Building an annotated English-Vietnamese parallel corpus

DINH Dien
National University of HCM City, Vietnam

Abstract

Corpus-based Natural Language Processing (NLP) tasks for such popular languages as English, French, etc. have been well studied with satisfactory achievements. In contrast, corpus-based NLP tasks for lesser known languages (e.g. Vietnamese) are at a deadlock due to absence of annotated training data for these languages. Furthermore, hand-annotation of even reasonably well-determined features such as part-of-speech (POS) tags has proved to be labor intensive and costly. In this paper, we present our building an annotated English-Vietnamese parallel Corpus named EVC, a corpus consisting of over 5 million words of English and Vietnamese. This EVC has been automatically word-aligned and POS-tagged by semantic-class model and Transformation-Based Learning (TBL) method. This annotated parallel corpus has been exploited to serve Vietnamese-related NLP tasks such as Vietnamese Word Segmentation, Vietnamese POS-tagger, English-to-Vietnamese Word Order transfer, Word Sense Disambiguation in English-to-Vietnamese Machine Translation, etc.

1. Introduction

NLP tasks are interesting and difficult. The highest difficulty which computers had to face, is the built-in ambiguity of Natural Languages. To disambiguate it, formerly, they based it on human-devised rules. Building such a complete rule-set is time-consuming and a labor-intensive task, whilst it doesn't cover all the cases. Besides, when the scale of system increases, it is very difficult to control that rule-set. So, recently, many NLP tasks have changed from rule-based approaches into corpus-based approaches with large annotated corpora.

Nowadays more and more people are interested in extracting information about language from very large annotated corpora. Such annotated corpora have been built for popular languages (e.g. Penn Tree Bank for English, French, Japanese, etc.) and these corpora have been used to effectively serve such well-known NLP tasks as POS-Tagger, Phrase Chunker, Parser, Structural Transfer, Word Sense Disambiguation (WSD), etc. Unfortunately, so far, there has been no such annotated corpora available for Vietnamese NLP tasks. Furthermore, building manually annotated corpora is very expensive (e.g. Penn Tree Bank invested over one million dollars and many person-years). To overcome this drawback, we have indirectly built such an annotated corpus for Vietnamese by taking advantage of annotated English corpora available, mutual disambiguation and projection via automatic word-alignments in an English-Vietnamese parallel corpus with five million words. For example:

> *D02:01323: *Jet planes fly about nine miles high.*
> +D02:01323: *Các phi cơ phản lực bay cao khoảng chín dặm.*

In this paper, we will present our experiences in collecting, building, and semi-automatically annotating the five-million word parallel corpus (named EVC). The rest of this paper will be organized as follows:

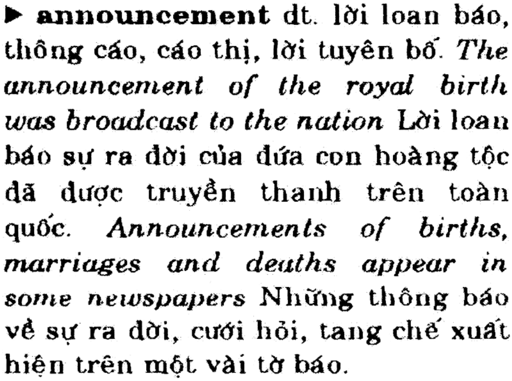
- Collecting English-Vietnamese parallel corpus: resources of raw parallel texts, its styles, etc.
- Normalizing English-Vietnamese parallel corpus: converting from different formats, types, spelling, etc. into unique ones. Sentence alignment of EVC (Dinh Dien 2001b).
- Word alignment of EVC: to automatically word-align EVC by Semantic-Class approach (Dinh Dien et al. 2002a). Manually correcting word-alignments. Problems of Vietnamese Word Segmentations (Dinh Dien et al. 2001a).
- Annotating EVC with POS-tags: Automatically POS-tagging by TBL method of Eric Brill (1995) for the English side first, then projecting the English side to the Vietnamese one. POS-Tagsets of English and Vietnamese.
- Conclusion: limitations of current EVC and its future developments, etc.

