On the Effect of the Label Bias Problem in Part-Of-Speech Tagging

Phuong Le-Hong
University of Science
Vietnam National University, Hanoi
phuonglh@vnu.edu.vn

Xuan-Hieu Phan
University of Engineering and Technology
Vietnam National University, Hanoi
hieupx@vnu.edu.vn

The-Trung Tran
FPT University
FPT Corporation, Vietnam
trung@fpt.edu.vn

***Abstract*—This paper investigates the effect of the label bias problem of maximum entropy Markov models for part-of-speech tagging, a typical sequence prediction task in natural language processing. This problem has been underexploited and underappreciated. The investigation reveals useful information about the entropy of local transition probability distributions of the tagging model which enables us to exploit and quantify the label bias effect of part-of-speech tagging. Experiments on a Vietnamese treebank and on a French treebank show a significant effect of the label bias problem in both of the languages.**

***Index Terms*—label bias problem, machine learning, MEMM, CRF, part-of-speech tagging, Vietnamese, French, treebank**

I. Introduction

In the sequence prediction problem, we attempt to predict elements of a sequence on the basis of the preceding elements. Many statistical sequence models have been developed for sequence prediction, for example hidden Markov models (HMM) [1], [2], maximum entropy Markov models (MEMMs) [3] or conditional random fields (CRFs) [4]. These are all powerful probabilistic tools for modeling sequential data and have been applied to many text-related tasks such as part-of-speech tagging, named entity recognition, text segmentation and information extraction.

These typical models are all finite-state automata with stochastic state transitions and observations. In addition, these models all use the Markov property in that the decisions about the state at a particular position in the sequence can depend only on a small local structure. Without this property, the inference of the models is intractable. Among them, MEMMs are widely used in practice because of their efficiency and accuracy. On the one hand, MEMMs are a discriminative counterpart of HMMs which offers many advantages and usually outperforms HMMs in a wide range of sequence prediction tasks. On the other hand, MEMMs are much more rapid than CRFs and HMMs in both of the training and the decoding phases. In CRFs and HMMs we need to use some version of the forward-backward algorithm in training. In contrast, in MEMMs, estimating the parameters of the transition probability distributions can be done for each local model in isolation. Although the prediction accuracy of MEMMs is usually reported to be inferior than that of CRFs, the difference is sometimes not significant; the two discriminative models may be competitive in some tasks.

The main drawback of MEMMs which makes their accuracy inferior than CRFs is "the label bias problem". This problem is mainly due to the uselessness of the observation at a particular position in predicting the most probable state at that position of the sequence. Although this is a well-known problem in the machine learning community, it has been underexploited and underappreciated. To the best of our knowledge, we are not aware of any study which investigates in detail the label bias problem in sequence prediction in general and in part-of-speech tagging in particular.

In this paper, we investigate the effect of the label bias problem in part-of-speech tagging, a particular sequence prediction task in natural language processing. We show the evidence that the entropy of transition probability distribution of a local model in MEMMs is log-normally distributed. Based on this observation, we propose a method for quantifying the label bias effect which directly uses the entropy of transition probability distribution. We find that half of the times that the MEMMs for Vietnamese part-of-speech tagging does not need the current word identity for predicting its tag but it can achieve a ratio of about 68% of its maximal accuracy. This is indeed show the significant effect of the label bias problem in Vietnamese part-of-speech tagging. It is also confirmed by a CRF model which shows a large improvement of accuracy over MEMMs for part-of-speech tagging of Vietnamse. In addition, we see that the label bias effect on French tagging is also significant and slightly stronger than on Vietnamese tagging.

The paper is structured as follows. Section II introduces some preliminaries on MEMMs for sequential tagging and its inherent label bias problem. Section III describes the methodology for investigating and quantifying the label bias effect in tagging. Section IV presents the evaluation results and discussion. Finally, section V concludes the paper.

