An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts

Phuong Le-Hong, Azim Roussanaly, Thi Minh Huyen Nguyen, Mathias Rossignol

An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts

Phuong Le-Hong[1, 2] Azim Roussanaly[1]
Thi Minh Huyen Nguyen[2] Mathias Rossignol[2]
(1) LORIA, Nancy, France
(2) Hanoi University of Science, Hanoi, Vietnam
lehong@loria.fr, azim@loria.fr,
huyenntm@vnu.edu.vn, mathias.rossignol@gmail.com

Résumé. Nous présentons dans cet article une étude empirique de l'application de l'approche de l'entropie maximale pour l'étiquetage syntaxique de textes vietnamiens. Le vietnamien est une langue qui possède des caractéristiques spéciales qui la distinguent largement des langues occidentales. Notre meilleur étiqueteur explore et inclut des connaissances utiles qui, en terme de performance pour l'étiquetage de textes vietnamiens, fournit un taux de précision globale de 93.40% et de 80.69% pour les mots inconnus sur un ensemble de test du corpus arboré vietnamien. Notre étiqueteur est nettement supérieur à celui qui est en train d'être utilisé pour développer le corpus arboré vietnamien, et à l'heure actuelle c'est le meilleur résultat obtenu pour l'étiquetage de textes vietnamiens.

Abstract. This paper presents an empirical study on the application of the maximum entropy approach for part-of-speech tagging of Vietnamese text, a language with special characteristics which largely distinguish it from occidental languages. Our best tagger explores and includes useful knowledge sources for tagging Vietnamese text and gives a 93.40% overall accuracy and a 80.69% unknown word accuracy on a test set of the Vietnamese treebank. Our tagger significantly outperforms the tagger that is being used for building the Vietnamese treebank, and as far as we are aware, this is the best tagging result ever published for the Vietnamese language.

Mots-clés : Etiqueteur syntaxique, entropie maximale, texte, vietnamien.

Keywords: Part-of-speech tagger, maximum entropy, text, Vietnamese.

1 Introduction

After decades of research in natural language processing (NLP) mostly concentrated on English and other well-studied languages, recent years have seen an increased interest in less common languages, notably because of their growing presence on the Internet. Vietnamese, which belongs to the top 20 most spoken languages and is employed by an important community all over the world, is one of those new focuses of interest. Obstacles remain, however, for NLP research in general and part-of-speech (POS) tagging in particular : Vietnamese does not yet have vast and readily available manually constructed linguistic resources upon which to build effective statistical models, nor acknowledged reference works against which

new ideas may be experimented.

Moreover, most existing research so far has been focused on testing the applicability of existing methods and tools developed for English or other Western languages, under the assumption that their logical or statistical well-foundedness guarantees cross-language validity, when in fact assumptions about the structure of a language are always made in such tools, and must be amended to adapt them to different linguistic phenomena. For an isolating language such as Vietnamese, techniques developed for flexional languages cannot be applied "as is". It is with this idea in mind that we have devised the Vietnamese POS tagger presented in this paper.

The primary motivation for the development of an efficient tagger for the Vietnamese language is the need of a fast and highly accurate tagger which may contribute more effectively in the development of basic linguistic resources and tools for automatic processing of Vietnamese text (Nghiem *et al.*, 2008; Tran *et al.*, 2009). Recently, a group of Vietnamese computational linguists has been involved in developing a treebank for Vietnamese (Nguyen *et al.*, 2009). Among tools for annotation of treebanks, a POS tagger is an essential tool that must be robust and minimize errors. As reported in (Nguyen *et al.*, 2009), the tagger that is currently being used in the development of the treebank has the F_1 score of 90.5%, a not very high accuracy, far from the state of the art on Western languages. One of the reasons for this modest performance is that the tagger employs only generic information sources used for tagging and does not take into account some specific characteristics of the Vietnamese language.

We present in this article a more accurate and efficient tagger for Vietnamese. Based on the classical principle of maximum entropy, it however explores and includes new useful knowledge sources for tagging Vietnamese text and achieves state-of-the-art result for this language. It yields a 93.40% overall accuracy and a 80.69% unknown word accuracy on a test set of the Vietnamese treebank.

