HLT-NAACL 2003 Workshop: Building and Using Parallel Texts
Data Driven Machine Translation and Beyond , pp. 88-95
Edmonton, May-June 2003

POS-Tagger for English-Vietnamese Bilingual Corpus

Dinh Dien
Information Technology Faculty of
Vietnam National University of HCMC,
20/C2 Hoang Hoa Tham, Ward 12,
Tan Binh Dist., HCM City, Vietnam
ddien@saigonnet.vn

Hoang Kiem
Center of Information Technology
Development of
Vietnam National University of HCMC,
227 Nguyen Van Cu, District 5, HCM City,
hkiem@citd.edu.vn

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 unpopular 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 suggest a solution to partially overcome the annotated resource shortage in Vietnamese by building a POS-tagger for an automatically word-aligned English-Vietnamese parallel Corpus (named EVC). This POS-tagger made use of the Transformation-Based Learning (or TBL) method to bootstrap the POS-annotation results of the English POS-tagger by exploiting the POS-information of the corresponding Vietnamese words via their word-alignments in EVC. Then, we directly project POS-annotations from English side to Vietnamese via available word alignments. This POS-annotated Vietnamese corpus will be manually corrected to become an annotated training data for Vietnamese NLP tasks such as POS-tagger, Phrase-Chunker, Parser, Word-Sense Disambiguator, etc.

1 Introduction

POS-tagging is assigning to each word of a text the proper POS tag in its context of appearance. Although, each word can be classified into various POS-tags, in a defined context, it can only be attributed with a definite POS. As an example, in this sentence: "*I can can a can*", the POS-tagger must be able to perform the following: "I_{PRO} can_{AUX} can_{V} a_{DET} can_{N}".

In order to proceed with POS-tagging, such various methods as Hidden Markov Models (HMM); Memory-based models (Daelemans, 1996); Transformation-based Learning (TBL) (Brill, 1995); Maximum Entropy; decision trees (Schmid, 1994a); Neural network (Schmid, 1994b); and so on can be used. In which, the methods based on machine learning in general and TBL in particular prove effective with much popularity at present.

To achieve good results, the abovementioned methods must be equipped with exactly annotated training corpora. Such training corpora for popular languages (e.g. English, French, etc.) are available (e.g. Penn Tree Bank, SUSANNE, etc.). Unfortunately, so far, there has been no such annotated training data available for Vietnamese POS-taggers. Furthermore, building manually annotated training data is very expensive (for example, Penn Tree Bank was invested over 1 million dollars and many person-years). To overcome this drawback, this paper will present a solution to indirectly build such an annotated training corpus for Vietnamese by taking advantages of available English-Vietnamese bilingual corpus named EVC (Dinh Dien, 2001b). This EVC has been automatically word-aligned (Dinh Dien et al., 2002a).

Our approach in this work is to use a bootstrapped POS tagger for English to annotate the English side of a word-aligned parallel corpus, then directly project the tag annotations to the second language (Vietnamese) via existing word-alignments (Yarowsky and Ngai, 2001). In this work, we made use of the TBL method and SUSANNE training corpus to train our English POS-tagger. The remains of this paper is as follows:

- POS-Tagging by TBL method: introducing to original TBL, improved fTBL, traditional English POS-Tagger by TBL.
- English-Vietnamese bilingual Corpus (EVC): resources of EVC, word-alignment of EVC.
- Bootstrapping English-POS-Tagger: bootstrapping English POS-Tagger by the POS-tag of corresponding Vietnamese words. Its evaluation
- Projecting English POS-tag annotations to Vietnamese side. Its evaluation.
- Conclusion: conclusions, limitations and future developments.

2 POS-Tagging by TBL method

The Transformation-Based Learning (or TBL) was proposed by Eric Brill in 1993 in his doctoral dissertation (Brill, 1993) on the foundation of structural linguistics of Z.S.Harris. TBL has been applied with success in various natural language processing (mainly the tasks of classification). In 2001, Radu Florian and Grace Ngai proposed the fast Transformation-Based Learning (or fTBL) (Florian and Ngai, 2001a) to improve the learning speed of TBL without affecting the accuracy of the original algorithm.