II. Background

A. Tagging with MEMMs

Part-of-speech (POS) tagging is a typical sequence prediction task in natural language processing. In POS tagging we are interested in building a model that reads text in some language and assigns parts of speech to each tokens, such as noun, verb, adjective. In general, POS taggers in computational applications use more fine-grained POS tags like common noun

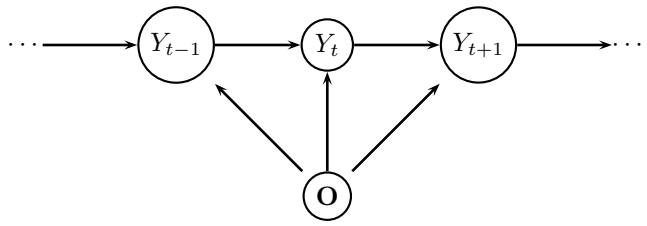

Fig. 1. A local model of a MEMM at position t.

or proper noun. In MEMM for POS tagging, we model the conditional probability of a tag sequence $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ given a word sequence $\mathbf{o} = (o_1, o_2, \ldots, o_T)$ as follows:

$$P(\mathbf{y} \mid \mathbf{o}) = \prod_{t=1}^{T} P(y_t|y_1, \ldots, y_{t-1}, \mathbf{o}) \approx \prod_{t=1}^{T} P(y_t|y_{t-2}, y_{t-1}, \mathbf{o}) \approx \prod_{t=1}^{T} P(y_t|y_{t-1}, \mathbf{o}),$$

where the observations o_t are words and y_t are their corresponding tags. The first approximation of the conditional probability is used in a second order MEMM while the second approximation is used in a first order MEMM[1]. Let $h_t = \langle \mathbf{o}, t, y_{t-2}, y_{t-1} \rangle$ denote the tagging context at position t. A MEMM can be considered a product of local models which are chained together as shown in Figure 1 where an uppercase bold character represents a random field which is a collection of random variables.

Each local model of MEMM is defined by a maximum entropy model (also called multinomial logistic regression model), defined as

$$P(y_t|h_t) = \frac{\exp(\theta \cdot f(h_t, y_t))}{\sum_{s \in \mathcal{S}} \exp(\theta \cdot f(h_t, s))}, \tag{1}$$

where $f(h_t, s) \in \mathbb{R}^D$ is a real-valued feature vector, $\mathcal{S}$ is the set of possible tags and $\theta \in \mathbb{R}^D$ is the parameter vector to be estimated from training data. Note that we use the same parameter vector for all local models in a MEMM. This form of distribution corresponds to the maximum entropy probability distribution satisfying the constraint that the empirical expectation for the feature is equal to its true expectation given the model:

$$\widehat{\mathbb{E}}(f_j(h,t)) = \mathbb{E}(f_j(h,t)), \qquad \forall j = 1, 2, \ldots, D.$$

The parameter $\theta \in \mathbb{R}^D$ can be estimated using iterative scaling algorithms [5], [4], [6] or some more efficient gradient-based optimization algorithms like conjugate gradient or quasi-Newton methods [7], [8], [9]. In the decoding phase, the optimal tag sequence $\mathbf{y}$ for a given word sequence $\mathbf{o}$ can be found using a Viterbi-like algorithm as the one used for HMMs. A detail presentation of this model for POS tagging can be found in [3].

[1]To make the term $P(y_t|y_{t-2}, y_{t-1}, \mathbf{o})$ meaningful for $t \leq 2$, one may pad the beginning of the sentence with a distinguished token marking the beginning of sentence.

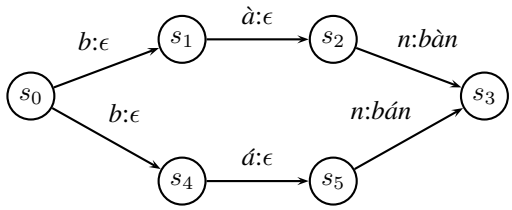

Fig. 2. An example of the label bias problem

B. The Label Bias Problem

As mentioned in the previous section, the major shortcoming of MEMMs is that they potentially suffer from "the label bias problem". This problem is due to two possible sources. The main source is that there are cases that a given observation is not useful in predicting the next state of the model.

