Journal of Computer Science and Cybernetics, V.30, N.2 (2014), 177–188

A MODEL FOR EXPLOITING THE TARGET LANGUAGE CHARACTERISTICS TO EXTRACT BILINGUAL BASE NOUN PHRASES

NGUYEN CHI HIEU

Faculty of Information Technology, Industrial University of Ho Chi Minh City;
nchieu@hui.edu.vn

Tóm tắt. Rút trích cụm danh từ song ngữ là một trong những bài toán quan trọng trong xử lý ngôn ngữ tự nhiên (NLP). Bài toán này càng trở nên khó khăn hơn với cặp song ngữ Anh-Việt do thiếu vắng nguồn tài nguyên tiếng Việt bao gồm các công cụ xử lý ngôn ngữ tự nhiên như treebanks, part-of-speech taggers, parsers và dữ liệu huấn luyện có chú giải. Trong bài báo này, chúng tôi đề xuất một mô hình tổ hợp sử dụng đặc tính ngôn ngữ đích để rút trích cụm danh từ song ngữ qua phương pháp chiếu trên kết quả đối sánh từ bằng phương pháp thống kê. Đặc tính ngôn ngữ đích được sử dụng trong mô hình này là phân đoạn từ, trật tự từ và phân lớp từ [1]. Mô hình của chúng tôi không những khắc phục được sự thiếu vắng nguồn tài nguyên cho xử lý ngôn ngữ tự nhiên tiếng Việt mà còn cải thiện được kết quả do đối sánh rỗng, đối sánh lỗi, vấn đề chồng chéo và xung đột của phương pháp chiếu. Mô hình đề xuất có thể được áp dụng cho các cặp ngôn ngữ khác. Thực nghiệm trên 66.646 cặp câu song ngữ Anh-Việt, mô hình đề xuất cho kết quả rất khả quan.

Từ khóa. Npbase, từ phân lớp, trật tự từ, NLP

Abstract. Bilingual Base Noun Phrase (BaseNP) extraction is one of the key tasks of Natural Language Processing (NLP). This task is more challenging for the pair of English-Vietnamese due to the lack of available Vietnamese language resources such as treebanks, part-of-speech taggers, and parsers. In this paper, we propose a combination model that uses language characteristics based on statistics and projection method to extract BaseNP correspondences from a bilingual corpus. The language characteristics used in this model include the word segmentation, word order and word classification [1]. Our model not only overcomes the lack of resources of Vietnamese but also improves the performance of miss-alignment, null-alignment, overlap and conflict projection of the existing methods. The proposed model can be easily applied to another language pairs. Experiment on 66,646 pairs of sentences in the English-Vietnamese bilingual corpus shows that our proposed model is very satisfactory.

Key words. Npbase, classifiers, word order, NLP.

1. INTRODUCTION

Natural language processing (NLP) is a research field that helps computer system to understand and process human language. Recently, many applications in NLP, such as information extraction, cross-language information retrieval, document summary, automatic question-answer and automatic machine translation, have strongly developed and brought practical

benefits. In these applications, base noun phrases (BaseNP) play an important role. Thus, monolingual and bilingual BaseNP extraction from the corpus attracts many researchers, for example: [2-5]. In [2], Kupiec used expectation maximum (EM) algorithm with hidden Markov model. In this algorithm, the author calculated the result only based on simultaneous appearance value and did experimentation with 2,600 English-French pairs of sentences in order to identify English-French BaseNP correspondence. In [3], Yarowsky proposed a new approach, which projected based on word alignment result and did experimentation with 40 pairs of sentences. However, the challenges of this approach are the null-alignment problem, overlap and conflict projection problem. In [4], E.Riloff and colleagues presented a new method for creating an information extraction system for the target language by exploiting the existing information extraction system (source) with the cross-language projection direction. This group did one way projection from English to French and used transfer learning in order to generate French rules. In [5], N.P.Thai used source syntax analysis program with probability and used Giza++ program to align English-Vietnamese word into English-Vietnamese machine translation. However, identification and extraction of Vietnamse noun phrases in particular and English-Vietnamese bilingual BaseNP in general are still open problems. These problems become more difficult when we lack resources for Vietnamese language processing, such as Vietnamese treebank, Vietnamese part of speech (POS) tagging (only obtaining the accuracy of 85% for Vietnamese POS tagging as the report of Nguyen Thi Minh Huyen in [6]) and the parser...

