Patrick BELLOT, Vu DUONG, Marc BUI (Editors)

Proceedings of the
4th IEEE International Conference on Computer Sciences
Research, Innovation & Vision for the Future

February 12-16, 2006 Ho Chi Minh City, Vietnam.

Proceedings of the 4th IEEE International Conference on
Computer Sciences Research, Innovation and Vision for the Future

Additional copies of this publication are available from

IEEE Operations Center
P. O. Box 1331
445 Hoes Lane
Piscataway, NJ 08855-1331 USA

+1 800 678 IEEE
+1 732 981 1393
+1 732 981 0600
+1 732 981 9667 (FAX)
email: customer.service@ieee.org

PREFACE

Welcome to the RIVF'06 Conference!

This year there were 104 qualified submissions to RIVF'06 and 34 acceptances, for an acceptance rate of just 33%. As RIVF is still a young conference, we had to face a very difficult decision to ensure the quality of the conference and to encourage participations from young researchers. Beside these 34 papers, selected as long papers to be included in our IEEE Proceedings, 26 papers were accepted as short papers to be included in the Conference's Addendum Contributions, totaling an overall acceptance rate of less than 60%. All selected papers, long and short, are of high quality, and we are very proud of the professionalism of all authors, reviewers and program committee members. Thank you so much for your contributions.

This was also the first year in which electronic submission of all materials was supported. 100% of submissions were made electronically, and the referee process was fully performed on-line.

The proceedings and addendum contributions you are handling are the result of much hard work from many people. We would like to thank:

- The authors and co-authors of the paper submissions. They are, of course, what makes the RIVF Conference Program great.
- The RIVF'06 Scientific Program Committee. There were 72 committee members, over the half of whom were serving on the committee for just the first or second time. Youthful enthusiasm was complemented by the experience of several veterans to whom we are especially grateful.
- The tertiary reviewers, who often supply the most expert and informed comments on their review.
- The logistic team: Long Duc Pham, Minh-Dung Dang, Frederic Ferchaud, Lien Pham who worked hard to ensure the on-line processes, and to compile and edit the final proceedings.
- The Steering Committee members who helped with some difficult decisions.
- The Local Organizing Committee members and volunteers, for the local arrangements, the design of the cover pages, and for the printing of the Proceedings and the Addendum Contributions.
- The various institutions that provided the support for the paper process. The list includes the employers of all the reviewers and committee members. IEEE Region X, in particular Prof. Cheng Tee Hiang, helped not only with the technical co-sponsoring but also the reviews of the submissions; Ecole Nationale Superieure des Telecoms Paris hosts the website; Eurocontrol Joint Research Lab CSMC supports the IEEE Proceedings. In addition, the following companies and universities provided financial support: Ho Chi Minh City University of Technology (HCMUT), Groupement des Ecoles de Telecoms (GET), EPITA, Quaternove S.A., and IFI-Solution Vietnam. This fund will be used to sponsor students from Vietnam to be attending RIVF'06.

Next year conference will take place at Hanoi University of Technology at the same time frame. Your contributions will be again more than expected.

Thank you all again, authors and committee members and reviewers, for your contribution to RIVF'06 that surely be a success.

Patrick Bellot, Conference Chair.
Marc Bui, Publication Chair.
Vu Duong, Program Chair.

Conference Co-Chairs:

George Donohue – George Mason University, USA
Patrick Bellot – Ecole Nationale Superieure des Telecoms, France
Dinh-Tri Nguyen – Institut de la Francophonie pour l'Informatique, Vietnam

Program committee:

Romain Alléaume - Ecole Nationale Superieure des Telecoms, France
Giovanni Andreatta – University of Padova, Italy
Alexandre d'Aspremont –Princeton Universit., USA
Philippe Baptiste – Ecole Polytechnique, France
Leopoldo Bertossi - Carleton University, Ottawa, Canada
Lorne Bouchard – University of Quebec at Montreal, Canada
Marc Bouisset - University of Quebec at Montreal, Canada
Jean-Pierre Briot – Laboratoire Informatique de l'Univeriste Paris 6, France
Peter Brucker – University of Osnabrueck, Germany
Marc Bui – Ecole Pratique des Hautes Etudes, France
Tru Hoang Cao - HCM University of Technology, Vietnam
Jacques Carlier – University of Technology of Compiegne, France
Edwin Cheng - Polytechnic University, HongKong
Gérard Cohen - Ecole Nationale Superieure des Telecoms, France
Stefan Conrad – University of Düsseldorf, Germany
Kevin Corker – San Jose State University, USA
Marc Dacier - Eurecom, France
Wolfgang Deiters - Fraunhofer ISST, Germany
Akim Demaille, EPITA, France
An-Hai Doan – University of Illinois at Urbana Champaign, USA
Bich-Thuy T. Dong – HCM University of Natural Sciences, Vietnam
Alexis Drogoul – University Paris 6 & IRD, France
Nicolas Durand – Ecole National de l'Aviation Civile & CENA, France
Duc Anh Duong - HCM University of Natural Sciences, Vietnam
Vu N. Duong - Program Chair - Eurocontrol, EU
Henrik Eriksson - Linköping University, Sweden
Laurent El Ghaoui – University of California at Berkeley, USA
Louisette Emirkanian - University of Quebec at Montreal, Canada
Roberto Gomez - ITESM-Mexico, Mexico
Oliver Günther - Humboldt-University Berlin, Germany
Quoc Trung Ha – Hanoi University of Technology, Vietnam
Willi Hasselbring – University of Oldenburg, Germany
Solange Ghernaouti-Hélie - University of Lausanne, Switzerland
Bao Tu Ho – Japan Advanced Insittute for Science and Technology, Japan
Vinh Tuong Ho - Institut de la Francophonie pour l'Informatique, Vietnam
Siu Cheung Hui – Nanyang Technical University, Singapore
Toshihide Ibaraki - Kwansei Gakuin University, Japan
Toru Ishida – University of Kyoto, Japan
Francois Jouen - Ecole Pratique des Hautes Etudes, France
Ernst Kesseler - NLR, Netherlands
Peter Kropf - Neuchâtel University, Switzerland
Pierre Kuonen - Friboug University, Switzerland
Ralf Detlef Kutsche – Technical University of Berlin, Germany
Ivan Lavallee - University Paris 8, France
Jonathan Lawry - University of Bristol, UK
Thang Q. Le - Can-Tho University, Vietnam
Marie-Noelle Lepareux, Thalès Group, France
Marc Lobelle - University Catholic de Louvain, Belgium

John Lygeros – University of Patras, Greece
Quang-Tuan Luong - SRI, USA
Wendy Mackay - INRIA, France
Jean-Frédéric Myoupo - Amiens University, France
David Naccache – Ecole Normale Superieure, Paris, France
Hung Q. Ngo - University of Buffalo, USA
Nhan Thanh Ngo – New York University, USA
Hoang-Lan Nguyen - Hanoi University of Technology, Vietnam
Hong Quang Nguyen - Institut de la Francophonie pour l'Informatique, Vietnam
Phong Nguyen - Ecole Normale Superieure, Paris, France
Son Thanh Nguyen - HCM University of Technology, Vietnam
Thuy Thanh Nguyen - Hanoi University of Technology, Vietnam
Cong Duc Pham - INRIA RESO/LIP/CNRS/UCB Lyon, France
Dinh-Dieu Phan – Vietnam National University at Hanoi, Vietnam
Alain Pirotte - University Catholic de Louvain, Belgium
Michel Riguidel - Ecole Nationale Superieure des Telecoms, France
Kurt Sandkuhl - Jönköping University, Sweden
Alexander Smirnov - SPIIRAS, Russia
Monica Tavanti - DeepBlue, Italy
Nam Thoai - HCM University of Technology, Vietnam
Lang Van Tran – Institute of Information Technology, Vietnam
Van Hoai Tran - HCM University of Technology, Vietnam
Hong Linh Truong - University of Innsbruck, Austria
Putchong Uthayopas - Kasetsart University, Thailand
Van Vu – University of California at San Diego,USA
Anders Ynnerman - Linkoping University, Sweden
Vilas Wuwongse – Asian Insitute of Technology, Thailand
Yakov Zinder – University of Technology of Sydney, Australia

Steering Committee:
Marc Bui - EPHE, France
Patrick Bellot - ENST, France
Tee Hiang Cheng - IEEE Region X/NTU, Singapore
Vu N. Duong - EEC, Europe
Nhan Thanh Ngo - NYU, USA
Son Thanh Nguyen - HCMUT, Vietnam

