12-2019

Punctuation prediction for Vietnamese texts using conditional random fields

Hong Quang PHAM
Singapore Management University, hqpham.2017@phdis.smu.edu.sg

Binh T. NGUYEN

Nguyen Viet CUONG

Citation

PHAM, Hong Quang; NGUYEN, Binh T.; and CUONG, Nguyen Viet. Punctuation prediction for Vietnamese texts using conditional random fields. (2019). *SoICT '19: Proceedings of the 10th International Symposium on Information and Communication Technology, Hanoi, December 4-6*. 322-327.
Available at: https://ink.library.smu.edu.sg/sis_research/7816

Punctuation Prediction for Vietnamese Texts Using Conditional Random Fields

Quang H. Pham
School of Information System
Singapore Management University

Binh T. Nguyen[*]
AISIA Research Lab
VNU HCM - University of Science

Nguyen Viet Cuong
Department of Engineering
University of Cambridge

ABSTRACT

We investigate the punctuation prediction for the Vietnamese language. This problem is crucial as it can be used to add suitable punctuation marks to machine-transcribed speeches, which usually do not have such information. Similar to previous works for English and Chinese languages, we formulate this task as a sequence labeling problem. After that, we apply the conditional random field model for solving the problem and propose a set of appropriate features that are useful for prediction. Moreover, we build two corpora from Vietnamese online news and movie subtitles and perform extensive experiments on these data. Finally, we ask four volunteers to insert punctuations into a small sample of our dataset. The experimental results show that this problem is challenging, even for a human, and our model can achieve near performance in comparison to a human.

CCS CONCEPTS

• **Computing methodologies** → *Information extraction*;

KEYWORDS

punctuation prediction, Vietnamese language, conditional random field, sequence labeling

ACM Reference Format:
Quang H. Pham, Binh T. Nguyen, and Nguyen Viet Cuong. 2019. Punctuation Prediction for Vietnamese Texts Using Conditional Random Fields. In *The Tenth International Symposium on Information and Communication Technology (SoICT 2019), December 4–6, 2019, Hanoi - Ha Long Bay, Viet Nam.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3368926.3369716

1 INTRODUCTION

Punctuation prediction [1, 6] has played a special role in language and speech processing and had a lot of applications in the industry. Models from punctuation prediction systems are often used to annotate machine-transcribed speeches which usually do not come with punctuation information. Investigating efficient algorithms to correctly predict punctuation marks in texts is a need to construct high-quality automatic speech recognition frameworks.

For the last two decades, punctuation prediction has been extensively studied in major languages such as English and Chinese. One of the first punctuation prediction systems was developed by Beeferman et al. [1] to insert commas into texts automatically. The system uses a finite state transition model and a Viterbi decoder to predict the positions of commas in a sentence. Huang and Zweig [6] propose a maximum entropy model for the task with three punctuations: period, comma, and question mark. Using CRF models, Lu and Ng [9] achieve better performances for punctuation prediction on both the English and Chinese data sets of the IWSLT corpus [12]. Notably, they show that using a dynamic CRF to jointly model word-level and sentence-level labeling tasks and thus capture some long-range dependencies is useful for punctuation prediction. Similarly, Cuong et al. [3] use high-order semi-Markov CRFs to capture long-range dependencies among punctuations and achieved better prediction performance than linear-chain CRFs.

Zhao et al. [21] investigate Chinese punctuation prediction by formulating the problem as a multiple-pass labeling task and applying the CRF model. Cho et al. [2] study a segmentation and punctuation prediction problem for German-English with a monolingual translation system and demonstrate their results in the oracle experiments. Zhang et al. [20] investigate a new method in punctuation prediction for the stream of words in transcribed speech texts with excellent accuracy in both test datasets of IWSLT [13] and TDT4 [17]. Peitz and colleagues [14] transform the punctuation prediction problem into a machine translation problem as an alternative to using a language model based punctuation prediction technique. The experimental results show that it can gain additional improvement in BLEU points on the dataset IWSLT 2011 English French Speech Translation of Talks. Nonetheless, to the best of our knowledge, there exist few studies related to punctuation prediction for the Vietnamese language.

