A method for word segmentation in Vietnamese

Le An Ha
Research Group in Computational Linguistics
School of Humanities, Languages and Social Sciences
University of Wolverhampton
Stafford Street, Wolverhampton,
WV1 1 SB, UK
L.A.Ha@wlv.ac.uk

Abstract

Word segmentation is the very first step in natural language processing for languages such as Vietnamese. Given the fact that un-annotated corpora are the only widely available resources, we propose a method of word segmentation for Vietnamese, which only uses n-gram information. We calculate the probabilities of different combinations of n-grams in a chunk, and choose the one that produces maximum probability. In order to calculate these probabilities, we build a 10M-word corpus of two year newspaper article. The results, while not very impressive, show that the method works.

1. Introduction

It has long been known that word segmentation is a vital problem in natural language processing for certain Asian languages such as Chinese, Japanese, and Vietnamese. Without knowing the boundaries of words in a sentence, we cannot move anywhere further. The problem is that in these languages, each sound unit, when written, is separated by a space, and there is nothing to be used to identify word boundaries. Of courses, native speakers do not have any problem using their language, they can still communicate efficiently and precisely in both spoken and written form without explicitly identifying word boundaries. But a computer is more confused by word boundary in these languages. But "what is a word?" itself is also a problem. This question cannot be answered in a way to satisfy everybody, so they continue to have their own definitions. The evidence is that there were efforts in the past, which tried to group sound units into "words" in written form, such as "ky-thuat", or "kythuat" instead of "ky thuat" (technology), and failed, because not everybody were happy with this kind of notation, and they are also still confusing. Thus the problem of word boundary continues to exist. A lot of methods have been introduced to solve the segmentation problem in Japanese and Chinese, but none for Vietnamese. Another problem is that these methods tend to use lexical resources, which is still underdeveloped for Vietnamese. These issues lead to an attempt to develop a method of word segmentation in Vietnamese, using only pure statistics. In order to do that, we build a corpus of 10M tokens, and use it to calculate different scores needed for the method. In this paper, different issues arising when we carry out the experiment will be also discussed.

2. About Vietnamese language and the problem of word boundary

Vietnamese, although still debatable, is a monosyllabic language, and belongs to the southeast Asian language family. It has different phonetical, grammatical and semantic features from Indo-European languages, which make it difficult for the Vietnamese, not only to learn European languages, but also to develop techniques for natural language processing. The options are either try to make the Vietnamese language fit into the framework of other well-investigated European languages, or to develop from scratch a new framework. Both are proved to be very painful, where the former is not very successful, given the fact that Vietnamese is significantly different from Indo-European languages, and the latter needs a lot of resources, both human and materials, which may not be available for a developing country such as Vietnam.

Vietnamese, traditionally, is a spoken language, thus to the native speaker, the problem of identifying word boundary is not a very serious problem. They know what it is, and use it naturally, even if there is some difference of opinion between them on the word boundary issue, it does not make the communication process more difficult, because they all agree on what is a sound, the more basic unit of Vietnamese. When written forms of the language were developed either using Chinese characters or Latin characters, these written forms are only the extensions of the spoken form, where each sound was represented by a (sequence of) character(s), and then separated by a space. This also causes no difficulty for native speakers to understand each other and the problem of word boundary did not arise yet. At the beginning and middle of the 20th century, when Vietnamese scholars were introduced to Western grammar schools, some changes in written form of Vietnamese had been proposed, to make it more "word oriented", using different marks for the word boundary to be more explicit, and the language look more like European ones. These changes included the elimination of space between syllables that may form "word", and using the hyphen marks. These attempts were unsuccessful, maybe because of the nature of Vietnamese language, whereas the identification of words may not be that important at all. The discussion about what is a word in Vietnamese goes on, and, up to now, there is still no general agreement on the issue.

The development of computers, in general, and computational linguistics, in particular, does not permit researchers to avoid the problem of word boundary anymore. Unlike native speakers, computers (at least until now), cannot easily identify word boundaries in electronic texts. And it is the bottleneck of natural language processing for Vietnamese, because without knowing word boundaries, the computer cannot do anything else (part-of-speech tagging, parsing etc.), all it can do is n-grams counting.

