiminh city, Vietnam
(2) National Institute of Informatics, Chiyoda-ku, Tokyo, Japan
quynt@uit.edu.vn, ngan@nii.ac.jp, yusuke@nii.ac.jp

Abstract

Syntactically annotated corpora have become important resources for natural language processing due in part to the success of corpus-based methods. Since words are often considered as primitive units of language structures, the annotation of word segmentation forms the basis of these corpora. This is also an issue for the Vietnamese Treebank (VTB), which is the first and only publicly available syntactically annotated corpus for the Vietnamese language. Although word segmentation is straight-forward for space-delimited languages like English, this is not the case for languages like Vietnamese for which a standard criterion for word segmentation does not exist. This work explores the challenges of Vietnamese word segmentation through the detection and correction of inconsistency for VTB. Then, by combining and splitting the inconsistent annotations that were detected, we are able to observe the influence of different word segmentation criteria on automatic word segmentation, and the applications of word segmentation, including text classification and English-Vietnamese statistical machine translation. The analysis and experimental results showed that our methods improved the quality of VTB, which positively affected the performance of its applications.

Title and Abstract in another language, L_2 (optional, and on same page)

So sánh các tiêu chí tách từ khác nhau thông qua ứng dụng

Trong bài báo này, chúng tôi khảo sát những nhãn ranh giới từ được đánh dấu trong ngữ liệu cây cú pháp tiếng Việt gọi tắt là VTB. Từ việc khảo sát những trường hợp bị gán nhãn không nhất quán, chúng tôi xác định một số trường hợp khó khăn của bài toán tách từ. Dựa trên những trường hợp này, chúng tôi xây dựng và khảo sát một số tiêu chí tách từ khác nhau. Cụ thể, chúng tôi đã đánh giá các các tiêu chí này thông qua bộ tách từ tự động, và hai ứng dụng: dịch tự động Anh-Việt theo phương pháp thống kê và phân loại văn bản tiếng Việt. Kết quả thí nghiệm cho thấy: (1) các tiêu chí tách từ khác nhau có ảnh hưởng đến độ chính xác của ứng dụng, (2) việc nâng cao chất lượng cho VTB là cần thiết để xây dựng ứng dụng có chất lượng cao.

Keywords: treebank, inconsistency detection, word segmentation, Vietnamese.

Keywords in L_2: ngữ liệu gán nhãn cú pháp, phát hiện nhãn không nhất quán, tách từ, tiếng Việt.

1 Introduction

Treebanks, which are corpora annotated with syntactic structures, have become more and more important for language processing. In order to strengthen the automatic processing of the Vietnamese language, the Vietnamese Treebank has been built as a part of the national project, "Vietnamese language and speech processing (VLSP)" (Nguyen et al., 2009c). However, in our preliminary experiment with VTB, when we trained the Berkeley parser (Petrov et al., 2006) and evaluated it by using the corpus, the parser achieved only 65.8% in F-score. This score is far lower than the state-of-the-art performance reported for the Berkeley parser on the English Penn Treebank, which reported 90.3% in F-score (Petrov et al., 2006). There are two possible reasons to explain this outcome. One reason for this outcome is the difficulty of parsing Vietnamese, which requires new parsing techniques. The second reason is the quality of VTB, including the quality of the annotation scheme, the annotation guidelines, and the annotation process.

The Vietnamese Treebank (VTB) contains 10.433 sentences (274.266 tokens) annotated with three layers: word segmentation, POS tagging, and bracketing. This paper focuses on the word segmentation, since *words* are the most basic unit of a treebank[1] (Di Sciullo and Edwin, 1987), and defining words is the first step (Xia, 2000b,a; Sornlertlamvanich et al., 1997, 1999). For languages like English, defining words is almost trivial, because the blank spaces denote word delimiters. However, for an isolating language like Vietnamese, for which blank spaces play a role of syllable delimiters, defining words is not a trivial problem. For example, the sentence "*Học sinh học sinh học (students learn biology)*[2]" is composed of three words "*học sinh (student)*", "*học (learn)*," and "*sinh học (biology)*". Word segmentation is expected to break down the sentence at the boundaries of these words, instead of splitting "*học sinh (student)*" and "*sinh học (biology).*" Note that the terminology *word segmentation* also refers to the task of extracting "words" statistically without concerning a gold-standard for segmentation, as in (Seng et al., 2009; Ha, 2003; Le et al., 2010). In such a context, the extracted "words" are more appropriate for building a dictionary, rather than for corpus-based language processing, which are outside of the scope of this paper. Because of the discussed characteristics of the language, there are challenges in establishing a gold standard for Vietnamese word segmentation.