For example, Figure 2 represents a simple model which is designed to discriminate between two Vietnamese words *bàn* and *bán*[2]. Suppose that the observation sequence is *bán*. In the first step, both of the two transitions from the state s_0 are *b*, therefore the transition probability is divided equally likely to the two out-going transitions. Next, the observation *á* is given. Both of the states s_1 and s_4 have unique out-going transition. The state s_1 has seen this observation multiple times and the state s_4 has never seen it in the training data. However, the state s_4 transfers all of the probability mass that it received to its only out-going transition since it does not generate the observation but being conditioned on the observation. Thus, if a state has a unique out-going transition, the given observation is useless in predicting the next state of the automaton. In general, a state whose transition probability distribution has a small entropy does not make use of the current observation in predicting the next state. In the example above, both of the two paths from state s_0 to state s_3 have the same probability regardless of the current observation. If an observation is seen more than the other in the training data, the priority will be given to it and its state sequence will be chosen regardless of the current observation, which could lead to a wrong prediction.

Another source of label bias is the use of previous tags in training and testing. In training, the model always uses known previous tags so they may decide a wrong tag at test time when there is uncertainty in the previous tag.

III. Methodology

A. Entropy of Transition Probability Distribution

We propose a method for investigating the label bias problem in sequence prediction which makes use directly the transition probability distribution of each local model in MEMMs. Consider the transition probability distribution (1) of a local model given tagging context h:

$$P(y|h) = \frac{\exp(\theta \cdot f(h, y))}{\sum_{s \in \mathcal{S}} \exp(\theta \cdot f(h, s))}, \qquad \forall y \in \mathcal{S}.$$

[2]In Vietnamese, *bàn* means either *table* (noun) or *to discuss* (verb) and *bán* means either *to sell* (verb) or *semi-* (adjective), depending on context.

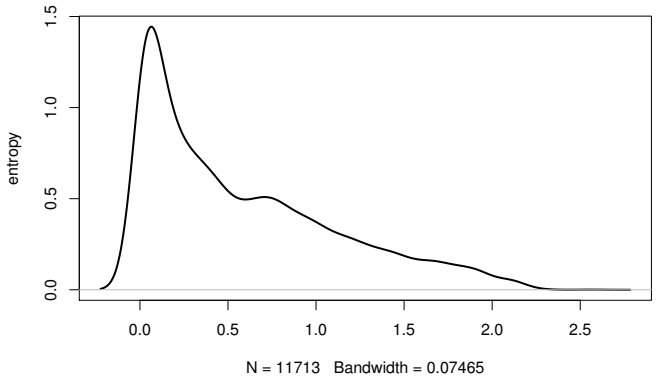

Fig. 3. Estimated density of an empirical transition entropy distribution

The entropy of a distribution $P(\cdot|h)$ is given by

$$\mathcal{E}(h) = -\sum_{y \in \mathcal{S}} P(y|h) \log P(y|h) \quad (2)$$

where $P(x) \log P(x)$ is understood to be zero whenever $P(x) = 0$. This quantity can be seen as a random variable which is a function depending on context. For short, we call this quantity the *transition entropy*. In general, entropy is a measure of the uncertainty or the average unpredictability in a random variable [10]. In this problem, transition entropy is a measure of the uncertainty of a transition probability distribution. The smaller the transition entropy is, the greater predictable the transition is.

We have seen in the previous section that the first source for the label bias problem is due to a small entropy of transition probability distribution at a state. Therefore, a natural question to ask is whether the transition distribution entropy can be statistically modeled so that its related properties can be drawn. With this in mind, we first try to find an approximate probability distribution for the transition entropy. Figure 3 shows the estimated density of the transition entropy distribution computed on a sample of 100 sentences. The entropy density looks similar on larger data sets.

An empirical study on the transition entropy reveals the evidence that this random variable is distributed log-normally.[3] This is conform to the fact that in statistical modelling, a variable might be modeled as a log-normal distribution if it can be thought of as the multiplicative product of many independent positive random variables as observed in many examples found in economics and finance.

In the following experiments, we characterize the effect of the label bias problem in first order MEMMs using the transition entropy information. In first order MEMMs, each tagging context at position t is denoted by $h_t = \langle \mathbf{o}, t, y_{t-1} \rangle$ where $\mathbf{o}$ is the entire word sequence. In all sequence models, the current word o_t in this sequence is the most important information to infer its tag y_t given the context h_t. The current word is used to extract many useful features which help predict its tag, including word form features such as the prefix, the suffix, the length or alphanumeric properties. Table I shows the accuracy of a typical MEMM tagger which is trained and tested on the Vietnamese Treebank corpus [11]. The parameters of the model are estimated using the limited memory BFGS optimization method [8] with L_2 regularization technique. Notice that this first order tagger has a modest accuracy in comparison with the result previously published in [12], [**?**]. This is due to three reasons. First, we do not focus here on improving the accuracy of the tagger but on investigating the label bias problem of the tagging model. We hence use a simple first order Markov model instead of the second order model which has been comprehensively demonstrated its superior. Second, we do not exploit the word form features in tagging which are crucial when predicting the tag of a word in general and of an out-of-vocabulary word in particular. Third, we use the most simple decoding algorithm – the greedy algorithm for tagging instead of a Viterbi-like algorithm which obviously results in an inferior performance.