The central idea of TBL is to start with some simple (or sophisticated) solution to the problem (called baseline tagging), and step-by-step apply optimal transformation rules (which are extracted from a annotated training corpus at each step) to improve (change from incorrect tags into correct ones) the problem. The algorithm stops when no more optimal transformation rule is selected or data is exhausted. The optimal transformation rule is the one which results in the largest benefit (repairs incorrect tags into correct tags as much as possible).

A striking particularity of TBL in comparison with other learning methods is perceptive and symbolic: the linguists are able to observe, intervene in all the learning, implementing processes as well as the intermediary and final results. Besides, TBL allows the inheritance of the tagging results of another system (considered as the baseline or initial tagging) with the correction on that result based on the transformation rules learned through the training period.

TBL is active in conformity with the transformational rules in order to change wrong tags into right ones. All these rules obey the templates specified by human. In these templates, we need to regulate the factors affecting the tagging. In order to evaluate the optimal transformation rules, TBL needs the annotated training corpus (the corpus to which the correct tag has been attached, usually referred to as the golden corpus) to compare the result of current tagging to the correct tag in the training corpus. In the executing period, these optimal rules will be used for tagging new corpora (in conformity with the sorting order) and these new corpora must also be assigned with the baseline tags similar to that of the training period. These linguistic annotation tags can be morphological ones (sentence boundary, word boundary), POS tags, syntactical tags (phrase chunker), sense tags, grammatical relation tags, etc.

POS-tagging was the first application of TBL and the most popular and extended to various languages (e.g. Korean, Spanish, German, etc.) (Curran, 1999). The approach of TBL POS-tagger is simple but effective and it reaches the accuracy competitive with other powerful POS-taggers. The TBL algorithm for POS-tagger can be briefly described under two periods as follows:

* <u>The training period</u>:

- Starting with the annotated training corpus (or called *golden corpus*, which has been assigned with correct POS tag annotations), TBL copies this golden corpus into a new unannotated corpus (called *current corpus*, which is removed POS tag annotations).
- TBL assigns an inital POS-tag to each word in corpus. This initial tag is the most likely tag for a word if the word is known and is guessed based upon properties of the word if the word is not known.
- TBL applies each instance of each candidate rule (following the format of templates designed by human beings) in the current corpus. These rules change the POS tags of words based upon the contexts they appear in. TBL evaluates the result of applying that candidate rule by comparing the current result of POS-tag annotations with that of the golden corpus in order to choose the best one which has highest mark. These best rules are repeatedly extracted until there is no more optimal rule (its mark isn't higher than a preset threshold). These optimal rules create an ordered sequence.

* <u>The executing period</u>:

- Starting with the new unannotated text, TBL assigns an inital POS-tag to each word in text in a way similar to that of the training period.
- The sequence of optimal rules (extracted from training period) are applied, which change the POS tag annotations based upon the contexts they appear in. These rules are applied deterministically in the order they appear in the sequence.

In addition to the above-mentioned TBL algorithm that is applied in the supervised POS-tagger, Brill (1997) also presented an unsupervised POS-tagger that is trained on unannotated corpora. The accuracy of unsupervised POS-tagger was reported lower than that of supervised POS-tagger.

Because the goal of our work is to build a POS-tag annotated training data for Vietnamese, we need an annotated corpus with as high as possible accuracy. So, we will concentrate on the supervised POS-tagger only.

For full details of TBL and FTBL, please refer to Eric Brill (1993, 1995) and Radu Florian and Grace Ngai (2001a).

