A Novel Template-based Architecture for Spoken Language Translation on Personal Digital Assistants

Jhing-Fa Wang **Shun-Chieh Lin** **Hsueh-Wei Yang** **Fan-Min Li**
Dept. Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C.
email wangjf@csie.ncku.edu.tw

Abstract
Two major problems for an automatic spoken language translation on a handheld device are limited resources and real-time constraints. In this paper, we focus on developing techniques to overcome these problems based on our novel template-based approach. First is to derive demand translation templates for memory saving and translation capacity promoting. Second is to develop the fast template-based translation approach composed of crude template candidate selection, latent grammar understanding of templates, score normalization and ranking, and translation result generation. According to the experimental results, the proposed approach can averagely achieve about 76.5% translation understanding rate and 6.23 sec response time on 200MHz PDAs compared with 5 sec in traditional 1~2GHz PC-based approaches.

Keywords
Spoken Language Translation, Personal Digital Assistant, Template-based Architecture

INTRODUCTION

Automatic spoken language translation (SLT) has been one of prospective applications of the speech and language technology [1][2][3][4]. One common target application has been communication assistance for travelers in a foreign travel situation. In this situation, the user will want to carry the translation device and use it in an open space, rather than sit in a computer room in front of a display. An obvious approach is to build a stand-alone speech translation device that is compact enough to carry around, like personal digital assistants (PDAs) [4][5][6].

Currently, there are two main architectures of speech translation: conventional sequential architecture and fully integration architecture [7]. In the integrated architecture, speech feature models are integrated into translation models in the similar way as for speech recognition. According to this integration, the translation process can be efficiently performed by searching for an optimal path of states through the integrated network [8]. Therefore, we adopt the concept of the integrated architecture and revise to provide better quality and real-time response for spoken language translation that can run on resource-limited PDAs.

For translation models in the integrated architecture, one of the data-driven approaches, which has been proposed for the integrated architecture, is Statistical-based Machine Translation (SBMT)[9][10]. However, when the training data of SBMT is insufficient, the results obtained by the sequential architecture are better than the results obtained by the integrated architecture [8]. In addition, word reordering is still a thorny problem in SBMT, and present alignment models of SBMT for suboptimal solutions seem to be insufficient [11]. Another remarkable data-driven approach, Example-based Machine Translation (EBMT), does not require the database to be as large as in SBMT. Furthermore, the example-based approach can explore the alignments between word sequences and syntactic grammars for language translation. Various example-based speech translation methods have been widely used and their effectiveness for spoken language processing has been confirmed with the help of domain specificity [12]. For that reason, we adopt and further systematize the example-based approach for SLT.

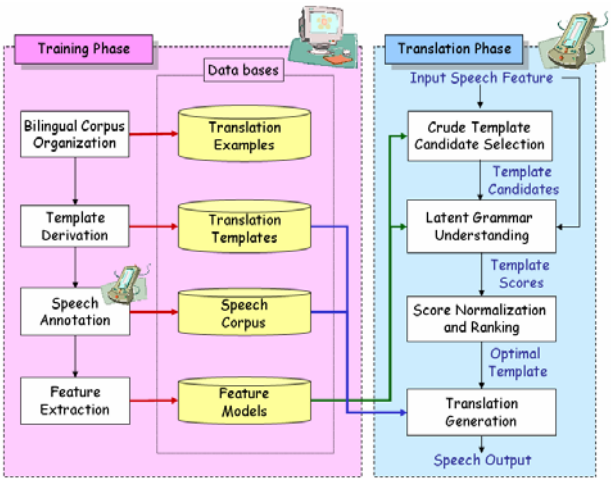

Figure 1. Architecture of the proposed system

The proposed framework involves two key components, the training phase and the translation phase, as shown in Fig. 1. The training phase is to derive four major required databases. During the translation phase, we use the derived databases to implement the spoken language translation on PDAs. The rest of the paper is organized as follows. First, discussing the construction of translation templates. Then, the proposed speech translation approach and the experimental results are presented in turn. Finally, we draw a generalized conclusion.

CONSTRUCTION OF TRANSLATION TEMPLATES

In this section, we will discuss three issues of our translation template construction underlying the example-based

approach. Starting with the organization for a database of examples, i.e. bilingual text corpora, we then derive the demand translation templates from the organized corpora, and finally, the related speech corpora are developed to apply for PDAs.