In fact, a way to process texts without knowing about word boundary may exist after all. But given the reality that a lot of developed methods of automatic natural language processing are for word-oriented languages, the time of developing word segmentation techniques, and after that, applying other available methods of natural language processing, maybe a lot less than the time spent on developing whole new theories and techniques for Vietnamese language processing.

Why identifying word boundary in Vietnamese is so difficult?

As discussed in the previous section, Vietnamese, at first, is a spoken language. In a spoken language, the most important unit, is the syllable, not the word. The word boundary can vary from person to person, and still does not affect the communication process. Another reason is that the combination of different syllable units is the only way to construct new lexical units to describe new concepts in Vietnamese. There is no prefix, suffix in Vietnamese, just syllable, and it makes everything look confusing. The fact that the part-of-speech system of Vietnamese, like that of Chinese or Japanese, is not very well-defined leads to differences among dictionaries, thus contributes its part in the difficulty of word boundary identification. Just another problem of word boundary is that a large part of vocabulary of Vietnamese come from (ancient) Chinese, and these units seem to bound together more strongly than pure Vietnamese syllable units, for example, "cong nhan" (worker), "thuong nhan" (businessman). ("nhan" roughly means "person" in Chinese). Shall we treat these units differently or the same? Note that their word order are different from natural Vietnamese word order (modifier head in comparison to head modifier as in pure Vietnamese (nguoi lao dong, nguoi buon ban).

Yet there is another type of lexical units in Vietnamese which is problematic. It is reduplication, usually two syllables, in which only one, or even none, has some meanings, the other one is just a variant of the sound of the other. This type of units is very popular, especially for adjectives, or in fact, almost every adjectives have their reduplication forms. It can be explained, for as a spoken language, the sound is very important. The question is that how we will treat this kind of unit in word boundary identification.

3. The construction of the corpus and some of its properties.

With so many problems, the only solution seems to be the use of corpora to discover the nature of word boundary in Vietnamese. Until recently, due to different reasons, it is very difficult to construct a Vietnamese corpus. It is either because the lack of human or materials resources, where there are other more important things to do in a developing country. But the rapid development of the Internet and World-wide web in Vietnam makes it possible to build an inexpensive corpus. News, stories, books, etc. become more and more available on the web. Finding resources to build a corpus is not a difficult task any more. Of course, if we consider the task of building a well-balanced corpus, maybe the web is still not a very good resource, but if our task is just to evaluate some techniques in their early days, the web is the right place. And for Vietnamese, methods of natural language processing are in their early days, and researchers cannot wait until a standard corpus is built, to try their techniques. For these reasons, we choose to harvest news from online sources to construct our corpus. In particular, two years news articles from http://vnexpress.net are collected. These articles came from different newspaper sources, thus somehow balanced in styles and genres. These articles are classified into different topics: society, world, business, health, science, and way-of-life. Different perl scripts are used to get these articles, and put them together in a corpus. The size of the final corpus is difficult to define, depending on what we will use to measure the size of the corpus. The Vietnamese characters are encoded using the unicode UTF-8 standard, and stored in a html-friendly format (in order to display the content easily on different applications). If we measure the size of the corpus, using every thing between two spaces (and other punctuation marks) as a token, (except punctuation marks) the number of tokens in the corpus is about 10M. Of courses, one can argue that this is a rather small size corpus, (where different segmentation techniques in Chinese use 40M to 60M corpus), but we think that this is enough for our experiment. The information harvested from the web site allows us to mark the author (sources) of the articles, a short summary, and the date of the articles. These will not be used in our project, but maybe useful for future development, thus they are recorded in the corpus using sgml tags. Paragraphs are also marked according to the html tags.

Statistical figures of the corpus.

There are quite a few problems counting n-gram in Vietnamese. As discussed above, natural language processing in Vietnamese is still in its early days, and the definition for an n-gram is not very well discussed. For example, personal names in Vietnamese, like in Chinese, have some meanings, and unlike Chinese, they are in upper case, thus distinguishable from other syllable. This leads to a problem: shall we convert everything into lower-case before counting n-gram? Of course, there are certain possible solutions for the problem, such as a name-entity recognition module, but we do not have it. In the end, we decide that we will count n-grams in both cases: original, and lower case converted ones. Observing the different between two approaches will give us certain ideas of the effects of lower and upper case in n-gram frequency in Vietnamese.

