ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2000)
Fragrant Hill Hotel, Beijing
October 13-15, 2000

STATISTICAL APPROACH TO CHINESE-ENGLISH SPOKEN-LANGUAGE TRANSLATION IN HOTEL RESERVATION DOMAIN [1]

CHENG Wei and XU Bo

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing
{wcheng, xubo}@nlpr.ia.ac.cn

ABSTRACT

This paper investigates a preliminary translation system from Chinese to English based on the statistical approach and tests its performance on a limited-domain spoken-language task: hotel reservation. A bilingual corpus is available for the task, which exhibits some typical phenomena of spontaneous speech. The experiments are performed on both the text transcription and the speech recognizer output. The word error rate is about 14%. Some analyses present a great potential for improving the translation quality. From the results and analyses, a broad prospect is showed on the statistical approach to the spoken-language translation.

1. INTRODUCTION

Spoken-language translation faces extra challenges besides those of written-language translation [1]. The main problems are 1) ungrammatical expressions, 2) speech recognition errors, and 3) the real-time requirements. These run traditional rule-based methods into many troubles. Thus, in recent years, some data-driven methods are taken as the effectual ways such as the example-based machine translation [2] (EBMT, advanced by ATR) and the statistical machine translation (SMT).

The statistical approach is an adequate framework for introducing automatic learning techniques in spoken-language translation [3]. It was first suggested by Warren Weaver in 1949 and was greatly developed by P. F. Brown in 1990 [4]. Moreover, it has been used in some western-language translation systems such as Head Transducers of AT&T and Verbmobil of BMBF. However, Chinese language is quit different from those western languages. It has some special structural features such as the open vocabulary nature and the flexibility in word ordering [5]. Hence, in this paper, we investigate a preliminary translation system from Chinese to English based on the statistical approach and test its performance on a limited-domain spoken-language task: hotel reservation. The organization of the paper is as follows.

After reviewing the principle of SMT, we present a translation model and a search algorithm, which is adapted for the translation model. Then the preliminary system is evaluated on a bilingual corpus including about 3000 sentences. This corpus exhibits some typical phenomena of spontaneous speech such as the flexibility in word ordering and high variability of hesitations. After defining our performance measure, we report the experimental results and analyze inherent problems. From these analyses, a conclusion can be drawn to show a broad prospect of the statistical approach applied for the Chinese-English spoken-language translation.

2. THE STATISTICAL APPROACH TO SPOKEN-LANGUAGE TRANSLATION

2.1. Principle

In statistical opinions, translation task can be described as follow. Given a source (" Chinese ") string $c_1^m = c_1 \cdots c_j \cdots c_m$, we choose the string $\hat{e}_1^l$ among all possible target (" English ") strings $e_1^l = e_1 \cdots e_i \cdots e_l$ with the highest probability that is given by Bayes' decision rule [4]

$$\begin{aligned} \hat{e}_1^l &= \arg\max_{e_1^l} \{\Pr(e_1^l \mid c_1^m)\} \\ &= \arg\max_{e_1^l} \{\Pr(e_1^l) \cdot \Pr(c_1^m \mid e_1^l)\} \end{aligned} \qquad (1)$$

$\Pr(e_1^l)$ is the probability of the language model produced by the target language. $\Pr(c_1^m \mid e_1^l)$ is the probability of the string translation model from the target language to the source language. The argmax operation demotes the search problem, i.e. the generation of the output sentence in the target language. The overall architecture of the statistical translation approach can be summarized in Fig 1.

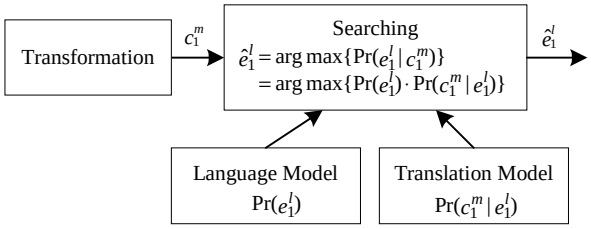

Fig1. Architecture of the translation approach based on Bayes decision rule

In general, as shown in Fig1, transformations are added to make the translation simpler for the algorithm. The transformations

[1] The research work described in this paper was supported by the National Key Fundamental Research Program (the 973 Program) of China under the grant G1998030504, the National Natural Science Foundation of China under the grant 69835030 and also the National 863 Hi-Tech Program under the grant 863-306-ZT03-02-2

may range from the categorization of single words and word groups to more complex preprocessing steps that require some parsing of the source string [6]. However, in this paper, we use only sentence segment, word segment and word categorization as explicit transformation step. Obviously, it can not preserve the effect of spontaneous speech.

2.2. Modeling and Training

In the translation experiments, the trigram language model [4] is used as the standard language model (LM). Moreover, the translation model (TM) of our system is a word-based translation model called as IBM-1 model [7]. This model assumes an uniformed alignment probability, which only depends on l, the length of the English string. Therefore, we can avoid the problems caused by the flexibility in word ordering of spoken language.

According to IBM-1, the probability of TM is calculated by

$$\Pr(c_1^m \mid e_1^l) = \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} t(c_j \mid e_i) . \quad (2)$$

ε presents some small, fixed number. $t(c_j \mid e_i)$ is the translation probability of c_j given e_i. It subjects to the constraints that for each English word **e**,

$$\sum_{\mathbf{C}} t(\mathbf{c} \mid \mathbf{e}) = 1 . \quad (3)$$

C is the finite Chinese lexicon. Using EM algorithm, the parameter $t(\mathbf{c} \mid \mathbf{e})$ can be estimated as follows:

1. Expectation: The parameter $c(\mathbf{c} \mid \mathbf{e}; c_1^m, e_1^l)$ is introduced to denote the expected number of times that the English word **e** connects to the Chinese word **c** in the translation (c_1^m, e_1^l). It can be estimated by

$$c(\mathbf{c} \mid \mathbf{e}; c_1^m, e_1^l) = \frac{t(\mathbf{c} \mid \mathbf{e})}{t(\mathbf{c} \mid e_0) + \cdots + t(\mathbf{c} \mid e_l)} \sum_{j=1}^{m} \delta(\mathbf{c}, c_j) \sum_{i=0}^{l} \delta(\mathbf{e}, e_i) . \quad (4)$$

   where δ is the Kronecker delta function, equal to one when both of its arguments are the same and equal to zero otherwise.