Section 2 gives an introduction to the specific properties of Vietnamese and presents the tagset used in our POS tagging task. Section 3 briefly presents the classical maximum entropy model which has been widely used in POS tagging. The main contribution of the paper is presented in Section 4, where experimental results are given and the best performing system is described in detail. We conclude this paper in Section 5 with a discussion of further possible improvements.

2 Vietnamese tagset

In this section, we first briefly introduce the Vietnamese language, bringing into focus the specific characteristics which make automatic processing tasks difficult for this language. We then introduce a tagset currently being used in an ongoing project involving the development of a treebank for the Vietnamese language, which we have also adopted in our work.

2.1 Vietnamese language

Vietnamese belongs to the Austro-Asiatic language family. It is the national language of Vietnam. The majority of the speakers of Vietnamese are spread over the South East Asia area.

Vietnamese is a quite fixed order language, with general word order SVO (subject-verb-object). As for most languages which have relatively restrictive word orders, Vietnamese relies on the order of constituents to

No.	Category	Description	No.	Category	Description
1.	Np	Proper noun	10.	M	Numeral
2.	Nc	Classifier	11.	E	Preposition
3.	Nu	Unit noun	12.	C	Subordinating conjunction
4.	N	Common noun	13.	CC	Coordinating conjunction
5.	V	Verb	14.	I	Interjection
6.	A	Adjective	15.	T	Auxiliary, modal words
7.	P	Pronoun	16.	Y	Abbreviation
8.	R	Adverb	17.	Z	Bound morpheme
9.	L	Determiner	18.	X	Unknown

TAB. 1 – Vietnamese tagset

convey important grammatical information. Although Vietnamese text is written in a variant of the Latin alphabet (a script that exists in its current form since the 17th century, and has become the official writing system since the beginning of the 20th), there are three principal characteristics of Vietnamese which distinguish it from occidental languages.

First, Vietnamese is an inflectionless language in which its word forms never change. Since there is no inflection in Vietnamese, all the grammatical information is conveyed through word order and tool words. The inflectionless characteristic makes a special linguistic phenomenon common in Vietnamese : type mutation, where a given word form is used in a capacity that is not its typical one (a verb used as a noun, a noun as an adjective. . .) without any morphological change. This leads to the fact that Vietnamese word forms are usually highly ambiguous in their part-of-speech. The same word may be a noun in a context while it may be a verb or a preposition in other contexts. For example, the word *yêu* may be a noun (*the devil*) or a verb (*to love*) depending on context.

Second, Vietnamese is an isolating language, the most extreme case of an analytic language, in which each morpheme is a single, isolated syllable. Lexical units may be formed of one or several syllables, always remaining separate in writing. Although dictionaries contain a majority of compound words, monosyllabic words actually account for a wide majority of word occurrences. This is in contrast to synthetic languages, like most Western ones, where, although compound words exist, most words are composed of one or several morphemes assembled so as to form a single token.

The third difference, a consequence of the latter, is that there is no word delimiter in Vietnamese : spaces are used to separate syllables, and no specific marker distinguishes the spaces between actual words. For example, a simple sequence of three syllables *a b c* can constitute three words *(a) (b) (c)*, two words *(a b) (c)*, two words *(a) (b c)* or even a single one *(a b c)*. This characteristic leads to many ambiguities in word segmentation, the foremost basic processing task which influences part-of-speech tagging and higher levels of annotation for Vietnamese text. As an example, the sequence of three syllables *học sinh học*[1] may constitute the following word segmentations *(học) (sinh) (học)*, *(học sinh) (học)*, or *(học) (sinh học)*. This constitutes, however, a distinct task from POS tagging *per se*, and we shall therefore in this paper work on a corpus that has undergone automatic segmentation (several tools for this task are presented in (Nguyen *et al.*, 2008; Le-Hong *et al.*, 2008)) with manual correction.

[1]This phrase can be roughly translated as *the pupil learns* or *study biology* depending on context.

2.2 Vietnamese tagset

Because of its inflectionless nature, Vietnamese does not have morphological aspects such as gender, number, case... such as in occidental languages. Vietnamese words are classified based on their combination ability, their syntactic functions and their general meaning. The complete tagset which was designed for use in the Vietnamese treebank (Nguyen *et al.*, 2009) is shown in Table 1.