Related to the Vietnamese language processing, there have been various works in different directions such as word segmentation [5, 11] and part-of-speech (POS) tagging [19]. Using a weighted finite state transducer and neural network, Dien et al. [5] build a Vietnamese word segmentation system with high precision. Nguyen et al. [11] also investigate the Vietnamese word segmentation problem using CRF and SVM models. The Vietnamese POS tagging task is studied by Tran et al. [19] with three different models: CRF, SVM, and the maximum entropy.

In this study, we report results for the first Vietnamese punctuation prediction system. Our proposed method is based on linear-chain conditional random fields (CRFs) [8], a powerful sequence labeling model that has been used in many applications such as part-of-speech tagging [8], phrase chunking [16], or named entity recognition [10]. This model has previously been applied to punctuation prediction for both English [3, 9] and Chinese [9, 21]

[*] Corresponding author: ngtbinh@hcmus.edu.vn

SoICT 2019, December 4–6, 2019, Hanoi - Ha Long Bay, Viet Nam

ACM ISBN 978-1-4503-7245-9/19/12…$15.00
https://doi.org/10.1145/3368926.3369716

languages with promising results. To this aim, we first describe our datasets for the Vietnamese punctuation prediction task. We collect our data from different online Vietnamese news sources and movie subtitles. These data can be considered as the first corpora for Vietnamese punctuation prediction. After that, we explain how to model the problem as a sequence labeling task and employ a suitable CRF model. More specifically, we introduce our label sets and features for the CRF model. Subsequently, we measure the corresponding performance of CRF models in two different datasets and compare results with human evaluation.

The rest of this paper can be organized as follows. In Section 2, we provide a brief introduction to the conditional random field model. We illustrate our approach for the Vietnamese punctuation prediction task and present the corresponding experimental results in Section 3 and 4. Finally, the paper ends with our conclusion and future work.

2 CONDITIONAL RANDOM FIELDS

Conditional random fields [8] are discriminative, undirected Markov models which can capture various dependencies between a structured observation $\mathbf{x}$ and its corresponding formal label $\mathbf{y}$. In this section, we briefly introduce linear-chain CRFs, a particular type of CRFs that is widely used in practice, especially for sequence labeling tasks. As we only focus on linear-chain CRFs, we use the term CRF to refer to a linear-chain CRF throughout this paper.

To formally define what is a CRF model, we consider the following notations. First, let $\mathbf{x} = (x_1, x_2, \ldots, x_T)$ be an input sequence and Y be a random vector of the corresponding output sequence with values of the form $\mathbf{y} = (y_1, y_2, \ldots, y_T)$. Suppose we have a set of real-valued feature functions

$$\{f_k(y_t, y_{t-1}, t, \mathbf{x})\}_{k=1}^{K}$$

for the CRF model and λ_k is the corresponding weight of f_k.

A CRF defines the conditional distribution $p(\mathrm{Y} \mid \mathbf{x})$ as:

$$p(\mathrm{Y} = \mathbf{y} \mid \mathbf{x}) = p(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{k=1}^{K}\sum_{t=1}^{T} \lambda_k f_k(y_t, y_{t-1}, t, \mathbf{x})\right),$$

where $Z(\mathbf{x}) = \sum_{\mathbf{y}} \exp\left(\sum_{k=1}^{K}\sum_{t=1}^{T} \lambda_k f_k(y_t, y_{t-1}, t, \mathbf{x})\right)$ is the normalization function, and also called the partition function.

Given the training data $\mathcal{D} = \{(\mathbf{x}^i, \mathbf{y}^i)\}_i$, the CRF model is trained by choosing the parameters $\vec{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_K)$ that maximize the following regularized conditional log-likelihood of the data:

$$\mathcal{L}(\vec{\lambda}) = \sum_i \ln p(\mathrm{y}^i | \mathrm{x}^i) - \sum_{k=1}^{K} \frac{\lambda_k^2}{2\sigma^2},$$

where σ is a parameter that controls the degree of regularization. This function is concave and thus the global optimum can be found using any convex optimization algorithm. Optimization algorithms for $\mathcal{L}(\vec{\lambda})$ usually require inference on CRFs. Similar to hidden Markov models [15], inference on CRFs is made by defining a set of the forward and backward variables and using dynamic programming to compute them efficiently. During testing, the label sequence for a new test input is determined by a Viterbi-like algorithm [18], which returns the label sequence with the highest probability according to the trained model. We use these algorithms in our punctuation prediction system.