[3]Due to space constraint, we do not present in detail the tests of log-normality here. This will be appended in an extended version of this paper.

TABLE I
ACCURACY OF A FIRST ORDER MEMM TAGGER

Feature templates	Accuracy
o_t, y_{t-1}, o_{t-1}	88.59%
y_{t-1}, o_{t-1}	36.43%

Algorithm 1: Compute $P(\cdot|h_t)$ using a set of feature templates F

$z \leftarrow 0$;
for $y \in S$ **do**
  $u_t \leftarrow \theta \cdot \texttt{featureVector}(h_t, y, F(t))$;
  $\texttt{score}(y) = \exp(u_t)$;
  $z \leftarrow z + \texttt{score}(y)$;
for $y \in S$ **do**
  $P(y|h_t) \leftarrow \texttt{score}(y)/z$;
return $P(\cdot|h_t)$;

The above tagging results demonstrate the importance of the identity of the current word in tagging. Without using the current word, the the accuracy drops heavily from 88.59% to 36.43%. Howerver, one might be surprised that we can correctly guess the tag of a word given only its previous word and its previous tag in more than one third of the times. At first glance, one may think that the label bias problem could not be significant in this task given this fact. However, we shall prove that this intuition is wrong. To show this, we exploit the transition entropy as follows: if the transition entropy of a local model is less than a fixed threshold, we omit the identity of a current word when predicting its tag, otherwise the word identity is used.

Let $F_1(t)$ denote the typical set of feature templates used in each local models of first order MEMMs consisting of the identity of current word, the tag of the previous word, and the identity of the previous word, that is $F_1(t) = \{o_t, y_{t-1}, o_{t-1}\}$; and let $F_2(t) \equiv F_1(t) \setminus \{o_t\} = \{y_{t-1}, o_{t-1}\}$.

Suppose that the function `featureVector(`h_t, y, F`)` computes the feature vector using a template set F. According

Algorithm 2: Compute $P(\cdot|h_t)$

$P(\cdot|h_t) \leftarrow P(\cdot|h_t, F_1(t))$;
$\mathcal{E}(h_t) \leftarrow \texttt{entropy}(P(\cdot|h_t))$;
if $\mathcal{E}(h_t) < \epsilon$ **then**
  $P(\cdot|h_t) \leftarrow P(\cdot|h_t, F_2(t))$;
return $P(\cdot|h_t)$;

to the standard MEMMs, each local transition probability distribution $P(y|h_t)$ is computed by the procedure $\texttt{prob}(\cdot|h_t, F)$ as shown in Algorithm 1.

Following the idea of exploiting the transition entropy, the real distribution $P(\cdot|h_t)$ used in the decoding phase is computed as in Algorithm 2, where `entropy` is a function calculating the entropy of a transition distribution and ϵ is a fixed threshold. The main idea is that if the entropy of a tagging context is less than the threshold then the "biased" feature template set F_2 is used, otherwise the original feature template set F_1 is used when computing each local transition probability distribution. If a context h_t satisfies $\mathcal{E}(h_t) < \epsilon$, it is called a biased context.

B. Conditional Random Fields

To overcome the label bias problem while reaping the benefits of using discriminative model for labeling sequential data, Lafferty *et al.* [4] introduced CRFs, a form of undirected graphical model that defines a single log-linear distribution over an entire label sequence given the observation sequence. CRFs are designed to avoid the label bias problem of MEMMs and other discriminative Markov models based on directed graphical models. Therefore, a comparison of accuracy between CRFs and MEMMs in a specific sequence prediction task would be useful to quantify the advantage of CRFs over MEMMs and the performance gain may partially give an empirical evidence about the effect of the label bias problem on the task. In this section, we briefly introduce CRFs and their use in sequence tagging.