Beyond the classical POS used in Western languages (noun, verb, ...), one may notice the presence of classifiers, which are commonly found in Asian languages, and of modal words, which convey some of the nuances borne by flection in synthetic languages. "Bound morphemes" designate syllables that are not supposed to appear alone and should only be seen as part of a compound word, and this tag is normally only ever used to deal with cases when the segmentation of the corpus has been done improperly—which is not our concern here.

3 Maximum Entropy model

The principle of Maximum Entropy states that given a set of observations, the most likely underlying probability distribution is that which has minimal bias—that is, maximal entropy—while verifying the statistical properties measured on the observation set. Those enforced properties are referred to as the constraints imposed on the distribution.

This principle provides a guide for the acquisition of conditional probability distributions from training data, but since it is quite computationally taxing, algorithms applying it have only become widely used in recent years. Maximum entropy models have been applied successfully in many problems in NLP with state-of-the-art results. Maximum entropy models are also known as log-linear models in NLP literature. These models have been 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 (Berger *et al.*, 1996).

Maximum entropy models have been used extensively in numerous tasks of NLP for English and French including machine translation (Deyi *et al.*, 2006; Chiang, 2005), statistical parsing (Clark & Curran, 2007; Finkel *et al.*, 2008; Tsuruoka *et al.*, 2009), part-of-speech tagging (Ratnaparkhi, 1996; Toutanova & Manning, 2000; Toutanova *et al.*, 2003; Denis & Sagot, 2009).

In part-of-speech tagging, the maximum entropy model tagger learns a log-linear conditional probability model from tagged text using a maximum entropy method. One of its most appreciable strengths compared with other methods is that it potentially allows a word to be tagged with a POS which has never been observed in the training corpus.

Given a word w and its context h, the model assigns a probability for every tag t in the set T of possible tags of w. The context h is usually defined as the sequence of several words and tags surrounding the word. The probability of a tag sequence $t_1 t_2 \dots t_n$ given a sentence $w_1 w_2 \dots w_n$ can be estimated as

$$p(t_1 \dots t_n | w_1 \dots w_n) \approx \prod_{i=1}^{n} p(t_i | h_i). \quad (1)$$

The tagging is the process of assigning the maximum likelihood tag sequence to a sequence of words. In the maximum entropy modeling, the distribution p is chosen so that it has the highest entropy out of those distributions that satisify a set of constraints. These constraints restrict the model to match a set of empirical statistics collected from the training data. The statistics are expressed as the expected values of feature functions defined on the context h and tag t.

In the maximum entropy framework it is possible to define and incorporate complex statistics, not restricted to n-gram statistics as in a Markov model. We can easily define and incorporate for example statistics about the shape of a certain word (its prefix or suffix, whether it contains a number or not, whether it is capitalized or not...).

The constraints of the model are that the expectations of these feature functions according to the joint distribution p are equal to the emprical expectations of the feature functions in the training data distribution $\widehat{p}$:

$$\mathbb{E}_p f_i(h,t) = \mathbb{E}_{\widehat{p}} f_i(h,t), \forall i = 1,\ldots,k. \tag{2}$$

In order to make the model tractable, the joint distribution of contexts and tags $p(h,t)$ is usually approximated as follows

$$p(h,t) \approx \widehat{p}(h) \cdot p(t|h).$$

Then the constraints (2) could be rewritten as

$$\sum_{h\in\mathcal{H},t\in\mathcal{T}} \widehat{p}(h)p(t|h)f_i(h,t) = \sum_{h\in\mathcal{H},t\in\mathcal{T}} \widehat{p}(h,t)f_i(h,t),$$

$\forall i = 1,\ldots,k$, where $\mathcal{H}$ is the space of possible contexts h when predicting a part-of-speech tag t.

The model that is the solution to this constrained optimization task is a log-linear model with the parametric form :

$$p(t|h;\lambda) = \frac{1}{Z(h,\lambda)} \exp \sum_{i=1}^{k} \lambda_i f_i(h,t), \tag{3}$$

where $Z(h)$ is the normalizing term :

$$Z(h,\lambda) = \sum_{h\in\mathcal{H}} \exp \sum_{i=1}^{k} \lambda_i f_i(h,t).$$

There exists efficient algorithms for estimating the parameter $\lambda = (\lambda_1,\ldots,\lambda_k)$ of the model. Some of widely used methods are Improve Iterative Scaling (Lafferty *et al.*, 2001) and quasi-Newton unconstrained optimization methods like L-BFGS and conjugate gradient. Readers interested in a more extensive discussion of parameter estimation methods for maxium entropy models may refer to (Gao *et al.*, 2007; Malouf, 2002).