Bilingual text corpus organization

To deal with the speech translation, we have obtained the Taiwanese/English bilingual corpus[1]. By utilizing the corpus, we can acquire the alignments between word sequences and translation templates for language translation. However, there are a number of divergences, which make the straightforward mapping between languages impractical [13]. With regard to resource-limited handheld devices, these divergences also raise the cost of translation process and reduce the translation performance.

For Mandarin-English bilingual corpus, some divergence types, called thematic, structural, morphological, and conflational, are discussed [14]. On the other hand, Taiwanese and Mandarin both belong to the family of Chinese language [15], these divergences are also existed in the Taiwanese/English text corpus. Therefore, if the sentence pairs contain above a certain number of null mappings, the degree of alignment divergence comes into notice and the content of the sentence pairs has to be updated in order to improve the accuracy and effectiveness of the alignment execution and translation template derivation.

Translation template derivation

Figure 2 is an example of a simple translation template derived from one translated example indicating how a sentence in English (source language, SL) containing "Is there…" may be translated by a sentence in Taiwanese (target language, TL) containing "u…but"[2][16].

Example part: SL: *Is there* laundry service?
↔ TL: "*u* sea svaar e hogbu *but*"?
Alignment part: Is there E_p? ↔"u" T_p "but"?
alignments
Variable part: If $E_p \leftrightarrow T_p$,
laundry service ↔"sea svaar e hogbu"

Figure 2. A simple translation template

Functionality of such a simple template is like phrase books for travelers, which have fixed expressions and intentions, are often used as examples for translation with a replacement of the phrase variables. Besides, duplicate expressions can be projected into a unique template for memory saving. Therefore, we expect to expand more variable mappings for template flexibility and replacement rationality. An expansion strategy for multiform translation template generation [14] was developed and is adopted here. There are two levels of expansion: expansion within the intention class and expansion between intention classes. For the expansion within the intention class, the iterative procedure is shown as follows:

$$C_i \leftarrow C_i \oplus E_w(\chi_j, \chi_k), \quad \text{if } \tilde{\chi}_j = E_w(\chi_j, \chi_k) \text{ is classified correctly by } C_i\text{'}$$

where $E_w(\chi_j, \chi_k)$ expands template χ_j within template χ_k in class C_i with the variable mappings of template χ_k. For the expansion between the intention classes, the iterative procedure is as below:

$$C_i \leftarrow C_i \oplus E_b(\chi_j, C_k), \quad \text{if } \tilde{\chi}_j = E_b(\chi_j, C_k) \text{ is classified correctly by } C_i\text{'}$$

where $E_b(\chi_j, C_k)$ expands template χ_j between each template of its variable mappings in class C_k. After the expansion within and between the intention classes, the variable mappings in each template become more plentiful and reasonable for translation.

Speech data systemization

According to the derived translation templates, we now focus on how to efficiently record the information including speech data and feature models for promoting capacity of the templates. Within our proposed approach, we plan to unify the template-based linguistic translation knowledge and PDA-based speech features for spoken language translation. For feature extraction, we obtain the example speech and variable speech and are in chorus to reference model construction. The reference model v_i of one language is redefined as a two-layer model $v_i = \{v_i', v_i''\}$, where v_i' is an intention subspace of v_i and v_i'' is a variable subspace of v_i. The example of English is shown in Fig. 3. On the other hand, the obtained example speech and variable speech are also used for small-scale corpus-based speech generation. And the alignment information of translation templates is embedded in generation rules for speech replacement.

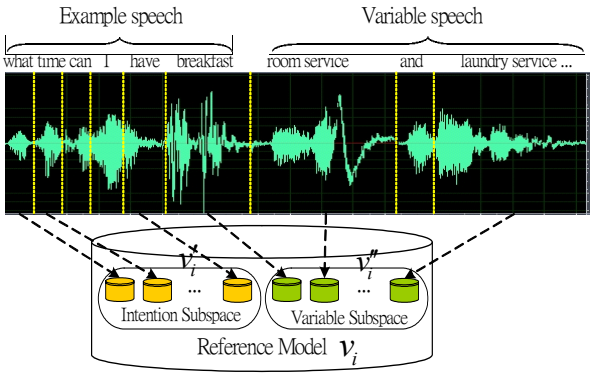

Figure 3. Speech annotation of data systemization

[1] The bilingual corpus is sentence-aligned and sentences are tokenized.

