The Parallel Corpus Approach to Building the Syntactic Tree Transfer Set in the English-to-Vietnamese Machine Translation

Dinh Dien, Thuy Ngan, Xuan Quang, Chi Nam
Faculty of Information Technology, University of Natural Sciences,
Vietnam National University of HCM City
Email: ddien@fit.hcmuns.edu.vn, nltngan@fit.hcmuns.edu.vn, dxquang@fit.hcmuns.edu.vn

Abstract:
Recently, with the machine learning trend, most of the machine translation systems on over the world use two syntax tree sets of two relevant languages to learn syntactic tree transfer rules. However, for the English-Vietnamese language pair, this approach is impossible because until now we have not had a Vietnamese syntactic tree set which is correspondent to English one. Building of a very large correspondent Vietnamese syntactic tree set (thousands of trees) requires so much work and take the investment of specialists in linguistics.
To take advantage from our available English-Vietnamese Corpus (EVC) which was tagged in word alignment, we choose the SITG (Stochastic Inversion Transduction Grammar) model to construct English-Vietnamese syntactic tree sets automatically. This model is used to parse two languages at the same time and then carry out the syntactic tree transfer. This English-Vietnamese bilingual syntactic tree set is the basic training data to carry out transferring automatically from English syntactic trees to Vietnamese ones by machine learning models.
We tested the syntax analysis by comparing over 10,000 sentences in the amount of 500,000 sentences of our English-Vietnamese bilingual corpus and first stage got encouraging result (analyzed about 80%)[5]. We have made use the TBL algorithm (Transformation Based Learning) to carry out automatic transformations from English syntactic trees to Vietnamese ones based on that parallel syntactic tree transfer set[6].

Keywords: SITG, EVC, syntactic tree transfer, bilingual corpus, TBL.

1. INTRODUCTION

In machine translation, the syntactic transfer stage is one of the most importance stages since it takes a role in making translated sentences smooth. There are many approaches proposed in syntactic transfer stage and shown their advantages. It, however, belongs to the features of source and target languages. Among these approaches, the ones that use two sets of syntactic tree in syntactic transferring get significant results (Stuart et al, 1990). It is simple and rapid if we have both source and target tree sets. In Vietnamese, the syntactic tree set has not fully made due to the fact that the research in Vietnamese syntactic parsing has not developed adequately. The English syntactic one, on the contrary, has been researched and proposed many efficient models in syntax such as TAG (Stuart et al, 1990), CONTEX... and there are several English syntactic tree sets that have been developed. From that, we are very interested in constructing the tree set for Vietnamese. This is the first step in extracting the transfer rules for structural transfer component in the machine translation system, which have the corpus-based approach, like English-Vietnamese one. In the context of Vietnamese, we have to do from the beginning. We surveyed and chose the SITG model, which has been applied for the English-Chinese pair successfully (Wu, 1995). Chinese and Vietnamese have the same linguistic type, isolation, so we hope to succeed when apply for English-Vietnamese pair. The most advance feature of this model is that its input is only the bilingual corpus pre-aligned in sentence level. This condition is reasonable for our paradigm, a corpus-based one had over 500,000 sentence pairs pre-aligned. After modifying the algorithm in order to be adapting with Vietnamese we applied it for our parsing component, the practical result shows that it is the suitable model for English-Vietnamese translation system. Over eighty percent sentence pairs were parsed correctly (Dien et al, 2003). For the higher quality of transfer stage, however, we did not come to a halt. The outputs of SITG model, the bilingual syntactic tree sets, were used as the basis training data for the progress of machine learning in extracting transfer rules to transfer automatically from English syntactic trees to Vietnamese ones (Dien et al, 2003). In this paper we will present the SITG model applied for English-Vietnamese pair and the phase of learning transfer rules using TBL (Transformation Based Learning) model.

2. STOCHASTIC INVERSION TRANSDUCTION GRAMMARS (SITG)

We present some basic concepts before we can go on the main contents.
Transduction grammar is the grammar used to describe a language pair having the correlation in structures. The most useful feature of this grammar is in the feature of producing transduction. It means that we will have two outputs for two relative languages simultaneously.