The difficulties in Vietnamese word segmentation have been recognized by many researchers (Ha, 2003; Nguyen et al., 2004, 2006; Le et al., 2010). Although most people agree that the Vietnamese language has two types of words: single and compound, there is little consensus as to the methodology for segmenting a sentence into words. The disagreement occurs not only because of the different functions of blank spaces (as mentioned above), but also because Vietnamese is not an inflectional language, as is the case for English or Japanese, for which morphological forms can provide useful clues for word segmentation . While similar problems also occur with Chinese word segmentation (Xia, 2000b), Vietnamese word segmentation may be more difficult, because the modern Vietnamese writing system is based on Latin characters, which represent the pronunciation, but not the meaning of words. All these characteristics make it difficult to perform word segmentation for Vietnamese, both manually and automatically, and have thus resulted in different criteria for word segmenation. However, so far there have been few studies on the challenges in word segmentation, and the comparison of different word segmentation criteria.

[1] In this paper, the terminology *word* is used with the meaning *the most basic unit of a treebank.*

[2] The English translation for a Vietnamese example text is given in parentheses following the text.

In this paper, a brief introduction of the Vietnamese Treebank (VTB) and its annotation scheme are provided in Section 2. Then, we described our methods for the detection and correction of the problematic annotations in the VTB corpus (Section 4.2). We classified the problematic annotations into several patterns of inconsistency, part of which were manually fixed to improve the quality of the corpus. The rest, which can be considered as the most difficult and controversial instances of word segmentation, were used to create different versions of the VTB corpus representing different word segmentation criteria. Finally, we evaluated these criteria in automatic word segmentation, and its application in text classification and English-Vietnamese statistical machine translation in Section 4.

This study is not only beneficial for the development of computational processing technologies for Vietnamese, a language spoken by over 90 million people, but also for similar languages such as Thai, Laos, and so on. This study also promotes the computational linguistic studies on how to transfer methods developed for a popular language, like English, to a language that has not yet intensively studied.

2 Word segmentation in VTB

Word segmentation in VTB aims at establishing a standard for word segmentation in a context of multi-level language processing. VTB specifies 12 types of units that should be identified as words (Table 1) (Nguyen et al., 2009b), which can be divided up into three groups: single, compound, and special "words." Single words contain only one token. The terminology *tokens* refers to text spans that are separated from each other by blank spaces. Compound words have two or more tokens, and are divided into four types: compound words composed by semantic coordination (semantic-coordinated compound), compound words composed by semantic subordination (semantic-subordinated compound), compound words with an affix, and reduplicated words. Special "words" include idioms, locutions, proper names, date times, numbers, symbols, sentence marks, foreign words, or abbreviations. The segmentation of these types of words forms a basis for the POS tagging, with 18 different POS tags, as shown in Table 2 (Nguyen et al., 2009d).

Each unit in Table 1 goes with several example words; English translations are given in parentheses. Furthermore, we added a translation for each token, where possible, so that readers who are unfamiliar with Vietnamese can have an intuitive idea as to how the compound words are formed. The subscript of a token translation is the index of that token in the compound word. However, for some tokens, we could not find any appropriate English translation, so we gave it an empty translation marked with an asterisk. Note that a Vietnamese word or a token in context can have other meanings in addition to the given translations.

A classifier noun, denoted by the part-of-speech Nc in Table 2, is a special type of word in Vietnamese. One of the functions of classifier nouns is to express the definiteness. For example, the common noun "*bàn*" generally means tables in general, while "*cái bàn*" means a specific table, similar to "the table" in English.

3 Inconsistency detection for word segmentation annotation of VTB

In this section, we analyzed the VTB corpus to determine whether the difficulties in Vietnamese word segmentation affected the quality of VTB annotations. The analysis revealed several types of inconsistent annotations, which are also problematic cases for

Type	Example
Simple word	*ba (father), cá (fish)*
Semantic-coordinated compound	*quần áo / trousers*$_1$ *shirt*$_2$ *(clothes)*
Semantic-subordinated compound	*xe đạp / vehicle*$_1$ *pedal*$_2$ *(bicycle)*
Compound word with affix	*bất lương / not*$_1$ *honest*$_2$ *(dishonest)*
Reduplicated word	*long lanh /* $*_1$ $*_2$ *(glistening)*
Idiom	*có thực mới vực được đạo* *(a hungry belly has no ears)*
Locution	*nói tóm lại (in short)*
Proper name	*Việt Nam (Vietnam)*
Date time, number, symbol	*30-4-1975 (April 30, 1975),* *15% (fifteen percent)*
Sentence marks	. , !
Foreign word	*internet, chat*
Abbreviation	*WTO*