Like other languages, sentence limit is also a problem in processing Vietnamese, and a simple approach has been used, whereas everything between two periods is a sentence. Hopefully this will not make a lot of errors.

Observing the unigram list gives us the idea of how different Vietnamese is from English. Some of the highest frequency words are given in table 1. Comparing the unigram list when upper case is taken into account and not taken into account shows that it is better to ignore case in this particular application.

When looking at the number of unique unigrams, we find out that it is 60637, which is a very high number, and almost half of them only appear once in the corpus. The explanation is that the corpus contains a lot of foreign words, which will be another issue in NLP applications in Vietnamese.

và (and) 88000	theo 44716	để 28936	số 22936
các 84380	này 43040	trên (on) 27762	năm (year) 22902
của (of) 83915	những 41430	nước (water, nation) 26209	nhân 22726
có 81828	với 38816	vào (into) 25539	động 22471
một (one) 65332	ở 38767	từ (from) 25109	việc 22446
là (to be) 59392	công 37175	hiện 24467	thành 20947
trong (in) 58964	sẽ (will) 34160	nhà 24439	phải 20721
cho 56378	ra 31654	đến 24050	như 20711
người 55388	bị 31093	về 24015	nhiều 20173
được 54722	khi 29587	đó 23843	chính 19952
đã 53700	thể 29308	tại 23172	
không (negative) 48899	ông 28955	cũng 22980	

Table 1: unigram frequency, and their meaning in English (where available).

4. Methods used for word segmentations in different languages.

Word segmentation is a common problem for Asian languages, such as Chinese, Japanese, Korean, etc., although the root of the problem maybe different. Numerous methods have been introduced to solve the problem, which can also be divided into two sub problems: 1)disambiguating word boundary, whereas a lexicon database is available, and 2) identifying word boundary for new lexical items. In order to solve the first problem, dictionaries and statistical methods have been used. Usually, the longest possible string will be matched, using a dictionary, and statistical scores will be calculated to disambiguate when ambiguities arise. These approaches have some problems. Firstly, when based on a dictionary, it implies that the dictionary is a reliable source for natural language processing, in the sense that it is consistent and complete. But in reality, such a dictionary is very hard to find, and different dictionaries seem to be inconsistent, especially with languages which have not been extensively investigated. Furthermore, these days, vocabulary of a developing country language is also developing, and dictionary cannot be up-to-date any more. In particular, the problem with Vietnamese dictionaries is that they are often built on different standards, and have different sets of entries. The grammar is also partly to be blamed, where there is still no concrete grammar theory for Vietnamese, and every attempt to apply Indo-European grammars for Vietnamese have failed, making the dictionary compiler confused about which system of category (s)he should classify her/his words into. Nouns are generally agreed between linguists, but verbs, adjectives and adverbs are arguable. In the past, these problems were tolerant, because, as discussed previously, it does not affect the quality of communication much, but we cannot avoid them any more, if we want to use computer to process natural language in Vietnamese. The computer is not human, and without explicitly showing it the exact word boundary, it will refuse to do anything else.

The main approach for identifying word boundary of new lexical units is the statistical one, where different statistical scores have been calculated, to determine the level of "bound" between different smaller units (sound). The assumption lies behind this approach is that certain units appearing together in the text frequently form a bigger, steadier unit that we call "word". The problem of this approach is to find a good score that really reflects the phenomenon of "word" in the subject language. This problem seems to be a difficult one, where a solid statistical measure cannot guarantee the success of the segmentation process. Mutual information and t-score are examples. Although these scores have solid statistical theories behind them, the results using them are not that impressive. In the t-score case, it maybe because the assumption of normal distribution has failed, or natural occurring texts are not very suitable for information theory, in the case of mutual information. There are still a lot of works to do in the field to establish the relation between the phenomenon of "word" in monosyllabic languages, and statistics, or even disprove the phenomenon. But it does not mean that we should not do anything. Different methods should be tried to solve smaller problems, thus may give hints to the development of a more robust theory of "word".