Simple transduction grammar which is the limitation of the transduction grammar has the feature of syntactic orientation and freedom in context. In this paper, we focus on the property of freedom in context due to the fact that we concern on the structural knowledge of neither English nor Vietnamese.

The inversion transduction grammar model (ITG) is the expansion of simple transduction grammar (Wu, 1995). The ITG can increase the system ability in producing rules. In ITG, the operator [ ] outputs a string pair in a normal order from a string input. For example, apply operator [ ] on the string "*xy*", where "*x*" and "*y*" are two constituents. We have:

$$[xy] = (xy, xy)$$

Whereas, the operator $\langle\ \rangle$ combines the constituents in the first output string in a normal order, but invert the order in the second output string. We can see that in the following example:

$$\langle xy \rangle = (xy, yx)$$

This model proved that it is effective for language pair having the inversion in word order like English-Vietnamese.

Parsing in the context of ITG means to take and input of a sentence pair in L_1 and L_2 then output a parsed tree where the structure of the syntactic tree in L_1 and the other in L_2 are both imposed. For example, let (E) is the English sentence and (V) is the corresponding Vietnamese sentence, we have a English-Vietnamese sentence pair:

(E) I have read that interesting book.

(V) Tôi đã đọc quyển sách thú vị đó rồi.

Figure 1 shows the parse tree of this sentence pair in the English-Vietnamese translation system. The English constituents are read in normal order (left-node-right), but for the Vietnamese ones, the horizontal lines show the inversion between the left and the right sub-trees.

