Vietnamese Text Accent Restoration With Statistical Machine Translation

Luan-Nghia Pham
Department of Information Technology
Haiphong University
Haiphong,Vietnam
nghialuan@gmail.com

Viet-Hong Tran
University of Economic
And Technical Industries
Hanoi,Vietnam
thviet@uneti.edu.vn

Vinh-Van Nguyen
University of Engineering and Technology
Vietnam National University
Hanoi,Vietnam
vinhnv@vnu.edu.vn

Abstract

Vietnamese accentless texts exist on parallel with official vietnamese documents and play an important role in instant message, mobile SMS and online searching. Understanding correctly these texts is not simple because of the lexical ambiguity caused by the diversity in adding diacritics to a given accentless sequence. There have been some methods for solving the vietnamese accentless texts problem known as accent prediction and they have obtained promising results. Those methods are usually based on distance matching, n-gram, dictionary of words and phrases and heuristic techniques. In this paper, we propose a new method solving the accent prediction. Our method combine the strength of previous methods (combining n-gram method and phrase dictionary in general). This method considers the accent predicting as statistical machine translation (SMT) problem with source language as accentless texts and target language as accent texts, respectively. We also improve quality of accent predicting by applying some techniques such as adding dictionary, changing order of language model and tuning. The achieved result and the ability to enhance proposed system are obviously promising.

1 Introduction

Accent predicting problem refers to the situation where accents are removed (e.g. by some email preprocessing systems), cannot be entered (e.g. by standard English keyboards), or not explicitly represented in the text (e.g. in Arabic). We resolve the languages using Roman characters in writing together with additional accent and diacritical marks. These languages include European languages such as Spanish and French and Asian languages such as Chinese Pinyin and Vietnamese.

Vietnamese accentless texts coexist with official Vietnamese texts and it is relatively common texts on the internet. Official Vietnamese language is a complex language with many accent (including acute, grave, hook, tilder, and dot-below) and Latinh alphabets. These are two inseparable components in Vietnamese. However, many Vietnamese choose to use accentless Vietnamese because it is easier and quickly to type. For example, a official Vietnamese sentence: *chúng tôi sẽ bay tới Hà Nội vào chủ nhật* ('We will fly to Hanoi on Sunday') will be written as an Vietnamese accentless sentence as *chung toi se bay toi Ha Noi vao chu nhat*. Decoding such a sentence could be quite hard for both human and machine because of lexical ambiguity. For instance, the accentless term "*toi"* can easily lead to confusion between the original Vietnamese "*tôi"* ('we') and the plausible alternative "*tới*" ('to').

Nowadays, the application of information technology to exchanging information is more and more popular. We daily receive many of emails, SMS but the majority of them are without accents which may cause troubles for interpreting the meaning. Therefore, automatic accent restoration of accentless Vietnamese texts have many of applications such as automatically inserting accent to emails, instant message, SMS are written without diacritics Vietnamese, or assistant for website administration in which accent Vietnamese is required. Therefore, it is essential to develop supporting tools which can automatically insert accent to Vietnamese texts.

Accent predicting problem is the particular problem of lexical disambiguation. The recent approach to lexical disambiguation is corpus-based such as n-gram, dictionary of phrases, ...

Copyright 2013 by Luan-Nghia Pham, Viet-Hong Tran, and Vinh-Van Nguyen
27th Pacific Asia Conference on Language, Information, and Computation pages 423－429

In this paper, we propose the method for automatic accent restoration using Phrase-based SMT. Vietnamese accentless sentence and Vietnamese accent sentence (office Vietnamese sentence) will be source and target sentence in Phrase based SMT, respectively. We also improve quality of accent predicting by applying some techniques such as adding dictionary, changing size of n-gram of language model. The experiment results with Vietnamese corpus showed that our approach achieves promising results.

The rest of this paper is organised as follows. Related works are mentioned in Section 2. The methods for accent restoration using SMT are proposed in Section 3. In Section 4, we describe the experiments and results for evaluating the proposed methods. Finally, Section 5 concludes the paper.

2 Related works

In the recent years, several different methods were proposed to automatically restore accent for Vietnamese texts.