4 Experiments

This section presents experiments and results of a number of tagging models which are acquired using a training method based on the maximum entropy model presented in the previous section. The difference between them is in the use of different feature sets. All the models are trained and tested on the same training and test corpus.

4.1 Corpus constitution

The models are trained and tested on the part-of-speech tagged section of the Vietnamese treebank. The treebank is currently composed of $10,165$ sentences which are manually segmented, tagged and parsed. The raw texts of the treebank are collected from the social and political sections of the Youth online daily newspaper (Nguyen *et al.*, 2009). The minimal and maximal sentence lengths are 2 words and 105 words respectively.

Evaluation of the trained models is performed classically using 10-fold cross-validation, where the corpus is randomly split into 10 parts which are each used for testing a model built using data from the 9 others, after which the computed accuracies for each run are averaged.

4.2 Baseline model

We first develop a baseline model similar to a simple trigram conditional Markov model in that the context available for predicting the part-of-speech of a word w_i in a sentence of words $w_i w_2 \dots w_n$ with tags $t_1 t_2 \dots t_n$ is $\{t_{i-1} t_{i-2} w_i w_{i+1}\}$. The features for this model are automatically produced by instantiation of feature templates which are shown in Table 2. We call this baseline model Model 0.

No.	Template
1.	$w_i = X, t_i = T$
2.	$w_{i-1} = X, t_i = T$
3.	$w_{i+1} = X, t_i = T$
4.	$t_{i-1} = T_1, t_i = T$
5.	$t_{i-1} = T_1, t_{i-2} = T_2, t_i = T$

TAB. 2 – Features used in the baseline model

This baseline model presents an accuracy of 90.23%, which is comparable with the precision of the tagger used by (Nguyen *et al.*, 2009) (90.5%) ; in particular, it performs very poorly on words that have never been seen in the training data ("unknown words"), with a precision of only 47.08% in that case. For that reason, we have first decided to focus on the extension of the feature set to allow a better guess of the POS of unknown words.

4.3 Unknown words features

In order to increase the model's prediction capacity for unknown words, some special feature templates are included in the baseline model and the resulting model is called Model 1a. Based on the idea that unknown words are likely to be rare words, if they have not been observed before, the additional features selected to deal with unknown words are therefore only instanciated during training for the rare words of the training corpus ; we talk henceforth of "rare features". The rare feature templates are shown in Table 3[2]. We define rare words to be words that appear less than six times in the training data (threshold selected empirically).

[2]The actual general and rare feature templates for the model are a subset of the features used in (Ratnaparkhi, 1996; Toutanova & Manning, 2000) for tagging English text.

No.	Template
1.	w_i contains a number, $t_i = T$
2.	w_i contains an uppercase character, $t_i = T$
3.	w_i contains all uppercase characters, $t_i = T$
4.	w_i contains a hyphen, $t_i = T$

TAB. 3 – Features used for rare words

It should be noted that, unlike in the taggers for occidental languages, we do not include prefix and suffix features when predicting rare words, since the concept of affixes is undefined for isolating languages.

Nevertheless, in the mechanics of syllable combination that allow the creation of compound words in Vietnamese, some syllables play roles similar to the affixes of synthetic languages. For example, with *biên* meaning "to write, to compile" and *biên tập* meaning "to edit", the syllable *viên* may be appended to produce *biên tập viên*, meaning "editor". It is used similarly, with the same effect on part of speech, in *nghiên cứu viên* ("researcher") *diễn viên* ("actor"), *etc*. Similarly, the syllable *hóa* is usually appended to the end of a noun or adjective to transform it into a verb, for example the word *công nghiệp* ("industry") can be transformed into the word *công nghiệp hóa* ("to industrialize"), or the word *hiện đại* ("modern") into the word *hiện đại hóa* ("to modernize"). The syllables *viên* and *hóa* in these examples can be considered as a "suffix" of related composed words.