Table 1: Word types in VTB word segmentation guidelines

	Label	Name	Example
1	N	Noun	*tiếng (syllable), nhân dân (people), chim muông (birds)*
2	Np	Proper noun	*Việt Nam, Nguyễn Du*
3	Nc	Classifier noun	*con, cái, bức*
4	Nu	Unit noun	*mét (meter), nhúm (pinch), đồng (VND)*
5	V	Verb	*ngủ (sleep), ngồi (sit), suy nghĩ (think)*
6	A	Adjective	*tốt (good), đẹp (beautiful), cao (high)*
7	P	Pronoun	*tôi (I), hắn (he), nó (it)*
8	L	Determiner	*mỗi (every), những, mấy*
9	M	Number	*một (one), vài (a few), rưỡi (half)*
10	R	Adverb	*đã, sẽ, đang*
11	E	Preposition	*trên (on), dưới (under), trong (in)*
12	C	Conjunction	*và (and), tuy nhiên (however)*
13	I	Exclamation	*ôi, chao, a ha*
14	T	Particle	*ạ, ấy, chăng*
15	B	Foreign word	*internet, email, video, chat*
16	Y	Abbreviation	*APEC, WTO, HIV*
17	S	Affix	*bất, vô, đa*

Table 2: VTB part-of-speech tag set

Vietnamese word segmentation. Our analysis is based on two types of inconsistencies: variation and structural inconsistency, which are defined below.

Variation inconsistency: is a sequence of tokens, which has more than one way of segmentation in the corpus. For example, "*con gái/girl*" can remain as one word, or be segmented into two words, "*con*" and "*gái*". A variation can be an annotation inconsistency, or an ambiguity in Vietnamese. While ambiguity cases reflect the difficulty of the language, annotation inconsistencies are usually caused by the confusion in the decision of annotators, which should be eliminated in annotation. We use the term *variation instance* to refer to a single occurrence of a variation.

Structural inconsistency: happens when different sequences have similar structures, thus should be split in the same way, but are segmented in different ways in the corpus. For example, "*con gái/girl*" and "*con trai/boy*" have similar structures: a combination of a classifier noun and a common noun Nc + N, so when "*con gái/girl*" is split, and "*con trai/boy*" is not, it is considered as a structural inconsistency of Nc. It is likely that structural inconsistency at the word segmentation level complicates the higher levels of processing, including POS tagging and bracketing.

3.1 Variation inconsistency detection

3.1.1 Detection method

N-gram	Number of variations	Number of variation instances
2-gram	157	2686 (92.9%)
3-gram	31	177 (6.1%)
4-gram	7	28 (1.0%)
Total	195	2891 (100.0%)

Table 3: Statistics of N-gram variations

POS sequences	Count	Examples
N-N	83	*vụ việc/* $*_1$ *job*$_2$ *(event),* *quê nhà/ native place*$_1$ *house*$_2$ *(hometown)*
V-N	33	*nói chuyện/ say*$_1$ *story*$_2$ *(say),* *cho phép/ give*$_1$ *permission*$_2$ *(permit)*
V-V	25	*ra vào/ go out*$_1$ *go in*$_2$ *(go in and out)*
N-A	22	*đường mòn/ path*$_1$ *worn*$_2$ *(trail),* *năm xưa/ year*$_1$ *old*$_2$ *(long ago)*
N-V	20	*nhà ở/ house*$_1$ *live*$_2$ *(house),* *câu hỏi/ sentence*$_1$ *question*$_2$ *(question)*
Nc-N	16	*niềm tin/* $*_1$ *believe*$_2$ *(belief),* *bà mẹ/ Mrs.*$_1$ *mother*$_2$ *(the mother)*
A-A	13	*đen trắng/ black*$_1$ *white*$_2$ *(black and white),* *đúng mức/ suitable*$_1$ *level*$_2$ *(moderate)*
V-R	11	*trở lại/ go*$_1$ *back*$_2$ *(return)*
N-P	9	*trước đây/ before*$_1$ *now*$_2$ *(previous)*
A-N	8	*cao tầng/ high*$_1$ *storey*$_2$ *(multi-storey)*

