A Maximum Entropy Approach to Sentence Boundary Detection of Vietnamese Texts

Hong Phuong Le, Tuong Vinh Ho

A Maximum Entropy Approach to Sentence Boundary Detection of Vietnamese Texts

Lê Hồng Phương
LORIA, Nancy, France
Email: lehong@loria.fr

Hồ Tường Vinh
IFI, Hanoi, Vietnam
Email: ho.tuong.vinh@auf.org

***Abstract*—We present for the first time a sentence boundary detection system for identifying sentence boundaries in Vietnamese texts. The system is based on a maximum entropy model. The training procedure requires no hand-crafted rules, lexicon, or domain-specific information. Given a corpus annotated with sentence boundaries, the model learns to classify each occurrence of potential end-of-sentence punctuations as either a valid or invalid sentence boundary. Performance of the system on a Vietnamese corpus achieved a good recall ratio of about 95%. The approach has been implemented to create a software tool named `vnSentDetector`, a plug-in of the open source software framework `vnToolkit` which is intended to be a general framework integrating useful tools for processing of Vietnamese texts.**

I. Introduction

Sentence is a fundamental and relatively well understood unit in theoretical and computational linguistics. Many linguistic phenomena such as collocations, idioms, and variable binding, to name a few are constrained by the abstract concept "sentence" in that they are confined by sentence boundaries. The successful determination of these boundaries is thus a prerequisite for proper sentence processing.

Sentence boundary detection is not a trivial task, though. On first glance, it may appear that using a short list of sentence-final punctuation marks, such as ".", "?", and "!", is sufficient. However, graphemes often serve more than one purpose in writing systems. The punctuation marks are not used exclusively to mark sentence breaks. For example, embedded quotations may contain any of the sentence ending punctuation marks and a period is also used to mark abbreviations, initials, ordinal numbers, decimal points, dates, email, Internet addresses and ellipses. Moreover, punctuation marks may be used to mark an abbreviation and a sentence boundary at the same time, or they may be used multiple times for emphasis to mark a single sentence boundary. Sentence boundary detection thus could be considered as an instance of ambiguity resolution.

Recently, there has been a throughout investigation of systems for sentence boundary detection in the journal *Computational Linguistics* which gives a detailed description of methods and systems for identification of sentence boundaries for occidental languages [9]. In general, methods used in sentence boundary detection systems can be roughly classified into three approaches, namely rule-based, supervised machine-learning and unsupervised machine-learning approaches. The state-of-the-art results can be achieved by systems which combine these approaches.

In this article, we present for the first time a sentence boundary detection system for Vietnamese texts that makes use of a maximum entropy model. The model is trained on a corpus of $4,800$ sentences which are manually segmented by linguists at the Vietnam Lexicography Center (Vietlex[1]). The model is then used to identify sentence boundaries of Vietnamese texts with a high accuracy (about 95%).

The rest of this article is organized as follows: Section II discusses the motivation of our work. Section III introduces the maximum entropy model; Section IV describes the selection of features for Vietnamese sentence boundary detection; Section V reports the results and analysis of the experiment; finally Section VI presents the conclusion.

II. Motivation

Like other alphabetic script languages, Vietnamese sentence identification is an important task which has a significant effect on higher levels of Vietnamese text processing tasks. Although, to our knowledge, there does not exist any published work about identifying sentence boundaries for Vietnamese texts. In practice, Vietnamese linguists are segmenting sentences by simply using appearance of sentence-final marks and then manually correct the wrong segmentations which are relatively numerous. Several papers and tools have recently discussed results on Vietnamese text processing tasks required their input to be divided into sentences, but make no mention of how to accomplish this [6], [14]. Some Vietnamese text segmentation tools, like `vnTokenizer` [15] which dissects a Vietnamese text into lexical units by processing the entire text without segmenting it into sentences. This may reduce the time efficiency and the accuracy of the system. We believe that the performance of the system can be improved further if its input are highly accurate pre-segmented sentences. These reasons are the first motivation of our work.