Following is a short review of methods used for word segmentation.

Maosong el. al. use mutual information and t-scores to identify word boundary in Chinese, and the reported results, although challenged by other authors, are very high (>90%). The problem with this method is the use of t-score implies the normal distribution. Wong and Chan employ a lexicon of 80.000 entries and a corpus of 63M character for word segmentation based on maximum matching and word binding force. This algorithm relies heavily on the lexicon. Sproat et. al. introduce a stochastic finite-state word-segmentation algorithm, which also relies on lexicon resources. Sornlertlamvanich el. al. use C4.5 learning algorithm for word segmentation in Thai. Information used for the algorithm includes string length, mutual information, frequency, entropy.

5. The proposed method

Given the situation that, for Vietnamese, there is a lack of large lexicographic resources, and annotated corpora are also rare, we want to develop a method that will not rely on these resources, and will use only raw corpus. From a raw corpus, frequency is the only statistical score that can be calculated reliably. Our approach begins with other units that native speaker will agree on, which are sentences, and/or chunks separated by punctuation marks (commas, hyphens, quotes, etc.). These units are less ambiguous than words, in both spoken and written forms. We then try to maximise the probability of the chunk, using different segmentations. The final segmentation is the one that gives the chunk the maximum probability. Of course, we are not talking about real probability of an n-gram, or a chunk. The used "probability" of an n-gram is its maximum likelihood estimation, and the probability of a chunk is the product of its n-gram "probabilities". The implicit idea behind this calculation is that given a chunk, it is most likely that it is combined by the segmentation that gives the maximum probability. But in this method of calculation, we face certain problems. The first one is combinatorial explosion, where the number of ways of segmentation is an exponential function of the length of the chunk. Other problems include the difficulty of calculating probability, and sparse data. In order to solve the problem of combinatorial explosion, dynamic programming is employed, whereas the maximum probability of a smaller chunk is only calculated once and then reused in other calculations. By using dynamic programming, the complexity of the problem is reduced to $O(n^3)$. The detailed calculation is given below.

P(i,j): maximised probability of a segment begin at i, end at i+j.
p(i,j): probability of the n-gram begin at i, and end at i+j.
The calculation of maximum probability:

$$P(i,0)=p(i,0);$$
$$P(i,j)=\max \{P(i,0)* P(i+1,j-1), P(j+i,0)*P(i,j-1), P(i,1)*P(i+2,j-2), P(j+i-1,1)*P(i,j-2),..., p(i,j)\}$$

The actual calculation using dynamic programming method is following:

```
for(i=0;i<chunk_length;i++){
        P(i,0)=p(i,0);
        BackTrack(i,0)="this n-gram";
};
for(i=1;i<chunk_length;i++){
        for(j=0;j<chunk_length-i;j++){
                max=p(i,j);
                backtrack="this n-gram";
                k=0;
                while(1){
                        if((i-1-k)<k){
                                last;
                        };
                        if(max< P(k,j)*P(i-1-k,j+1+k)){
                                max= P(k,j)*P(i-1-k,j+1+k);
                                backtrack="(k,j):(i-1-k,j+1+k)";
                        };
                        if((i-1-k)<=k){
                                 last;
                        };
                        if(max< P(k,j+i-k)*P(i-1-k,j)){
                                max= P(k,j+i-k)*P(i-1-k,j);
                                backtrack="(k,j+i-k):(i-1-k,j)"
                        };
                }
                k++;
        };
        P(i,j)=max;
        BackTrack(i,j)=backtrack;
};
```

At each step, the combination that maximises the probability will be recorded for backtracking. In the end, these backtracking information will be used to reconstruct the combination of n-grams that give the chunk the maximum probability. In order to avoid problem of multiplying probability, we use logarithm and convert multiply operations into add operations.

In the experiment, we will stop at the tri-gram level. The reason is that, our corpus is still a small size one, and if we go beyond this tri-gram level, the probability will become biased toward the whole n-gram rather than the combination of smaller n-grams, making the results less reliable. As we can see, the complexity of the algorithm is only $O(n^3)$ (where n is the length of the chunk).