Not all suffixes, however, define the type of the resulting word as in those examples. That is for example the case of *phó* ("vice-") in *phó giám đốc* ("vice-director"), whose POS tag is the same as that of *giám đốc* ("director").

To account for all those cases, we consider as additional feature when trying to find out the tag of unknown compound words their first and last syllables (potential affixes) as well as the groups of the first two and last two syllables (potential "semantic kernel" of the compound word). Model 1a augmented with these additional features, shown formally in Table 4, is referred to as Model 1b.

No.	Template
1.	$\sigma(1, w_i) = X, t_i = T$
2.	$\sigma(m, w_i) = X, t_i = T$
3.	$\sigma(1, w_i) = X_1, \sigma(2, w_i) = X_2, t_i = T$
4.	$\sigma(m, w_i) = X_1, \sigma(m-1, w_i) = X_2, t_i = T$
5.	Number of syllables of w_i

TAB. 4 – Features for syllables of a word. $\sigma(j, w_i)$ is a function which returns the j-th syllable of an m-syllable composed word.

We also found the usefulness of the length of a word measured in syllables when predicting unknown words. Adding template 5 of Table 4 to Model 1b slightly increased the accuracy of the resulting Model 1c. In our opinion, the number of syllables of a word can help predict long unknown words since they are often classified correctly as locutions ; in addition, it is predictable that monosyllabic words are hardly

unknown words.

The results on the test set of the four presented models are shown in Table 5.

Model	**Overall Accuracy**	**Unknown Word Accuracy**
Model 0	90.23%	47.08%
Model 1a	92.64%	68.92%
Model 1b	92.85%	73.23%
Model 1c	92.92%	76.92%

TAB. 5 – Accuracy of the first four models.

The results show the importance of lexical information to the tagger by a large reduction of unknown word error of models 1* compared with Model 0. The results also show the benefit of incorporating syllable-specific features to the baseline model.

4.4 Discussion of problematic cases

There are many words which can have more than one syntactic category. In addition, the syntactic category mutation presented earlier is a frequent phenomenon in Vietnamese. This introduces many ambiguities that the tagger has to resolve. The confusion matrix of Model 1c on the test set is given in Table 6. The row labels indicate the correct tags, and the column labels give the assigned tags. For example, the number 41 in the (N,V) position is the number of common nouns (N) that have incorrectly been tagged as verbs (V). We see that the ambiguities between syntactic pairs (N,A), (N,V) and (A,V) are the harder for the tagger to resolve. This result reflects and illustrates the frequent mutation between these syntactic categories in Vietnamese. These particular confusions make up a large percentage of the total error ($\approx 65\%$).

There are several types of tagger errors. First, some errors are the result of inconsistency in labeling in the training data. Like in most other treebank data, this is inevitable, especially for an initial version of the Vietnamese treebank. For example, the word *ông* is sometimes labeled Nc (classifier), sometimes N (common noun) or P (pronoun). Second, some errors are due to systematic tag ambiguity patterns in Vietnamese, for instance the patterns P/N, A/N, V/N and A/V are ambiguous pairs highly context-dependent and sometimes difficult to disambiguate even by human annotators. Finally, it seems that the

	N	Nc	Np	V	A	R	E	P
N	0	16	9	**41**	**20**	2	6	11
Nc	**24**	0	0	1	0	0	0	0
Np	12	0	0	4	1	0	0	0
V	**45**	0	2	0	**21**	14	12	0
A	**33**	0	0	**29**	0	6	2	1
R	5	0	0	16	3	0	4	1
E	2	0	0	10	0	0	0	0
P	5	0	0	0	0	1	0	0

TAB. 6 – Confusion matrix Model 1c.

tagger has some difficulties when resolving ambiguities between proper nouns, classifiers and common nouns while human annotators do not frequently make this type of errors. With this error classification in mind, the best way at the moment for tagging accuracy improvement appears to come from minimizing errors in the third class of this classification.

In the following subsection, we discuss how we include additional information sources to help disambiguate proper nouns and thus improve the overall accuracy of the tagger.