Another reason that motivates us to develop a sentence detection system which makes use of log-linear models is the successful application of these models to a number of problems in natural language processing (NLP). *Log-linear models* are also referred to as *maximum entropy models* and *random fields* in the NLP literature. These models have been

[1] *http://www.vietlex.com/*

seen as a successful supervised machine-learning approach to linguistic classification problems, in which contexts are used to predict linguistic classes. They offer a clean way to combine diverse pieces of contextual evidence in order to estimate the probability of a certain linguistic class occurring with a certain linguistic context [1], [16]. Log-linear models have been adopted in NLP for English, including machine translation [1], statistical parsing [21], [12], language modeling [20], text classification [13], part-of-speech tagging [17], named entity recognition [2], [8], and sentence boundaries identification [19]. Recently, log-linear models have been also successfully applied to use with languages other than English, for example for Chinese NLP tasks [10], [11], [22] and [26]; for Japanese NLP tasks [23], [24]. The application of this approach has achieved state-of-the-art results. Therefore, we believe that this approach may achieve similar successes in a wide range of NLP tasks for Vietnamese, a language whose linguistic mechanism is somehow close to that of the Chinese language.

Despite the fact that Vietnamese is an alphabetic script and the sentence-final punctuation marks are the same as those of occidental languages, the simple application of a model developed for occidental languages does not give the best result for Vietnamese because of a number of reasons. Firstly, the use of sentence boundaries in a language depends on the language nature in general and on its grammatical context in particular. Secondly, as stated in [9] in sentence boundary detection systems, the determination of abbreviation types already yields a large percentage of the overall accuracy of these systems because all periods occurring after non-abbreviation types can be classified as end-of-sentence markers. Such a disambiguation on the type level, however, is insufficient by itself because it still has to be determined for every period following an abbreviation whether it serves as a sentence boundary marker at the same time. The detection of initials and of ordinal numbers, which are represented by digits followed by a period in several languages, also requires the application of token-based methods because these subclasses of abbreviations are problematic for type-based methods. The abbreviation list of the Vietnamese language is obviously different from that of other languages. Finally, one of the most important facts that may have a significant influence on the good result of a log-linear model is the careful selection of useful contextual features for use in the model. As shown in the subsequent sections, we have selected 13 useful features for our sentence boundary detection system in comparison with seven features of the model used for English.

We present in the next section the probability model of the maximum entropy framework.

III. The Probability Model

Many problems in NLP can be viewed as tasks of classification in which the task is to estimate the probability of class b occurring with context c, or $p(b, c)$. Contexts in NLP tasks usually include words, and the exact context depends on the nature of the task; for some tasks, the context c may consist of just a single word, while for others, c may consist of several words and their syntactic labels. The training data usually contain some information about the cooccurence of b's and c's but never enough to completely specify $p(b, c)$ for all possible $p(b, c)$ pairs, since the words in c are typical sparse [16]. Maximum entropy is a method for using the sparse evidence about the b's and c's to reliably estimate a probability model $p(b, c)$.

The model used here for sentence boundary detection is based on the maximum entropy model used for part-of-speech tagging of English text which is first introduced by Ratnaparkhi [17]. For each potential sentence boundary token (".", "?", and "!"), we estimate a joint probability of the token and its surrounding context, both of which are denoted by c, occurring as an actual sentence boundary. The model's probability is defined as:

$$p(b, c) = \pi \prod_{j=1}^{k} \alpha_j^{f_j(b,c)}, \tag{1}$$

where $b \in \{\texttt{no}, \texttt{yes}\}$, where the α_j are the unknown parameters of the model, and where each α_j corresponds to a f_j – a feature of the model, π is the normalization constant. Let $\mathcal{B} = \{\texttt{no}, \texttt{yes}\}$ the set of possible classes, and $\mathcal{C}$ the set of possible contexts. Features are binary-valued functions on events: $f_j = \mathcal{B} \times \mathcal{C} \rightarrow \{0, 1\}$ which are used to encode contextual information that is of interest. The probability of seeing an actual sentence boundary in the context c is given by $p(\texttt{yes}, c)$. The parameters α_i are chosen to maximize the likelihood of the training data using the *Generalized Iterative Scaling* algorithm [4] or the *Improved Iterative Scaling* algorithm [5].

The model also can be viewed under the maximum entropy framework, in which we choose a probability distribution p that maximizes the entropy $H(p)$:

$$H(p) = -\sum_{\mathcal{B} \times \mathcal{C}} p(b, c) \log p(b, c) \tag{2}$$

under the k constraints of the model's expectations of $f_j, \forall 1 \leq j \leq k$:

$$\sum_{\mathcal{B} \times \mathcal{C}} p(b, c) f_j(b, c) = \sum_{\mathcal{B} \times \mathcal{C}} \widehat{p}(b, c) f_j(b, c), \tag{3}$$

where $\widehat{p}(b, c)$ is the observed distribution of sentence boundaries and contexts in the training data. As a result, the model in practice tends not to commit towards a particular outcome (`yes` or `no`) unless it has seen sufficient evidence for that outcome; that is, the model satisfies all the evidence but it is as uniform as possible.