This paper presents a solution to overcome the lack of resources as mentioned above, based on the projection solution of Yarowsky, through a resource-rich language for natural processing such as English in order to indentify English-Vietnamese bilingual noun correspondence. In this solution, we propose "a model for exploiting the target language charateristics to extract bilingual base noun phrases". Target language characteristics used in this paper are the word segmentation, word order and word classification, extraction technique based on the result of word alignmment by projection approach with statistical method, that specifically applied hidden Markov model using open source software Giza++ [7].

Thus, the key point that affects the getting result with projection approach through word alignment is the result of English-Vietnamese word alignment process using Giza++ and the result of English syntax parsing. In English structure parsing, English POS tagging and BaseNP identification are quite complete and achieved high accuracy: Florian reached the accuracy of 96.87% in English POS tagging[8]; Tjong Kim Sang showed the result of English BaseNP indentification up to 94% [9]. However, word alignment had a modest result. Hwa [10] projected to obtain the POS label of Chinese using the result of word alignment with Giza++ for English-Chinese, the percentage of error is 40% - 50%. N. P. Thai and colleagues [5] used Giza++ to align for English-Vietnamese machine translation. The result is indirectly evaluated through English-Vietnamese machine translation with the accuracy of 36.79% to 47.16%.

In this paper, we propose solutions to improve the result of word alignment with Giza++ and reduce the percentage of error in the process of Vietnamese noun phrase correspondence indentification by exploiting the characteristics of Vietnamese classification word to "a model for exploiting the target language charateristics to extract bilingual base noun phrases". The proposed model can be applied to other pairs of languages. The experiment of this model was done on 66,646 pairs of bilingual English-Vietnamese sentences and achieved satisfactory results. The remain of this paper is organized as follows: section 2 presents the target language characteristics; the model for exploiting target language characteristics is presented in section 3; In section 4, experimental results are showed; and finally, section 5 is our conclusion.

2. TARGET LANGUAGE CHARACTERISTICS

Vietnamese is an isolating language, that is, each syllable is pronounced separately and displayed by a written word. This feature is evident in all aspects of pronunciation, vocabulary, grammar. A syllable is a base unit of meaningful units system of Vietnamese. From it, other lexical units are created to identify things, phenomena, etc by word combination or reiteration method. The creation of lexical unit using combination method is always dominated by semantic association rules, for example: *đất nước (land-water = nation), máy bay (fly-machine = airplane), nhà lầu xe hơi (house-floor steam-vihicle = building and car), nhà tan cửa nát (house-crumble door-ruined = broken family), etc.* Combination method is the main one in Vietnamese language. Thus, Vietnamse NLP systems have to go through a word segmentation step.

Vietnamese word segmentation

Word segmentation is a process that split a sentence into the smallest phrases (can be one syllable or some syllables with space bar apart) which have a particular meaning in dictionary and can be tagged with POS types so that they carry a particular grammar title. Different from English and some other European languages, Vietnamese words may include space bars. Vietnamese word can contain one syllable (monosyllabic) such as *đi (go), làm (work), ăn (eat), yêu (love), nhớ (miss)*, etc or two syllables such as *băn khoăn (anxious), lo lắng (worry), cá nhân (personal), hợp tác hóa (co-operative)*, etc. For this reason, Vietnamese word segmentation has its own characteristics [11].

Word order

Vietnamese words do not change their complexion. This characteristic will dominate other grammar features. When a word combines with other words to become some structures such as syntactic group or sentence, word and expletive order method is respected. The arrangement of words in a certain order is primary way to express syntax relationships. In Vietnamese, "*anh ta lại đến*" (he comes again) is different from "*lại đến anh ta*" (his turn again). When the words of the same POS types are combined following principal and accessory relation, the previous word keeps main role while the next word has the auxiliary. Rely on combination in order, " *củ cải*" (beet) is different from "*cải củ*" (white radish), "*tình cảm*" (sentiment) is different from "*cảm tình*" (sympathy).

There are few similarities for word order in English and Vietnamese, but basically they are very different. To regconize their differences, we use the research results of Vu Ngoc Tu [12] and Tuong Hung Nguyen [13] to build a model for "Transfer of English noun phrase syntax structure from Vietnamese". Details of this model are presented in [14].

