A New Two-Layer Approach for Spoken Language Translation

Jhing-Fa Wang, Shun-Chieh Lin, and Hsueh-Wei Yang

Department of Electrical Engineering, National Cheng Kung University
wangjf@csie.ncku.edu.tw

Abstract. This study proposes a new two-layer approach for spoken language translation. First, we develop translated examples and transform them into speech signals. Second, to properly retrieve a translated example by analyzing speech signals, we expand the translated example into two layers: an intention layer and an object layer. The intention layer is used to examine intention similarity between the speech input and the translated example. The object layer is used to identify the objective components of the examined intention. Experiments were conducted with the languages of Chinese and English. The results revealed that our proposed approach achieves about 86% and 76% understandable translation rate for the Chinese-to-English and the English-to-Chinese translations, respectively.

1 Introduction

With the growing of globalization, people now often meet and do business with those who speak different languages, on-demand spoken language translation (SLT) has become increasingly important (See JANUS III [6], Verbmobil [9], EUTRANS [3] and ATR-MATRIX [1]). Currently, there are two main architectures of SLT: conventional sequential architecture and fully integrated architecture [1]. For the sequential architecture, a spoken language translation is composed by a speech recognition system followed by a linguistic (or non-linguistic) text-to-text translation system. In the integrated architecture, acoustic-phonetic models are integrated into translation models in the similar way as for speech recognition.

Recently, an integrated architecture based on stochastic finite-state transducer (SFST) has been presented in [3,4]. The SFST approach integrated three models in a single network where the search process takes place. The three models are Hidden Markov Models for the acoustic part, language models for the source language and finite state transducers for the transfer between the source and target language. The output of this search process is the target word sequence associated to the optimal path. Fig. 1 shows an example of the SFST approach. λ denotes the empty string. The source sentence "*una habitación doble*" can be translated to either "*a double room*" or "*a room with two beds*". The most probable translation is the first one with probability of 0.09.

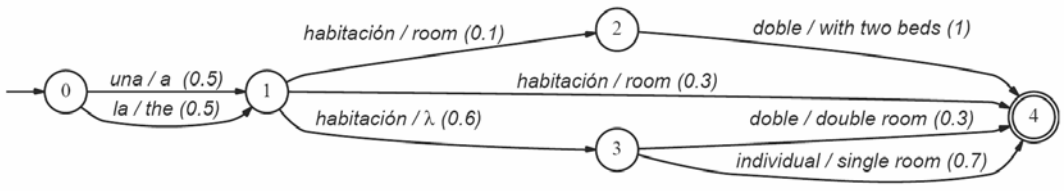

Fig. 1. Examples of the stochastic finite-state transducer

However, when the training data of SFST is insufficient, the results obtained by the sequential architecture are better than the results obtained by the integrated architecture [4]. In addition, word reordering is still a thorny problem in SFST which is based on statistical-based translation methods [5]. Therefore, we propose adopting example-based approaches for better integration. Such the adopted approach does not require the database to be as large as in SFST and can utilize word mappings between source-target language of a chosen translated example for word reordering [2,8]. In this paper, we further propose a new two-layer approach for the example-based spoken language translation. First, we develop translated examples and transform them into speech signals. Second, to properly retrieve a translated example by analyzing speech signals, we expand the translated example into two layers: an intention layer and an object layer. The intention layer is used to examine intention similarity between the speech input and the translated example. The object layer is used to identify the objective components of the examined intention.

The rest of this paper is organized as follows. Section 2 discusses the proposed two-layer approach. Score normalization is presented in Section 3. The experimental results are given in Section 4. Concluding remarks are finally made in Section 5.

2 The Proposed Two-Layer Approach

Referring to Fig. 2, the first step of the proposed two-layer approach is to expand translated examples, which have intention components and object components. After expanding the translated examples, the second step is to adapt the two-layer search plan composed of an intention layer and an object layer. At last, measurement modification is used to modify similarity measurement between the intention layer and the object layer. This study further discusses *translated example expansion*, *two-layer search plan adaptation*, and *measurement modification*.