3 PUNCTUATION PREDICTION FOR VIETNAMESE TEXTS: THE MODEL AND DATA SETS

In this section, we describe our sequence labeling model and data sets for punctuation prediction for Vietnamese texts. Our contributions are a useful set of CRF features and two new data sets, which are the first data sets for the task for the Vietnamese language.

3.1 Punctuation Prediction as Sequence Labeling

Similar to previous works for both English and Chinese languages [9, 21], we formulate the punctuation prediction task as a sequence labeling problem. For this reason, one can naturally apply CRF models for solving it. Notably, we treat each sentence as a sequence and aim to label each word by the punctuation that immediately follows the word. In the simple case, we use the label O to indicate that a word is not followed by any punctuation.

For instance, the following sentence in Vietnamese[1]

> Khu vực Đà Nẵng - Bình Định có tần số bão ít hơn và bão thường tập trung tháng 10 và 11.
> (The area of Da Nang - Binh Dinh has a lower storm frequency and storms usually occur in October and November.)

can be labeled as follows:

> khu/O vực/O đà/O nẵng/O -/O bình/O định/O có/O tần/O số/O bão/O ít/O hơn/O và/O bão/O thường/O tập/O trung/O tháng/O 10/O và/O 11/Period

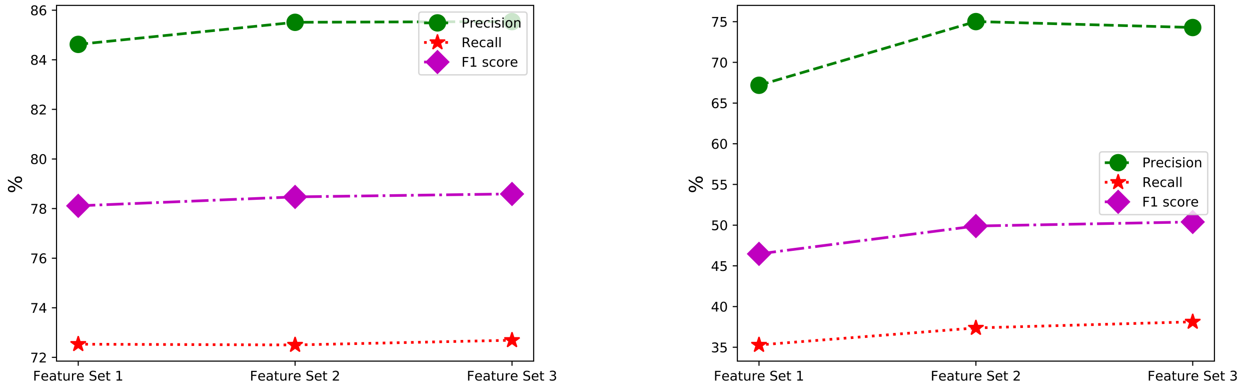
Note that all the words are in lower case because the word case information is usually not available for the punctuation prediction task. For instance, when the texts are transcribed from speeches, we do not have the case information for the words.

3.2 Features for CRFs

To apply CRF models for punctuation prediction in Vietnamese texts, we need to construct a set of features that is useful for the punctuation prediction task. Our baseline zero-th order CRF features include the following unigram features: the current label without any word, the current word itself, the words within four positions preceding the considering words, and words within two places succeeding the present words. Next, we add all pairs of consecutive words within a window of size 2 before and after the current position as bigram features. For extracting the first-order CRF features, we include transitions between two labels as features. Table 1 describes all the details of our set of features. The table follows the template of the CRF++ toolkit [7], which we use to train and test our model.

Usually, including other words surrounding the current position as unigram features (such as words at the 3rd or 4th position succeeding the current location) and the trigram features is helpful

[1] http://vnexpress.net/tin-tuc/thoi-su/bao-xuat-hien-nhieu-nhat-o-quang-ninh-thanh-hoa-3077937.html.



(a) Movie Subtitle.

(b) News Dataset.

Figure 1: The performance among different sets of features in two datasets: Movie Subtitle and News Dataset.