Classifier (CL)

In Vietnamese, CL is used along with noun, located before noun, for example "cái" in "cái bút" (the pen), "con" in "con cá" (the fish), "chiếc" in "chiếc lá" (the leaf), "quyển" in "quyển sách" (the book), "tờ" in "tờ giấy" (the paper), "bức" in "bức thư" (the letter), etc. However, most of CL have no corresponding meaning in English and will be null-alignment, as in example 1(a). Thus, studies on CL in order to find a solution for identifying and extracting

noun phrase by computer is very necessary. In practical language, CL can be used as in example 1(a) or not used as in example 1(b).

Example 1:

(a)	cuốn/CL sách/NN *the/DT book/NN*	(b)	Tôi/PRP mua/VB sách/NN *I/PRP buy/VB the/DT book/NN*

CL does not usually appear alone in noun phrase, as in example 2.

Example 2:

con/CL trâu/NN hay/CC cái/CL nhà/NN ?/?

the/DT buffalo/NN or/CC the/DT house/NN ?/?

In some special cases, CL can appear alone in answer sentence, as in example 3(a), 3(b), 3(c), if the noun is determined in the question.

Example 3:

Anh/PRP cần/VBP cuốn/CL sách/NN nào/PRP ?/?

Which/WDT book/NN do/VBP you/PRP need/VBP ?/?

(a)	cuốn/CL (sách/NN) kia/DT	*that/DT one/PRP*
(b)	cuốn/CL (sách/NN) mới/JJ	*the/DT new/JJ one/PRP*
(c)	cuốn/CL (sách/NN) (mà/CC) anh/PRP vừa/RB mua/VB	*the/DT one/PRP you/PRP just/RB bought/VBD*

CL is divided into three categories: *unit-classifiers* as in example 1(a), example 2, *kind-classifiers* as in example 4, 5 and *event-classifiers* as in example 6, example 7.

Example 4:

(a)	hai/CD loại/CL chó/NN *two/CD kinds/NNS of/IN dogs/NNS*	(b)	hai/CD thứ/CL chanh/NN *two/CD kinds/NNS of/IN lemons/NNS*

Example 5:

(a)	hai/CD loại/CL đường/NN *two/CD kinds/NNS of/IN sugar/NN*	(b)	ba/CD thứ/CL sữa/NN *three/CD types/NNS of/IN milk/NN*

Example 6:

(a)	một/DT trận/CL mưa/NN *an/DT outburst/NN of/IN rain/NN*	(b)	một/CD cuộc/NN họp/VB *a/DT meeting/NN*

Example 7:

(a)	niềm/CL hạnh phúc/NN *feeling/NN of/IN happy/NN*	(b)	một/CD vụ/NN trộm/VB *a/DT housebreaking/NN*

The modifiers cannot be inserted between CL and central noun. It should be "*một cuốn sách mới*" instead of "*một cuốn mới sách*".

In Table 1, we use the word classification (Appendix A) to tag POS for English word in the first column (1), tag POS for Vietnamese in the second column (2). The third column (3) shows POS strings in Vietnamese noun phrase. NN is abbriviated for classification of noun.

CL is abbriviated for POS of classifiers. PL is abbriviated for POS of plural article such as "những", "các", "nhiều". Like this, CL appears immediately before central noun [CL-NN]. PL can be added to front of [CL-NN] to be [PL-CL-NN], but PL cannot stand immediately before noun.

Table 1. An example of CL in English-Vietnamese translation

English phrase (1)	*Vietnamese phrase (2)*	*Notes (3)*
I/PRP buy/VBP a/DT [book/NN]	Tôi/PRP mua/VB [sách/NN]	[NN]
[rare/JJ books/NNS]	[những/PL cuốn/CL sách/NN hiếm/JJ]	[PL CL NN JJ]
[the/DT black/JJ horses/NNS]	[các/PL con/CL ngựa ô/NN]	[PL CL NN]
[a/DT dog/NN]	[một/CD con/CL chó/NN]	[CD CL NN]

3. A MODEL FOR EXPLOITING TARGET LANGUAGE CHARACTERISTICS

In this section, we present the model for exploiting target language characteristics. As mentioned in the instruction section, this paper proposes a solution to improve the result in word alignment with Giza++ and reduce the percentage of error in identifying correspondence noun phrase with projection method of Yarowsky [3]. With word alignment, we exploit two characteristics of target language that are word order and word segmentation factor. With correspondence noun phrase identification, we exploit one more target language characteristic, that is classifiers.