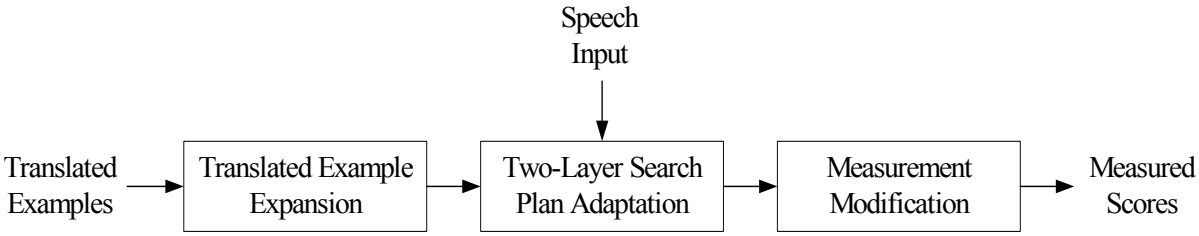

Fig. 2. Framework of the proposed two-layer approach

2.1 Translated Example Expansion

The process of translated example expansion is to group similar translated examples and compare their differences for expanding objects. Table 1 shows an example of fours pairs of grouped translated examples. For these grouped translated examples, the similar constitutes "*Is … still available for …*" ↔ "… 還 有 … 嗎" are defined into an intention sequence translation, which would conduct the meaning of a translation. And the differences compared with the intention sequence are regarded as expanded objects.

Table 1. Fours pairs of grouped translated examples

	Translated examples	Word mappings
1	Is room service still available? ↔ 還 有 客房 服務 嗎?	⟨Is↔嗎, room↔客房, service↔服務, still↔還, available↔有⟩
2	Is breakfast available for tomorrow? ↔ 明天 有 早餐 嗎?	⟨Is↔嗎, breakfast↔早餐, available for↔有, tomorrow↔明天⟩
3	Is laundry service still available? ↔ 還 有 洗滌 服務 嗎?	⟨Is↔嗎, room↔洗滌, service↔服務, still↔還, available↔有⟩
4	Is a single room available for tonight? ↔ 今晚 有 一間 單人房 嗎?	⟨Is↔嗎, a↔一間, single↔單人, room↔房, available for↔有, tonight↔今晚⟩

For example, a new expanded translated example, denoted by *ExTrans*, derived from the translated examples in Table 2 is shown below.

Table 2. An example of expanded translated example

Expanded translated example: *ExTrans*	
The intention sequence translation:	
Is ⟨V^1⟩ still available for ⟨V^2⟩ ?	
↔⟨V^3⟩ 還 有 ⟨V^4⟩ 嗎?	
where ⟨V^1⟩=⟨room service, breakfast, laundry service, a single room⟩,	
⟨V^2⟩=⟨tomorrow, tonight⟩,	
⟨V^3⟩=⟨客房 服務, 早餐, 洗滌 服務, 一間 單人 房⟩,	
⟨V^4⟩=⟨明天, 今晚⟩	
Object translations:	
⟨V^1⟩↔⟨V^4⟩	⟨V^2⟩↔⟨V^3⟩
⟨room,service⟩↔⟨客房,服務⟩	⟨tomorrow⟩↔⟨明天⟩
⟨breakfast⟩↔⟨早餐⟩	⟨tonight⟩↔⟨今晚⟩
⟨laundry,service⟩↔⟨洗滌,服務⟩	
⟨a,single,room⟩↔⟨一間,單人,房⟩	

where *ExTrans* comprises an intention translation, and six object translations. The six object translations are "*room service*↔客房 服務," "*breakfast*↔早餐," "*laundry service*↔洗滌 服務," "*a single room*↔一間 單人房," "*tomorrow*↔明天," and "*tonight*↔今晚".

2.2 Two-Layer Search Plan Adaptation of Expanded Translated Examples

After expanding translated examples, each translated example has two parts: an intention part and an object part. While measuring the speech signals of *i-th* translated example v_i, the speech signals of v_i need to be redefined two layers $v_i = \{v_i', v_i''\}$, where v_i' is an intention layer component of v_i and v_i'' is an object layer component of v_i. Each two-layer searching plan is generated by the translated example and the speech input and the object layer is used to identify the objective components of the examined intention. In terms of searching for an optimal path of states through the two-layer search plan, the issue now is to measure the pair (s, v_i') of a fixed number, says N_i, of v_i'.