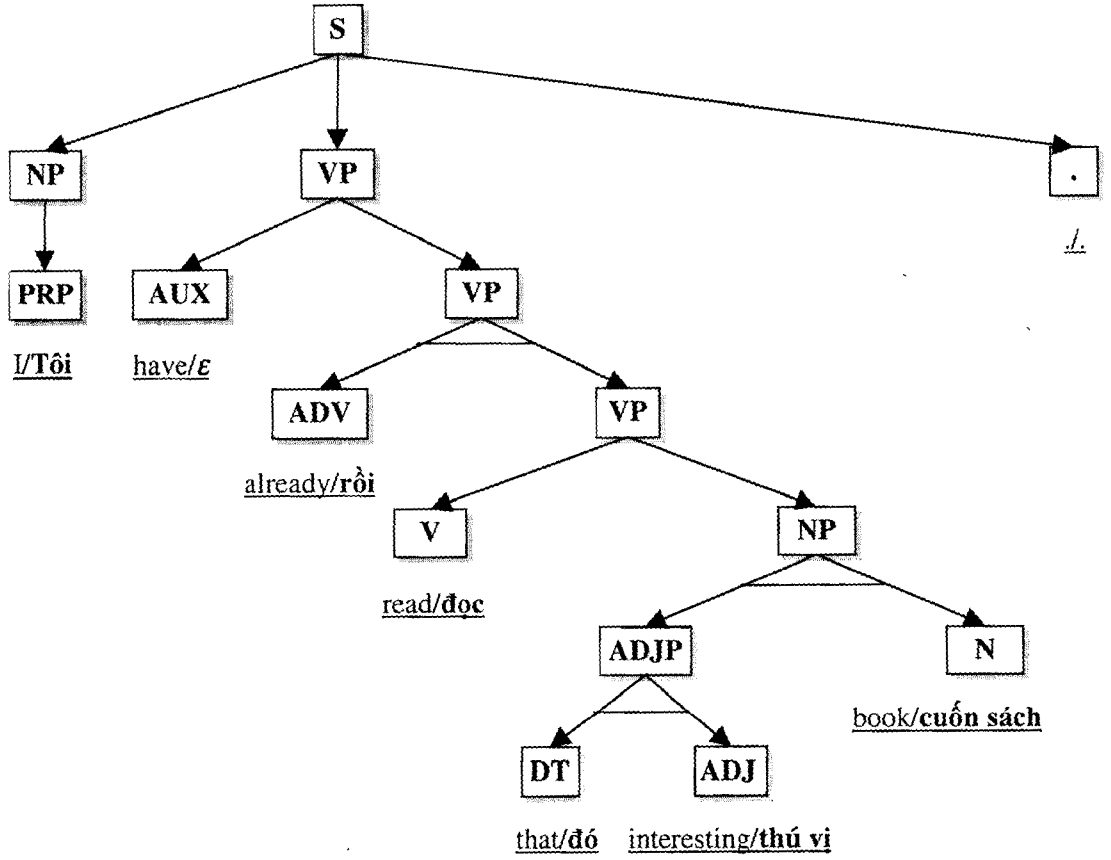

Figure 1 The parsed tree in inversion transduction grammar

We can also write this parsed tree in the bracket type like the following:

[[I/tôi]NP [have/ε <already/rồi [read/đọc <<that/đó interesting/thú vị>ADJP book/cuốn sách>NP]VP >VP]VP .]

The stochastic inversion transduction grammar (SITG) attaches each grammar rule to a probability. For example, the probability of the grammar rule $NP \xrightarrow{0.4} [ADJ\ NN]$ is $a_{NP \to [ADJ\ NN]} = 0.4$; the probability of the lexical rule $A \xrightarrow{0.001} u \mid v$ is $b_A(u, v) = 0.001$, where u and v are symbols of source and target languages referred as L1 and L2 (so a and b is the probability of the grammar and lexical rule respectively).

The SITG guarantees that the probability of the grammar and lexical rules has to satisfy the constraint:

$$\sum_{1\le j,k\le N}\left(a_{i\to[jk]}+a_{i\to\langle jk\rangle}\right)+\sum_{\substack{1\le u\le w_1\\1\le v\le w_2}} b_i(u,v)=1$$

where

$a_{i\to[jk]}=P(i\to[jk]\mid i)$ and $b_i(u,v)=P(i\to u/v\mid i)$, N, W1 and W2 are the number of non-terminal and the size of vocabulary in two languages respectively.

The algorithm for parallel parsing based on the SITG calculates the best parsed tree by using the linear programming. In the parallel parsing, the calculating probability for grammatical and lexical rules will give us the clues to solve the ambiguities by choosing the parsed tree with highest probability. The following is the algorithm in parallel parsing.

Let $\mathbf{e}_1\ldots\mathbf{e}_X=\mathbf{e}_{1\ldots X}$ is the English input sentence, and $\mathbf{v}_1\ldots\mathbf{v}_Y=\mathbf{v}_{1\ldots Y}$ is the corresponding Vietnamese input sentence. Let $\mathbf{e}_{s+1}$, $\mathbf{e}_{s+2}$, ..., $\mathbf{e}_t$ is $\mathbf{e}_{s\ldots t}$ and $\mathbf{v}_{u+1}$, $\mathbf{v}_{u+2}$, ..., $\mathbf{v}_v$ is $\mathbf{v}_{u\ldots v}$. The empty string is denoted by $\mathbf{e}_{s\ldots s}=\mathbf{v}_{u\ldots u}=\varepsilon$

For every node q in parsed tree we denote $q=(s, t, u, v)$ for the derivation of sub-strings $\mathbf{e}_{s\ldots t}$ and $\mathbf{v}_{u\ldots v}$. We denote the non-terminal label for q as $\ell(q)$. Then the maximum probability of any derivation from i that allows to parse both $\mathbf{e}_{s\ldots t}$ and $\mathbf{v}_{u\ldots v}$ successfully is:

$$\delta_q(i)=\delta_{stuv}(i)=\max_{subtrees\ of\ q} P[\text{ subtree of } q,$$

$$\ell(q)=i, i\overset{*}{\Rightarrow}\mathbf{e}_{s..t}/\mathbf{v}_{u..v}]$$

The best parsing for the English-Vietnamese sentence pair will have the probability: $\delta_{0,T,0,V}(S)$.