Studying on Giza++ [7], we found that training diagram of Giza++ is executed with the sequence: $1^5 2^5 3^3 4^3$, $1^5 H^5 4^3$ và $1^5 H^5 3^3 4^3$. Characters and digits show the traning model in which exponential number is the number of passes, for example the sequence $1^5 H^5 3^3 4^3$ means: training with model 1 for 5 times, after that process Hidden Markov model (H) for 5 times, model 3 for 3 times and finally use model 4 for 3 times.

Hidden Markov Model (HMM) [15] predicts the distance between the position of words in source language, and the model 4 predicts the distance between the position of words in target language. Thus, word order factor will affect the training result of Giza++. This was proved by the experimental results of Och và Ney [16] in English-Germany and English-French word alignment. The pair of English-Germany gave a lesser error rate than pair of English-French because English and Germany are closer language family than English and French. Hence, we propose a method to tranfer order of source language (English) according to the order of target language (Vietnamese) before applying Giza++ to train (so that it is appropriate with the distance of model 4).

Similar to word segmentation factor, model 2, among the word alignment models using statistics of Brown [17], hypothesized that difference in length of the sentence affects the alignment result. Consequently, we find a way to reduce the difference in length between source and target languages by doing word segmentation before aligning. Figure 2 shows our proposed model, the detailed algorithms are presented in Algorithms 1 and 2.

With two specific factors of Vietnamese, word order and word segmentation (compound) factors, we combined four empirical models as described in Table 2. In empirical diagram (Table 2), we used the results of word order transfer in [14], anchor point alignment in [18] and exploited one more specific factor of Vietnamese language, that is word classification. The

proposed model is simulated in Figure 2. The detailed algorithms are described in Figures 3 and 4.

Table 2. Empirical models

No.	*Model*	*Notes*
1	WAP	Projection on the result of word alignment with Giza++
2	WAP-WS	Vietnamese word segmentation before alignment (Giza++)
3	WAP-STT	English word order transfer before alignment (Giza++)
4	WAP-LCC	Word order transfer and word segmentation before alignment with Giza++; use of Vietnamese CL characteristics in noun phrase corresponding identification.