3 English – Vietnamese Bilingual Corpus

The bilingual corpus that needs POS-tagging in this paper is named EVC (English – Vietnamese Corpus). This corpus is collected from many different resources of bilingual texts (such as books, dictionaries, corpora, etc.) in selected fields such as Science, Technology, daily conversation (see table 1). After collecting bilingual texts from different resources, this parallel corpus has been normalized their form (text-only), tone marks (diacritics), character code of Vietnam (TCVN-3), character font (VN-Times), etc. Next, this corpus has been sentence aligned and checked spell semi-automatically. An example of unannotated EVC as the following:

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

Where, the codes at the beginning of each line refer to the corresponding sentence in the EVC corpus. For full details of building this EVC corpus (e.g. collecting, normalizing, sentence alignment, spelling checker, etc.), please refer to Dinh Dien (2001b).

Next, this bilingual corpus has been automatically word aligned by a hybrid model combining the semantic class-based model with the GIZA++ model. An example of the word-alignment result is as in figure 1 below. The accuracy of word-alignment of this parallel corpus has been reported approximately 87% in (Dinh Dien et al., 2002b). For full details of word alignment of this EVC corpus (precision, recall, coverage, etc.), please refer to (Dinh Dien et al., 2002a).

The result of this word-aligned parallel corpus has been used in various Vietnamese NLP tasks, such as in training the Vietnamese word segmenter (Dinh Dien et al., 2001a), word sense disambiguation (Dinh Dien, 2002b), etc.

Remarkably, this EVC includes the SUSANNE corpus (Sampson, 1995) – a golden corpus has been manually annotated such necessary English linguistic annotations as lemma, POS tags, chunking tags, syntactic trees, etc. This English corpus has been translated into Vietnamese by English teachers of Foreign Language Department of Vietnam University of HCM City. In this paper, we will make use of this valuable annotated corpus as the training corpus for our bootstrapped English POS-tagger.

No.	Resources	The number of pairs of sentences	Number of English words	Number of Vietnamese morpho-words	Length (English words)	Percent (words/ EVC)
1.	Computer books	9,475	165,042	239,984	17.42	7.67
2.	LLOCE dictionary	33,078	312,655	410,760	9.45	14.53
3.	EV bilingual dictionaries	174,906	1,110,003	1,460,010	6.35	51.58
4.	SUSANNE corpus	6,269	131,500	181,781	20.98	6.11
5.	Electronics books	12,120	226,953	297,920	18.73	10.55
6.	Children's Encyclopedia	4,953	79,927	101,023	16.14	3.71
7.	Other books	9,210	126,060	160,585	13.69	5.86
	Total	250,011	2,152,140	2,852,063	8.59	100%

Table 1. Resources of EVC corpus

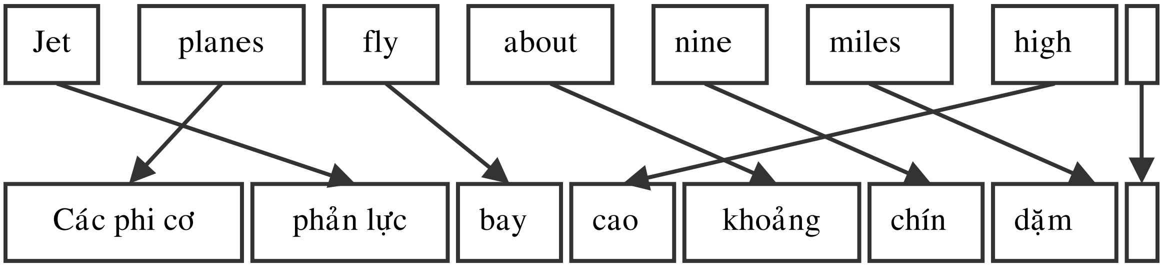

Figure 1. An example of a word-aligned pair of sentences in EVC corpus

4 Our Bootstrapped English POS-Tagger

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 (table 4). 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 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 word' s 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 that 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 (section 2) with the SUSANNE training corpus (section 3). 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.