Local organization Committee:
Anh-Vu Dinh-Duc - HCMUT, Vietnam
Son T. Nguyen - IFI Solution, Vietnam
Huan Viet Tran - IBM, Vietnam

Publications & Tutorial chair:
Marc Bui - EPHE, France

TABLE OF CONTENTS

A Maximum Entropy Approach for Vietnamese Word Segmentation

Dinh Dien, Vu Thuy

Abstract **In this paper, we introduce a new approach for Vietnamese Word Segmentation. The word segmentation problem is restated into the morpho-syllable position-in-word (PIW) tagging problem. We used the Maximum Entropy with the Generalized Iterative Scaling (GIS) to train on the annotated corpora. The result of the training process was used to tag all the morpho-syllables of the input sentence. With the output sentence tagged, we can convert it into a segmented sentence for evaluation. The results on a lot of tagged-corpora show that this approach is suitable for Vietnamese Word Segmentation. The performance achieves precision and recall rates of 94.87% and 94.08% respectively, and the F-measure of 94.44%.**

Index Terms **word segmentation, maximum entropy**

I. Introduction

WORD SEGMENTATION is an important step in every natural language processing system, especially in some isolated Oriental languages. In these languages, the separators between each word are not available. So far, a great variety of strategies for the word segmentation problem have been explored, yielding a large volume of literature on both linguistic and computational sides. In general, these strategies can be divided into two directions, namely, dictionary-based and statistical-based approaches.

The dictionary-based approach usually uses some basic mechanical segmentation methods based on string matching. The most common method in this approach is *maximum matching* method. In order to increase the performance of this method, some heuristics are included. The simple *maximum matching* algorithm gives only one result.

The pure statistical-based approach has a lot of drawbacks because it does not use any dictionary and the word boundaries are decided only by the statistics of the corpora.

In recent times, the combination of these two approaches is usually used, so called the statistics-aided approach. In general, this involved the evaluation among possible segmentations and choosing the best case by some statistical models with a dictionary. One of these approaches for the Vietnamese word segmentation is [1]. This is the combination of WFST (Weighted Finite State Transducer) and Neural Network using Vietnamese dictionary.

Manuscript received October 12, 2005.

Dinh Dien and Vu Thuy are with the Faculty of Information Technology, University of Natural Sciences, Vietnam National University, Ho Chi Minh City, Vietnam.

(e-mail:ddien@fit.hcmuns.edu.vn; email:vthuy@fit.hcmuns.edu.vn).

In this paper, we introduce a statistical model which trained from an annotated corpus with word boundary tags and uses a Vietnamese dictionary to find some possible segmentation cases of an input Vietnamese sentence.

The annotated corpus we use in this paper is the result of project [2]. With the average balance F-measure reaching 98.59%, we can certainly use the list of Vietnamese tagged sentences as a training corpus for the Maximum Entropy model. This Maximum Entropy model has been effectively applied to the Part Of Speech Tagger of Adwait Ratnaparkhi [3]. The state-of-the-art accuracy of this tagger (96.6%) shows that Maximum Entropy is a suitable model for tagging problem.

This paper also describes briefly the Maximum Entropy properties of the model. Finally, the results in this paper are compared to some previous models for Vietnamese Word Segmentation.

II. Vietnamese Morphology

To define exactly what is a word is not a simple matter. In linguistics, hundreds of definitions of the word have been brought out. But there is no definition can embrace all aspects of words. Therefore, by the goal of automatic processing natural language, the satisfaction of the definitions about word in general linguistics, and the specific characteristics of isolated language like Vietnamese, we use the viewpoint of Dinh Dien s thesis [4]:

A. Vietnamese Morpheme

Following the idea in [4], morpho-syllable is the basic unit in Vietnamese, because it can be identified easily by the native speakers, and also automatically by computer.

B. Vietnamese Word

In this paper, we use the word definition in [4]: *A Vietnamese word is composed of Vietnamese morphemes* .

Vietnamese words include *single words*, *compound words*, *duplicative words*, and *fortuitous concurrence words*.