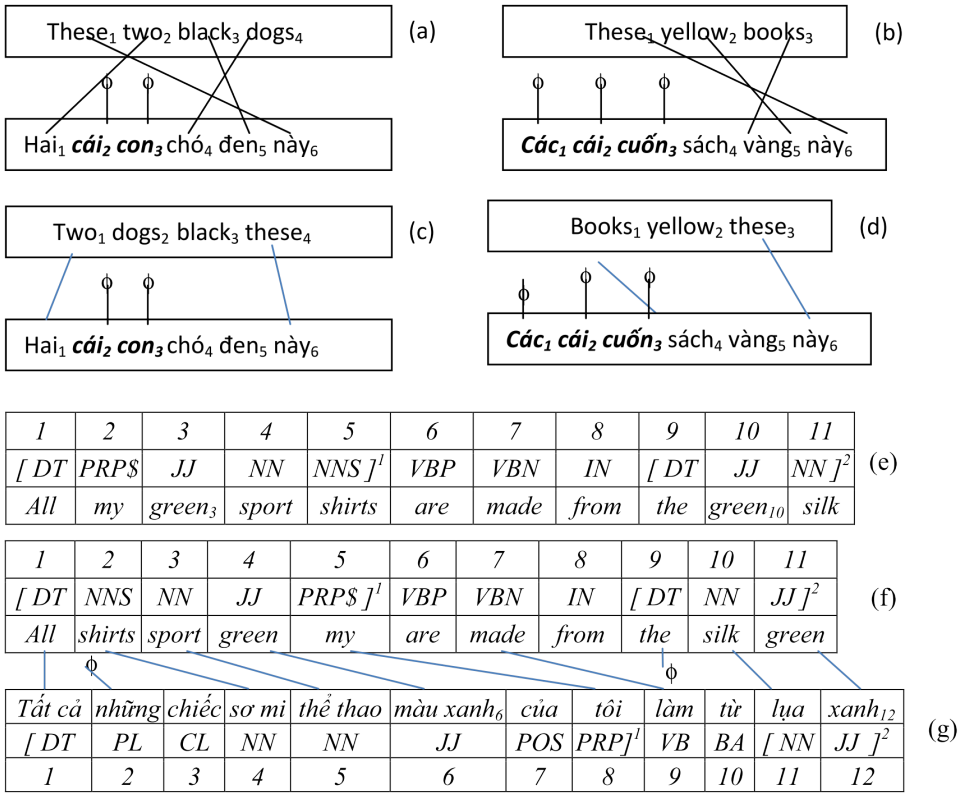

Figure 1. An example of alignment in WAP-LCC model

The simulation example is presented in Figure 1. In Figure 1(c, d), English noun phrases are transferred based on Vietnamese word order. CLs such as "cái", "con" in Figure 1(c) are automatically taken into Vietnamese BaseNP relying on anchor point alignment concept [9]. Figure 1(d) shows that the principle to get the left and right points (anchor points) of Vietnamese noun phrase based on the left and right points of English noun phrase does not have enough information to identify exactly in this situation. Hence, we proposed solution to extend

Vietnamese BaseNP by recognizing Vietnamese CL as presented in Algorithm 2.

The examples in Figure 1: Vietnamese sentence (g) is the translation of English sentence (e), English sentence (f) is transferred word order in BaseNP of Vietnamese sentence (g). We do the alignment with English sentence (f) and Vietnamese sentence (g) instead of aligning (e) and (g) sentence.

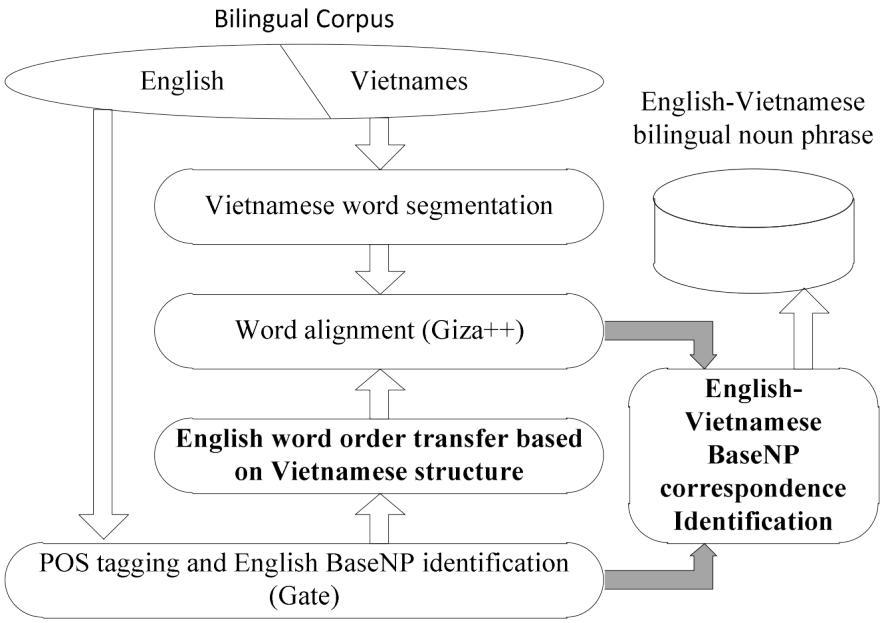

Figure 2. WAP-LCC model

Figure 2 is a diagram that illustrates the process of WAP-LCC model. Its sequence is as follows:
Step 1: Word order transfer in English noun phrase based on Vietnamese structure
Step 2: Vietnamese word segmentation
Step 3: English Vietnamese word alignment using Giza++
Step 4: Projection for identifying Vietnamese noun phrase.

The structure of WAP model is similar to WAP-LCC model, but it does not execute step 1 and 2, WAP-WS model does not execute step 1 and WAP-STT model does not execute step 2. The algorithm identifying BaseNP correspondence is presented in Figure 3 and 4.

An important module in WAP, WAP-WS WAP-STT and WAP-LCC models is Vietnamese BaseNP identification module for a pair of English-Vietnamese sentences. The input of this module is a pair of English-Vietnamese sentences, where English sentence is POS tagged and BaseNP identified while Vietnamese sentence is word segmented and English-Vietnamese word aligned. The output of this module is Vietnamese BaseNP and English BaseNP, respectively. Detail process of this module is presented in Algorithm 2 (Figure 4).