[2] Taiwanese words in phonetic transcription

SPOKEN LANGUAGE TRANSLATION

So far we have constructed the required translation templates with related speech data and feature models. In this section, according to the translation database, we propose to handle source speech input for real-time spoken language translation in the following way:

Crude template candidate selection

In order to speed up the determination of an optimal template, to preselect possible templates can achieve the requirement. Therefore, in the temporal fluctuation region between two speech spectra (source speech and intention speech), the range-limited dynamic time warping (DTW) solution is applied to measure efficiently the pattern dissimilarity with embedded time-normalization and alignment. In this paper, the pre-selection accuracy is estimated by the criterion that if the intention of source speech is located in the set of best N preselected templates. The criterion is best known as "TopN".

Latent grammar understanding for templates

Once the template candidates are retrieved, a one-stage based approach is applied to further segment the source speech input and to decide the optimal template with the two-layer model of each template candidate. In terms of searching for an optimal path of states through the two-layer model, the objective now is to measure the dissimilarity $d_i'^*$ of pair (s, v_i') of a fixed number of reference patterns, says N_i, for understanding the latent grammar in the intention subspace. (See Fig. 4) However, another problem is how to properly measure the dissimilarity score of pair (s, v_i') while adjudging the variable subspace v_i'' for segmenting the variable patterns of user's speech input.

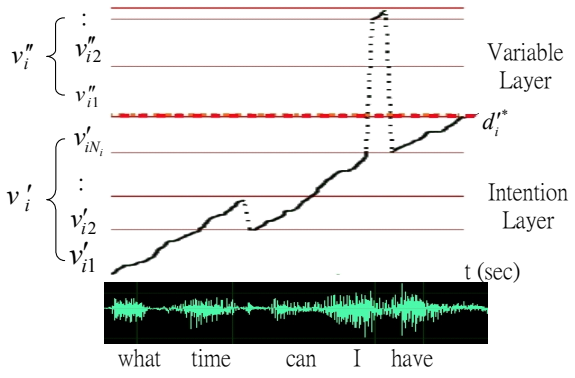

Figure 4. A one-stage based dissimilarity measurement

Referring to Fig 5., given two optimal matching paths of different preselected templates v_i and v_j, the scores used for comparing are the intention regions $D_i'^*$ and $D_j'^*$, where $D_i'^*$ is the dissimilarity measurement region of pair (v_i', s) and $D_j'^*$ is the dissimilarity measurement region of pair (v_j', s).

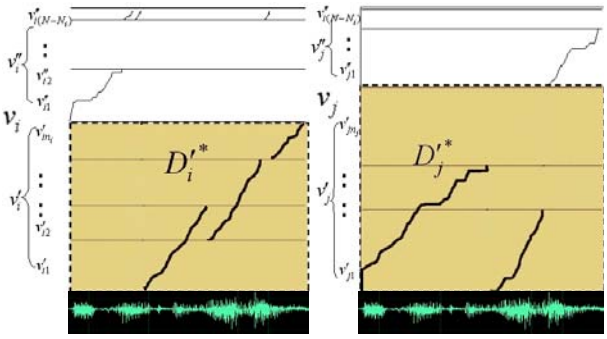

Figure 5. Search results of various reference models for speech input

For the consideration of the dissimilarity measurement between the intention layer and the variable layer, there are two additional types of search path registers for one-stage approach: 1) paths *between* v_i' and v_i'' and 2) paths *within* v_i' or v_i''. For the paths *between* v_i' and v_i'', a search block Z in the variable layer, which will be referred to a score-skip block while backtracking, contains more than one such path e connected by node u and node v. (See Fig. 6) Therefore, according to our proposed two-layer model, the one-stage based algorithm not only computes best paths to every reference pattern frame at every input frame, but also backtrack the desired score with giving the best word sequence.

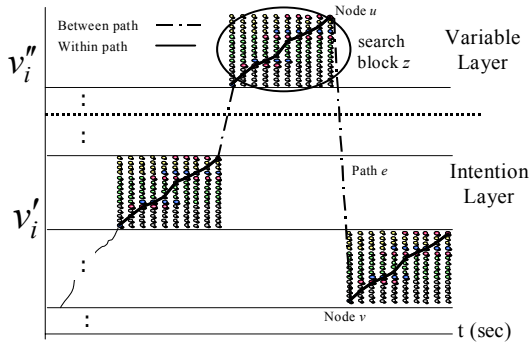

Figure 6. Additional search path registers for one-stage algorithm

Score normalization and ranking

Since the length of matching sequence can seriously affect the accumulative score of the dissimilarity measurement, a time-conditioned weight concept is further adopted to compensate for this defect. There are a number of similar ways to define the scoring methods with the length measurement of $\Delta(s, v_i')$ [17]:

$$\Delta(s, v_i') = \max(\|s\|, \|v_i'\|) (\text{or } \min(\|s\|, \|v_i'\|)) \quad (2)$$

$$\Delta(s, v_i') = \|s\| * \|v_i'\| \quad (3)$$

$$\Delta(s, v_i') = N(L, N_i) + F(L, N_i)/3, \quad (4)$$

where $\|s\|$ is the number of frames in speech input s, $\|v_i'\|$ is the number of total search frames in v_i', $N(L, N_i)$ is the number of frame comparison, and $F(L, N_i)$ is the number of frames that fail to get matched. In order to improve the flexibility of the dissimilarity measurement, an exponential definition for $\Delta(s, v_i')$ is described as follows:

$$\Delta(s, v_i') = \partial^{w_{v_i',s}}, \quad (5)$$

where ∂ is a weight factor, $w_{v_i',s} = \left(\|v_i'\| - \|s\|\right) \cdot \|s\|^{-1}$. Therefore, the modified dissimilarity measurement is given thereafter:

$$\tilde{d}_i'^{*} = d_i'^{*} \cdot \partial^{w_{v_i',s}} \quad (6)$$

The experimental analysis shows that the proper interval of ∂ is $[1.3-\delta, 1.3+\delta]$ to obtain the best accuracy of the dissimilarity measurement. Therefore, the value of ∂ chosen here is 1.3. After ranking all template candidates, the optimal template is decided by the one with minimum dissimilarity score.

Translation result generation

As soon as deciding the optimal template and smoothing the matching sequence of word patterns, we can exploit the gap between source speech input and the optimal template to generate the target speech. The generation process is straightforward as a corpus-based speech generation approach [15] by synthesizing example speech and variable speech. The process contains complete match, speech replacement, speech insertion and deletion, and composition in the example-based translation method [17].

EXPERIMENTAL RESULTS

A precision comparison test was performed to assess if the proposed template-based method translated on PDAs as well as on PC-based platform. Furthermore, the related experimental analysis of the proposed spoken language translation system is shown as follows.

The task and the corpus

The divergence between language pairs of bilingual corpus on unknown domain may be quite immeasurable when processing bilingual text corpus organization. Therefore, the bilingual corpus extended from specific domains is selected for experiments, covering simple travel conversations, about accommodation, transportation, restaurants, shopping, and other topics of interest to travelers. Additional important information was tagged to these sentences. From this bilingual corpus, we have derived basic travel mulitpotent templates that can be used for utterance recognition, understanding, and translation. In order to evaluate the system performance, a collection of 1,885 utterances for template derivation is speaker-dependent trained and 30 additional utterances of each language is collected for inside testing. Table 1 and Table 2 show the basic characteristics of the organized bilingual text corpus and the derived templates. All the utterances were sampled by a Taiwanese speaker and an English speaker at an 8 kHz sampling rate with a 16-bit precision on PCs and PDAs. The Taiwanese words in the corpus were obtained automatically by a Taiwanese morphological analyzer [15] and the English words were automatically tagged by LinkGrammar [18].

Table 1. Basic characteristics of the organized corpus

	Taiwanese	English
Number of sentences	2,084	2,114
Total number of words	14,317	11,648
Number of word entries	6,291	5,118
Average number of words per sentence	6.87	5.51

Table 2. Basic characteristics of the derived multipotent templates

Number of utterances	1,885
Number of templates	1,050
Total number of variables	5,500
Number of variable entries	1,260
Average number of variables per template	5.24

User interface

Figure 7 shows an example of the speech translation result, with: 1) the translation result by the decided optimal template and (2) the target speech output. Also, an agent is showing the status of the language templates in the upper windows. The user, first, selects the translation mode (the choices are Taiwanese=>English or English=>Taiwanese) and speaks to the system while pushing the "Speech Input" button. Then, the user can see the hypothesized template in the middle window and possible target text generation in the bottom window. The generated speech comes out when the user touches the "Output Replay" button. In Figure 7, there is an error in the speech segmentation result. However, even if the system makes the error, the window shows a possible generation because the system is using only the reliable keywords by measuring the intention region of the two-layer reference models.