In essence, in CRFs, one considers a global feature vector $F(\mathbf{o}, \mathbf{y})$ defined as a sum of T local feature vectors defined over the entire sequence:

$$F(\mathbf{o}, \mathbf{y}) = \sum_{t=1}^{T} f(h_t, y_t), \tag{3}$$

The probability of the label sequence given an observation sequence is modeled as

$$P(\mathbf{y} \mid \mathbf{o}) = \frac{\exp(\theta \cdot F(\mathbf{o}, \mathbf{y}))}{\sum_{\mathbf{s} \in \mathcal{S}^T} \exp(\theta \cdot F(\mathbf{o}, \mathbf{s}))}. \tag{4}$$

This is a maximum entropy model but it is much bigger than the local model of MEMMs in that the set of possible values of $\mathcal{S}^T$ is large if T is big and the normalizing constant is a sum on this set.

Although this model has a high complexity, we can still use a Viterbi-like decoding algorithm to find the best tag sequence of a given observation sequence.

TABLE II
ACCURACY OF THE MODIFIED MEMM ON THE VIETNAMESE TREEBANK

ϵ	Percentage	Accuracy	ϵ	Percentage	Accuracy
0.01	7.43%	82.29	0.30	43.52%	58.72
0.02	9.89%	78.41	0.40	50.45%	55.92
0.03	13.38%	76.36	0.50	56.22%	53.10
0.04	15.70%	73.99	1.00	79.16%	43.43
0.05	17.89%	72.88	1.50	92.40%	38.28
0.06	19.70%	71.32	2.00	98.87%	36.45
0.10	25.39%	67.92	2.50	100.0%	36.43
0.20	35.07%	62.80	2.56	100.0%	36.43

Given an observation sequence $\mathbf{o} = (o_1, o_2, \ldots, o_T)$, we need to find $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ such that $P(\mathbf{y} \mid \mathbf{o}) \to \max$. Since the normalizing constant does not depend on $\mathbf{y}$, we have

$$\begin{aligned}
\arg\max_{\mathbf{y} \in S^T} P(\mathbf{y} \mid \mathbf{o}) &= \arg\max_{\mathbf{y} \in S^T} [\theta \cdot F(\mathbf{o}, \mathbf{y})] \\
&= \arg\max_{\mathbf{y} \in S^T} \left[\theta \cdot \sum_{t=1}^{T} f(h_t, y_t) \right] \\
&= \arg\max_{\mathbf{y} \in S^T} \left[\sum_{t=1}^{T} \theta \cdot f(h_t, y_t) \right].
\end{aligned}$$

For details of the linear chain CRFs model for sequence prediction, see [4]. In the CRFs model, the parameters interact with each other on a global scope in order to give the probability of the sequence via the global normalization constant. As a result, all parts of the training data will affect the parameters. This helps avoid the label bias problem exposed by MEMMs. In order to quantify the label bias effect of MEMMs, we implement a CRFs model for tagging and compare the difference of accuracy between the two models.

IV. EVALUATION

In this section, we report the results of the experiments implementing the methodology presented in the previous section. The experiments are carried out on a Vietnamese treebank and a French treebank.

A. Results on a Vietnamese Treebank

The Vietnamese treebank containing $10,165$ manually tagged sentences, of which $9,665$ sentences are used as training set and 500 sentences are used as test set [11].

Table II shows the accuracy of the sparse MEMM which is trained and tested on the treebank. In this table, ϵ is the threshold of transition entropy used in Algorithm 2. The second and fifth columns indicate the percentage of transition entropy values that are less than ϵ in the test data. We use the greedy decoding algorithm to predict tag sequences.

We see that when fixing the transition entropy threshold $\epsilon = 0.4$, half of the times that we do not need the current word for tag prediction but we can achieve about 67.95% of the maximal accuracy obtainable by the model (55.92% over 82.29%. If $\epsilon = 0.1$, 25% of the times that the model can achieve 82.53% of the maximal accuracy without using the current word (67.92% over 82.29%). Figure 4 presents the dependence of the accuracy and the percentage of biased contexts on threshold ϵ.