We use a simple decision rule to classify each potential sentence boundary: a potential sentence boundary is an actual sentence boundary if and only if $p(\texttt{yes}\,|c) > 0.5$, where

$$p(\texttt{yes}\,|c) = \frac{p(\texttt{yes}, c)}{p(c)} = \frac{p(\texttt{yes}, c)}{p(\texttt{yes}, c) + p(\texttt{no}, c)}, \tag{4}$$

and where c is the context including the potential sentence boundary.

We present in the next section the selection of features f_j to use in the above-mentioned model to detect sentences boundaries in Vietnamese texts.

IV. Features Selection

The features used in the maximum entropy model should encode any information that might be helpful for predicating sentence boundaries. If the feature exists in the feature set, its corresponding coefficient will contribute towards the conditional probability $p(b|c)$.

Potential sentence boundaries are identified by scanning the text for sequences of characters separated by white space (or tokens) containing one of the symbol ".", "?", or "!" (without quotes). We use information about the token containing the potential sentence boundary, as well as contextual information about the tokens immediately to the left and to the right.

We call the token containing a potential end-of-sentence (eos) character the `candidate`. The portion of the `candidate` preceding the potential eos character is called the `prefix`, and the portion following it is called the `suffix`. We also take into account the `previous` token, that is the space delimited token preceding the `candidate`, and the `next` token, the space delimited token following the `candidate`. The positions (offsets) of the eos characters in training sentences are also encoded. Furthermore, it is noted that in Vietnamese writing, the eos characters "?" and "!" might be preceded by a space character or not. We thus need to add features about space characters surrounding a candidate to improve the performance of the system. The system that focused on maximizing performance used the following list of features:

1) Whether there is a space preceding the eos candidate character
2) Whether there is a space following the eos candidate character
3) The eos character
4) The prefix feature
5) The length of the prefix if it is greater than zero
6) Whether the first character of the prefix is a uppercase letter
7) Whether the prefix is in a list of induced abbreviations
8) The suffix feature
9) The previous feature
10) Whether the first character of the previous is a uppercase letter
11) Whether the previous is in a list of induced abbreviations
12) The next feature
13) Whether the candidate is in capital

For example, some contextual information used by the maximum entropy model for the potential eos character marked by "." in *10.000USD* in a sentence extracted from a corpus in preparation for the construction of a Vietnamese treebank:

Những hacker máy tính sẽ có cơ hội chiếm giải thưởng trị giá 10.000USD và 10.000 đôla Singapore (5.882USD) trong một cuộc tranh tài quốc tế mang tên "Hackers Zone" được tổ chức vào ngày 13.5.1999 tại Singapore.

Table I
Performance of the system

R	P	F
94.96	92.48	93.70

would be `previous`=*giá*, `next`=*và*, `prefix`=*10*, `suffix`=*000USD*.

The abbreviation list is automatically produced from the training data, and the contextual information is also automatically generated by scanning the training data with the above question features. By analyzing the training results, we found a considerable amount of abbreviations which are commonly used, especially in Vietnamese newspapers, for instance GS. (Professor), TT (Prime Minister), UBND (People Committee)...

V. Evaluation

A. Corpus Constitution

Although up to now there have existed debates going on about several linguistic phenomena among Vietnamese linguists community, almost all Vietnamese linguists accord with the definition of a Vietnamese sentence, in that, it is the minimal linguistic unit of speech characterized by its ability to make a statement about a fact, an opinion or a sentiment and by its independence in a text [3], [7] and [25]. This definition of sentence is not only concise but also general enough and semantically agree on that of other languages. Therefore, despite of complex classifications of Vietnamese sentences, by following this guideline, the determination of what constitutes a sentence in a Vietnamese text is not a burdened work for linguists.

The corpus upon which we evaluate the performance of the system is a collection of $4,800$ Vietnamese sentences (about $113,000$ syllables, ≈ 23.54 syllables per sentence) which are manually segmented by linguists at the Vietnam Lexicography Center. About 80% of the corpus belongs to the domain of Vietnamese history texts, the rest is taken from a corpus of many domains which is a part of an ongoing Vietnam national project whose aim is to build a treebank for the Vietnamese language.