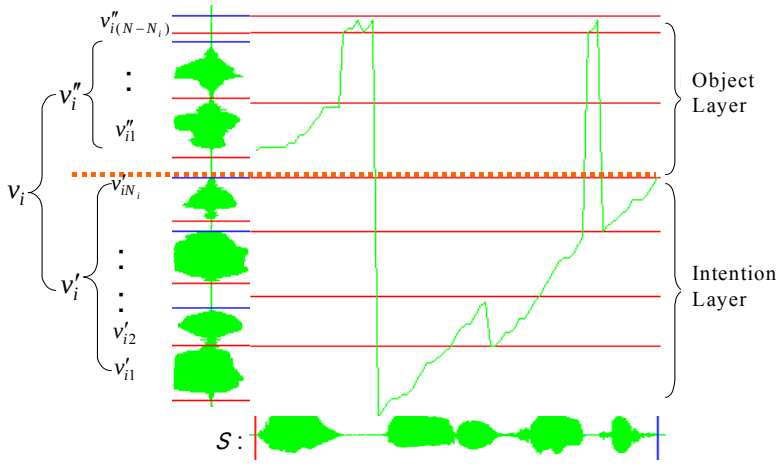

Fig. 3. The proposed two-layer search plan

2.3 Measurement Modification

After adapting the two-layer search plan, another problem is how to measure the similarity of pair (s, v'_i) while adjudging the object frames of v''_i for identifying the other object patterns. Referring to Fig 4., given two similarity measurement scores of pair (s, v_i) and pair (s, v_j), the scores used for comparing the two pairs are D'^*_i and D'^*_j, where D'^*_i is the similarity measurement of pair (v'_i, s) and D'^*_j is the similarity measurement of pair (v'_j, s).

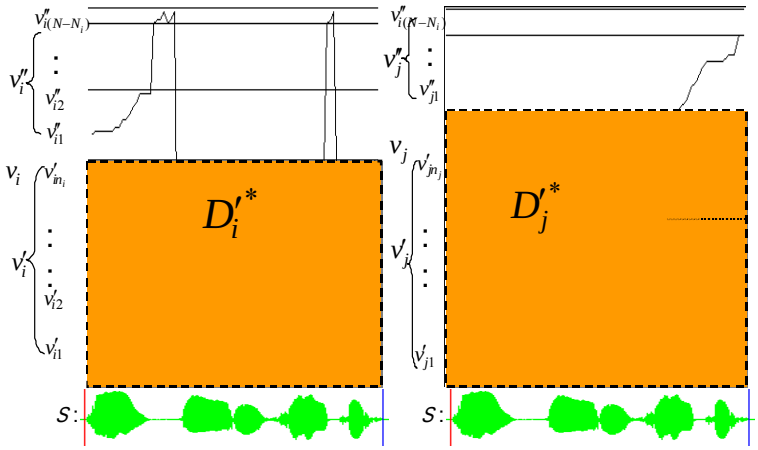

Fig. 4. Search results of various translated examples

For the modification of similarity measurement between the intention layer and the object layer, there are two additional types of search paths in this research: 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 object layer, which will be referred to a score skip level block, contains more than one path connected by $node_{start}$ (or $node_{end}$). And D'^*_i is computed in the intention layer. (See Fig. 5)