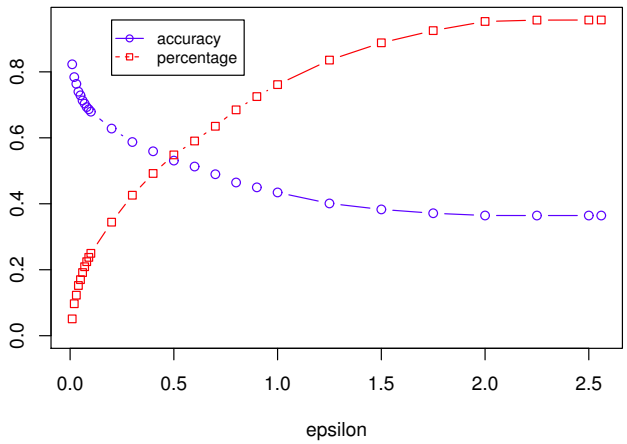

Fig. 4. The dependence of the tagging accuracy and the percentage of biased contexts on entropy threshold ϵ on a test set of the Vietnamese treebank

We then build a CRF tagging model which makes use of the same feature template as that of MEMMs. The CRF model is trained by the limited-memory BFGS optimization algorithm [8] and L_2 regularization technique with smooth constant fixed at 1.0. The resulting CRF model gives an accuracy of 90.36% on the test set, which has a net gain of 1.77% over the best MEMM reported in Table I. This result reconfirms a significant label bias phenomenon in Vietnamese part-of-speech tagging.

B. Results on a French Treebank

This paragraph reports the experimental results on the Sequoia treebank of French. Sequoia is a freely available French treebank comprising of $3,204$ sentences ($69,246$ tokens), from the French Europarl, the regional newspaper LEst Rpublicain, the French Wikipedia and documents from the European Medicines Agency [13].

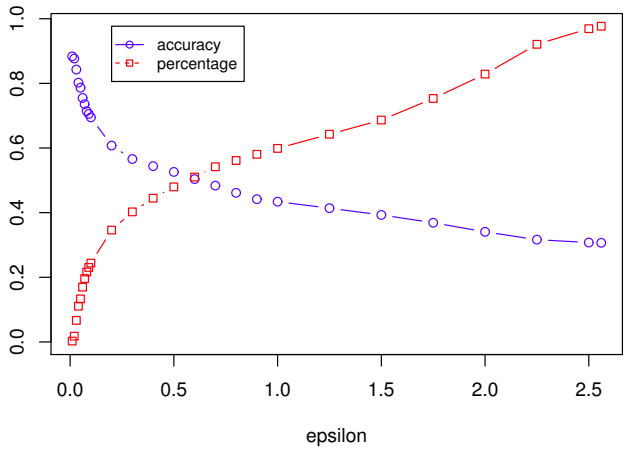

Fig. 5. The dependence of the tagging accuracy and the percentage of biased contexts on entropy threshold ϵ on a test set of the French treebank

Figure 5 presents the dependence of the accuracy and the percentage of biased contexts on threshold ϵ. We see a similar behavior of the effect of biased contexts on the accuracy of

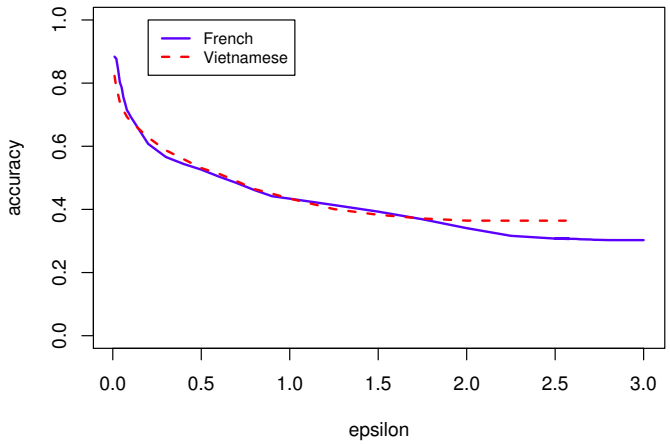

Fig. 6. Comparison of the label bias effect on French and Vietnamese part-of-speech tagging

the tagging model. In particular, when the transition entropy is fixed at 0.7, we can achieve about 57% of the maximal accuracy obtainable by the model without using the current word identity when predicting its tag.