B. System Performance

We perform a 10-fold cross validation on the test corpus. In each experiment, we take 90% of the gold test set as the training set ($4,320$ sentences), and 10% as test set (480 sentences). Table I reports the average values of precisions, recalls and F-measures of the system. Recall is computed as the count of common sentences over sentences of the manually segmented files, precision as the count of common sentences over sentences of the automatically segmented files, and F-measure, the weighted harmonic mean of precision and recall, is computed as usual, $F = \frac{2RP}{R+P}$. As can seen, the average recall ratio of the system is relatively good, approximately 95%.

Table II
PERFORMANCE AS A FUNCTION OF TRAINING SET SIZE

Size	R	P	F
$1,000$	92.25	91.12	91.68
$2,000$	92.67	91.91	92.29
$3,000$	94.75	92.65	93.69
$4,000$	94.78	92.55	93.65
$4,320$	94.96	92.48	93.70

In order to evaluate the effect of training set size to the results, we also experimented the model with the quantity of training data required to maintain performance. Table II shows the dependence of system's performance on the training set size measured in sentences. It can be seen that performance degrades as the quantity of training data decreases, but even with only $1,000$ example sentences, the performance is much better than the baseline of about 60% if a sentence boundary is simply guessed at every potential end-of-sentence punctuation.

It can be noted that the overall accuracy of the system only slightly improves starting from the training set corpus of size $3,000$ sentences. We thus think that this is a good starting size for training the proposed model on this genre of Vietnamese text.

C. Software Tools

We have developed a software tool named `vnSentDetector` that implements the presented approach for automatic sentence detection of Vietnamese texts. We facilitate the development of this tool by reusing the `opennlp.maxent` package, a mature open source package for training and using maximum entropy models[2].

The tool is written in Java and is bundled as an Eclipse plug-in. It has already been integrated into `vnToolkit`, an Eclipse Rich Client[3] application which is intended to be a general framework integrating tools for processing of Vietnamese text. `vnSentDetector` plug-in, `vnTokenizer` plug-in [15], `vnToolkit` framework and related resources, including the test corpus are freely available for download[4]. These software tools are distributed under the GNU General Public License[5].

VI. CONCLUSION

We have presented in this paper a system for sentence boundary detection of Vietnamese text which is based on a maximum entropy approach. The approach is implemented to create an open source software tool and the experimental results show that the approach can achieve a high accuracy (95%). Because of the simplicity of the maximum entropy model when retraining for new text domains, it is easy to re-train the presented model on any genre of Vietnamese text. It is also believed that the model can improve further the accuracy if it is retrained on a larger corpus. It is worth mentioning that the state-of-the-art sentence boundary detection system for English performs at the accuracy of 98% when be trained on a large corpus of about $40,000$ sentences.

Together with the adaptation of a maximum entropy model for segmentation of Vietnamese text that has been presented in [6], it is proved that the maximum entropy principle has a good performance for Vietnamese linguistic modeling if task-specific features are well exploited and combined.

[2] *http://maxent.sourceforge.net/*
[3] *http://www.eclipse.org/rcp/*
[4] *http://www.loria.fr/∼lehong/projects.php*
[5] *http://www.gnu.org/copyleft/gpl.html*

ACKNOWLEDGMENT

The authors would like to thank all the linguists at the Vietnam Lexicography Center for their enthusiastic collaboration in data preparation, especially thank to Mr. Vũ Xuân Lương for providing us a ready-to-use corpus and for suggesting helpful comments to improve the tools.