Table 1: The set of features for the CRF model.
The feature template %x[r,0] denotes the words at the r^{th} position relatively to the current position

ID	CRF++ feature template	Feature definition
U	N/A	The current label without any word
U00	%x[0,0]	The current word and its label
U01	%x[-1,0]	The preceding word and the current label
U02	%x[-2,0]	The word at position -2 and the current label
U03	%x[1,0]	The succeeding word and the current label
U04	%x[2,0]	The word at position +2 and the current label
U05	%x[-3,0]	The word at position -3 and the current label
U06	%x[-4,0]	The word at position -4 and the current label
U07	%x[-2,0]/%x[-1,0]	Bigram of words at positions -2 and -1, plus the current label
U08	%x[-1,0]/%x[0,0]	Bigram of preceding and current words, plus the current label
U09	%x[0,0]/%x[1,0]	Bigram of current and succeeding words, plus the current label
U10	%x[1,0]/%x[2,0]	Bigram of words at positions 1 and 2, plus the current label
B	N/A	Transitions between labels

for punctuation prediction in the English language [3, 9]. However, our preliminary experiments indicate that they are not useful for the Vietnamese language. Thus, we do not include these features in our model.

In Table 2, we describe three different sets of features we choose for doing experiments with CRF models.

3.3 Datasets

As there was no standard dataset available in the Vietnamese language for the punctuation prediction task, we create our data from online news sources. More specifically, we collect 500 online newspaper articles and 100 transcribed movie subtitles to build two datasets and release them publicly as a contribution to the research community.

As a pre-processing step, we clean the data by manually fixing common writing errors and non-standard uses of punctuations such as two or more punctuations at the end of a sentence. For each data set, we also remove rare punctuations whose occurrence is less than 0.1% of the total punctuations in the data. Because the nature of each dataset is different, the collection of punctuations for each dataset is also disparate. For the news article data set, we have three labels: PERIOD, COMMA, and O. For the movie subtitle data set, we have five classes: PERIOD, COMMA, QMARK (question mark), EXCLAM (exclamation mark), and O.

Table 3 shows the statistic of each data set, including the number of each punctuation and its percentage. Note that we do not show the default punctuation O in the table.

4 EXPERIMENTAL RESULTS

To evaluate the performance of CRF models in the punctuation prediction problem, we consider several experiments on two datasets. In what follows, we introduce details of these experiments and the corresponding results.

4.1 Setup

For the experiments, we run our punctuation prediction models on each dataset separately. Each dataset is divided into two parts: 2/3 for training and the rest for testing. We use three metrics

Table 2: The list of all feature sets used in the experiments

Type	Description
Feature Set 1	U, U00, U01, U02, U03, U04, U05, U06
Feature Set 2	U, U00, U01, U02, U03, U04, U05, U06, U07, U08, U09, U10
Feature Set 3	U, U00, U01, U02, U03, U04, U05, U06, U07, U08, U09, U10,

Table 3: Distribution of punctuations in training and testing data sets. The rest of the data sets contain the empty punctuation (label O)

Punctuation	News Dataset		Movie Subtitle	
	Number	Percentage (%)	Number	Percentage (%)
Comma	12713	5.33	1867	3.64
Period	7860	3.29	5544	10.81
Question mark	N/A	N/A	1547	0.95
Exclamation mark	N/A	N/A	488	3.02

Table 4: Micro-averaged metrics on two datasets

Punctuation	News Dataset			Movie Subtitle		
	P	R	F	P	R	F
Feature set 1	67.18	35.28	46.27	84.62	72.53	78.11
Feature set 2	**75.01**	37.37	49.89	85.51	72.50	78.47
Feature set 3	74.28	**38.12**	**50.39**	**85.54**	**72.69**	**78.59**

to measure the performance of our system: precision (denoted by P), recall (denoted by R), and F_1 (denoted by F). From Table 3, the punctuations are not equally distributed, hence we use micro-averaged scores [4] instead of macro-averaged scores for the overall performance of the system. The formula of the micro-averaged precision and recall can be given as follows:

$$P = \frac{\sum_j tp_j}{\sum_j (tp_j + fp_j)} \qquad R = \frac{\sum_j tp_j}{\sum_j (tp_j + fn_j)}$$

where tp_j is the number of punctuations correctly classified as class j (true positive), fp_j is the number of punctuations incorrectly classified as class j (false positive), and fn_j is the number of punctuations in class j that are misclassified as another class (false negative). The micro-averaged F_1 score is computed as:

$$F = \frac{2PR}{P + R}.$$

In our experiments, we shall illustrate the effects of different combinations of features to the performance of our Vietnamese punctuation prediction system. We begin with the unigram words features and subsequently add the bigram word features as well as the label transition features.