Beginning with the demand of automatic processing Vietnamese corpus by the computer, Dinh Dien [4] has offered these methods to formalize the conception about the Vietnamese morphemes and Vietnamese words as follows:

1) Because a Vietnamese morpheme is also a *dictation word*[1] (separated syllable), the method to formalize Vietnamese morpheme is very simple. In English corpus

[1] Following [4], a Vietnamese morpho-syllable is a *dictation word*.

or Vietnamese corpus, the basic units are also the *dictation words*.

2) To represent word boundaries, we use the definition *word in dictionary* (dictionary-word) in [4]. *Dictionary-word* is defined: *the unit that has been set into a dictionary based on the meaning feature and has been tagged as the unit of the language.* The selection of which word set in a dictionary depends on the linguistics, or the corpus builder decision, following the opinions above. In this paper, we use the Vietnamese dictionary of Hoang Phe [5].

III. Restating the Word Segmentation Problem

The word segmentation problem is restated to the morpho-syllable PIW tagging problem as follows:

Given a Vietnamese sentence $S = c_1 c_2 ... c_n$ with n morpho-syllables. We segment S by tagging the PIW tag t_i to each morpho-syllable. There are four PIW tags:

- LL: when the syllable is on the left of the word.
- RR: when the syllable is on the right of the word.
- MM: when the syllable is in the middle of the word.
- LR: when the syllable is the only syllable of the word (single word).

After all the morpho-syllables c_i in S have been tagged, we can convert this result to the word boundary tags by their positions in word.

For example:

S = *Tôi tư duy nghĩa là tôi tồn tại*

T = LR LL RR LL RR LR LL RR

= *Tôi # tư duy # nghĩa là # tôi # tồn tại*

IV. Vietnamese Word Segmentation Model

Our model for Vietnamese word segmentation is shown in **Figure 1**. The model includes the following components:

- Sentence preprocessing: standardize the input Vietnamese sentence to a unique form of spelling.
- Unknown word recognition: extract all the unknown word in the sentence.
- GIS training on tagged corpus: this process includes two sub-components:
  - Extracting all the features in the tagged corpus.
  - Using Generalized Iterative Scaling (GIS) to evaluate the model parameters.
- Evaluation using Maximum Entropy: choose the best segmentation case by using the Maximum Entropy.

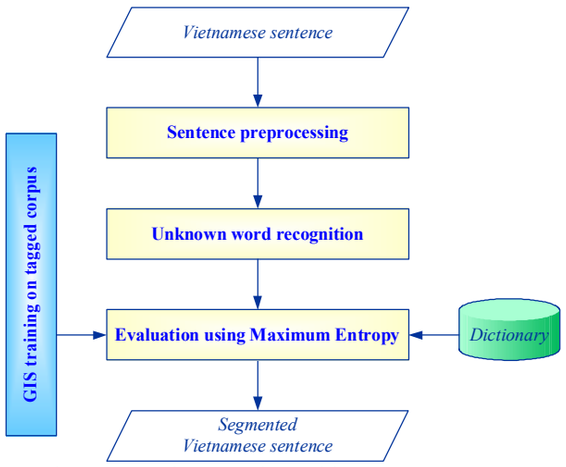


Figure 1. Vietnamese word segmentation model

A. Tagged corpus preparation

The corpus used in this model is extracted from the CADASA bilingual corpus, the result of the project in building corpus for Vietnamese English Machine Translation System [2]. Here are some statistics about this corpus:

Parameter	Value
Number of sentences	24,240
Number of pair-sentences	12,120
Number of English words	226,953
Number of different English words	4,555
Number of Vietnamese dictation words	229,357
Number of Vietnamese words	181,192
Average English sentence length	18.73 w/s
Average Vietnamese sentence length	14.95 w/s

Table 1. Some statistics about Cadasa corpus

B. Sentence preprocessing

In Vietnamese, the preprocessing plays an important role with the performance of the system. There are two steps in this process:

1) Spelling Standardization: Vietnamese has two types of variant spelling:
   a. Tone rule: there are two ways to put the sign of tone:
      i. Aesthetic rule: the sign is put in the middle phoneme. Ex: hòa
      ii. Phonetics rule: the sign is put in the main phoneme. Ex: hoà

      We choose the phonetics rule to standardize sign because of convenience. For example: hòa → hoà
   b. Letter variant: standardize letter variants to a unique form. The letter variant is a variant form of a morpho-syllable, but the sound and the meaning is the same.
      For example: thời kỳ → thời kì.
2) Atom Segmentation: to segment the sentence into atom-units (cannot be separated into smaller units). The atom

can be: morpho-syllable, sign, symbol, abbreviation, foreign string, factoids[2].