4.5 Features for proper noun disambiguation

One of the significant source of tagging errors for our model are the Np/N ambiguity. However, in many cases, it is trivial for humans to determine the correct category in that case, using very simple heuristics. In Vietnamese, a proper noun having multiple syllables is written in a consistent form in that the first letter of every syllable is always capitalized, for instance *Nguyễn Tấn Dũng*, *Hà Nội*. To help resolve a Np/N ambiguity, we can add a feature template that looks at all the syllables of the current word and activates if their first characters are uppercase ones. Table 7 shows the result of the new Model 1d incorporating this feature. We see a considerable improvement of the accuracy on unknown words and the overall accuracy.

Model	**Overall Accuracy**	**Unknown Word Accuracy**
Model 1d	93.13%	80.62%

TAB. 7 – Accuracy when adding the syllable capitalization feature.

4.6 Best overall model

It has been shown that the broad use of lexical features including jointly conditioning on multiple consecutive words produces a superior level of tagger performance. Notably, in the case of Vietnamese, since many grammatical nuances borne in Western languages by inflections of the considered word itself are instead indicated by separate tool words surrounding it, we need to consider a wider context in order to extract a comparable amount of information. By combining all the useful features of Model 1d and adding two more word feature templates that look at positions ± 2 of the current word, we obtain Model 2, which constitutes the best model in our experiments. The complete list of feature templates used in the best model is shown in Table 8.

To our knowledge, this model gives the state-of-the-art accuracy of tagging result for Vietnamese text. Table 9 shows the token accuracy of the best model together with sentence accuracy.

It is worth noting that our tagger also seems to outperform the tagger vnQTAG, a hidden Markov model tagger for Vietnamese (Nguyen *et al.*, 2003) whose average result on four test sets is 92.5%. However these results are not directly comparable since the models are trained and tested on different corpora and, more importantly, different tagsets.

No.	Template	No.	Template
	Word and tag contexts		*Syllable contexts*
1.	$w_i = X, t_i = T$	8.	$\sigma(1, w_i) = X, t_i = T$
2.	$w_{i-1} = X, t_i = T$	9.	$\sigma(m, w_i) = X, t_i = T$
3.	$w_{i+1} = X, t_i = T$	10.	$\sigma(1, w_i) = X_1, \sigma(2, w_i) = X_2, t_i = T$
4.	$t_{i-1} = T_1, t_i = T$	11.	$\sigma(m, w_i) = X_1, \sigma(m-1, w_i) = X_2, t_i = T$
5.	$t_{i-1} = T_1, t_{i-2} = T_2, t_i = T$	12.	Number of syllables of w_i
6.	$w_{i-2} = X$	13.	Particular features for proper nouns
7.	$w_{i+2} = X$		

TAB. 8 – Features used in the best model.

Overall Accuracy	Unknown Word Accuracy	Sentence Accuracy
93.40%	80.69%	31.40%

TAB. 9 – Accuracy of the best model.

4.7 Software package

We have developed a software package named `vnTagger` that implements the presented models for tagging Vietnamese text. The software is written in the Java programming language. The development of the software has been greatly facilitated thanks to the open source implementation of a maximum entropy part-of-speech tagger of the Stanford Natural Language Processing Group[3]. This software implements a maximum entropy classifier which uses a conjugate gradient procedure and a Gaussian prior to maximize the data likelihood (Toutanova *et al.*, 2003). Our software is also freely distributed under the GNU/GPL license, available online at our website[4].

It is worth noting that our tagger has been used in a number of national and international research projects, for example in a project which builds very large lexicographic corpora for many languages (Kilgarriff *et al.*, 2009).

5 Conclusion

The work presented in this paper explores the use of specially selected, language-specific sources of information used for tagging Vietnamese text. These particular descriptive features are incorporated into a maximum entropy model for developing a state-of-the-art tagger for the Vietnamese language.

Since Vietnamese is an isolating language, the most extreme case of an analytic language, it is worth noting that our study may suggest methodology for the part-of-speech tagging of other analytic languages in mainland southest Asia.

[3] *http ://nlp.stanford.edu/software/tagger.shtml*
[4] *http ://www.loria.fr/ lehong/tools/vnTagger.php*

It was shown that the use of bidirectional dependency networks with a series of local maximum entropy models produces better result when tagging English text (Toutanova *et al.*, 2003). It is hopeful that an appropriate application of this approach for tagging Vietnamese text may help resolve frequent ambiguities in syntactic category mutation, and thus improve the tagging result.