Chunks separated by punctuation marks will be extracted from the corpus and the above calculation will be performed. The results will be given to a native speaker, who will judge whether or not the segmentations make sense.

6. Evaluation: results and some issues

As discussed in various papers, evaluation for word segmentation is very problematic. We think that the root of the problem is that there is no concrete definition for “word” in languages such as Vietnamese. This lead to the phenomenon that even native speakers can not agree with each other about word boundary (Sproat and Chang). And it makes all evaluations of word segmentation, either based on native speakers or on available lexicon, only have relative meanings. Word segmentation is only the first step of natural language processing, further steps should be made before we can look back and say whether or not a word segmentation method is reliable.

In our evaluation scheme, we ask a native speaker to look at the results from the method, and say 1) whether or not he will identify word boundaries like the system, 2) whether or not he thinks the segmentation produced by the system is reasonable. The two questions asked are to make sure that we have to accept the differences in word segmentation, as far as it is reasonable. Of course, one can argue that “reasonable” is rather vague, but so is “word”. An ideal evaluation for word segmentation should be an extrinsic one, where word segmentation is incorporated in an NLP application, but when such an evaluation still cannot be carried out, the above evaluation scheme still gives us a fair idea how the method perform.

We extract randomly 100 chunks from the corpus. From these chunks, the system identify 614 “words”, of which, the evaluator agrees on 315 (51%), and thinks that 402 (65%) are reasonable. The percentage seems to be low, but when taking into account that the method neither uses lexicon nor annotated corpus, we think that the result is not very disappointing. The reason why we do not extend our evaluation is that the we feel with this method alone, we cannot get a higher number of successful. Hopefully, when combining with other methods of word segmentation, the results will become and more acceptable.

Conclusion and future work

The introduced method has yield a promising results, when we take into account its advantage, which is only using un-annotated corpus for word segmentation. But we also raise a few problems of word segmentation, not only for Vietnamese, but also for other languages. A lot of discussions are still needed, including 1) definition of what is a “word”; 2) definition concrete evaluation schemes for word segmentation methods, including extrinsic evaluation. And for Vietnamese, a lot of works also have to be done to improve the performance of word segmentation, including the use of other available resources and/or the development of resources, and the combination of different methods for word segmentation.

Reference

Cao, X H. 1985. Ve cuong vi ngon ngu hoc cua Tieng. *Ngon Ngu* 1. 25-53.

Cao, X H. 1990. Some Preliminaries to the Syntactic Analysis of the Vietnamese Sentence. *In Proceeding of the Prague Congress of 1990.*

Constantine P. Papageorgiou. 1994. Japanese word segmentation by hidden markov model. *In Proceedings of the Human Language Technology Workshop.* 283-288.

Doan, T. T. Dong gop vao viec gioi dinh tu da tiet bang tieu chi trong am trong tieng Viet. 1965. Thong bao khoa hoc 2. 124-125.

Hoang, V. H. Ve hien tuong lay trong tieng Viet. *Ngon ngu* 2. 5-15.

Li Hai-Zhou and Yuan Bao-Sheng 1998. Chinese word segmentation. In *Proceedings of the 12th Pacific Asia Conference on Language, Information and Computation, PACLIC-12*, 1998. 212-217.

Maosong Sun, Dayang Shen, and Benjamin K. Tsou. 1998. Chinese word segmentation without using lexicon and hand-crafted training data. *In Proceeding. of COLING-ACL 98.* 1265-1271.

Meknavin, S., Charoenpornsawat,P. and Kijsirikul, B. 1997. Feature-based Thai Word Segmentation. *In Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'97).* 41-46.

Palmer, David. 1997. A Trainable Rule-based Algorithm for Word Segmentation. *In Proceedings of ACL.* Madrid. 321-328.

Sornlertlamvanich, V., T. Potipiti and T. Charoenporn. 2000. *Automatic Corpus-Based Thai Word Extraction with the C*4.5 Learning Algorithm. In *Proceedings of COLING 2000.*

Sproat, R., Shih, C., Gale, W. and Chang, N. 1996. A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics*, 22 (3), 377-404.