The VietPad (Quan, 2002) used a dictionary file and this one is stored all of words in Vietnamese. The idea of VietPad is based on the use of dictionary file and each non-diacritic word is mapped 1-1 into diacritic word. However, the dictionary file also is stored many words which are rarely used so these words is incorrectly restored accent. Therefore, VietPad can only restore Vietnamese accent texts with accuracy about 60-85% and this accuracy is depended on content of text.

The AMPad (Tam, 2008) is an efficiency Vietnamese accent restoration tool and texts can be restored online. The idea of AMPad is based on the statistical frequency of diacritics words which correspond with non-diacritics word. The program used selection algorithm and the most appropriate word is chosen. AMPad can restore Vietnamese accent texts with accuracy about 80% or higher for political commentary and popular science domain. However, It also restore Vietnamese accent with accuracy about 50% for specialized documents or literature and poetry documents which have complex sentence structure and multiple meaning.

The VietEditor (Lan, 2005) is based on the idea of VietPad but it is improved. This program used a dictionary file and this file is stored all of phrase which are often used in Vietnamese texts. This file is called phrase dictionary. After each non-diacritic word is mapped 1-1 into diacritic word, the program check the phrase dictionary to find the most appropriate word. VietEditor restore Vietnamese accent texts more flexibly and accurately than VietPad.

The viAccent (Truyen et al., 2008) allows you to restore Vietnamese accent texts online and with several different speed. Generally, the slower speed is the better result is. The idea of program used n-gram language model and it is reported at the conference PRICAI 2008 (The Pacific Rim International Conference on Artificial Intelligence).

The VnMark (Toan, 2008) used n-gram language model to create a dictionary file. It is VNMarkDic.txt file. This file shows occurrence probability of phrase in Vietnamese texts and it will build the case restoration for word or phrase. This combination will create sentences which are restored Vietnamese accents. When the weight of each sentence is identified, the best way will be selected accent restoration for Vietnamese text. However, Accuracy depends on the sentences of the dictionary file. Because the number of sentence is few so the result is still limited.

3 Our method

As mentioned in the section 2, the studies about accent restoration for Vietnamese text are based on experience. These studies used the dictionary file (such as VietPad) or n-gram language model with phrase (such as VnMark, AmPad, viAccent ...) but They are not yet generalized because this combination has some limitation as following:

- The number of phrases are few (each phrase is only 2 words, 3 words) and there are no priority when each phrase is chosen.
- The combination of language model and phrase dictionary is simple and It is mainly based on experience.

To overcome the above limitation, we present a general approach to restore accent for Vietnamese text. This method is viewed as machine translation from non-diacritical Vietnamese language (source language) into diacritical Vietnamese language (target language). This method has solved two above limitation by the use of phrase-table with the priority levels (the length of the phrase is arbitrary and the corresponding translation probability value) and It is combined language model

with phrase-table by log-linear model (adding and turning of parameters for combination).

3.1 Overview of Phrase-table Statistical Machine Translation

In this section, we will present the basics of Phrase-based SMT toolkit(Koehn et al., 2003). The goal of the model is translating a text from source language to target language. As described by (1), we have sentences in source language (English) $e_1^I = e_1, \ldots e_j, \ldots, e_I$ which are translated into target language (Vietnamese) $v_1^J = v_1, \ldots, v_j, \ldots v_J$. Each sentence can be found in the target text then the we will choose sentence so that:

$$V_1^J = argmax_{v_1^I} \Pr(v_1^J / e_1^J) \quad (1)$$

The figure below illustrates the process of phrase-based translation:

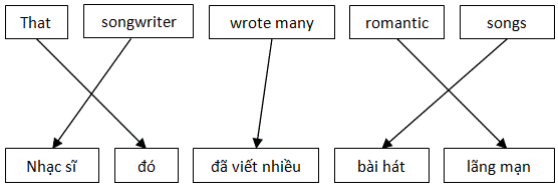

Figure 1: Phrase-based translation model

In phrase translation model, sequence of consecutive words (phrases) are translated into the target language. The length of source phrase can be different from target phrase. We divide source sentence into several phrases and each phrase is translated into a target language, then it reorder the phrases so that the target sentence to satisfy the formula (1) and then they are grafted together. Finally, we get a translation in the target language.

Figure 2 shows the phrase-based statistical translation model. There are many translation knowledge which can be used as language models, translation models, etc. The combination of component models (language model, translation model, word sense disambiguation, reordering model...) is based on log-linear model (Koehn et al., 2003).