2. Collecting English-Vietnamese bitexts

Firstly, due to no official English-Vietnamese bilingual corpus available up to now, we have had to build it by ourselves by collecting English-Vietnamese bilingual texts from selected sources. Secondly, as most of these sources are not electronic forms, we must convert them into electronic form. During the process of electronic conversion, we have met a drawback. That is, there is no effective OCR (Optical Character Recognition) software available for Vietnamese characters. Compared with English OCR softwares, the Vietnamese OCR one is lower just because Vietnamese characters have tone marks (acute, breve, question, tilde, dot below) and diacritics (hook, caret,..). So, we must manually input most of the Vietnamese texts (low-quality hardcopies). Only OCR of high-quality hardcopies have been used and manually revised.

During collecting English-Vietnamese bilingual texts, we choose only the following materials:

- Science or technical materials.
- Conventional examples in dictionaries (Figure 1).
- Bilingual texts whose translations are exact (translated by human translator and published by reputable publishers) and not too diversified (no "one-to-one" translation).



▶ **announcement** dt. lời loan báo, thông cáo, cáo thị, lời tuyên bố. *The announcement of the royal birth was broadcast to the nation* Lời loan báo sự ra đời của đứa con hoàng tộc đã được truyền thanh trên toàn quốc. *Announcements of births, marriages and deaths appear in some newspapers* Những thông báo về sự ra đời, cưới hỏi, tang chế xuất hiện trên một vài tờ báo.

Figure 1. An example collected from English-Vietnamese dictionary

So far, we have collected a 5,000,000-word corpus containing approx. 500,000 sentences (Dinh Dien 2001b) and most of them are texts in science and conventional fields (Table 1).

Table 1. Collection of English-Vietnamese parallel Corpus (EVC)

No.	Resources	The number of pairs of sentences	Number of English words	Number of Vietnamese morpho-words[1]	Length (English words)	Percent (words/EVC)
1	Computer books[2]	9,475	165,042	239,984	17.42	7.67
2	LLOCE dictionary[3]	33,078	312,655	410,760	9.45	14.53
3	EV bilingual dictionaries[4]	174,906	1,110,003	1,460,010	6.35	51.58
4	SUSANNE corpus[5]	6,269	131,500	181,781	20.98	6.11
5	Electronics books[6]	12,120	226,953	297,920	18.73	10.55
6	Children's Encyclopedia[7]	4,953	79,927	101,023	16.14	3.71
7	Other books[8]	9,210	126,060	160,585	13.69	5.86
	Total	250,011	2,152,140	2,852,063	8.61	100%

[1] Vietnamese "word" is a special linguistic unit in Vietnamese language only, which is often called "tiếng". This lexical unit is lower than traditional words but higher than traditional morphemes.

[2] The set of 12 volumes of bilingual books titled "Come to the World of Microcomputers" compiled by Mr. Nguyen The Hung, published by CADASA Center.

[3] Examples in the Longman Lexicon Of Contemporary English compiled by Arthur in 1997. The Vietnamese version of LLOCE is edited by Tran Tat Thang and published by the Education Publisher.

[4] English-Vietnamese Dictionary of Foreign University, VNU-Hanoi in 2000 and Vietnamese-English Dictionary of Bui Phung in 2001, published by the World Publisher of HCM City.

3. Normalization of EVC

However, after the collection, we must convert them into unified forms (normalization) as follows: type of files: text only; code: TCVN3 (Standard of Vietnamese character codes); Vietnamese spelling, etc. Then, we proceed with sentence alignment.

3.1 Sentence-alignment of bilingual corpus

During inputting this bilingual corpus, we have aligned sentences manually under the following format:

**D02:01323: The announcement of the royal birth was broadcast to the nation.*

+D02:01323: Lời loan báo sự ra đời của đứa con hoàng tộc đã được truyền thanh trên toàn quốc.

**D02:01324: Announcements of births, marriages and deaths appear in some newspapers.*

+D02:01324: Những thông báo về sự ra đời, cưới hỏi, tang chế xuất hiện trên một vài tờ báo.

in which, first characters are reference numbers indicating its source and the position of the sentence in a text.

Because most of our bilingual corpus is manually typed, we haven't used automatic sentential alignment. Automatic sentential alignment would be necessary if we had already had online bilingual texts.