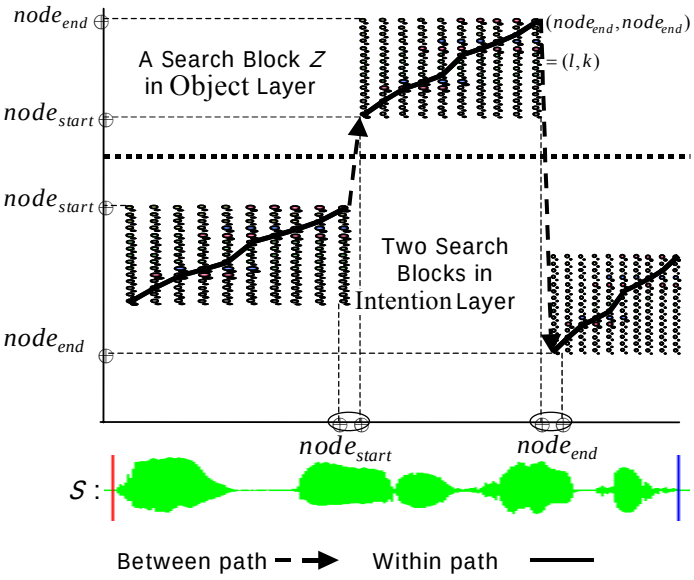

Fig. 5. Additional types of two-layer search paths

3 Score Normalization

The intention sequence in the translated example is an important identification part, where the intention sequence would conduct the meaning of a translation. Therefore, the dissimilarity measurement of the part of the intention sequence is used to rank all the translated examples. However, the cumulative measured dissimilarity score is propagated to the length of the intention sequence. In this study, a length-conditioned weight concept is adopted to compensate this defect. The normalized measured dissimilarity ($\Delta(s, v_i')$) is determined as follows:

$$\Delta(s, v_i') = \partial^{w_{v_i,s}} \tag{1}$$

where ∂ is a weight factor, $w_{v_i',s} = \left(\left\|\left\|v_i'\right\| - \left\|s\right\|\right\|\right) \cdot \left\|s\right\|^{-1}$. The weight ∂ is decided by an interval [1.0, 2.0]. Experimental analysis shown in Fig. 6 indicates that the interval ∂, which yields the most accurate retrieval results, is $[1.3-\delta, 1.3+\delta]$. Therefore, the ∂ is set to 1.3 in this study.

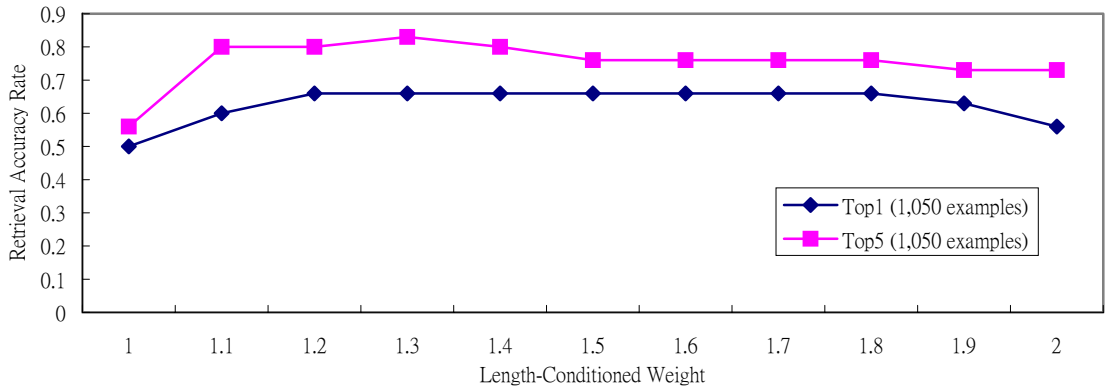

Fig. 6. Retrieval accuracy rate comparison with different setting of weight ∂

4 Experimental Results

4.1 The Task and the Corpus

This study built a collection of English sentences and their Chinese translations that frequently appear in phrasebooks for foreign tourists. Because the translations were made on a sentence-by-sentence basis, the corpus was sentence-aligned after being collated. Table 3 lists a summary of the corpus used in the experiments. The corpus comprises two parts: a training set of 11,885 translated examples for the training phase, and a test set of 105 translated examples for the translation phase (the test set differs from the training set).