The structure of WAP model is similar to WAP-LCC model, but it does not execute step 1 and 2, WAP-WS model does not execute step 1 and WAP-STT model does not execute step 2. The algorithm identifying BaseNP correspondence is presented in Figure 3 and 4. An important module in WAP, WAP-WS WAP-STT and WAP-LCC models is Vietnamese BaseNP

Algorithm 1: WAP-LCC Algorithm
Input: English-Vietnamese bilingual corpus that have been aligned at sentence level
Output: Pairs of English-Vietnamese bilingual BaseNP
Process:
Step 1: Tag POS and identifyEnglish BaseNP (Gate)[19]
Step 2: Transfer word order of English BaseNP based on Vietnamese [14]
Step 3: Do Vietnamese word segmentation [11]
Step 4: Do English-Vietnamese word alignment (Giza++) [7]
Step 5: Identify English-Vietnamese BaseNP correspondence (Algorithm 2)

Figure 3. WAP-LCC Algorithm

Algorithm 2: Identify and extract BaseNP
Input:
- English sentence which POS tagged and BaseNP identified
- Vietnamese sentence which is translation of English sentence
- Alignment result of these two sentences
Output: English-Vietnamese Base NP correspondence
Process:
For $(k = 1; k < m; k++)$ *do*
Begin
Find iL and iR of NPVk corresponding with NPEk where alignment index is:
$iL = \min(a(i,j))$; $iR = \max(a(i,j))$
If the $(iL-1)-th$ *word belongs in CL (Classifiers) then* $iL = iL\breve{}1$
If the $(iL-1)-th$ *word is copulative "cái" then* $iL = iL\breve{}1$
If the $(iL-1)-th$ *word belongs in Ar class then* $iL = iL\breve{}1$, $Ar = \{$*một, những, các*$\}$
End;

Figure 4. Algorithm for Vietnamese BaseNP identification

identification module for a pair of English-Vietnamese sentences. The input of this module is a pair of English-Vietnamese sentences, where English sentence is POS tagged and BaseNP identified while Vietnamese sentence is word segmented and English-Vietnamese word aligned. The output of this module is Vietnamese BaseNP and English BaseNP, respectively. Detail process of this module presents in Algorithm 2 (Figure 4).

In algorithm 2:

- m = the number of BaseNP in English sentence

- iL = left point of NPVk

- iR = right point of NPVk

- $NPEk = k - th$ English BaseNP

- $NPVk = k - th$ Vietnamese BaseNP

- $a(i,j)$: the result of alignment between $j - th$ English word and i-th Vietnamese word, for example$a(i,j) = (2,3)$ means 3rd English word is aligned with 2nd Vietnamese word

- $j := 1$ to n (n is the number of words in English sentence)

- $i := 1$ to t (t is the number of words in Vietnamese segmented sentence).

4. EXPERIMENTAL RESULT

Training data and evaluation

Corpus was studied and evaluated by linguists from the '50s of past century [20]. The meaning of "*corpus*" term is considered as a collection of documents. Corpus is data, data of language, i.e. the empirical evidence of the use of language. These evidences of the use of language can be spoken or written language. Bilingual corpus is a pair of corpus where one corpus is the translation of another corpus in another language.

Collecting standard [21]

Language standard: The collecting source must be acknowledged by many people and the sentence must have standard grammar.

Translation style: The translation sentences must be 1-1, that means it is closely related to the meaning, not recapitulative, approximate, reverse translation.

Construction of training corpus

In Vietnam, although there are some corpus collecting authors such as D. Dien with 7 million word corpus [21, page 183]; N.P.Thai and Akira Shimazu [5] with more than 25,000 pairs of sentences; lexicography center with 100,000 pairs of bilingual sentences [22], but it is not easy to share these sources for research. Hence, we have to collect by ourselves. The total and the source of collection are shown in Table 3.

Table 3. Bilingual corpus collection

No.	*Source of English-Vietnamese documents*	*Number of pairs of sentence*	*Number of English words*	*Number of Vietnameses words*
1	Life in Australia	2,144	17,016	26,329
2	Dictionnary (E-V)	30,367	128,959	164,876
3	Children's Encyclopedia	6,118	56,036	65,579
4	Labour code	800	20,184	23,397
5	Network Encyclopedia	19,948	257,866	394,145
6	Stories	10,034	99,403	106,132
7	WSJ (Wall Street Journal)	1,235	32,772	44,275
Total:		**66,646**	**612,236**	**824,733**

Creation of sample corpus

Sample corpus is pairs of English-Vietnamese bilingual sentences which are extracted from Penn Treebank [23], translated by language research center – Ho Chi Minh Institute of Social Sciences.