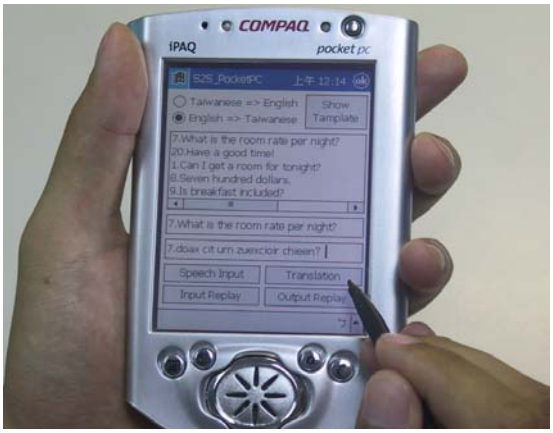

Figure 7. An example of the speech translation result

Translation evaluations

The translation experiments were performed for both a PC-based platform and iPAQ PDAs. For the PC-based platform, the software simulations were done using Windows CE 3.0 on a Pentium® IV 1.8GHz, 1GB RAM, Windows® XP PC. For COMPAQ iPAQ PDAs, the system was implemented on StrongARM SA-1110 (200MHz), 32 MB SDRAM. Speech feature analysis was performed using 10 linear prediction coefficient cepstrum (LPCC) on a 32ms frame overlapped every 8ms. Figure 8 shows the precision results of crude template candidate selection, which intention segments of the source speech can be preselected correctly among the top 10 template candidates. According to the template candidates of top 10, the latent grammar understanding and score normalization are applied for further obtaining the target generation.

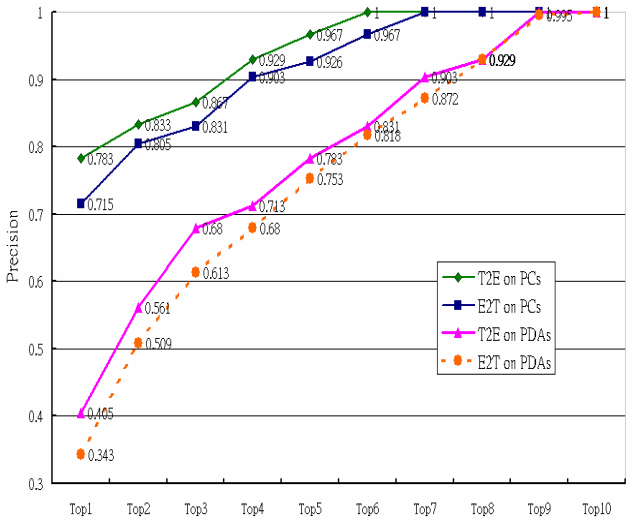

Figure 8. The results of template candidate selection

A bilingual evaluator classified the translation generation results into three categories [6]: *Good*, *Understandable*, and *Bad*. A *Good* generation should have no syntactic errors, and its meaning has to be correctly understood. *Understandable* generations may have some errors, but the main intention of source speech must be conveyed without misunderstanding. Otherwise, the translations are classified as *Bad*. With this subjective measure, the ratio of *Good* or *Understandable* generations with COMPAQ iPAQs was 79% for the Taiwanese-to-English (T2E) translation and 74% for the English-to-Taiwanese (E2T) translation. The ratio of *Good* generations was 70% for the T2E translation and 63% for E2T translation. Table 3 shows the average response time of 5 utterances with a duration of 1~3 seconds by the proposed template-based translation method on personal digital assistants. The proposed approach can achieve about 6.23 sec response time per source speech input on average compared with 5 sec in traditional 1~2GHz PC-based approaches [12].

Table 3. Average response time in the case of accommodation conversations for PDAs. (sec.)

Speech	Time
Yes, sir.	5.18
Is breakfast included?	5.42
What is the room rate per night?	6.30
Please wake me up at seven o'clock	7.02
I want to have breakfast tomorrow morning.	7.04

CONCLUSIONS

In this paper, we propose a novel template-based architecture for spoken language translation efficiently run on handheld devices. The proposed architecture is conducted with a mixture of utterance recognition, understanding, and translation. For the multipotent templates construction, which derives the translation database of spoken language, we blend with text-form translation templates and speech feature model for memory saving and translation capacity promoting. With the constructed templates, the proposed spoken language translation can be efficiently performed by crude template candidate selection, latent grammar understanding for templates, score normalization and ranking, and target speech generation. Experiments are conducted for the languages of Taiwanese and English on COMPAQ iPAQ Pocket PCs. According to the experimental results, our system can achieve about 76.5% translation understanding rate and 6.23 sec response time on average.

ACKNOWLEDGMENTS

The authors would like to thank the National Science Council, Taiwan, Republic of China, for its financial support of this work, under Contract No. NSC90-2215-E-006-009.