We have some initial probabilities:

$$\delta_{t-1,t,v-1,v}(i)=b_i(\mathbf{e}_t/\mathbf{v}_v),\ \begin{cases}1\le t\le X\\1\le v\le Y\end{cases};$$

$$\delta_{t-1,t,v,v}(i)=b_i(\mathbf{e}_t/\varepsilon),\ \begin{cases}1\le t\le X\\0\le v\le Y\end{cases}$$

$$\delta_{t,t,v-1,v}(i)=b_i(\varepsilon/\mathbf{v}_v),\ \begin{cases}0\le t\le X\\1\le v\le Y\end{cases};$$

and the recursive expressions:

For every i,s,t,u,v that $\begin{cases}1\le i\le N\\0\le s\le t\le X\\0\le u\le v\le Y\\t-s+u-v>2\end{cases}$ we have

$$\delta_{stuv}(i)=\max[\delta^{[]}_{stuv}(i),\delta^{\langle\rangle}_{stuv}(i)]$$

$$\theta_{stuv}(i)=\begin{cases}[] & khi\ \ \delta^{[]}_{stuv}(i)\ge\delta^{\langle\rangle}_{stuv}(i)\\ \langle\rangle & khi\ \ \delta^{[]}_{stuv}(i)<\delta^{\langle\rangle}_{stuv}(i)\end{cases}$$

where

$$\delta^{[]}_{stuv}(i)=\max_{\substack{1\le j\le N\\1\le k\le N\\s\le S\le t\\u\le U\le v\\(S-s)(t-S)+(U-u)(v-U)\ne0}} a_{i\to[jk]}\delta_{sSuU}(j)\delta_{StUv}(k)$$

$$\begin{bmatrix}\iota^{[]}_{stuv}(i)\\ \kappa^{[]}_{stuv}(i)\\ \sigma^{[]}_{stuv}(i)\\ \upsilon^{[]}_{stuv}(i)\end{bmatrix}=\operatorname*{argmax}_{\substack{1\le j\le N\\1\le k\le N\\s\le S\le t\\u\le U\le v\\(S-s)(t-S)+(U-u)(v-U)\ne0}} a_{i\to[jk]}\delta_{sSuU}(j)\delta_{StUv}(k)$$

and

$$\delta^{\langle\rangle}_{stuv}(i)=\max_{\substack{1\le j\le N\\1\le k\le N\\s\le S\le t\\u\le U\le v\\(S-s)(t-S)+(U-u)(v-U)\ne0}} a_{i\to\langle jk\rangle}\delta_{sSuU}(j)\delta_{StUv}(k)$$

$$\begin{bmatrix}\iota^{\langle\rangle}_{stuv}(i)\\ \kappa^{\langle\rangle}_{stuv}(i)\\ \sigma^{\langle\rangle}_{stuv}(i)\\ \upsilon^{\langle\rangle}_{stuv}(i)\end{bmatrix}=\operatorname*{argmax}_{\substack{1\le j\le N\\1\le k\le N\\s\le S\le t\\u\le U\le v\\(S-s)(t-S)+(U-u)(v-U)\ne0}} a_{i\to\langle jk\rangle}\delta_{sSuU}(j)\delta_{StUv}(k)$$

In practical, the optimized sub-node q = (s, t, u, v) of the parsed tree is compute by the recursive expression:

$$\text{LEFT}(q)=\begin{cases}\text{NIL} & \text{if}\ \ t-s+v-u\le2\\ (s,\sigma^{[]}_q(\ell(q)),u,\upsilon^{[]}_q(\ell(q))) & \text{if}\ \ \theta_q(\ell(q))=[]\\ (s,\sigma^{\langle\rangle}_q(\ell(q)),\upsilon^{\langle\rangle}_q(\ell(q))) & \text{if}\ \ \theta_q(\ell(q))=\langle\rangle\end{cases}$$