3.2 Spelling checker of bilingual corpus

After aligning sentences, we check the spell of English words and Vietnamese words automatically. Here, we have met another drawback in processing the Vietnamese word segmentation because Vietnamese words (similar to Chinese words) are not delimited by spaces (Dien Dinh 2001b). However, our spelling checker is able to detect non-existent words in English or Vietnamese only. So, we must review this corpus manually. In fact, Vietnamese "word" here is only "tiếng", which is equivalent to Vietnamese "spelling word" or "morpheme" (due to features of isolated language typology).

[5] SUSANNE (Surface and Underlying Structural Analyses of Naturalistic English) is constructed by Geoffrey Sampson (1995) at Sussex University, UK. Vietnamese translation is performed by English teacher of VNU-HCMC.

[6] The set of 12 volumes of Telecommunication TextBooks of University, compiled by Vietnam-Korea Committee.

[7] The set of three volumes of bilingual Children's encyclopedia, published by Education Publisher.

[8] Other books of computers, English-Vietnamese sentence patterns, English for Computer Sciences, etc.

4. Word alignment of EVC

Before describing this algorithm briefly, we have the following conventions:

S stands for the English sentence and T stands for the Vietnamese one. We have sentence pairs translated by each other as (S,T): s is the word in S, t is the word in T which is translated by s in S in context. DTs is the set of dictionary meanings for s entry, each meaning is represented by d.

$W_S = \{ s \}$, set of English real words and idioms presented in S.

$W_T = \{ t \mid t \in T \wedge t \in VD \}$, set of Vietnamese possible words presented in T.

where : VD is the Vietnamese Dictionary containing Vietnamese possible words and phrases.

The problem is how computers can recognise which t in T will be aligned with which s in S. Relying on W_T, we can solve the case resulting in the wrong definitions of words in Vietnamese sentences when we only carry out word segmentation relying on VD. Our algorithm is in conformity with the following steps.

4.1 Dictionary-based word alignment

We mainly calculate the similarity on morphemes between each word d in DTs with all t in W_T based on formula calculating Dice coefficient (Dice 1945) as follows:

$$\boxed{\mathrm{Sim}(d, t) = \frac{2 \text{ x } | d \cap t |}{| d | + | t |}}$$

where: | d | and | t | : the number of morphemes in d and in t.
| d ∩ t | : the number of morphemes in the intersection of d and t.

Next, for each word pair (s, t) obtained from Descartes product (W_S x W_T), we calculate the value of DTSim(s, t) presenting the likelihood of a connection as follows:

$$\boxed{\mathrm{DTSim}(s, t) = \max \mathrm{Sim}(d, t)}$$

Examining a sample on following sentence pairs:
S = "The old man goes very fast"
T = "Ông cụ đi quá nhanh"

We will have:
W_S = { the, old, man, go, very, fast }
W_T = { ông, ông cụ, cụ, đi, nhanh, quá }

Suppose that we are examining "man",
DT (man) = { người, đàn ông, nam nhi }

So, we have:

DTSim(man, ông) = max{ Sim(người, ông), Sim (đàn ông, ông), Sim (nam nhi, ông) } = max{(2x0)/(1+1),(2x1)/(2+1),(2x0)/(2+1)} = 0.67

DTSim (man, ông cụ) = max{ Sim(người, ông cụ), Sim(đàn ông, ông cụ), Sim(nam nhi, ông cụ)} = max{(2x0)/(1+2),(2x1)/(2+2),(2x0)/(2+2)} = 0.5

Then, we choose candidate translation pairs of greatest likelihood of connection.

4.2 Calculating the correlation between two classes of two languages

The correlation ratio of class X and class Y can be measured using the Dice coefficient as follows:

$$\boxed{ClassSim(X,Y) = \frac{\sum_{a\in X} From(a,Y) + \sum_{b\in Y} To(X,b)}{|X|+|Y|}}$$

Where |X| = the total number of the words in X, |Y| = the total number of the words in Y, From(a,Y) = 1,if $(\exists y \in Y)(a, y) \in ALLCONN$,
= 0, otherwise

To(X,b) = 1, if $(\exists x \in X)(x,b) \in ALLCONN$,
= 0, otherwise,

ALLCONN: a list of initial connections obtained by running above dictionary-based word alignment over the bilingual corpus.