Evaluation standard

We use the standard of chunker evaluation of Jurafsky and Matin [24] including measurement of precision Pre, recall Rec, reconcile F_β with formulas (1), (2), (3). We also compute the alginment error ratio AER by using the formular of Och [16] (formular (4)) to evaluate

the research efficiency.

$$\text{Pre(A,B)} = \frac{\text{A}}{\text{B}} \tag{1}$$

$$\text{Rec(A,C)} = \frac{\text{A}}{\text{C}} \tag{2}$$

$$\text{F}_\beta(\text{Pre, Rec}) = \frac{(\beta^2 + 1) * \text{Pre} * \text{Rec}}{\beta^2 * (\text{Pre} + \text{Rec})} \tag{3}$$

$$\text{AER}(A, B, C) = 1 - \frac{2 * \text{A}}{\text{B} + \text{C}} \tag{4}$$

where:

- A: number of BaseNP re-evaluated by people from the result of computer ; B: number of BaseNP by computer; C: number of BaseNP by people from evaluation data;

- Pre: precision ; Rec: Recall ; AER: alginment error ratio; F_β: reconcile weght; β: constant, $\beta = 1$ for this study.

Experimental result

Table 4 presents the word alignment and BaseNP corresponding identification result with WAP, WAP-WS, WAP-STT and WAP-LCC model. The precision (Pre) is ascending from the WAP-WS to WAP-LCC model. This result is appropriate with the hypothesis of Brown [17] that is the length of sentences affects the alignment result. The result of WAP-STT model is better than WAP-WS model and it shows that word order factor affect the result more than word segmentation factor. WAP-LCC model gets the best result because it combines both word order and word segmentation factor.

Table 4. Comparison of word alignment and BaseNP identification

Model	*Word alignment*				*English-Vietnamese BaseNP correspondence*			
	Pre	**Rec**	**AER**	$\mathbf{F}_\beta$	**Pre**	**Rec**	**AER**	$\mathbf{F}_\beta$
WAP	38,4%	33,7%	64,1%	35,9%	38,4%	33,7%	64,1%	35,9%
WAP-WS	57,1%	53,1%	44,9%	55,0%	52,1%	50,4%	48,4%	52,1%
WAP-STT	65,6%	62,2%	36,2%	63,8%	61,1%	59,3%	41,3%	60,2%
WAP-LCC	81,8%	77,3%	20,5%	79,5%	77,3%	77,8%	32,5%	77,5%

5. CONCLUSION

This paper proposes a model for exploiting target language characteristics to extract bilingual noun phrase with the projection approach on the result of word alignment by statistical method. In this proposed model, studied target language characteristics (Vietnamese) are word segmentation, word order in noun phrase structure, and classifier factor. Word segmentation and word order transfer are preprocessing before doing word alignment by statistical model using open source Giza++ [7]. CL characteristic is used in post-processing step of noun phrase corresponding identification algorithm. The proposed model can be applied for other pairs of languages.

This study achieved a good preliminary result through exploiting target language characteristics on four above models. However, statistical method needs a large enough bilingual

corpus to get a better result (for example, the English-French corpus has 1.5 million pairs of bilingual sentences [15]). In the future, we hope to do experiments on larger and more diversity corpus source.

REFERENCES

[1] Nguyen Chi Hieu, *A model for for exploiting target language characteristics to extract English-Vietnamese bilingual noun phrase correspondence,* Ph.D. dissertation, 2008, Ho Chi Minh City University of Technology.

[2] J.Kupiec, An Algorithm for finding Noun phrase Correspondences in Bilingual Corpora, *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, Columbus, Ohio, USA, 1993, 17 - 22.

[3] D.Yarowsky, G.Ngai and R.Wicentowski, *Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora*, Proceedings of NAACL-2001. (ISBN: 1-55860-786-2), 2001, 161-168.

[4] E.Riloff, C.Schafer and D.Yarowsky, Inducing Information Extraction Systems for New Languages via Cross-Language Projection, *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, 2002, 1-7.

[5] N.P.Thai and A.Shimazu, Improving Phrase-Based SMT with Morpho-Syntactic Analysis and Transformation, *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, 2006, 138-147, Cambridge, August.

[6] Nguyen Thi Minh Huyen, Laurent Romary, and Vu Xuan Luong, *A Case Study in POS Tagging of Vietnamese Texts, TALN 2003*, *Batz-sur-Mer, 11-14 June.* (2003)

[7] F.J.Och (2003), *Giza++: Training of statistical translation models*, http://www.isi.edu/∼och/ GIZA++.html.