It was also shown that the integration of an external lexical resource to the training corpus help improve tagging performance (Denis & Sagot, 2009). Obviously, the use of an external dictionary has a potential advantage of better handling unknown words in case words are not present in training corpus but they may be present in the external dictionary. In future work, we plan on integrating a Vietnamese lexicon (Nguyen *et al.*, 2006) into our system to improve further its performance.

Finally, we believe that the accuracy of the tagger can be further improved when it is trained on larger corpora, as the Vietnamese treebank grows.

One weakness of this POS tagger, which it shares with all POS-taggers for Vietnamese, is that it depends on the segmentation accuracy of the input text. However, proper segmentation, cannot be achieved with high accuracy without considering potential POS tags, and future developments of the system will require to integrate those two tasks.

Références

BERGER A., PIETRA S. D. & PIETRA V. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, **22**(1), 39–71.

CHIANG D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the ACL'05*, p. 263–270.

CLARK S. & CURRAN J. R. (2007). Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, **33**(4), 493–552.

DENIS P. & SAGOT B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of PACLIC 2009*.

DEYI X., QUN L. & SHOUXUN L. (2006). Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 44th Annual Meeting of the ACL*, p. 521–528.

FINKEL J. R., KLEEMAN A. & MANNING C. (2008). Efficient, feature-based, conditional random field parsing. In *ACL*, p. 959–967.

GAO J., ANDREW G., JOHNSON M. & TOUTANOVA K. (2007). A comparative study of parameter estimation methods for statistical natural language learning. In *ACL*, p. 824–831.

KILGARRIFF A., REDDY S. & POMIKÁLEK J. (2009). Corpus factory. In *Proceedings of Asialex*, Bangkok, Thailand.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *ICML*, p. 282–289.

LE-HONG P., NGUYEN T. M. H., ROUSSANALY A. & HO T. V. (2008). A hybrid approach to word segmentation of Vietnamese texts. In M.-V. CARLOS, Ed., *Proceedings of the 2nd International Conference on Language and Automata Theory and Applications*. Tarragona, Spain : Springer, LNCS 5196.

MALOUF R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In *CONLL*.

NGHIEM Q. M., DINH D. & NGUYEN T. N. M. (2008). Improving Vietnamese POS-tagging by integrating a rich feature set and support vector machines. In *Proceedings of the 6th IEEE International Conference in Computer Science : Research, Innovation and Vision of the Future, RIVF*, HCMC, Vietnam.

NGUYEN P. T., XUAN L. V., NGUYEN T. M. H., NGUYEN V. H. & LE-HONG P. (2009). Building a large syntactically-annotated corpus of Vietnamese. In *Proceedings of the 3rd Linguistic Annotation Workshop, ACL-IJCNLP*, Singapore.

NGUYEN T. M. H., ROMARY L., ROSSIGNOL M. & VU X. L. (2006). A lexicon for Vietnamese language processing. *Language Resources and Evaluation*, **40**(3–4).

NGUYEN T. M. H., ROSSIGNOL M., LE-HONG P., DINH Q. T., VU X. L. & NGUYEN C. T. (2008). Word segmentation of Vietnamese texts : a comparison of approaches. In *Proceedings of the 6th Language Resources and Evaluation Conference, LREC*, Marrakech, Morroco.

NGUYEN T. M. H., VU X. L. & LE-HONG P. (2003). A case study of the probabilistic tagger QTAG for tagging Vietnamese texts. In *Proceedings of the 1st National Conference ICT RDA*.

RATNAPARKHI A. (1996). A maximum entropy model for part-of-speech tagging. In *EMNLP*, p. 133–142.

TOUTANOVA K., KLEIN D., MANNING C. D. & SINGER Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL*.

TOUTANOVA K. & MANNING C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *EMNLP/VLC*, p. 63–71.

TRAN T. O., LE A. C., HA Q. T. & LE Q. H. (2009). An experimental study on Vietnamese POS tagging. In *Proceedings of the International Conference on Asian Language Processing, IALP*, Singapore.

TSURUOKA Y., TSUJII J. & ANANIADOU S. (2009). Fast full parsing by linear-chain conditional random fields. In *EACL*, p. 790–798, Athens, Greece.