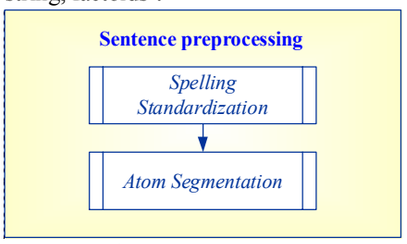


Figure 2. Sentence preprocessing

C. Unknown word recognition

Unknown word recognition is one of the two most important steps of a word segmentation system beside the ambiguity segmentation.

We classify the unknown word into three types:

1) Vietnamese proper name:
   a. Person names: *Nguyễn Văn Tuấn*, *Trần Thị Tâm*,
      With this type, we use a list of proper names in the GIS training. Because of the flexibility of the Maximum Entropy features, we can easily use this list a part of the training corpus.
   b. Place names: *Hà Nội*, *Sài Gòn*,
      With this type, we use the Gazetteer dictionary of the place name to recognize them. Our Gazetteer dictionary is built manually and has about 1,500 names of places.
2) Foreign proper names: in Vietnamese documents, we can easily recognize these proper name, because all the tokens in a name are usually not in the dictionary.
3) Factoids: being recognized in the preprocessing step.

Figure 3. Unknown word recognition

D. The probabilistic model

The probabilistic model is defined over $H \times T$ where H is the set of possible contexts or histories of an item, and T is the set of possible tags of an item. The model s joint probability of history h and a tag t is defined as:

$$p(h,t) = \pi \prod_{j=1}^{k} \alpha_j^{f_j(h,t)}$$

Where:

- π: normalization constant.
- $\{\alpha_1, ..., \alpha_k\}$: model parameters.
- $\{f_1, ..., f_k\}$: model features, $f_j(h,t) \in \{0,1\}$

Each feature f_j has its own parameter α_j.

In the training process, given a sequence of morpho-syllables $\{c_1, ..., c_n\}$ and their PIW tags $\{t_1, ..., t_n\}$ as training data, the aim is finding the parameters $\{\alpha_1, ..., \alpha_k\}$ that maximize the likelihood of the training data:

$$L(p) = \prod_{i=1}^{n} p(h_i, t_i) = \prod_{i=1}^{n} \pi \prod_{j=1}^{k} \alpha_j^{f_j(h_i, t_i)}$$

E. GIS training on tagged corpus

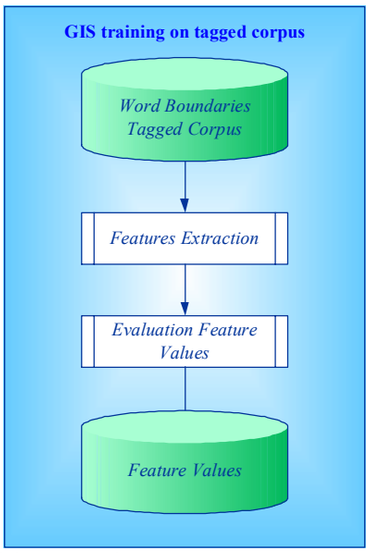
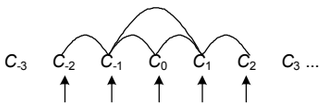


Figure 4. GIS training on tagged corpus

From a segmented corpus, we can produce the PIW corpus with perfect accuracy. After that, we extract the features of each morpho-syllable in the corpus.

The performance of the model depends on the way we select the features. Given (h,t), a feature has to encode every information that can help to get the tag t. In this paper, we use 5 main features as below:

- Current morpho-syllable.
- The previous and the next two morpho-syllables
- The previous, the next, and the current character morpho-syllables, the previous two morpho-syllables and the next two morpho-syllables.
- The previous and the next morpho-syllables.
- The tag of the previous two morpho-syllables before the current morpho-syllables.



[2] Factoid is a string that represents special information. The factoids in this paper are date, time, percent, money, number, measure, email, phone, and website.