4.3 Estimating the likelihood of candidate translation pairs

A coefficient, presented by Brown et al. (1993) establishing each connection is a probabilistic value Pr(s, t), showing translated probability of each pair (s, t) in (S, T), calculated by product of dictionary translated probability, t(s | t), and dislocated probability of words in sentences, d(i | j, l, m). However Sue J. Ker and Jason S. Chang did not agree with it completely. In their opinion, it is very difficult to estimate t(s, t) and d(i, j) exactly for all values of s, t, i, j in the formula:

$$\boxed{\Pr(s, t) = t(s, t) \text{ x } d(i, j)}$$

We have the same opinion with them. We can create functions based on dictionary, word concept and position of words in sentences to limit cases to be examined and computed.

The similar concept of word pair (s, t) function:

$$\boxed{\text{ConceptSim}(s, t) = \underset{s \in X, t \in Y}{\text{maxClassSim}}(X,Y)}$$

Then, combining with DTSim(s, t), we have four value of t(s, t). We have to combine with DTSim(s, t) because we are partially basing on the dictionary. Besides, we can solve the case that there are many words belonging to the same class in sentences.

Table 2. Constants in word alignment

DTSim(s, t)	ConceptSim(s, t)	
a) t1	≥ h1	≥ h2
b) t2	≥ h1	< h2
c) t3	< h1	≥ h2
d) t4	< h1	< h2

where h1 and h2 are thresholds chosen via experimental results. An example of word-alignment is as Figure 2 below.

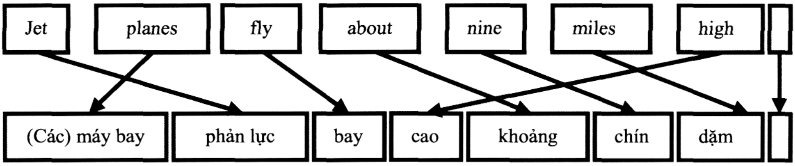


Figure 2. An example word-alignment of English-Vietnamese parallel Corpus

5. POS-Tagging for EVC

So far, existing POS-taggers for (mono-lingual) English have been well developed with satisfactory achievements, and it is very difficult (it is nearly impossible for us) to improve their results. Actually, those existing advanced POS-taggers have exhaustively exploited all linguistic information in English texts, and there is no way for us to improve English POS-tagger in case of such a monolingual English texts. By contrast, in the bilingual texts, we are able to make use of the second language's linguistic information in order to improve the POS-tag annotations of the first language.

Our solution is motivated by I. Dagan, I. Alon and S. Ulrike (1991); W. Gale, K. Church and D. Yarowsky (1992). They proposed the use of bilingual corpora to avoid hand-tagging of training data. Their premise is that "different senses of a given word often translate differently in another language (for example, *pen* in English is *stylo* in French for its *writing implement* sense, and *enclos* for its *enclosure* sense). By using a parallel aligned corpus, the translation of each occurrence of a word such as *pen* can be used to automatically determine its sense". This remark is not only true for word sense but also for POS-tag and it is more exact in such typologically different languages as English vs. Vietnamese.

In fact, POS-tag annotations of English words as well as Vietnamese words are often ambiguous, but they are not often exactly the same. For example, "can" in English may be "Aux" for *ability* sense, "V" for *to make a container* sense, and "N" for *a container* sense, and there is hardly an existing POS-tagger which can tag POS for that word "can" exactly in all different contexts. Nevertheless, if that "can" in English is already word-aligned with a corresponding Vietnamese word, it will be POS-disambiguated easily by Vietnamese words' POS-tags. For example, if "can" is aligned with "có thể", it must be *Auxiliary*, if it is aligned with "đóng hộp" then it must be a *Verb*, and if it is aligned with "cái hộp" then it must be a *Noun*.

However, not all Vietnamese POS-tag information is useful and deterministic. The big question here is when and how we make use of the Vietnamese POS-tag information? Our answer is to have this English POS-tagger trained by TBL method with the SUSANNE training corpus. After training, we will extract an ordered sequence of optimal transformation rules. We will use these rules to improve an existing English POS-tagger (as baseline tagger) for tagging words of the English side in the word-aligned EVC corpus. This English POS-tagging result will be projected to Vietnamese side via word-alignments in order to form a new Vietnamese training corpus annotated with POS-tags.