4.1 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 (2001b) 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, 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 TreeBank tagset and for Vietnamese tagset, we use the POS-tagset mapping table (see appendix A).
- $S = \chi$xC: 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 i^th^ 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, whereat, x will be changed into a new tag c.
- Giving a state s = (x,c) and rule r = (π, 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 = (π, c) is usually the difference in performance (on the training data) that results from applying the rule, as follows:

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

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

4.2 The TBL algorithm for POS-Tagging

The TBL algorithm for POS-tagging can be briefly described as follows (see the flowchart in figure 2):

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 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 k^{th}, 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.

* Convergence ability of the algorithm: call e_k the number of error (the difference between the tagging result in conformity with rule r and the correct tag in the golden corpus in time k^{th}), we have: $e_{k+1} = e_k -$ Score(r), since Score(r) > 0, so $e_{k+1} < e_k$ with all k, and $e_k \in N$, so the algorithm will be converged after limited steps.

* Complexity of the algorithm: O(n*t*c) where n: size of training set (number of words); t: size of possible transformation rule set (number of candidate rules); c: size of corpus satisfied rule applying condition (number of order satisfied predicate π).

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

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

Table 2. The result of Bootstrapped POS-tagger for English side in EVC.

It is thanks to exploiting the information of the corresponding Vietnamese POS that the English POS-tagging results are improved. If we use only available English information, it is very difficult for us to improve the output of Brown POS-tagger. Despite the POS-tagging improvement, the results can hardly said to be fully satisfactory due to the following reasons:

* The result of automatic word-alignment is only 87% (Dinh Dien et al., 2002a).

* It is not always true that the use of Vietnamese POS-information is effective enough to disambiguate the POS of English words (please refer to table 3).

Through the statistical table 3 below, the information of Vietnamese POS-tags can be seen as follows:

- Case 1,2,3,4: no need for any disambiguation of English POS-tags.
- Case 5, 7: Full disambiguation of English POS-tags (majority).
- Case 6, 8, 9: Partial disambiguation of English POS-tags by TBL-method.