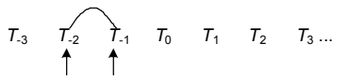


Figure 5. Features used in the system

Some of these features are similar to the features used in [6] for Chinese. To introduce more about these 5 features, we will examine a Vietnamese sentence:

$S =$ *tôi tư duy nghĩa là tôi tồn tại*

Morpho-syllable	tôi	tư	duy	nghĩa	là	tôi	tồn	tại
Tag	LR	LL	RR	LL	RR	LR	LL	RR

The features for *nghĩa* are:

- $C_0 =$ *nghĩa* & $t_i = LL$
- $C_{-2}C_{-1}C_1C_2 =$ *tư duy là tôi* & $t_i = LL$
- $C_{-2}C_{-1}C_0C_1C_2 =$ *tư duy nghĩa là tôi* & $t_i = LL$
- $C_{-1}C_1 =$ *duy là* & $t_i = LL$
- $T_{-2}T_{-1} = LL\ RR$ & $t_i = LL$

The model parameters are obtained via *Generalized Iterative Scaling* [7]. Besides the five features above, we have to use two more features to ensure that the iterative converges. The first feature, the *default feature*, will have value 1 when all other features have values 0. The second feature, the *correction feature*, will have a special value (can be greater than 1) defined by:

$$C = \max_{x \in E} \sum_{i=1}^{k} f_i(h,t)$$

In general, adding new features can affect the system. However, these two new features are completely dependent on other features. They add no new information, and therefore place no new constraints on the model. As a result, they will not affect the system.

F. Evaluation using Maximum Entropy

With the list of atoms and unknown words, we use a recursion based on the *maximum matching* algorithms to find *k* best cases for each sentence. The value of *k* depends on the length of the sentence. By experiment, we choose the value for *k* by the table below:

Length	<10	11-20	21-30	31-40	>40
k value	3	5	8	12	20

After the training process, the features and their corresponding parameters will be used to calculate the probability of the tagged sequence of a sentence input. For each case of segmentation, we have a sequence of morpho-syllables and a sequence of tags. With a morpho-syllable sequence $\{c_1,...,c_n\}$, the system will give the tag sequence $\{t_1,...,t_n\}$ with the highest probability:

$$P(t_1,...,t_n|c_1,...,c_n) = \prod_{i=1}^{n} P(t_i|h_i)$$

The conditional probability for each (h,t) is calculated by:

$$P(t|h) = \frac{p(h,t)}{\sum_{t' \in T} p(h,t')}$$

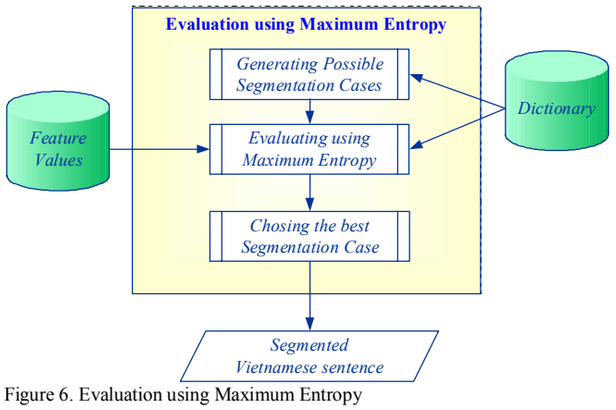


Figure 6. Evaluation using Maximum Entropy

V. RESULT

To evaluate the result of the system, we use Bakeoff s evaluation in [8]:

- Recall (R): number of correct hits divided by the number of items that should have been selected.

$$R = \frac{N_3}{N_1}$$

- Precision (P): number of correct hits divided by the total number of items selected.

$$P = \frac{N_3}{N_2}$$

- Balance F-measure (F):

$$F = \frac{(1+\beta)PR}{\beta P + R}$$

- Recall on Out Of Vocabulary (R_{OOV}): number of correct unknown word hits divided by the number of unknown word that should have been selected.

$$R_{OOV} = \frac{N_{31}}{N_{11}}$$

- Recall on In Vocabulary (R_{IV}): number of correct known word hits divided by the number of known word that should have been selected.

$$R_{IV} = \frac{N_{32}}{N_{12}}$$

Where:

- N_3: number of correct word hits.
- N_{31}: number of correct known word hits.
- N_{32}: number of correct unknown word hits.