5.1 The English POS-Tagger by TBL method

To make the presentation clearer, we re-use notations in the introduction to fnTBL-toolkit of Radu Florian and Grace Ngai (2001) as follows:

- χ : denotes the space of samples: the set of words which need POS-tagging. In English, it is simple to recognize the word boundary, but in Vietnamese (an isolate language), it is rather complicated. Therefore, it has been presented in another work (Dinh Dien et al. 2001a).
- C : set of possible POS-classifications c (or tagset). For example: *noun* (N), *verb* (V), *adjective* (A), ... For English, we made use of the Penn Tree Bank tagset and for the Vietnamese tagset, we use the POS-tagset mapping table (see Appendix A).
- $S = \chi \mathrm{x} C$: the space of states: the cross-product between the sample space (word) and the classification space (tagset), where each point is a couple (word, tag).
- π : predicate defined on S^+ space, which is on a sequence of states. Predicate π follows the specified templates of transformation rules. In the POS-tagger for English, this predicate only consists of English factors which affect the POS-tagging process, for example $\bigcup_{\exists i \in [-m,+n]} Word_i$ or $\bigcup_{\exists i \in [-m,+n]} Tag_i$ or $\bigcup_{\exists i \in [-m,+n]} Word_i \wedge Tag_j$.

where, *Word$_i$* is the morphology of the ith word from the current word. Positive values of i mean preceding (its left side), and negative ones mean following (its right side). i ranges within the window from $-m$ to $+n$. In this English-Vietnamese bilingual POS-tagger, we add new elements including $VTag_0$ and $\exists VTag_0$ to those predicates. $VTag_0$ is the Vietnamese POS-tag corresponding to the current English word via its word-alignment. These Vietnamese POS-tags are determined by the most frequent tag according to the Vietnamese dictionary.

- A rule r defined as a couple (π, c) which consists of predicate π and tag c. Rule r is written in the form $\pi \Rightarrow$ c. This means that the rule r = (π, c) will be applied on the sample x if the predicate π is satisfied on it, whereas, x will be changed into a new tag c.
- Giving a state $s = (x, c)$ and rule $r = (\pi, c)$, then the result state $r(s)$, which is gained by applying rule r on s, is defined as:

$$r(s) = \begin{cases} s & \text{if } \pi(s)=\text{False} \\ (x, c') & \text{if } \pi(s)=\text{True} \end{cases}$$

- T : set of training samples, which were assigned correct tag. Here we made use of the SUSANNE golden corpus (Sampson, 1995) whose POS-tagset was converted into the PTB tagset.

- The score associated with a rule $r = (\pi, c)$ is usually the difference in performance (on the training data) that results from applying the rule, as follows:

$$Score(r) = \sum_{s \in T} score(r(s)) - \sum_{s \in T} score(s)$$

$$score((x,c)) = \begin{cases} 1 & \text{if c = True(x)} \\ 0 & \text{if c} \neq \text{True(x)} \end{cases}$$

5.2 The TBL algorithm for POS-Tagging

The TBL algorithm for POS-tagging can be briefly described as follows:

Step 1: Baseline tagging: To initiatize for each sample x in SUSANNE training data with its most likely POS-tag *c*. For English, we made use of the available English tagger (and parser) of Eugene Charniak (1997) at Brown University (version 2001). For Vietnamese, it is the set of possible parts-of-speech tags (follow the appearance probability order of that part-of-speech in the dictionary). We call the starting training data as T_0.

Step 2: Considering all the transformations (rules) *r* to the training data T_k in time kth, choose the one with the highest Score(r) and applying it to the training data to obtain new corpus T_{k+1}. We have: $T_{k+1} = r(T_k) = \{ r(s) \mid s \in T_k \}$. If there are no more possible transformation rules which satisfies: Score(r) > β, the algorithm is stopped. β is the threshold, which is preset and adjusted according to reality situations.

Step 3: k = k+1.

Step 4: Repeat from step 2.

Step 5: Applying every rule *r* which is drawn in order for new corpus EVC after this corpus has been POS-tagged with baseline tags similar to those of the training period.