3.2 Accent prediction based on Phrase Statistical Machine Translation.

Vietnamese texts with accent are collected from newspapers, books, the internet, etc., and they are reprocessed to remove the excess characters. Then

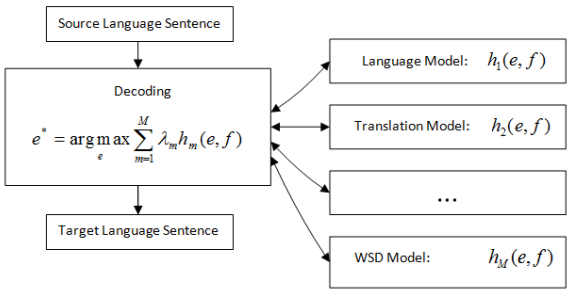

Figure 2: Diagram Phrase-based SMT translation based on log-linear model

vnTokenizer tool (Hong et al., 2008) is used to segment words in Vietnamese language. After that, the text file with accent and a corresponding accentless text fileis are created. Two text files are the same number of line and each line of accent text file corresponds with an accentless line in another text file. Figure 3 shows some sentences in corpus.

Accentless Vietnamese	Accent Vietnamese
cai ban_nay hinh ban_nguyet .	cái bàn_này hình bán_nguyệt .
toc_do truyen_thong se tang_cao .	tốc_độ truyền_thông sẽ tăng_cao .
toi_nay toi di_choi .	tối_nay tôi đi_chơi .
nhung van_de lien_quan toi nguoi dong_tinh luyen_ai duoc ban_bac soi_noi trong buoi_hop nhom toi hom_qua	những vấn_đề liên_quan tới người đồng_tính luyến_ái được bàn_bạc sôi_nổi trong buổi_họp nhóm tối hôm_qua .

Figure 3: Some sentences in corpus

The accents removal is processed by building mapping table between the accents words and corresponding accentless Vietnamese words . For example:

a = {a, á, à, ạ, ả, ă, ắ, ặ, ẳ, â, ầ, ấ, ậ, ẩ}

Mapping table of characters include uppercase and lowercase letters in Vietnamese. After preprocessing we have a corpus file and we processed training data. Finish, we received phrase table and language model for machine learning. This phrase table is stored phrases with the different length. Language model with n-grams is covered nearly all Vietnamese sentences and this training process is automatically executed.

The general architecture of our method is described on Figure 4:

Phrase-based Statistical Translation model has three important components. They include translation model, language model and decoder. Translation results are calculated with parameters in the

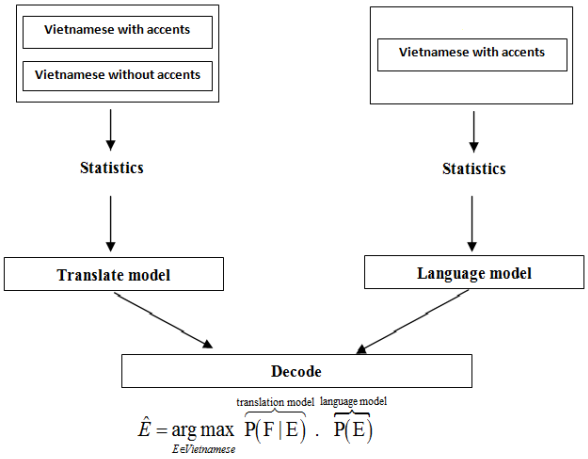

Figure 4: Accents restoration base on Phrase-based SMT

phrase table and language model. Example, accentless sentence restoration in Vietnamese:
neil Young da bieu dien nhieu the loai nhac rock va blue.
After the phrases is segmented. This sentence is:
neil_Young da bieu_dien nhieu the loai nhac rock va blue.
Results after the translation:
neil_Young đã biểu_diễn nhiều thể loại nhạc rock và blue .

First, the source sentence will be divided into phrases neil_Young, neil_Young da, neil_Young da bieu_dien, da bieu_dien, bieu_dien, bieu_dien nhieu,. . . .