Table 3. Basic characteristics of the collected translated examples

		English	Chinese
Training:	Translated Examples	11,885	
	Lexicons	80,699	66,915
	Vocabulary Size	6,278	5,118
	Average number of lexicons	6.79	5.63
Test:	Sentences	105	
	Lexicons	673	641

In order to evaluate the system performance, a collection of 1,050 utterances from the 11,885 examples were speaker-dependent trained, and 105 additional utterances of each language were collected by using one male speaker (Sp1) for inside testing and by using two bilingual male speakers (Sp2 and Sp3) for outside testing. All the utterances were sampled at an 8 kHz sampling rate with 16-bit precision on a Pentium® IV 1.8GHz, 1GB RAM, Windows® XP PC.

4.2 Translation Evaluations

For the spoken language translation system, we found that the recognition performance of 39-dimension MFCCs and 10-dimension LPCCs was close. Therefore, we adopted 10-dimension LPCCs due to their advantages of faster operation. Speech feature analysis of recognition was performed using 10 linear prediction coefficient cepstrums (LPCCs) on a 32ms frame that overlapped every 8ms.

When input speech is being translated, a major sub-problem in speech processing is determining the presence or absence of a voice component in a given signal, especially the beginnings and endings of voice segments. Therefore, the energy-based approach, which is a classic one and works well under high SNR conditions, was applied to eliminate unvoiced components in this research. The measurement results were divided into four parts: the dissimilarity measurement of linear prediction coefficient cepstrum (LPCC)-based (baseline), the baseline with unvoiced elimination (+unVE), the baseline with the score normalization (+ScN), and the combination of unVE and ScN considerations with the baseline (All). A given translated example is called a match when it contained the same intention as the speech input. The reason for adopting this strategy was that objects could be confirmed again while a dialogue was being processed, while wrong intentions could cause endless iterations of dialogue. The experimental results for proper translated example retrieval are shown in Table 4 and Table 5.

Table 4. Average retrieval accuracy of baseline and the improvement in English-to-Chinese(E2C) Translation

	1		2		3		4	
	Baseline		+unVE		+ScN		All	
Example Size	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
150	0.53	0.66	0.63	0.86	0.66	0.86	0.8	1
250	0.53	0.66	0.63	0.86	0.66	0.86	0.8	1
350	0.53	0.63	0.6	0.83	0.66	0.86	0.76	0.96
450	0.53	0.63	0.6	0.83	0.63	0.83	0.76	0.93
550	0.5	0.6	0.6	0.8	0.6	0.8	0.76	0.93
650	0.5	0.56	0.6	0.76	0.6	0.8	0.76	0.9
750	0.46	0.5	0.56	0.73	0.56	0.76	0.73	0.86
850	0.43	0.5	0.53	0.7	0.53	0.73	0.73	0.83
950	0.43	0.46	0.53	0.7	0.5	0.66	0.7	0.83
1050	0.4	0.43	0.46	0.66	0.46	0.66	0.66	0.8

Table 5. Average retrieval accuracy of baseline and the improvement in Chinese-to-English(C2E) Translation

	1		2		3		4	
	Baseline		+unVE		+ScN		All	
Example Size	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
150	0.46	0.6	0.63	0.8	0.6	0.76	0.76	1
250	0.46	0.6	0.6	0.76	0.6	0.73	0.76	0.96
350	0.46	0.56	0.6	0.76	0.56	0.7	0.73	0.93
450	0.43	0.56	0.56	0.73	0.53	0.66	0.7	0.9
550	0.43	0.53	0.56	0.7	0.53	0.63	0.7	0.86
650	0.43	0.53	0.53	0.7	0.5	0.6	0.66	0.83
750	0.4	0.5	0.53	0.66	0.5	0.6	0.63	0.8
850	0.4	0.5	0.5	0.66	0.46	0.56	0.63	0.8
950	0.4	0.46	0.46	0.63	0.46	0.56	0.6	0.76
1050	0.36	0.43	0.46	0.6	0.43	0.56	0.6	0.7

Based on the developed translated examples, when the example or vocabulary size increases, more examples would possibly lead to more feature models and more similarities in speech recognition, thus causing false recognition results and lower retrieval accuracy. Additionally, multiple speaker dependent results were obtained using three speakers. The first speaker's feature models were used to perform tests on the other two speakers, and the results are shown in Table 6. The experimental results show that although the feature models were trained by Sp1, the retrieval accuracy of Sp2 and Sp3 was only reduced by 10 to 15 percent.