5.3 Experiment and results of bootstrapped English POS-Tagger

After the training period, this system will extract an ordered sequence of optimal transformation rules under following format, for examples:

$$((tag_{-1} = TO) \wedge (tag_0 = NN)) \Rightarrow tag_0 \leftarrow VB$$

$$((Word_0 = "can") \wedge (VTag_0 = MD) \wedge (tag_0 = VB)) \Rightarrow tag_0 \leftarrow MD$$

$$((\exists i \in [-3,-1] \mid Tag_i = MD) \wedge (tag_0 = VPB)) \Rightarrow tag_0 \leftarrow VB$$

These are intuitive rules and easy to understand by humans. For examples: the 2nd rule will be understood as follows: *"if the POS-tag of current word is VB (Verb) and its word-form is "can" and its corresponding Vietnamese word-tag is MD (Modal), then the POS-tag of current word will be changed into MD."*

We have experimented with this method on EVC corpus with the training SUSANNE corpus. To evaluate this method, we held-back a 6,000-word part of the training corpus (which had not been used in the training period) and we achieved the POS-tagging results as follows:

Table 3. The result of bootstrapped POS-tagger for English side in EVC

Step	Correct tags	Incorrect Tags	Precision
Baseline tagging (Brown POS-tagger)	5724	276	95.4%
TBL-POS-tagger (bootstrapping by corresponding Vietnamese word)	5850	150	97.5%

For details of POS-Tagging for EVC, please refer to Dinh Dien and Hoang Kiem (2003).

5.4 Projecting English POStags to Vietnamese

After having English-POS-tag annotations with high precision, we proceed to directly project those POS-tag annotations from the English side into the Vietnamese side. Our solution is motivated by a similar work of David Yarowsky and Grace Ngai (2001). This projection is based on available word-alignments in the automatically word-aligned English-Vietnamese parallel corpus.

Nevertheless, due to typological difference between English (an inflected typology) vs. Vietnamese (an isolated typology), direct projection is not a simple 1-1 map, but it may be a complex m-n map:

- Regarding grammatical meanings, English usually makes use of inflectional facilities, such as suffixes to express grammatical meanings. For example: *-s* →plural, *-ed* →past, *-ing*→continuous, *'s* → possesive case, etc. Whilst Vietnamese often makes use of function words, word order facilities. For example: "các" "những" → plural, "đã" → past, "đang" → continuous, "của" → possessive cases, etc.

- Regarding lexicalization, some words in English must be represented by a phrase in Vietnamese and vice-versa. For example: "cow" and "ox" in English will be rephrased into two words "bò cái" (female one) and "bò đực" (male one) in Vietnamese; or "nghé" in Vietnamese will be rephrased into two words "buffalo calf" in English. The result of projecting is as Table 4 below. In addition, tagsets of the two languages are different due to characteristics of each language (please refer to the English-Vietnamese *consensus tagset map in* Appendix A).

Table 4. An example of English POS-tagging in parallel corpus EVC

English	**Jet**	**planes**	**fly**	**about**	**nine**	**miles**	**high**
E-tag	NN	NNS	VBP	IN	CD	NNS	RB
VN-ese	phản lực	(các) phi cơ	bay	khoảng	chín	dặm	cao
V-tag	N	N	V	IN	CD	N	R

6 Conclusion

We have just presented the building of an annotated English-Vietnamese parallel Corpus. This 5-million word EVC has been collected from selected sources, normalized into standard format, and word-aligned by semantic class-based approach. Finally, this EVC has been POS-tagged by POS-tagging English words first, and then projecting them to Vietnamese side later. The English POS-tagging is done in 2 steps: The basic tagging step is achieved through the available POS-tagger (Brown), and the correction step is achieved through the TBL learning method in which the information on the corresponding Vietnamese is used through available word-alignment in the EVC.

The result of word-alignment and POS-tagging of Vietnamese in the English-Vietnamese bilingual corpus has played a meaningful role in the building of the automatic training corpus for our Vietnamese NLP tasks, such as Vietnamese POS-taggers, WSD in English-to-Vietnamese MT (Dinh Dien and Hoang Kiem 2002b), etc. By making use of the language typology's differences and the word-alignments in bilingual corpus for the mutual disambiguation, we are still able to improve the result of the word-alignment and other linguistic annotation.

Currently, we are improving the quality of EVC by manually correcting linguistic annotations such as: word alignment, POS-tagging, etc. We are also tagging this EVC semantic-label by using semantic class names via available word-alignment in EVC.

REFERENCES

Brill, Eric. 1995. “Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging.” *Computational Linguistics* 21(4):543-565.

Brown, Peter F., Robert L. Mercer, Stephen A. Della Pietra and Vincent J. Della Pietra. 1993. “The mathematics of statistical machine translation: parameter estimation.” *Computational Linguistics* 19(2):263-311.