$$\text{RIGHT}(q)=\begin{cases}\text{NIL} & \text{if}\ \ t-s+v-u\le2\\ (\sigma^{[]}_q(\ell(q)),t,\upsilon^{[]}_q(\ell(q)),v) & \text{if}\ \ \theta_q(\ell(q))=[]\\ (\sigma^{\langle\rangle}_q(\ell(q)),t,u,\upsilon^{\langle\rangle}_q(\ell(q))) & \text{if}\ \ \theta_q(\ell(q))=\langle\rangle\end{cases}$$

$$\ell(\text{LEFT}(q))=\iota_q^{\theta_q(\ell(q))}(\ell(q))$$

$$\ell(\text{RIGHT}(q))=\kappa_q^{\theta_q(\ell(q))}(\ell(q))$$

For this algorithm, the complexity is $\mathbf{O}(N^3X^3Y^3)$, where N is the number of non-terminal, X and Y are the size of sentences in English and Vietnamese.

3. LEARNING TRANSFER RULES

After applying this algorithm on our corpus (with over 500,000 English-Vietnamese sentence pairs), we get the English-Vietnamese syntactic tree set that contain both English parsed tree and corresponding Vietnamese parsed tree. This English-Vietnamese bilingual syntactic tree set is the basic training data for our machine learning model to extract transfer rule. We use machine learning method TBL (Brill, 1993) to extract rules from the syntactic tree set we create above.