- N_2 : number of words recognized by the system.
- N_1 : number of words in the corpus.
- N_{11} : number of known words in the corpus.
- N_{12} : number of unknown words in the corpus.

The result of the system tested on the corpus of Vietnamese Lexicography Center (www.vietlex.com.vn) are shown below. The F measure reaches 94.44% on the total corpus.

No	Name	N_1	N_2	N_3	R	P	F	OOV	R_{oov}	R_{iv}
1	ChTinhTLRD1	14302	14232	13563	94.83	95.30	**94.79**	3.43	53.76	95.75
2	ChTinhTLRD2	12510	12402	11812	94.42	95.24	**94.83**	3.25	77.14	95.52
3	HoangTuBe	15643	15537	14843	95.47	95.53	**95.50**	1.26	39.60	95.84
4	Luocsuthoigian	10695	10551	9947	93.00	94.27	**93.63**	3.70	73.76	94.25
5	Congnghe	1036	1026	972	94.18	94.73	**94.46**	4.53	81.39	94.74
6	MuoiCuaRung	3143	3082	2978	94.78	96.62	**95.69**	3.62	90.35	94.78
7	NBaihocNT	6688	6651	6148	91.92	92.43	**92.17**	5.84	65.98	93.28
8	**Summary**	**64017**					**94.44**			

Table 2. The result of the system tested on the corpus of Vietnamese Lexicography Center

The result is compared with some other Vietnamese word segmentations:

No	Name	N_1	N_2	N_3	R	P	F	OOV	R_{oov}	R_{iv}
1	ChTinhTLRD1	14302	14297	11640	81.39	81.41	**81.40**	3.43	5.70	84.08
2	ChTinhTLRD2	12510	11640	10348	82.71	82.53	**82.62**	3.25	3.93	85.38
3	HoangTuBe	15643	15294	14072	89.96	92.01	**90.97**	1.26	75.75	90.14
4	Luocsuthoigian	10695	10559	9490	88.73	89.87	**89.30**	3.96	43.93	90.45
5	Congnghe	1036	1022	1022	92.27	93.54	**92.91**	4.53	42.55	94.64
6	MuoiCuaRung	3143	3094	2894	92.07	93.53	**92.80**	3.62	22.81	92.80
7	NBaihocNT	6688	6651	5843	87.36	87.85	**87.60**	5.84	3.32	92.58
8	**Summary**	**64017**					**88.28**			

Table 3. The result of Vietnamese Word Segmentation using Maximum Matching

No	Name	N_1	N_2	N_3	R	P	F	OOV	R_{oov}	R_{iv}
1	ChTinhTLRD1	14302	14154	13033	91.12	92.08	**91.60**	3.43	53.76	92.45
2	ChTinhTLRD2	12510	12372	11782	94.18	95.23	**94.70**	3.25	77.14	94.75
3	HoangTuBe	15643	15286	14691	93.92	96.11	**95.00**	1.26	82.32	94.06
4	Luocsuthoigian	10695	10533	9924	92.79	94.21	**93.49**	3.70	78.03	93.36
5	Congnghe	1036	1019	979	94.49	95.48	**95.27**	4.53	61.70	96.05
6	MuoiCuaRung	3143	3091	2974	94.62	96.21	**95.41**	3.62	90.35	94.78
7	NBaihocNT	6688	6634	6132	91.68	92.43	**92.05**	5.84	65.98	93.28
8	**Summary**	**64017**					**93.93**			

Table 4. The result of Vietnamese Word Segmentation using MMSEG rules [9]

VI. Conclusions and Future work

A. Conclusions:

The result shows that Maximum Entropy is suitable for the tagging problem. The performance of the word segmentation system is superior to other previous systems. Besides, this word segmentation is quite effective because it can recognize more correct unknown word, especially Vietnamese person names.

B. Future work:

There are many ways to increase the accuracy of the system. Firstly, because this model is the supervised machine-learning model, we can improve the performance of the system by using much more training data. Secondly, because the features impacts directly to the performance, we can apply linguistics

knowledge about the context of a word to obtain more useful features to the system.

REFERENCES

[1] D.Dien, H.Kiem, and N.V. Toan, Vietnamese Word Segmentation, Proceedings of NLPRS 01 (The 6th Natural Language Processing Pacific Rim Symposium), Tokyo, Japan, 11/2001, pp.749-756. 2001.