Table 4: Top 10 POS sequences of 2-gram variation inconsistencies

The detection method for variation inconsistency is based on N-gram sequences and the phrase structures in the VTB, following the definition for variation inconsistency, above. In detail, we counted N-gram sequences of different lengths in VTB that have two or more ways of word segmentation, satisfying one of the following two conditions:

- N tokens are all in the same phrase, and all have the same depth in phrase. For

POS pattern	Count
N-	148
V-	79
A-	27
Nc-	21
R-	17
E-	12
S-	11
C-	10
M-	10
P-	7
Np-	3
Nu-	2
L-	1
T-	1
Total	349

Table 5: Counts of POS sequences of 2-gram variation inconsistencies grouped by the first POS

POS pattern	Count
-N	166
-V	53
-A	45
-P	40
-R	16
-M	9
-Np	8
-C	4
-X	2
-T	2
-Nc	1
-S	1
-Nu	1
-Nb	1
Total	349

Table 6: Counts of POS sequences of 2-gram variation inconsistencies grouped by the second POS

example, the 3-gram "*nhà tình nghĩa (house of gratitude)*" in this structure "*(NP (Nc-H căn) (N nhà) (A tình nghĩa))*," OR

- N tokens are all in the same phrase, and some token can appear in an embedded phrase which contains only one word. For example, "*nhà tình nghĩa*" in this structure "*(NP (Nc-H căn) (N nhà) (ADJP (A tình nghĩa)))*," where the ADJP contains only one word.

3.1.2 Evaluation and results

Table 3 shows the overall statistics of the variation inconsistency detected by method described above. Most of the difficult cases of word segmentation occur in two-token variations, occupying the majority of variations (92.9%). This ratio of 2-gram variations is much higher than the average ratio of two-token words in Vietnamese, as reported in (Nguyen et al., 2009a), which is 80%. Variations that have lengths of three and four tokens occupy 6.1% and 1.0%, respectively.

We estimated the precision of our method by randomly selecting 130 2-gram variation instances, extracted from the method described above, and manually checked whether the inconsistencies are true. We found that 129 cases occupying 99.2% of all extracted 2-grams are true inconsistencies. Only one instance of inconsistency was an ambiguous sequence *giá cả*, which is one word when it means *price*, and two words *giá/price cả/all* in *đều có giá cả/all have (their own) price*. The precision of our method is high, so we can use the extracted variations to provide insights on the word segmentation problem.

3.1.3 Analysis of 2-gram variations

We further analyzed the 2-gram variations to understand what types of 2-grams were most confusing for annotators. The analysis results showed that compound nouns, compound verbs, and compound adjectives are the top difficult cases of word segmentation.

We classified the 2-gram variations according to their POS sequences in case the tokens in the 2-gram are split. There are a total of 54 patterns of POS sequences. The top 10 confusing patterns, their counts of 2-gram variations, and examples are depicted in Table 4. Table 5 and Table 6 show the POS patterns that are a specific POS tag appearing at the beginning or ending of the sequence.

Investigating the inconsistent 2-grams extracted, we found that most of them are compound words according to the VTB guidelines (Section 2). One of the reasons why the compound words are sometimes split, is because the tokens in those compound words have their own meanings, which seem to contribute to the overall meaning of the compounds. This can be seen through the examples provided in Table 4, where the meanings of tokens are given with a subscript. This scenario has proven to be problematic for the annotators of VTB.

Furthermore, by observing the POS patterns in Table 5 and Table 6, we can see the potential for structural inconsistency, particularly for closed-set POS tags. Among them, classifier nouns (Nc) and affixes (S) are two typical cases of structural inconsistency, which will be used in several settings for our experiments. The same affix or classifier noun can modify different nouns, so when they are sometimes split, and combined in the variations, we can conclude that classifier nouns and affixes involve in structural inconsistencies. In the following section, we present our detection method for structural inconsistency for classifier nouns and affixes.

3.2 Structural inconsistency detection for classifier nouns and affixes

3.2.1 Detection method

We collected all affixes and classifier nouns in the VTB corpus, and then extracted 2-grams containing these affixes or classifier nouns, which they are also structural inconsistencies. For example, since "*con*" is tagged as a classifier noun in VTB, we extracted all 2-grams of "*con*" including both "*con gái/girl*" and "*con trai/boy*".

Even though the sequence, "*con trai*" is always split into two words throughout the corpus, it can still be an inconsistency, if we consider similar structures such as "*con gái*". In other words, by this method, we extract sequences that may be consistent at the surface level, but are not consistent, if we consider the higher analysis levels, such as POS tagging.