After that, the system find the probability of each phrase in the phrase table and language model then the weight of sentence is computed . Example:
We found weight of above phrases in the phrase table and language model:
da bieu_dien ||| đã biểu_diễn ||| 1 1 1 0.857179 2.718 ||| 1 1
the loai ||| thể loại ||| 1 1 1 0.0362146 2.718 ||| 2 2

-3.436823 đá -0.3055001
-2.309609 đã -0.5276677
-4.109961 biểu_diễn -0.2860174
-4.168163 đã biểu_diễn
-2.227281 biểu_diễn nhiều
-1.628649 thể loại

Finally, the system is implemented by the decoder process. For each translation choice will have a hypothesis. Suppose first selection word is the neil_Young, this word is translated into neil_Young (unchanged) because it do not find corresponding word. For simplicity, we translate from left to right of sentence. Next, da word is translated, for example, it can be đá or đã. The probability of each hypothesis is calculated and updated for the each new hypothesis. Next, bieu_dien is translated, this phrase is found in the phrase table, language model and a phrase is chosen that is biểu_diễn phrase. Combining hypotheses can happen, da bieu_dien phrase is restored to đã biểu_diễn. Continue until all the words of sentence are translated.

4 Experiment Results

4.1 Corpus Statistics and Experiments

For evaluation, we used an accentless Vietnamese corpus with 206000 pairs, including 200000 pairs for training, 1000 pairs for tunning and 5000 pairs for development test set.

The corpus for experiments was collected from newspapers, books,. . . on the internet with topics such as social, sports, science (Nguyen et al., 2008). The Table 1 shows the summary statistical of our data sets. Several experiments are processed on the basis of Phrase-based Statistical Machine Translation model with MOSES open-source decoder (Koehn et al., 2007). For training data and turning parameters, we used standard settings in the Moses toolkit (GIZA++ alignment, grow-diagfinal-and, lexical reordering models, MERT turning). To build the language model, we used SRILM toolkit (Stolcke, 2002) with 3 and 4-gram. In this experiments, we evaluated the quality of the translation results by BLEU score (Papineni et al., 2002) and accuracy sentences.

We performed experiments on MT_VR system and MT_VR+Dict system:

- MT_VR is a baseline Vietnamese restoration system. This system uses phrase-based statistical machine translation with standard settings in the Moses toolkit.
- MT_VR+Dict is a baseline Vietnamese restoration combine with dictionary information.

4.2 Results

- We experimented on several different corpus and we evaluated translation quality.

Corpus Statistical		Vietnamese with accents	Vietnamese without accents
Training	Sentences	200000	
	Average Length	22.3	22.3
	Word	4474378	4474378
Development	Sentences	1000	
	Average Length	24,3	24,3
	Word	24343	24343
Test	Sentences	5000	
	Average Length	22,1	22,1
	Word	110729	110729

Table 1: The Summary statistical of data sets

- Translate model with the corpus include 50.000, 100.000, 150.000 and 200.000 sentence pairs. After successful training, we tested with 5.000 pairs of sentence.

To improve the quality of the system we need to build a corpus with better quality as well as greater coverage and we need to process accurately data. We have improved on some of the approach

4.2.1 Improved models using dictionary

The training from the raw corpus may have some limitations due to the size of the corpus. If the corpus is too small, the possibility of useful phrases are not learned when building phrase table. However, if corpus is too larger could in excess. In addition, we used the automatic segment of phrase tool so that it can be some errors in the analysis. We added Vietnamese dictionary of compound and syllable word into the phrase translation table and we assigned weight 1 into the each word, we solved this problem. Results as following:

System	BLEU score			
	Corpus 50.000	Corpus 100.000	Corpus 150.000	Corpus 200.000
MT_VR *(Baseline Vietnamese restoration)*	0.9744	0.9800	0.9830	0.9848
MT_VR+Dict *(Baseline Vietnamese restoration combine Dictionary)*	0.9748	0.9803	0.9832	0.9850

Table 2: The accuracy of experiment systems

Training	MT_VR		MT_VR+Dict	
	Complete correct sentences	Accuracy(%)	Complete correct sentences	Accuracy(%)
50.000	4120	82.40%	4533	90.66%
100.000	4291	85.82%	4558	91.16%
150.000	4300	86.00%	4585	91.70%
200.000	4352	87.04%	4628	92.00%

Table 3: The accuracy of experiment systems

4.2.2 Improved model using n-gram level changes

Changing n-gram level, the increased level of language model improved translation results. However, with level 4 or higher then the results has not been almost changed. Because in the Vietnamese language, the phrases including 3 or 4 words are more related than each other. The result with the weight assigned to 1 and level of language model 4: BLEU score=0.9850; accuracy when using MT_VR combine dictionary information: 92%.