After obtaining the best set of features, we use it to train a CRF model using the expanded label set described in Section 3.1. Then, we compare this model with the CRF model using the original label set. For all the experiments, we train CRF models on the training set with the regularizer $\sigma = 1$ and then test our models on the testing set. Our scores are computed on the token level.

4.2 Results

In Table 4, we show the micro-averaged metrics of each feature set on our datasets. As visualized in Figure 1, using unigram word features (the red line) alone achieves the lowest performance, 46.27% and 78.11% F_1 score on the news and movie subtitle datasets, respectively. Meanwhile, adding unigram further (the purple line) improves the models' performance and increases F_1 score to 49.89% and 78.47% on both datasets. Finally, the best model (the green line) is achieved by adding the label transition label. On the news dataset, it decreases the precision by 0.73%, but increases the recall by 0.75%, and hence overall increases the F_1 score by 0.5%. However, on the subtitle dataset, transition label increases all precision, recall and F_1 to 85.54%, 72.69%, and 78.59%, respectively.

4.3 Human Evaluation

We do another experiment by comparing the performance of our trained model with the human. In this case, we randomly sampled 100 sentences in the movie subtitle dataset; then, we removed the labels and asked *four* candidates to inserted the missing punctuations to the plain text. For fair comparisons, we also gave the candidates another 100 different and fully annotated sentences for training before the test. Subsequently, we used the best-trained model to predict punctuations on these 100 samples and compare its performance with those candidates.

Table 5 shows the F_1 scores of each candidate on this dataset; we report the F_1 score on each punctuation and micro-averaged score

Table 5: The performance of different candidates on the sample test

Candidate	Punctuation					
	PERIOD	QMARK	COMMA	EXCLAM	O	micro F_1
Candidate 1	**89.33**	83.33	**80.00**	21.05	**99.1**	**81.6**
Candidate 2	61.53	62.85	12.5	17.77	97.59	46.28
Candidate 3	64.46	**84.21**	51.16	14.81	98.72	54.68
Candidate 4	70.00	75.00	0.00	**23.53**	97.42	54.22
Average	71.33	76.3475	35.915	19.29	98.2075	59.195
Model	86.20	0.00	6.25	0.0	96.74	65.23

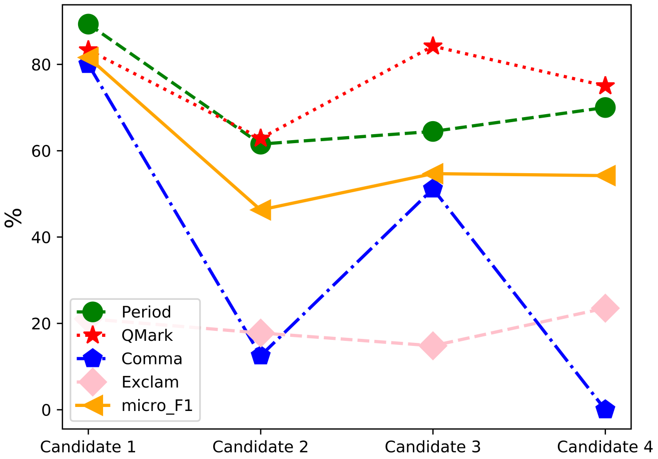


Figure 2: Validation results from different testing persons.

as the combined metric. In Figure 2, one can see that our candidates could achieve micro-averaged F_1 scores ranging from 46.28% to 81.6%. Although our model gets a higher F_1 score than the average score of all candidates, it fails to predict both question marks and exclamation marks. Besides, it has an abysmal performance on inserting commas.

The experiment results have shown that punctuation prediction task is quite difficult, even for a human. It turns out that building an appropriate model that can insert meaningful punctuation marks to Vietnamese texts is still challenging.