According to the VTB POS-tagging annotation guidelines (Nguyen et al., 2009d), classifier nouns should be separated from the words they modify. However, in practice, when a classifier noun can be standalone as a meaningful single word, it may be difficult for annotators to decide whether to split, or to combine it with the noun it modifies to form a semantic-subordinated compound. For example a classifier noun, e.g., "*con*" in "*con trai (boy)*", or "*con gái (girl)*", can also be a simple word, which means "*I (first person pronoun used by a child when talking to his/her parents)*", or part of a complex noun "*con cái (children)*". Therefore, in our experiments, we want to evaluate the "splitting" and "combining" of these cases, in order to see whether the solution is successful for applications of the corpus.

Type	Number of combinations	Number of instances
Affix	345	1289
Nc	2715	10445

Table 7: Statistics of targeted structural inconsistency

3.3 Correction of inconsistency in annotations of special characters

By examining the variations extracted by the variation inconsistency detection, we found that there are cases when a special character like a percentage (%) in "30%", is split or combined with "30". Such inconsistent annotations are manually fixed based on their textual context.

By checking structural inconsistencies of these special characters, including percentages (%), hyphens (-), and other symbols, we found quite a significant number of inconsistent annotations. For example, the character, %, in "30%" is split, but is combined with a number in "50 %", which is considered to be a structural inconsistency. Note that it can be argued that splitting "N%" into two words or combined in one word is dependent on the blank space in-between N and "%". Higher-levels of annotation such as POS tagging is significant, because we may need one or two different POS tags for the different methods of annotation. Therefore, we think that it is better to carefully preprocess text and segment these special characters in a consistent way.

To improve the quality of the VTB corpus, we extracted the problematic sequences using patterns of the special characters, and manually fixed this type of inconsistency. Automatic modification is difficult, since we must check the semantics of the special characters in their contexts. For example, hyphens in date expressions like “5-4-1975”, which refers to the date, "the fifth of April, 1975," are combined with the numbers. However, when the hyphen indicates “(from) to” or “around ... or”, as in “*2-3 giờ sáng*” meaning “around 2 or 3 o’clock in the morning”, we decided to separate it from the surrounding numbers. As a result, we have fixed 685 inconsistent annotations of 21 special characters in VTB.

4 Comparing different word segmentation criteria

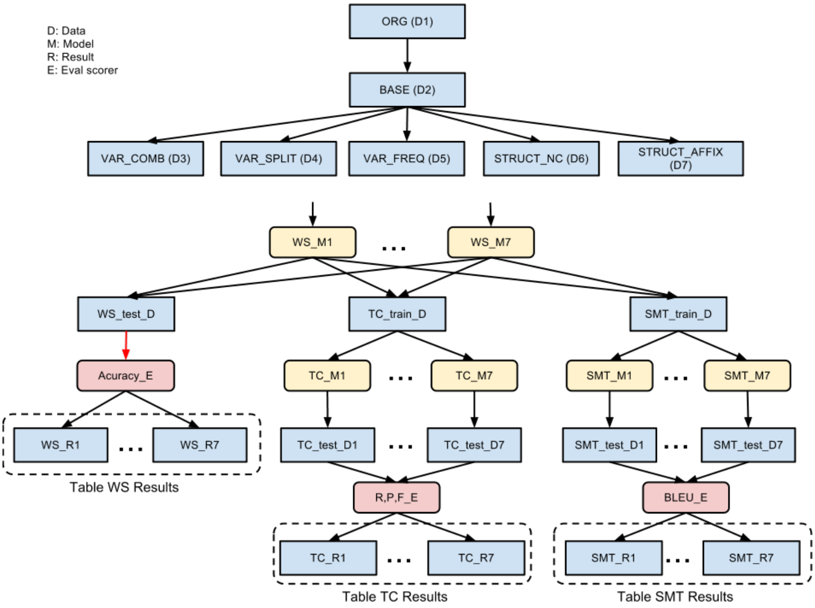

Figure 1: Experimental diagram showing how different word segmentation criteria are encoded in our experiments.

The variation inconsistency and structural inconsistency found in Section 3 can also be seen as representatives of different word segmentation criteria for Vietnamese. We organized the inconsistency detected in seven configurations of the original VTB corpus. Then, by using these data sets, we could observe the influence of the different word segmentation criteria on three tasks: automatic word segmentation, text classification, and English-Vietnamese statistical machine translation.

4.1 Data preparation for experiments on word segmentation criteria

Seven data sets corresponding to different segmentation criteria are organized as follows.