REFERENCES

[1] A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld, and P. Zahn, "JANUS III: Speech-to-speech translation in multiple languages," Proc. ICASSP97, pp.99-102 (1997).

[2] F. Sugaya, T. Takezawa, A. Yokoo, and S. Yamamoto, "End-to-end evaluation in ATR-MATRIX: speech translation system between English and Japanese," Proc. Eurospeech-99, pp. 2431-2434 (1999).

[3] T. Watanabe, A. Okumura, S. Sakai, K. Yamabana, S. Doi, and K. Hanazawa, "An automatic interpretation system for travel conversation," Proc. ICSLP-2000, pp. IV 444-447 (2000).

[4] W. Wahlster, "Mobile speech-to-speech translation of spontaneous dialogs: an overview of the final Verbmobil system," in "Verbmobil: foundations of speech-to-speech translation," Ed. W. Wahlster, pp. 3-21, Springer (2000).

[5] R. Isotani, K. Yamabana, S. Ando, K. Hanazawa, S. Ishikawa, T. Emori, H. Hattori, A. Okumura, and T. Watanabe, " An automatic speech translation system on PDAs for travel conversation," Proc. ICMI-02, pp. 211-216 (2002).

[6] K. Yamabana, K. Hanazawa, R. Isotani, S. Osada, A. Okumura, and T. Watanabe, "A speech translation system with mobile wireless clients," Proc. ACL03 (2003).

[7] ATR Spoken Language Translation Research Laboratories research available at <http://www.slt.atr.co.jp/>

[8] Francisco Casacuberta, Enrique Vidal, and Juan Miguel Vilar, "Architectures for speech-to-speech translation using finite-state models," Proc. of the Workshop on Speech-to-Speech Translation: Algorithms and Systems, Philadelphia, pp. 39-44, (2002).

[9] Enrique Vidal, "Finite-State Speech-to-Speech Translation," Proc. ICASSP97, Vol. 1, pp: 111-114 (1997).

[10] F. Casacuberta, D. Llorens, C. Martinez, S. Molau, F. Nevado, H. Ney, M. Pastor, D. Pico, A. Sanchis, E. Vidal, J.M. Vilar, "Speech-to-Speech Translation Based on Finite-State Transducers," Proc. ICASSP2001, Vol. 1, pp: 613 –616, (2001).

[11] Herrmann Ney, Sonja Nießen, Franz Josef Och, Hassan Sawaf, Christoph Tillmann, and Stephan Vogel, "Algorithms for Statistical Translation of Spoken Language," IEEE Transaction on Speech and Audio Processing, Vol. 8, pp. 24-36, (2000).

[12] Kenji Matsui, Yumi Wakita, Tomohiro Konuma, Kenji Mizutani, Mitsuru Endo, and Masashi Murata, "An experimental multilingual speech translation system," Proc. PUI2001 (2001).

[13] Bonnie J. Dorr, Pamela W. Jordan and John W. Benoit, "A Survey of Current Paradigms in Machine Translation," In Advances in Computers, vol. 49, Academic Press (1999).

[14] Jhing-Fa Wang and Shun-Chieh Lin, "Bilingual Corpus Evaluation and Discriminative Sentence Vector Expansion for Machine Translation," Int. Conf. on Artificial Intelligence in Engineering and Technology (ICAIET-2002), Universiti Malaysia Sabah, Kota Kinabalu, Malaysia, 17-18 June 2002.

[15] Jhing-Fa Wang, Bao-Zhang Houg, and Shun-Chieh Lin, "A study for Mandarin text to Taiwanese speech system," Proc. ROCLING XII, pp. 37- 53 (1999).

[16] Yung-Ji Sher, Kao-Chi Chung, and Chung-Hsien Wu, "Establish Taiwanese 7-tones syllable –based synthesis units database for the prototype development of text-to-speech system," Proc. ROCLING XII, pp. 15- 35 (1999).

[17] Liu, J.N.K.; Lina Zhou, "A hybrid model for Chinese-English machine translation," IEEE International Conference on Systems, Man, and Cybernetics, Vol. 2, pp: 1201-1206 (1998).

[18] Davy Temperley, Daniel Sleator, and John Lafferty, Link Grammar Parser 4.1, available at <http://www.link.cs.cmu.edu/link/>

[19] Lawrence Rabiner and Biing-Hwang Juang, "" in "Fundamentals of Speech Recognition," pp. 211, Prentice-Hall, Inc. (1993)