Corpus	BLUE score		
	MT_VR	MT_VR+Dict 3ngram	MT_VR+Dict4ngram
50.000	0.9744	0.9748	0.9743
100.000	0.9800	0.9803	0.9830
150.000	0.9830	0.9832	0.9826
200.000	0.9848	0.9850	0.9844

Table 4: Results with different n-grams levels

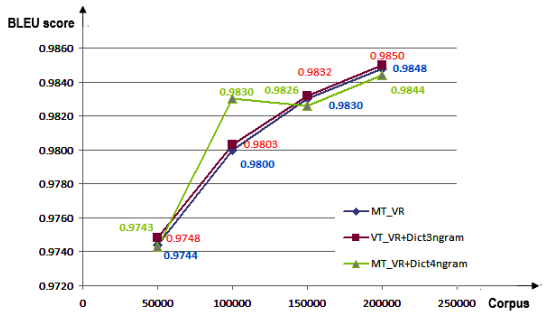

Figure 5: Compare BLUE score with experiment systems

Table 3 shows experimental results with different training corpus. Experimental results show that using MT_VR combine dictionary infomation with 3-gram have the best result. Table 4 and Figure 5 show that training with 200.000 pairs of sentence will have the best accuracy accents prediction.

Accentless Vietnamese	MT VR+Dict	viAccent	Reference Sentences
trong khu vuc choi game chinh , ban se su dung chuot de click vao bat ky nhan vat nao ban muon chon (co the chon mot luc mot hoac nhieu nhan vat den toi da la 9) .	trong khu vực chơi game chính , bạn sẽ sử dụng chuột để click vào bất kỳ nhân vật nào bạn muốn chọn (có thể chọn một lúc một hoặc nhiều nhân vật đến tối đa là 9) .	trong khu vực chơi game chính, bạn sẽ sử dụng chuột để click vào bất kỳ nhân vật nào bạn muốn chọn(có thể chọn một lúc một hoặc nhiều nhân vật đến tôi đã là 9).	trong khu vực chơi game chính , bạn sẽ sử dụng chuột để click vào bất kỳ nhân vật nào bạn muốn chọn (có thể chọn một lúc một hoặc nhiều nhân vật đến tối đa là 9) .
Tiep theo , ban se co ba cach khac nhau de dieu khien cac nhan vat nay .	tiếp theo , bạn sẽ có ba cách khác nhau để điều khiển các nhân vật này .	tiếp theo, bạn sẽ có ba cách khác nhau để điều khiển các nhân vật này.	tiếp theo , bạn sẽ có ba cách khác nhau để điều khiển các nhân vật này .
cach thu nhat : su dung cac nut de ra lenh cho nhan vat hanh dong bang cach nhap chuot trai vao chung .	cách thứ nhất : sử dụng các nút để ra lệnh cho nhân vật hành động bằng cách nhấp chuột trái vào chúng .	cách thứ nhất: sử dụng các nút đề ra lệnh cho nhân vật hành động bằng cách nhấp chuột trái vào chung.	cách thứ nhất : sử dụng các nút để ra lệnh cho nhân vật hành động bằng cách nhấp chuột trái vào chúng .
cach thu hai : ngoai ra , de tiet kiem thoi gian , co the nhan thang phim tat cua cac nut .	cách thứ hai : ngoài ra , để tiết kiệm thời gian , có thể nhấn thẳng phím tắt của các nút .	cách thứ hai: ngoài ra, để tiết kiệm thời gian, có thể nhận thẳng phim tật của các nút.	cách thứ hai : ngoài ra , để tiết kiệm thời gian , có thể nhấn thẳng phím tắt của các nút .
muon biet cac phim tat , ban hay de chuot tren nut tuong ung , lap tuc duoi day man hinh se xuat hien mot dong chu thong bao cho y nghia cua nut va phim tat .	muốn biết các phím tắt , bạn hãy để chuột trên nút tương ứng , lập tức dưới đáy màn hình sẽ xuất hiện một dòng chữ thông báo cho ý nghĩa của nút và phím tắt .	muốn biết các phim tật, bạn hãy để chuột trên nút tương ứng, lập tức dưới đây màn hình sẽ xuất hiện một dòng chữ thông báo cho ý nghĩa của nút và phim tất.	muốn biết các phím tắt , bạn hãy để chuột trên nút tương ứng , lập tức dưới đáy màn hình sẽ xuất hiện một dòng chữ thông báo cho ý nghĩa của nút và phím tắt .
cach thu ba : ngoai cach nhap chuot vao cac nut , chung ta co the su dung phim phai cua chuot nhu mot cach dieu khien tat .	cách thứ ba : ngoài cách nhấp chuột vào các nút , chúng ta có thể sử dụng phím phải của chuột như một cách điều khiển tắt .	cách thứ ba: ngoài cách nhấp chuột vào các nút, chúng ta có thể sử dụng phim phải của chuột như một cách điều khiển tật.	cách thứ ba : ngoài cách nhấp chuột vào các nút , chúng ta có thể sử dụng phím phải của chuột như một cách điều khiển tắt .
cai ban nay hinh ban nguyet.	cái bàn này hình bán nguyệt	cái bản này hình ban nguyệt	cái bàn này hình bán nguyệt
toi nay toi di choi.	tối nay tôi đi chơi	tới nay tôi đi chơi	tối nay tôi đi chơi
toc do truyen thong se tang cao	tốc độ truyền thông sẽ tăng cao	tốc độ truyền thống sẽ tăng cao	tốc độ truyền thông sẽ tăng cao