[2] D.Dien, H.Kiem, Building a training corpus for word sense disambiguation in the English Vietnamese bilingual corpus, Proceedings of Workshop on Machine Translation in Asia, COLING-02, Taiwan, 9/2002, pp.26-32. 2002.

[3] Adwait Ratnaparkhi, A Maximum Entropy Part-Of-Speech Tagger , In Proceedings of the Empirical Methods in Natural Language Processing Conference, University of Pennsylvania. 1996.

[4] Dinh Dien, Building an English Vietnamese Bilingual Corpus. Ph.D thesis in Comparative Linguistics, University of Social Sciences and Humanity of HCM City, Vietnam. 2005.

[5] Hoang Phe, Vietnamese Dictionary , Vietnam Lexicography Centre, Da Nang Publishing House.

[6] Nianwen Xue, Chinese Word Segmentation as Character Tagging , Computational Linguistics and Chinese Language Processing. 2003.

[7] J. N. Darroch and D.Ratcliff, Generalized Iterative Scaling for Log-Linear Model, The Annals of Mathematical Statistics, 43(5):1470-1480. 1972.

[8] Richard Sproat, Thomas Emerson, The First International Chinese Word Segmentation Bakeoff , In Proceedings of the second SIGHAN Workshop on Chinese Language Processing. ACL2003. 2003.

[9] Chih-Hao Tsai, MMSEG: A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm, 2000.

Dinh Dien was born on 18th January, 1966 in Saigon, Vietnam. He received the B.Sc. degree in Physics from the University of HCMCity in 1988, the Engineer degree in Electronics Techniques from the Polytechnics University of HCMC in 1993, the M.Sc. degree in Computer Science from the University of Natural Sciences in 1996, Ph.D. degree in Computer Science in 2003 under the co-supervision of Prof. Hoang Kiem (University of Natural Sciences) and Prof. Eduard Hovy (University of South California, ISI). He also received the M.A. degree in Comparative Linguistics in 2001 and Ph.D. degree in Linguistics in 2005 from the University of Social Sciences & Humanity, VNU-HCMC.

He took part the localization project of Vietnamese Windows 95 at Microsoft Corp. (Redmond, WA, USA) in 1996. He won the Gold Medal from the Vietnam Central Committee of the Blind and the Young Scientist Award from Prime Minister of Vietnam for his benevolent works for Vietnamese blinds (such as: the Braille English-Vietnamese Dictionary, Vietnamese Text-to-Speech for the blind, etc.). He is member of ACL (Association for Computational Linguistics).

His current research focuses on the Vietnamese Computational Linguistics, the English-to-Vietnamese Machine Translation and English-Vietnamese parallel corpora. He published a textbook titled Natural Language Processing (VNU-HCMC publisher). His articles in Machine Translation have been chosen as typical articles in English-Vietnamese Machine Translation by IAMT (International Association for Machine Translation) at website of MT-archives (www.MT-Archive.info).

At present, Dr. Dien is a lecturer of Computer Science and Computational Linguistics at University of Natural Sciences, VNU-HCMC. Dr. Dien is also deputy head of Natural Language Processing Department of Information Technology Faculty, University of Natural Sciences, VNU-HCMC.

Vu Thuy was born on January 4th, 1983 in Dong Nai province. He studied high school in Le Hong Phong High School in Ho Chi Minh City from 1998 to early 2001. He passed the entrance examination to the University of Natural Sciences, Vietnam National University of Ho Chi Minh City in 2001. From 2001 to 2005, he studied in Information Technology faculty with registered major on Knowledge Engineering department. In early 2005, he did the final thesis at the university named Morphological Tagger for English-Vietnamese Bilingual Corpus under the supervision of Dr. Dinh Dien at Vietnamese Computational Linguistics Group. He obtained a B.S with excellent rank in August, 2005.

Now he is a Teaching Assistant for Knowledge Engineering department of Information Technology faculty, University of Natural Sciences, Vietnam National University of Ho Chi Minh City. His research interests are Artificial Intelligent, Pattern Recognition and especially Natural Language Processing. He is currently a member of Vietnamese Computational Linguistics Group under the supervision of Dr. Dinh Dien.

Author Index