Table 6. Average retrieval accuracy in multiple speaker testing

			Example Size (Speech features of Sp1)									
		(Top5)	150	250	350	450	550	650	750	850	950	1050
All	Sp1	E2C	1	1	0.96	0.93	0.93	0.9	0.86	0.83	0.83	0.8
		C2E	1	0.96	0.93	0.9	0.86	0.83	0.8	0.8	0.76	0.7
	Sp2	E2C	0.9	0.86	0.83	0.8	0.76	0.73	0.73	0.7	0.66	0.66
		C2E	0.83	0.83	0.8	0.76	0.73	0.73	0.7	0.66	0.63	0.63
	Sp3	E2C	0.83	0.8	0.76	0.76	0.73	0.7	0.7	0.66	0.66	0.63
		C2E	0.76	0.76	0.73	0.73	0.7	0.66	0.66	0.63	0.6	0.6

A bilingual evaluator was used to classify the target generation results into three categories [10]: Good, Understandable, and Bad. A Good generation needed to have no syntactic errors, and its meaning had to be correctly understood. Understandable generations could have some syntactic errors and variable translation errors, but the source speech had to be conveyed without misunderstanding. Otherwise, the target generations were classified as Bad. With this subjective measure, the percentage of Good or Understandable generations for the Top 5 was 86% for English-to-Chinese (E2C) translation and 76% for Chinese-to-English (C2E) translation. The percentage of Good generations for the Top 1 was 60% for E2C translation, compared to 56% for C2E translation. We examined the translated examples in a specific domain and found that 100% translation accuracy could be achieved. In other words, translation errors occurred only as a result of speech recognition errors, such as word recognition errors and segmentation errors. Besides, these results also indicate that C2E performed worse than E2C. This difference may occur because Chinese is tonal, whereas English is not; thus, it is harder for C2E translation to obtain an appropriate translated example.

5 Conclusions

In this work, we have proposed a new two-layer approach for example-based spoken language translation. According to the proposed approach, the translated example can be properly retrieved by measuring the speech signals on the intention layer and the object layer. Experiments using Chinese and English were performed on Pentium® PCs. The experimental results reveal that our system can achieve an average understandable translation rate of about 81%. By collecting more speech databases, the system also applies speaker-dependent or speaker-independent HMM to the proposed two-layer approach for more robust speech translation.

References

[1] ATR Spoken Language Translation Research Laboratories research, http://www.slt.atr.co.jp/

[2] M. Carl. Inducing Translation pattern for Example-Based Machine Translation. In *Proc. of the 7th Machine Translation Summit*, pp.617–624, 1999.

[3] 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. In *Proc. of 26th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.613 –616, 2001.

[4] E. Vidal. Finite-State Speech-to-Speech Translation. In *Proc. of 22nd IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.111-114, 1997.

[5] H. Ney, S. Nießen, F. J. Och, H. Sawaf, C. Tillmann, and S. Vogel. Algorithms for Statistical Translation of Spoken Language. *IEEE Transaction on Speech and Audio Processing*(8), pp.24-36, 2000.

[6] 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. In *Proc. of 22nd IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.99–102, 1997.

[7] Rabiner, L. and B. H. Juang. Fundamentals of Speech Recognition. Prentice-Hall, Inc., 1993.
[8] J. Liu and L. Zhou. A hybrid model for Chinese-English machine translation. In *Proc. of IEEE International Conference on Systems, Man, and Cybernetics*, pp.1201-1206, 1998.
[9] Wahlster, W. Verbmobil: Foundations of Speech-to-Speech Translation. New York: Springer-Verlag Press, 2000.
[10] K. Yamabana, K. Hanazawa, R. Isotani, S. Osada, A. Okumura and T. Watanabe. A Speech Translation System with Mobile Wireless Clients. In *Proc. of the Student Research Workshop at the 41st Annual Meeting of the Association for Computational Linguistics*, pp.119–122, 2003.