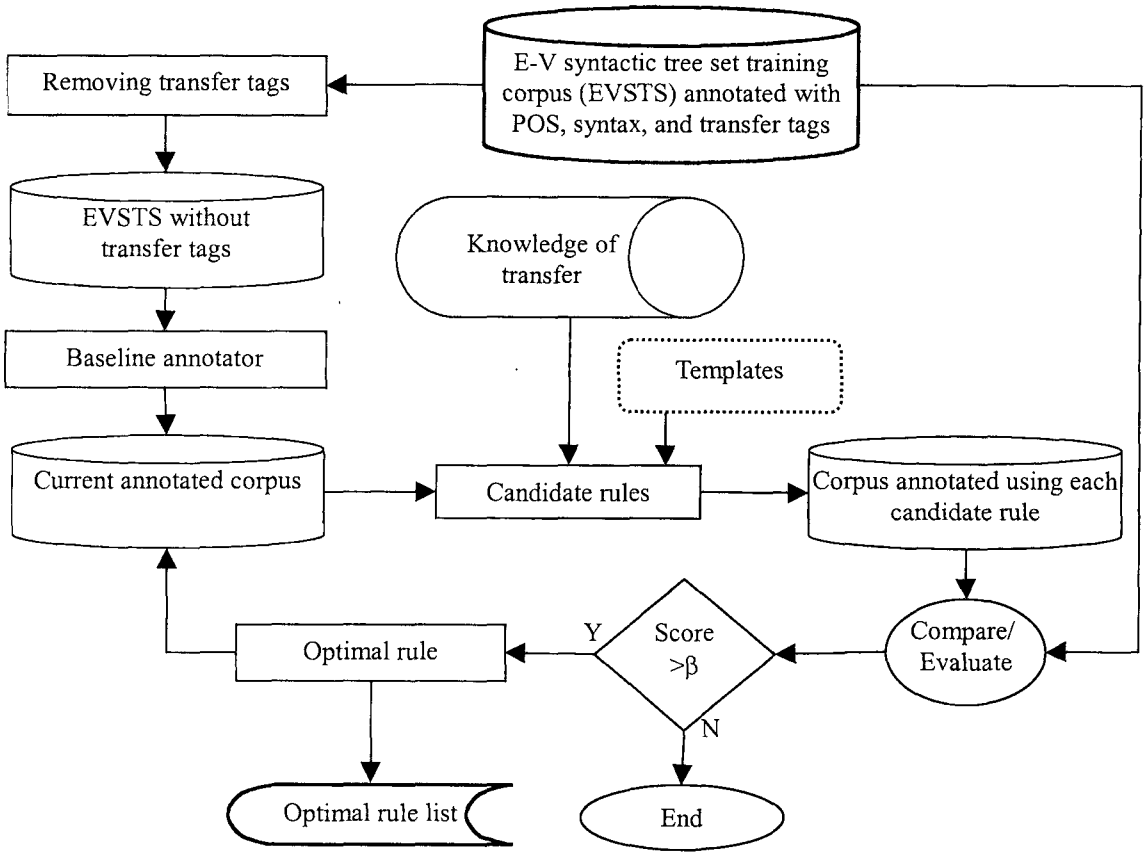

Figure 2 Flowchart of TBL- algorithm in learning transfer rules

4. EXPERIMENT

We carried out parsing parallel corpus by using SITG on our English-Vietnamese corpus with over 500,000 sentence pairs. This corpus is collect from many sources. However, almost pairs are in the science and technology field.

The accuracy of parallel parsing showed in the chart below. The accuracy seems to be decreased when the number of sentence pairs increased. The average accuracy, nevertheless, is about 80%.

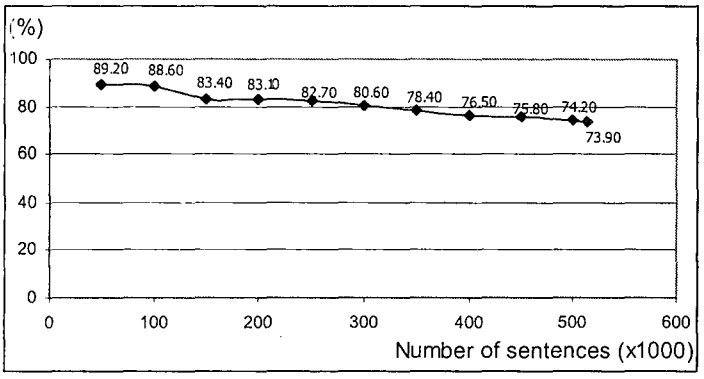

Figure 3 The accuracy of parsing with SITG

As we presented in [5], starting from the idea of Keh-Yih Su *et al.* (1992) about estimation measure of English-Chinese translation system, we change the problem of comparison between Vietnamese sentence transferred by machine with the one translated by human being into the problem of finding the shortest route between two points: from the departure point R and the target point Q. Where $R = \{c_{11}, c_{12}, ..., c_{1m}\}$ is the Vietnamese sentence translated by machine having m words, and $Q = \{c_{11}, c_{12}, ..., c_{1n}\}$ is the sentence translated by human being having n words.

Let $D = w_d*n_d + w_i*n_i + w_r*n_r + w_s*n_s$ be the cost of transform R into Q, where n_d, n_i, n_r and n_s are the number of deleting (DEL), inserting (INS), replacing (REP) and

swapping (SWAP) actions, and w_d, w_i, w_r and w_s are corresponding weights depending on language and experience. In our work w_d=1, w_i=5, w_r=5 and w_s=6.
From R to Q, we can have many routines with different costs.
These 2 routines (may have others) corresponding to two graphs (Figure 1). Where the horizontal line is DEL, vertical line is INS, diagonal is REP or SWAP. The accuracy of translation is computed as $k=m/D_{min}$ where D_{min} is the shortest routine (lowest cost) from R to Q and m is the length of Q. We use dynamic programming technique to compute D_{min}.
When the new text is translated, the system analyses each sentence of the text to get the structures and then the transfer rule will be applied to force the English structure to Vietnamese sentence structure. The source tree is matched against the left hand sides of the transfer rules which have been extracted. This process will looking for the rule which has the left hand side similarly to the English sentence structure, if the corresponding structure is found, the Vietnamese structure will be generated from the right hand side of that rule. But we may encounter the problem that there are more than one rule satisfying the English structure. In that case, the frequency of the English in training corpus will play an important role in choosing the best structure to transfer.
After applying transfer rules to our test data, we count the number of wrong alignments when compared with that golden corpus, and compute the accuracy of transfer task by dividing that number by the total of alignments. By this way of evaluation, the accuracy of our transfer module is 95%.