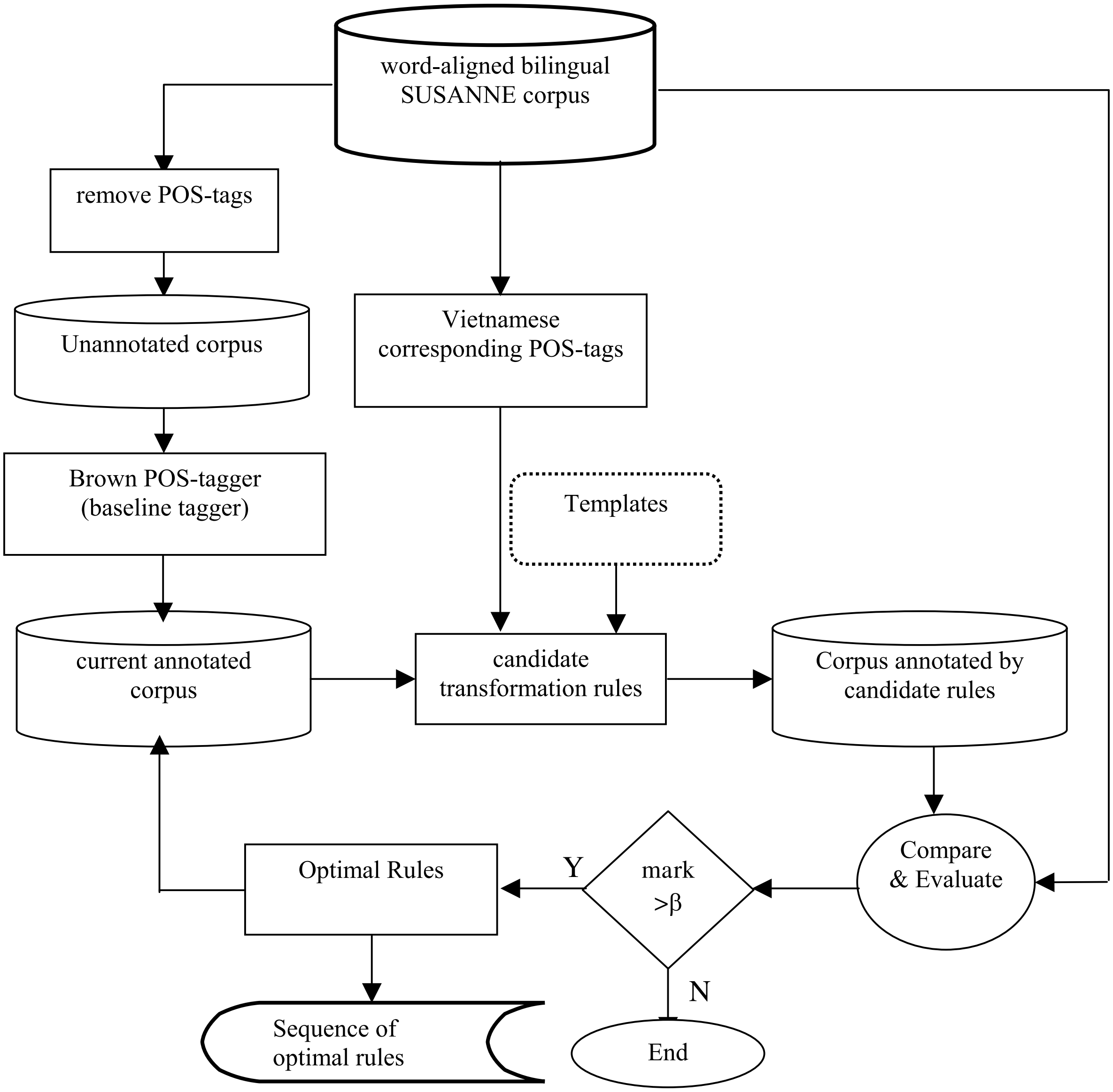
word-aligned bilingual SUSANNE corpus
remove POS-tags
Unannotated corpus
Brown POS-tagger (baseline tagger)
Vietnamese corresponding POS-tags
Templates
current annotated corpus
candidate transformation rules
Corpus annotated by candidate rules
Compare & Evaluate
mark >β
Y
N
Optimal Rules
Sequence of optimal rules
End

Figure 2. Flowchart of TBL-algorithm in POS-tagger for EVC corpus

No.	English POS-tags	Vietnamese POS-tags	Contrast English vs. Vietnamese POS-tags	Percent %
1.	One POS-tag only	One POS-tag only	Two POS-tags are identical	25.2
2.	One POS-tag only	One POS-tag only	Two POS-tags are different	1.2
3.	One POS-tag only	More than 1 POS-tag	One common POS-tag only	5.3
4.	One POS-tag only	More than 1 POS-tag	No common POS-tag	3.5
5.	More than 1 POS-tag	One POS-tag only	One common POS-tag only	50.5
6.	More than 1 POS-tag	One POS-tag only	No common POS-tag	2.8
7.	More than 1 POS-tag	More than 1 POS-tag	One common POS-tag only	6.1
8.	More than 1 POS-tag	More than 1 POS-tag	More than 1 common POS-tag	4.1
9.	More than 1 POS-tag	More than 1 POS-tag	No common POS-tag	1.3

Table 3. Contrast POS-tag of English and Vietnamese in the word-aligned EVC

5 Projecting English POS-Tags to Vietnamese

After having English-POS-tag annotations with high precision, we proceed to directly project those POS-tag annotations from English side into 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 two languages are different. Due characteristics of each language, we must use two different tagset for POS-tagging. Regarding English, we made use of available POS-tagset of PennTreeBank. While in Vietnamese, we made use of POS-tagset in the standard Vietnamese dictionary of Hoang Phe (1998) and other new tags. So, we must have an English-Vietnamese consensus tagset map (please refer to Appendix A).