Figure 6 shows the label bias effect on French and Vietnamese tagging. We see a stronger effect of the label bias problem on French tagging than on Vietnamese tagging. The accuracy curve of French tagger drops more quickly and deeply than that of Vietnamese tagger on the same scale of transition entropy ϵ.

V. Conclusion

The methodology described in this paper seems to be well suited to the investigation and quantification of the effect of the label bias problem in sequence prediction using MEMMs. To the best of our knowledge, this is the first attempt to quantify and evaluate the effect of this well-understood problem in sequence learning with MEMMs. We proposed to incorporate the entropy of the transition distribution at each local models of MEMMs into the decoding process. This enables us to detect biased contexts and examined their effect in sequence prediction. The method has been applied to part-of-speech tagging, a typical sequence labeling task in natural language processing.

Empirical experimentation on a Vietnamese corpus and a French corpus shows that the label bias problem is significant in part-of-speech tagging using MEMMs for both of the languages – The tagging models can achieve a high accuracy with respect to the maximal accuracy obtainable even when the identity of the current word is not used in tag prediction. The label bias effect in French tagging is slightly stronger than that in Vietnamese tagging. This observation may be explained by the difference in language nature, in that French is a moderately inflected language while Vietnamese is a typical isolating one.

This study also suggests that the use of conditional random fields for part-of-speech tagging improves significantly the accuracy of the taggers. This is an expected result since conditional random fields model is originally designed to

overcome the label bias problem encountered by MEMMs. The experiments shown in this paper reconfirm this result on part-of-speech tagging of the two different languages, a morphologically rich language and an isolating one.

Lastly, in this paper, we demonstrate and evaluate our approach on only the part-of-speech tagging problem. However, the proposed approach is general and it can be applied to other sequence labeling tasks whenever MEMMs are used, for example named-entity recognition and information extraction. The approach might be applied to investigate the effect of the label bias problem of sequence labeling tasks in natural languages other than Vietnamese and French as well. We plan to address these investigations in future works.

ACKNOWLEDGEMENTS

The authors would like to thank Hanoi University of Science, Vietnam National University for grants number TN-12-02. We are grateful to anonymous reviewers for helpful comments on the draft.

REFERENCES

[1] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257 –286, 1989.

[2] L. R. Welch, "Hidden Markov models and the Baum-Welch algorithm," *IEEE Information Theory Society Newsletter*, vol. 53, no. 4, December 2003.

[3] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy Markov models for information extraction and segmentation," in *Proceedings of ICML*, 2000.

[4] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001, pp. 282–289.

[5] J. N. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *Annals of Mathematical Statistics*, vol. 43, no. 5, pp. 1470–1480, 1972.

[6] J. Goodman, "Sequential conditional generalized iterative scaling," in *Proceedings of ACL*, 2002, pp. 9–16.

[7] J. Kazama and J. Tsujii, "Evaluation and extension of maximum entropy models with inequality constraints," in *EMNLP*, 2003.

[8] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York: Springer, 2006.

[9] G. Andrew and J. Gao, "Scalable training of l_1-regularized log-linear models," in *ICML*, 2007, pp. 33–40.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, Inc., 2006.

[11] P. T. Nguyen, L. V. Xuan, T. M. H. Nguyen, V. H. Nguyen, and P. Le-Hong, "Building a large syntactically-annotated corpus of Vietnamese," in *Proceedings of the 3rd Linguistic Annotation Workshop, ACL-IJCNLP*, Singapore, 2009.

[12] P. Le-Hong, A. Roussanaly, T. M. H. Nguyen, and M. Rossignol, "An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts," in *Proceedings of Traitement Automatique des Langues Naturelles (TALN-2010)*, Montreal, Canada, 2010.

[13] P. Le-Hong, T. M. H. Nguyen, A. Roussanaly, and T. V. Ho, "A hybrid approach to word segmentation of Vietnamese texts," in *Proceedings of the 2nd International Conference on Language and Automata Theory and Applications*, M.-V. Carlos, Ed. Tarragona, Spain: Springer, LNCS 5196, 2008.

[14] M. Candito and D. Seddah, "Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical," in *Proceedings of Traitement Automatique des Langues Naturelles (TALN-2012)*, Grenoble, France, 2012.