Charniak, Eugene. 1997. “Statistical parsing with a context-free grammar and word statistics.” *Proceedings of the 14th National Conference on AI*, AAAI Press/MIT Press, Menlo Park, pp. 598-603.

Dagan, Itai, I. Alon and S. Ulrike. 1991. “Two languages are more informative than one.” *Proceedings of the 29th Annual ACL*, Berkeley, CA, pp.130-137.

Dice, L.R. 1945. “Measures of the amount of ecologic association between species.” *Journal of Ecology* 26:297-302.

Dinh Dien, Hoang Kiem and Nguyen Van Toan. 2001a. “Vietnamese word segmentation.” *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, Tokyo, Japan, Nov. 2001, pp. 749-756.

Dinh Dien. 2001b. “Building the English-Vietnamese Parallel Corpus.” M.A. thesis in Comparative Linguistics in University of Social Sciences and Humanity of VNU-HCMC, Vietnam.

Dinh Dien and Hoang Kiem. 2002a. “Word alignment in English – Vietnamese bilingual corpus.” *Proceedings of East Asian Languages Processing’02*,
Ha Noi, Vietnam, pp. 3-11.

Dinh Dien and Hoang Kiem. 2002b. “Building a training corpus for word sense disambiguation in the English-to-Vietnamese machine translation.” *Proceedings of Workshop on Machine Translation in Asia, COLING-02*, Taiwan, Sept. 2002, pp.26-32.

Dinh Dien and Hoang Kiem. 2003. “POS-Tagging for English – Vietnamese Bilingual Corpus.” *Workshop on Parallel Texts at HLT-NAACL-03,* Edmonton, Canada.

Florian, Radu and Grace Ngai. 2001. “Fast transformation-based learning toolkit.” *Technical Report.*

Gale, W., K.W. Church and D. Yarowsky. 1992. “Using bilingual materials to develop word sense disambiguation methods.” *Proceedings of the International Conference on Theoretical and Methodological Issues in MT,* pp.101-112.

Jang, S.K. and J.S. Chang. 1997 “A class-based approach to word alignment.” *Computational Linguistics* 23(2):313-343.

Mc Arthur, Tom. 1997. *Longman Lexicon of Contemporary English.* (Vietnamese version by Tran Tat Thang), VN Education Publisher.

Sampson, Geoffrey. 1995. *English for the Computer: The SUSANNE Corpus and Analytic Scheme.* Clarendon Press (Oxford University Press).

Yarowsky, D. and Grace Ngai. 2001. "Induce, multilingual POS tagger and NP bracketer via projection on aligned corpora." *Proceedings of North America Association for Computational Linguistics '01,* pp. 200-207.

Received: 20 May 2003

National University of HCM City
Vietnam
<ddien@saigonnet.vn>

Appendix A. English-Vietnamese consensus POS-tagset mapping table

English POS	Vietnamese POS
CC (Coordinating conjunction)	CC
CD (Cardinal number)	CD
DT (Determiner)	DT
EX (Existential)	V
FW (Foreign word)	FW
IN (Preposition)	IN
JJ (Adjective)	A
JJR (Adjective, comparative)	A
JJS (Adjective, superlative)	A
LS (List item marker)	LS
MD (Modal)	MD
NN (Noun, singular or mass)	N
NNS (Noun, plural)	N
NP (Proper noun, singular)	N
NPS (Proper noun, plural)	N
PDT (Predeterminer)	DT
POS (Possessive ending)	“của”
PP (Personal pronoun)	P
PP$ (Possessive pronoun)	“của” P
RB (Adverb)	R
RBR (Adverb, comparative)	R
RBS (Adverb, superlative)	R
RP (Particle)	RP
SYM (Symbol)	SYM
TO ("to")	-
UH (Interjection)	UH
VB (Verb, base form)	V
VBD (Verb, past tense)	V
VBG (Verb, gerund or present participle)	V
VBN (Verb, past participle)	V
VBP (Verb, non-3rd person singular present)	V
VBZ (Verb, 3rd person singular present)	V
WDT (Wh-determiner)	P
WP (Wh-pronoun)	P
WP$ (Possessive wh-pronoun)	“của” P
WRB (Wh-adverb)	R