- ORG : The original VTB corpus.
- BASE : The original VTB corpus + Manual modification of special characters done in Section 3.3.
- VAR_SPLIT : BASE + split all variations detected in Section 3.1.
- VAR_COMB : BASE + combine all variations detected in Section 3.1.
- VAR_FREQ : BASE + select the segmentation with higher frequency among all variations detected in Section 3.1.
- STRUCT_NC : BASE + combine all classifier nouns detected in Section 3.2 with the words they modify.
- STRUCT_AFFIX : BASE + combine all suffixes detected in Section 3.2 with the words they modify.

These data sets are used in our experiments as illustrated in Figure 1. The names of the data sets are also used to label our experimental configurations.

4.2 Experimental settings

In this section, we briefly describe the task settings and the methods used for word segmentation (WS), text classification (TC), and English-Vietnamese statistical machine translation (SMT).

4.2.1 Word segmentation (WS)

We used YamCha (Kudo and Matsumoto, 2003), a multi-purpose chunking tool, to train our word segmentation models. The core of YamCha is the Support Vector Machine (SVM) machine learning method, which has been proven to be effective for NLP tasks. For the Vietnamese word segmentation problem, each token is labeled with standard B, I, or O labels, corresponding to the beginning, inside, and outside positions, respectively. The label of each token is determined based on the lexical features of two preceding words, and the two following words of that token. Since the Vietnamese language is not inflectional, we cannot utilize inflection features for word segmentation.

Each of the seven data sets is split into two subsets for training and testing our WS models. The training set contains 8443 sentences, and the test set contains 2000 sentences.

4.2.2 Text classification (TC)

Text classification is defined as a task of determining the most suitable topic from the predefined topics, for an input document. We implemented a text classification system similar to the system presented in (Nguyen et al., 2012). The difference is that we performed the task at the document level, instead of at the sentence level.

The processing of the system is summarized as follows. An input document is preprocessed with word segmentation and stop-word removals. Then, the document is represented in the form of a vector of weighted words appearing in the document. The weight is calculated using standard tf-idf product. An SVM-based classifier predicts the most probable topic for the vector, which also is the topic for the input document. In our experiment, for comparison of different word segmentation criteria in topic classification, we only vary the word segmentation model used for this task, while fixing other configurations.

News articles of five topics: music, stock, entertainment, education, and fashion are used. The sizes of the training and test data sets are summarized in Table 8.

Topic	Training (documents)	Test (documents)
Music	900	813
Stock	382	320
Entertainment	825	707
Education	821	707
Fashion	412	302
Total	3340	2849

Table 8: Data used in the text classification experiment

4.2.3 Statistical machine translation (SMT)

A phrase-based SMT system for English-Vietnamese translation was implemented. In this system, we used SRILM (Stolcke, 2002) to build the language model, GIZA++ (Och and Ney, 2003) to train the word-aligned model, and Moses (Holmqvist et al., 2007) to train the phrase-based statistical translation model. Translation results are evaluated using the word-based BLEU score (Papineni et al., 2002). Both training and test data are word-segmented using the word segmentation models achieved. For the experiment, we used the VCL_EVC bilingual corpus (Dinh and Hoang, 2005), 18000 pairs of sentences for training, and 1000 pairs for testing.

4.3 Experimental results and analysis

	Recall	Precision	F-score
ORG	95.89	95.44	95.66
BASE	96.00	95.60	95.80
VAR_COMB	96.05	95.69	95.87
VAR_SPLIT	96.53	96.27	**96.40**
VAR_FREQ	96.20	95.85	96.02
STRUCT_NC	95.08	94.79	94.93
STRUCT_AFFIX	96.03	95.59	95.81

Table 9: Evaluation results of automatic word segmentation with different WS criteria

Evaluation of word segmentation models trained on different versions of the VTB are given in Table 9. The experimental results with text classification and English-Vietnamese

	Recall	Precision	F-score
ORG	98.20	97.90	98.05
BASE	98.63	98.79	**98.71**
VAR_COMB	98.45	98.63	98.54
VAR_SPLIT	98.60	98.72	98.66
VAR_FREQ	98.68	98.65	98.67
STRUCT_NC	98.34	98.35	98.34
STRUCT_AFFIX	98.61	98.67	98.64

Table 10: Evaluation results of text classification with different word segmentation methods

	BLEU
ORG	36.36
BASE	36.44
VAR_COMB	36.03
VAR_SPLIT	**36.91**
VAR_FREQ	36.75
STRUCT_NC	35.41
STRUCT_AFFIX	36.36

Table 11: Evaluation results of SMT with different word segmentation methods

statistical machine translation are shown in Table 10 and Table 11, respectively. There are two important conclusions that can be drawn from these tables: (1) The quality of the treebank strongly affects the applications, since our BASE model and most of the other enhanced models improved the performance of TC and SMT systems; (2) "Splitting" seems to be a good solution for word segmentation of controversial cases, including the split of variations, affixes, and classifier nouns.