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

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

Regarding evaluation of POS-tag projections, because so far, there has been no POS-annotated corpus available for Vietnamese, we had to manually build a small golden corpus for Vietnamese POS-tagging with approximately 1000 words for evaluating. The results of Vietnamese POS-tagging is as table 5 below:

Method	Correct tags	Incorrect Tags	Precision
Baseline tagging (use information of POS-tag in dictionary)	823	177	82.3%
Projecting from English side in EVC	946	54	94.6%

Table 5. The result of projecting POS-tags from English side to Vietnamese in EVC.

6 Conclusion

We have just presented the POS-tagging for an automatically word-aligned English-Vietnamese parallel corpus 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 POS-tagging of Vietnamese in the English-Vietnamese bilingual corpus plays a meaningful role in the building of the automatic training corpus for the Vietnamese processors in need of parts of speech (such as Vietnamese POS-taggers, Vietnamese parser, 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 English POS-tagging of the currently powerful English POS-taggers.

Currently, we are improving the speed of training period by using Fast TBL algorithm instead of TBL one.

In the future, we will improve this serial POS-tagging to the parallel POS-tagging for both English and Vietnamese simultaneously after we obtain the exact Vietnamese POS-tags in the parallel corpus of SUSANNE.

Acknowledgements

We would like to thank Prof. Eduard Hovy (ISI/USC, USA) for his guidance as external advisor on this research.

References

E. Brill. 1993. *A Corpus-based approach to Language Learning*, PhD-thesis, Pennsylvania Uni., USA.

E. Brill. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics*, 21(4), pp. 543-565.

E. Brill. 1997. Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging. In *Natural Language Processing Using Very Large Corpora.* Kluwer Academic Press.

J. Curran. 1999. Transformation-Based Learning in Shallow Natural Language Processing, *Honours Thesis*, Basser Department of Computer Science, University of Sydney, Sydney, Australia.

E. Charniak. 1997. Statistical parsing with a context-free grammar and word statistics, in *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, AAAI Press/MIT Press, Menlo Park.

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

W. Daelemans, J. Zavrel, P. Berck, S. Gillis. 1996. MTB: A Memory-Based Part-of-Speech Tagger Generator. In *Proceedings of 4th Workshop on Very Large Corpora*, Copenhagen.

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

D. Dien. 2001b. *Building an English-Vietnamese bilingual corpus*, Master thesis in Comparative Linguistics, University of Social Sciences and Humanity of HCM City, Vietnam.

D. Dien, H.Kiem, T.Ngan, X.Quang, Q.Hung, P.Hoi, V.Toan. 2002a. Word alignment in English – Vietnamese bilingual corpus, *Proceedings of EALPIIT'02*, Hanoi, Vietnam, 1/2002, pp. 3-11.

D.Dien, H.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, 9/2002, pp. 26-32.

R. Florian, and G. Ngai. 2001a. Transformation-Based Learning in the fast lane, *Proceedings of North America ACL-2001*.

R. Florian, and G. Ngai. 2001b. Fast Transformation-Based Learning Toolkit. *Technical Report*.

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

H. Phe. 1998. *Từ điển tiếng Việt* (Vietnamese Dictionary). Center of Lexicography. Da Nang Publisher.

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

H. Schmid. 1994a. Probabilistic POS Tagging using Decision Trees, *Proceedings of International Conference on New methods in Language Processing,* Manchester, UK.

H. Schmid. 1994b. POS Tagging with Neural Networks, *Proceedings of International Conference on Computational Linguistics,* Kyoto, Japan, pp.172-176.

D. Yarowsky and G. Ngai. 2001. Induce, Multilingual POS Tagger and NP bracketer via projection on aligned corpora, *Proceedings of NAACL-01*.

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 (Whdeterminer)	P
WP (Wh-pronoun)	P
WP$ (Possessive wh-pronoun)	"của" P
WRB (Wh-adverb)	R