5 CONCLUSION AND FUTURE WORK

We have studied the punctuation prediction task for Vietnamese texts and presented our results for a system trained using the CRF model. We have also conducted a small scale experiment where we ask four candidates to fill the missing punctuations in a movie subtitle sample. The system shows promising results on the movie subtitle dataset, even better than the average score of our candidates. For future works, we plan to include more sophisticated types of features to achieve better results in prediction. These features may consist of POS tags, person name dictionary, etc. We also want to increase the scale of experiments by collecting more data with a focus on those rare punctuation marks that we omitted in our current investigation. Finally, our experiments are somehow limited since we constructed our corpus from online news texts, not from real transcribed speeches. As a result, in the future, we may need to consider the domain where data come from actual transcribed speeches. In that case, it will be interesting to investigate how to transfer the knowledge learned from the online news texts domain to this new domain.

REFERENCES

[1] Doug Beeferman, Adam Berger, and John Lafferty. 1998. Cyberpunc: A lightweight punctuation annotation system for speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

[2] Eunah Cho, Jan Niehues, and Alex Waibel. 2012. Segmentation and punctuation prediction in speech language translation using a monolingual translation system.. In *Workshop on Spoken Language Translation*.

[3] Nguyen Viet Cuong, Nan Ye, Wee Sun Lee, and Hai Leong Chieu. 2014. Conditional Random Field with High-order Dependencies for Sequence Labeling and Segmentation. *Journal of Machine Learning Research* 15 (2014), 981–1009.

[4] Manning D. Christopher, Raghavan Prabhakar, and Schacetzel Hinrich. 2008. *An Introduction to Information Retrieval.* Cambridge, UK: Cambridge University Press.

[5] Dinh Dien, Hoang Kiem, and Nguyen Van Toan. 2001. Vietnamese word segmentation. In *Natural Language Processing Pacific Rim Symposium*.

[6] Jing Huang and Geoffrey Zweig. 2002. Maximum entropy model for punctuation annotation from speech. In *International Conference on Spoken Language Processing*.

[7] Taku Kudo. 2005. CRF++: Yet another CRF toolkit. *Software available at http://crfpp. sourceforge.net* (2005).

[8] John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*.

[9] Wei Lu and Hwee Tou Ng. 2010. Better Punctuation Prediction with Dynamic Conditional Random Fields. In *Conference on Empirical Methods in Natural Language Processing*.

[10] Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Conference on Natural Language Learning*.

[11] Cam-Tu Nguyen, Trung-Kien Nguyen, Xuan-Hieu Phan, Le-Minh Nguyen, and Quang-Thuy Ha. 2006. Vietnamese word segmentation with CRFs and SVMs: An investigation. In *Pacific Asia Conference on Language, Information and Computation*.

[12] Michael Paul. 2009. Overview of the IWSLT 2009 evaluation campaign. In *Workshop on Spoken Language Translation*.

[13] Michael Paul. 2009. Overview of the IWSLT 2009 evaluation campaign. *International Workshop on Spoken Language Translation (IWSLT) 2009, 1-18* (2009).

[14] Stephan Peitz, Markus Freitag, Arne Mauser, and Hermann Ney. 2011. Modeling punctuation prediction as machine translation. In *IWSLT*.

[15] Lawrence Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77, 2 (1989), 257–286.

[16] Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.

[17] Strassel Stephanie, Junbo Kong, and David Graff. 2005. TDT4 Multilingual Text and Annotations LDC2005T16. *Web Download. Philadelphia: Linguistic Data Consortium*.

[18] Charles Sutton and Andrew McCallum. 2006. An introduction to conditional random fields for relational learning. *Introduction to Statistical Relational Learning* (2006), 93–128.

[19] Oanh Thi Tran, Cuong Anh Le, Thuy Quang Ha, and Quynh Hoang Le. 2009. An experimental study on Vietnamese POS tagging. In *International Conference on Asian Language Processing*.

[20] Dongdong Zhang, Shuangzhi Wu, Nan Yang, and Mu Li. 2013. Punctuation Prediction with Transition-based Parsing. In *ACL*.

[21] Yanqing Zhao, Chaoyue Wang, and Guohong Fu. 2012. A CRF sequence labeling approach to Chinese punctuation prediction. In *Pacific Asia Conference on Language, Information and Computation*.