According to the result in Table 9, the VAR_SPLIT criterion gives the highest WS performance. With the exception of STRUCT_NC, all of the modifications to the original VTB corpus increase the performance of WS. However, the word segmentation criterion with higher performance is not necessarily a better criterion, but a criterion should also be judged through applications of word segmentation. In both SMT and TC experiments, the BASE model, which is based on the manually-modified inconsistency of special characters, achieved better results than the ORG model. In particular, in the TC experiment, the BASE model achieved 0.66 point higher than ORG, which is a significant improvement. The results support the conclusion that the quality of the word-segmentation corpus is very important for building NLP applications.

The SMT results show that three out of six augmented models, VAR_SPLIT, VAR_FREQ and BASE, performed better than the ORG configuration. Among them, the best-performing model, VAR_SPLIT achieved 36.91 BLEU score, which is 0.55 higher than ORG. In TC results, all six augmented models achieved higher results than ORG. In general, the augmented models performed better than the ORG. Additionally, because our automatic methods for inconsistency detection could not cover all of the types of inconsistencies in word segmentation annotation, further improvement of corpus quality is demanded.

Comparing the results of STRUCT_AFFIX and STRUCT_NC with BASE in WS, TC, and SMT, we can observe that combining affixes with their head nouns resulted in slightly better results for WS and TC, and did not change the performance of SMT. However, the combination of classifier nouns with their head nouns had negative effects on WS and SMT.

Another part of the scope of our experiment is to compare two solutions for controversial cases of word segmentation, splitting and combining. Splitting and combining variations are reflected by VAR_COMB and VAR_SPLIT, while STRUCT_AFFIX and STRUCT_NC represent the combination of affixes or classifier nouns with the words that they modify. STRUCT_AFFIX and STRUCT_NC are contrasted with BASE where affixes and classifier nouns remain untouched. Comparing VAR_COMB and VAR_SPLIT in both the TC experiment and SMT experiment, we see that the VAR_SPLIT results are better in both cases. Since the ratio of combined variations in the ORG corpus is 60.9%, it can be observed that splitting seems to be better than combining for WS, TC and SMT.

5 Conclusion

In this paper, we have provided a quantitative analysis of the difficulties in word segmentation, through the detection of problematic cases in the Vietnamese Treebank. Based on the analysis, we automatically created data that represent the different word segmentation criteria, and evaluated the criteria indirectly through their applications.

Our experimental results showed that manual modification, done for annotation of special characters, and most other word segmentation criteria, significantly improved the performances of automatic word segmentation, text classification and statistical machine translation, in comparison with the use of the original VTB corpus. Since the VTB corpus is the first effort in building a treebank for Vietnamese, and is the only corpus that is publicly available for NLP research, this study contributes to further improvement of the corpus quality, which is essential for building efficient NLP systems in future.

References

Di Sciullo, A. M. and Edwin, W. (1987). On the definition of word. *The MIT Press*.

Dinh, D. and Hoang, K. (2005). Building an annotated english-vietnamese parallel corpus for training vietnamese-related nlps. *Mon-Khmer Studies: A Journal of Southeast, Asian Languages and Cultures*, 35:21–36.

Dinh, Q. T., Le, H. P., Nguyen, T. M. H., Nguyen, C. T., Rossignol, M., and Vu, X. L. (2008). Word segmentation of vietnamese texts: a comparison of approaches. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, B. M. J. M. J. O. S. P. D. T., editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Ha, L. A. (2003). A method for word segmentation in vietnamese. In *Proceedings of Proceedings of Corpus Linguistics*, pages –, Lancaster, UK.

Hoang, C. D. V., Dinh, D., Nguyen, L. N., and Ngo, Q. H. (2007). A comparative study on vietnamese text classification methods. In *IEEE International Conference: Research, Innovation and Vision for the Future*, pages 267 – 273.

Holmqvist, M., Stymne, S., and Ahrenberg, L. (2007). Getting to know moses: initial experiments on german–english factored translation. In *Proceedings of the Second Work-*

shop on Statistical Machine Translation, StatMT '07, pages 181–184. Association for Computational Linguistics.