REFERENCES

[1] Adam L. Berger, Stephen A. Della Pietra, Vincent J. Della Pietra, A Maximum Entropy Approach to Natural Language Processing, *Computational Linguistics*, Vol. 22, No. 1, Pages 39-71, 1996.
[2] A. Borthwick, A Maximum Entropy Approach to Named Entity Recognition, *Ph.D. Thesis*, New York University, 1999.
[3] Nguyễn Tài Cẩn, *Ngữ pháp Tiếng Việt*, NXB Đại học Quốc gia Hà Nội, Hà Nội, 1998.
[4] J. N. Darroch and D. Ratcliff, Generalized Iterative Scaling for Log-Linear Models, *The Annals of Mathematical Statistics*, Vol. 43, No. 5, Pages 1470-1480, 1972.
[5] S. Della Pietra, V. Della Pietra, and J. Lafferty, Inducing features of random fields, *IEEE Transactions on pattern analysis and machine intelligence*, 19(4), 380-393, April, 1997.
[6] D. Dien, V. Thuy, A maximum entropy approach for Vietnamese word segmentation , *In Proceedings of IEEE International Conference on Research, Innovation and Vision for the Future RIVF 2006*, Vietnam, 2006.
[7] C. X. Hạo, *Tiếng Việt - Sơ thảo Ngữ pháp Chức năng*, NXB Giáo dục, 2004.
[8] H. Chieu and H. Ng, Named Entity Recognition with a Maximum Entropy Approach, *In Proceedings of CoNLL-2003*, Edmonton, Canada, 2003.
[9] T. Kiss and J. Strunk, Unsupervised Multilingual Sentence Boundary Detection, *Computational Linguistics*, Vol. 32, No. 4, Pages 485-525, 2006.
[10] R. Li, X. Tao, L. Tang and Y. Hu, Using Maximum Entropy Model for Chinese Text Categorization, *LNCS - Advanced Web Technologies and Applications*, Pages 578-587, Springer, 2004.
[11] X. Luo, A Maximum Entropy Chinese Character-based Parser, *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, Vol. 10, Pages 192-199, USA, 2003.
[12] Miyao Yusuke, J. Tsujii, A Model of Syntactic Disambiguation based on Lexicalized Grammars, *Proceedings of the Seventh Conference on Natural Language Learning*, Edmonton, Canada, 2003.
[13] K. Nigam, J. Lafferty and A. McCallum, Using Maximum Entropy for Text Classification, *In IJCAI-99 Workshop on Machine Learning for Information Filtering*, Pages 61–67, 1999.
[14] C. T Nguyen, T. K. Nguyen, X. H. Phan, L. M. Nguyen and H. Q. Thuy, Vietnamese Word Segmentation with CRFs and SVMs: An Investigation, *Proceedings of PACLIC*, 2006.
[15] L. H. Phuong, N. T. M. Huyen, R. Azim, H. T. Vinh, A hybrid approach to word segmentation of Vietnamese texts, *Proceedings of the 2nd International Conference on Language and Automata Theory and Applications*, Tarragona, Spain, 2008.
[16] A. Ratnaparkhi, A Simple Introduction to Maximum Entropy Models for Natural Language Processing, *IRCS Report 97-98*, University of Pennsylvania, PA, USA, 1997.
[17] A. Ratnaparkhi, A Maximum Entropy Model for Part-of-Speech Tagging, *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, University of Pennsylvania, PA, USA, 1996.
[18] A. Ratnaparkhi, Learning to Parse Natural Language with Maximum Entropy Models, *Machine Learning*, Vol. 34, Pages 151-175, 1999.

[19] J. Reynar and A. Ratnaparkhi, A Maximum Entropy Approach to Identifying Sentence Boundaries, In Proceedings of the Fifth Conference on Applied Natural Language Processing, Pages 16-19, Washington D.C. 1997.

[20] R. Rosenfeld, A Maximum Entropy Approach to Adaptive Statistical Language Modeling, *Computer, Speech and Language* Vol. 10, 187-228, 1996

[21] Stephen Clark, James R. Curran, Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Models, *Computational Linguistics*, Vol. 33, No. 4, Pages 493-552, 2007.

[22] N. H. Tou, L. J. Kiat, G. Wenyua, A Maximum Entropy Approach to Chinese Word Segmentation, *In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Pages. 161-164, Korea, 2005.

[23] K. Uchimoto, S. Sekine and H. Isahara, Japanese Dependency Structure Analysis Based on Maximum Entropy Models, *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, Pages 196-203, Norway, 1999.

[24] K. Uchimoto, S. Sekine and H. Isahara, The Unknown Word Problem: A Morphological Analysis of Japanese Using Maximum Entropy Aided by Dictionary, *The 2001 Conference on Empirical Methods in Natural Language Processing*, Pittsburgh, USA, 2001.

[25] Uỷ ban Khoa học Xã hội Việt Nam, *Ngữ pháp Tiếng Việt*, Hà Nội, 1983.

[26] J. Zhao, X. L. Wang, Chinese POS Tagging Based on Maximum Entropy Model, *In Proceedings of the First International Conference on Machine Learning and Cybernetics*, Beijing, 2002.