[8] R.Florian (2002), *Transformation Based Learning and Data-Driven Lexical Disambiguation: Syntactic and Semantic Ambiguity Resolution*, the degree of Doctor of Philosophy, The Johns Hopkins University, Baltimore. Maryland.

[9] Tjong Kim Sang, Noun phrase representation by system combination, *Proceedings of ANLP-NAACL 2000*, Seattle, Washington, USA, Morgan Kaufman Publishers, 2000, 50-55.

[10] R.Hwa, *Breaking the resource bottleneck for multilingual processing*, University of Endinburgh IGK Summer School, 2004.

[11] Nguyen Quang Chau, Phan Thi Tuoi, Cao Hoang Tru, *Word segmentation and POS tagging for Vietnamese*, (a module of the national project KC01-21 "Semantic web"), 2005.

[12] Vu Ngoc Tu, *Research on English-Vietnamese word order projection on some basic structure*, Vice-doctoral dissertation, Hanoi National University.

[13] N.H.Tuong (2004), *The structure of the Vietnamese Noun Phrase*, Ph.D. dissertation, 1996, Boston University Graduate School of Arts and Sciences.

[14] Nguyen Chi Hieu, Vietnamese noun phrase syntax tree transfer based on English, *Journal of Information Technology & Communications* **9** (29) (2013) 3, 48-56, Vietnam, June 2013. ISSN: 1859 – 3526.

[15] J.Allen, *Natural Language Understanding,* The Benjamin/Cummings Publishing Company, ISBN 0-8053-0334-0, 1995.

[16] F.J.Och, H.Ney, *A Systematic Comparision of Various Statistical Alignment Models*, Association for Computational Linguistics, 2003.

[17] F.Brown, F.Peter, A.Stephen D.Pietra, J.Vincent and R. L. Mercer, The mathematics of statistical machine translation: Parameter estimation, *Computational Linguistics*, **19**(2) 1993 263–311.

[18] Hieu Chi Nguyen, A Combination System for Identifying Base Noun Phrase, *Advanced Methods for Computational Collective Intelligence, SCI* **57** (2012), 13-23, © Springer-Verlag Berlin Heidelberg.

[19] H.Cunningham, D.Maynard, K.Bontcheva and V.Tablan, GATE: A framework and graphical development environment for robust NLP tools and applications, *Proceedings of The 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, (2002), 168-175.

[20] Y.Deng, *Bitext Alignment for Statistical Machine Translation,* Ph.D. dissertation, 2005, Johns Hopkins University, Baltimore, Maryland

[21] Dinh Dien, Construction and exploitation an English-Vietnamese bilingual electronic corpus, Ph.D. dissertation, 2005, Ho Chi Minh city National University.

[22] http://www.vietlex.com/vncorpus/new.htm

[23] http://lcg-www.uia.ac.be/conll2000/chunking.

[24] D.Jurafsky and J.Matin, *Speech and Language Processing*, 2006, http://www.cs.colorado.edu/∼martin/slp2.html.

APPENDIX A
TABLE OF PARTS OF SPEECH

Label	**Description**	**Label**	**Description**
CC	Coordinating conjunction	RB	Adverb, comparative
CD	Cardinal number	RBS	Adverb, superlative
CD	Determiner	VB	Verb, base form
EX	Existential "there" ("có")	VBD	Verb, past tense
FW	Foreign word	VBG	Verb, gerund or present participle
IN	Preposition	VBN	Verb, past participle
JJ	Adjective	VBP	Verb, non 3^{rd} person singular present
JJR	Adjective, comparative	VBZ	Verb, 3rd person singular present
JJS	Adjective, superlative	WDT	Wh-determiner
NN	Noun, singular / mass	WP	Wh-pronoun
NNS	Noun, plural	WP$	Possessive Wh-pronoun
NP	Proper noun, singular	CL	Classifiers
NPS	Proper noun, plural	CA	Copulative "cái"
PDT	Pre-determiner	PL	"nhữ̃ng", "các"
POS	Possessive ending	BA	"bằng", "từ"
PRO	Personal pronoun	$	"đô la Mỹ"
PRP$	Possessive pronoun	#	"bảng Anh"

Received on January 15, 2014
Revised on April 23, 2014