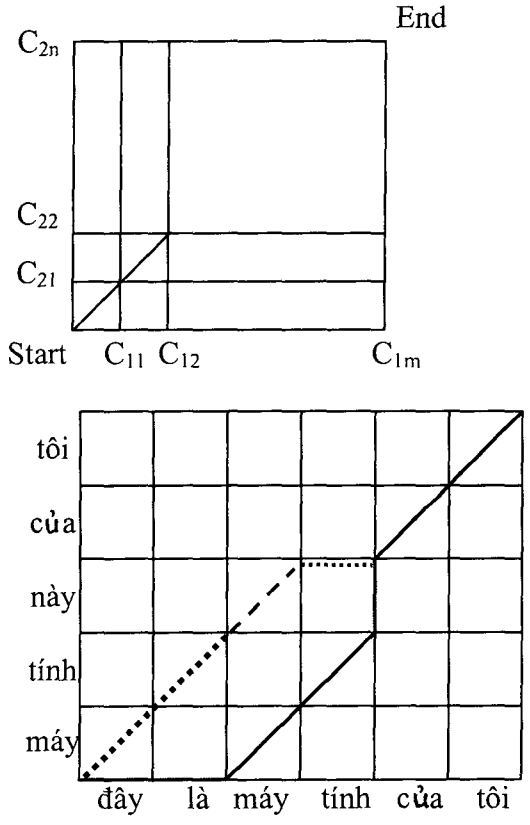

Figure 4 Routine from R to Q

5. CONCLUSION

So far, the transfer module in our English-Vietnamese translation system is proved to be better than other previous transfer methods that we experimented based on experts' hand-crafted transfer rules. But in our opinions, the quality of the transfer along with the transfer-based machine translation can be improved if we make use of our rich corpus resources for training. Although the algorithm based on the SITG presented above has been refined but the speed is still rather slow. In the future, we would like to decrease the training time as much as possible.

References

[1] Adam Meyers; Roman Yangarber; Ralph Grishman; Catherine Macleod; Antonio Moreno-Sandoval, 1998, "Deriving Transfer Rules from Dominance-Preserving Alignments", Proceedings of the 36th ACL, Monreal, Canada, pg. 843-847

[2] Can Nguyen Tai, "Ng pháp ti ng Vi t", The National University of Hanoi Publisher, 1998.

[3] Dekai Wu. 1995a. "An Algorithm for Simultaneously Bracketing Parallel Texts by Aligning Words". ACL-95. 33nd Annual Meeting of the Assoc. for Computational Linguistics, Cambridge, MA: Jun. 95

[4] Dekai Wu. 1994, "Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora", Computational Linguistics, 23(3): 377-403.

[5] Dinh Dien, Thuy Ngan, Xuan Quang, Chi Nam, 2003, "Automatic Tree Transfer in English-Vietnamese Machine Translation", Proceedings of CICT'03, 2/2003, Hanoi, Vietnam, pp. 7-12.

[6] Dien Dinh, Thuy Ngan, Xuan Quang, Chi Nam, 2003, "A hyrid approach to word-order transfer in the English – Vietnamese Machine Translation System", Proceedings of the Machine Translation Summit IX, Louisiana, USA, pp. 79-86.

[7] Dinh Dien, 2001, "Building English-Vietnamese bilingual corpus", Master thesis in Comparative Linguistics of University of Social Sciences and Humanity of VNU of HCM City.

[8] Dien Dinh, Kiem Hoang, 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, pg.26-32.

[9] Eric Brill, 1993, "A Corpus-based approach to Language Learning", PhD-thesis, Pennsylvania Uni., USA.

[10] Keh-Yih Su, Ming-Wen Wu, Jing-Shin Chang, 1992, "A New Quantitative Quality Measure for Machine Translation System", Proceedings of COLING-92, Nantes, France, pp. 433-439.

[11] Stuart Sheiber and Yves Schabes. 1990, "Synchronous Tree Adjoining Grammars", Proceedings of the 13th COLING-90, Helsinki, ACL.

[12] Y. Matsumoto, H. Ishimoto. T. Utsuro, and M. Nagao, 1993, "Structural Matching of Parallel Texts", ACL93.