Kudo, T. and Matsumoto, Y. (2003). Fast methods for kernel-based text analysis. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 24–31, Stroudsburg, PA, USA. Association for Computational Linguistics.

Le, H. P., Nguyen, T. M. H., Roussanaly, A., and Vinh, H. T. (2008). Language and automata theory and applications. chapter A Hybrid Approach to Word Segmentation of Vietnamese Texts, pages 240–249. Springer-Verlag, Berlin, Heidelberg.

Le, T. H., Le, A. V., and Le, T. K. (2010). An unsupervised learning and statistical approach for vietnamese word recognition and segmentation. In *Proceedings of the Second international conference on Intelligent information and database systems: Part II*, ACIIDS'10, pages 195–204, Berlin, Heidelberg. Springer-Verlag.

Nguyen, C. T., Nguyen, T. K., Phan, X. H., Nguyen, L. M., and Ha, Q. T. (2006). Vietnamese word segmentation with crfs and svms: An investigation. In *Proceedings of the 20th Pacific Asia Conference on Language, Information, and Computation (PACLIC)*.

Nguyen, D. (2009). Using search engine to construct a scalable corpus for vietnamese lexical development for word segmentation. In *Proceedings of the 7th Workshop on Asian Language Resources*, ALR7, pages 171–178, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nguyen, G. S., Gao, X., and Andreae, P. (2009a). Vietnamese document representation and classification. In *Proceedings of the 22nd Australasian Joint Conference on Advances in Artificial Intelligence*, AI '09, pages 577–586, Berlin, Heidelberg. Springer-Verlag.

Nguyen, P. T., Vu, X. L., and Nguyen, T. M. H. (2009b). Vtb word segmentation guidelines (vlsp project, report sp 8.2).

Nguyen, P. T., Vu, X. L., Nguyen, T. M. H., Dao, M. T., Dao, T. M. N., and Le, K. N. (2009c). Vtb bracketing guidelines (vlsp project, report sp 7.3).

Nguyen, P. T., Vu, X. L., Nguyen, T. M. H., Nguyen, V. H., and Le, H. P. (2009d). Building a large syntactically-annotated corpus of vietnamese. In *Proceedings of the Third Linguistic Annotation Workshop*, ACL-IJCNLP '09, pages 182–185, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nguyen, Q., Nguyen, A., and Dinh, D. (2012). An approach to word sense disambiguation in english-vietnamese-english statistical machine translation. In *The 9th IEEE - RIVF International Conference and Communication Technologies*, pages 125–129.

Nguyen, T. B., Nguyen, T. M. H., Romary, L., and Vu, X. L. (2004). Lexical descriptions for Vietnamese language processing. In *The 1st International Joint Conference on Natural Language Processing - IJCNLP'04 / Workshop on Asian Language Resources*, page 8 p, Sanya, Hainan Island, China. none. Colloque avec actes et comité de lecture. internationale. A04-R-031 || nguyen04b A04-R-031 || nguyen04b.

Nguyen, T. M. H., Hoang, T. H. L., and Vu, X. L. (2009e). Vtb part-of-speech tagging guidelines (vlsp project, report sp 7.3).

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 433–440, Stroudsburg, PA, USA. Association for Computational Linguistics.

Seng, S., Besacier, L., Bigi, B., and Castelli, E. (2009). Multiple text segmentation for statistical language modeling. In *10th International Conference on Speech Science and Speech Technology (InterSpeech 2009)*, pages 2663–2666.

Sornlertlamvanich, V., Charoenporn, T., and Isahara, H. (1997). Orchid: Thai part-of-speech tagged corpus. technical report orchid tr-nectec-1997-001. Technical report, National Electronics and Computer Technology Center.

Sornlertlamvanich, V., Takahashi, N., and Isahara, H. (1999). Building a thai part-of-speech tagged corpus (orchid). *The Journal of the Acoustical Society of Japan (E)*, 20(3):189–140.

Stolcke, A. (2002). Srilm - an extensible language modeling toolkit. pages 901–904.

Tran, T. O., Le, A. C., and Ha, Q. T. (2010). Improving vietnamese word segmentation and pos tagging using mem with various kinds of resources. *Information and Media Technologies*, 5(2):890–909.

Xia, F. (2000a). The part-of-speech tagging guidelines for the penn chinese treebank (3.0).

Xia, F. (2000b). The segmentation guidelines for the penn chinese treebank (3.0).

Xue, N., Xia, F., Huang, S., and Kroch, A. (2000). The bracketing guidelines for the penn chinese treebank (3.0).