Table 5: Accent prediction of some sentences

4.2.3 Comparison with other methods

We also compared our method with viAccent system(Truyen et al., 2008) because viAccent is the newest and efficient method for Vietnamese accent prediction. We conducted the experiment with the same test corpus (5000 sentences) for viAccent. Bleu scores of both MT_VR+Dict and viAccent system were showed on Table 6.

System	BLEU score
MT_VR+Dic	0.9850
viAccent	0.8875

Table 6: Compared our method with viAccent system

5 Conclusion

The experimental results showed that our approach achieves significant improvements over viAccent system. Performance of accent prediction with our method achieves better accuracy than that and some examples in test corpus was showed on Table 5.

In this paper, we introduced the issues of accents prediction for accentless Vietnamese texts and proposed a novel method to resolve this problem. The our idea is based on Phrase-based Statistical Machine Translation to develop a Vietnamese text accent restoration system.

We combined the advantage of previous approach such as n-gram languages model and phrase dictionary. In general, experimental results showed that our approach achieves promised performance. The quality of accents prediction can be improved if we have a better corpus or assigned appropriate weight to dictionary.

6 Acknowledgments

This work is funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2011.08

References

Phuong-Le Hong, Huyen-Nguyen Thi Minh, Azim Roussanaly, and Vinh-Ho Tuong. 2008. *A Hybrid Approach to Word Segmentation of Vietnamese Texts*. In Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, Springer, LNCS 5196.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Proceedings of HLT-NAACL 2003, pages 127–133. Edmonton, Canada.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. Proceedings of ACL, Demonstration Session.

Phan Quoc Lan. 2005. *Approach to add accents for Vietnamese text without accent*. Informatics Bachelor's thesis, VietNam National University of Ho Chi Minh City.

Thai Phuong Nguyen, Akira Shimazu, Tu Bao Ho, Minh Le Nguyen, and Vinh Van Nguyen. 2008. *A tree-to-string phrase-based model for statistical machine translation*. In Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL 2008), pages 143–150, Manchester, England, August. Coling 2008 Organizing Committee.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: A Method for Automatic Evaluation of Machine Translation*. ACL.

Nguyen Quan. 2002. *VietPad*. http://vietpad .sourceforge.net.

A. Stolcke. 2002. *"Srilm - an extensible language modeling toolkit," in Proceedings of International Conference on Spoken Lan-guage Processing*.

Tran Triet Tam. 2008. *AMPad*. http://www.echip .com.vn/echiproot/weblh/qcbg/duynghi/automark.

Nguyen Van Toan. 2008. *VnMark*.

Tran The Truyen, Dinh Q. Phung, and Svetha Venkatesh. 2008. *Constrained Sequence Classification for Lexical Disambiguation*. In Proceedings of PRICAI 2008, pages 430–441.