2. Maximization: The new value for $t(\mathbf{c}|\mathbf{e})$ is obtained according to the equation

$$t(\mathbf{c} \mid \mathbf{e}) = \frac{\sum_{\mathbf{S}} c(\mathbf{c} \mid \mathbf{e}; c_1^m, e_1^l)}{\sum_{\mathbf{C}} \sum_{\mathbf{S}} c(\mathbf{c} \mid \mathbf{e}; c_1^m, e_1^l)} . \quad (5)$$

   S is the training corpus.

2.3. Searching

Here a fast stack decoder [8] is adapted for IBM-1 model to perform the argmax operation. Its hypothesis is the prefix string of the English sentence, whose score is the likelihood of the Chinese sentence summed over all possible alignments. In this case, if $\tau_{kl}(j \mid i, H_l^k)$ denotes the contribution of the i^{th} position of the hypothesis $H_l^k = l : e_1 \cdots e_i \cdots e_k$ to the probability mass of the j^{th} Chinese word, we have

$$\tau_{kl}(j \mid i, H_l^k) = \begin{cases} t(c_j \mid e_i) & 0 \le i \le k \\ \sum_{n=0}^{|\mathbf{E}|} \Pr(w_n) \cdot t(c_j \mid w_n) & k < i \le l \end{cases} . \quad (6)$$

k is the length of the hypothesis, l is the assumption of the English sentence length. $|\mathbf{E}|$ is the size of the English lexicon. And $\Pr(w_n)$ is the prior probability of the English word w_n.

Thus the score of a hypothesis $H_k = e_1 \cdots e_i \cdots e_k$ is

$$S(H_k) = \sum_{i=k}^{L_m} [\Pr(i \mid m) \cdot \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} \tau_{kl}(j \mid i, H_l^k)] \times \Pr(e_1) \prod_{i=2}^{k} \Pr(e_i \mid e_{i-1}) . \quad (7)$$

Here L_m is the maximum sentence length allowed. $\Pr(i \mid m)$ is the length distribution of English sentences conditioned on the Chinese sentence length, which is modeled with Poisson distributions.

3. EXPERIMENTS AND DISCUSSION

3.1. Bilingual Corpus

Our experiments are carried out on a bilingual Chinese-English corpus. This corpus consists of spontaneous utterances from hotel reservation dialogs, which is originally produced in Chinese. Although this task is a limited-domain task, it is difficult for several reasons: first, the syntactic structures of the sentences are less restricted and highly variable; second, it covers a lot of spontaneous speech characters, such as hesitations, repetitions and corrections.

A summary of the corpus is given in Tables I and II. The corpus consists of the following three parts.

- Training set (TrainSet) of 3009 sentence pairs: Sentences in this set are all text input, which was obtained as the correct orthographic transcription of the spoken language. Thus, the effects of spontaneous speech are present in the training set although, in this way, the effect of speech recognition errors is not covered.

- English-Chinese dictionary (ECDic) of 850 items: Each item includes one English word, phrase or idiom in the training set and several Chinese translations. From Equation (4), we notice that $c(\mathbf{c} \mid \mathbf{e}; c_1^m, e_1^l)$ is not related to the other Chinese words in c_1^m. Therefore we can regard the dictionary as a part of corpus to train TMs.

- Test set: It consists of both text and speech input. The text input was obtained by manually transcribing the spontaneously spoken sentences. It does not contain the speech recognition errors. Moreover, it cannot include a word outside the training set for the experimental system is just a preliminary

system. In the case of speech input, the recognizer has a word error rate of about 22.8% and the experiments for speech input are performed on the top-one sentence without punctuation marks. The sentences in both text and speech input are outside the training set.

TABLE I TRAINING SET AND DICTIONARY

		Chinese	English
Training	Sentences	3009	
	Words	15547	16935
	Vocabulary Size	804	726
Dictionary	Item	850	

TABLE II TEST SET

		Chinese	English
Text	Sentences	100	
	Words	742	812
Speech	Item	51	
	Words	321	--

There were three optional preprocessing steps applied to the corpus:

1) *Sentence Segment*: Because the corpus was obtained by transcribing, there was no constraint on the length of the sentences. Some of the sentences were rather long. In this step, sentences were manually segmented as short as possible.

2) *Word Segment:* This work was only available for Chinese. It was also done by hand.

3) *Categorization:* To mitigate the sparse data problem, the words of the trained lexicon were subjected to a categorization step. There were three categories: personal names, city names and location names.

3.2. Performance Measures

It is difficult to define a suitable performance measure of translation approaches. In this paper each utterance was assigned one of three ranks for translation quality: (A) Fair: There is no problems in information and a few of flaws in grammar. (B) Acceptable: large part of important information can be get with effort. (C) Nonsense: Almost no information is translated correctly. Here we show samples for each rank containing information about 1) the Chinese input, 2) the system translation, and 3) a human translation.

- rank-A
  1. 我还不熟悉你们宾馆在什么地方。
  2. I do not know where the hotel.
  3. I do not know where your hotel is.
- rank-B
  1. 我想问一下，就是说，我想订四间。
  2. I want to inquire, I mean, do I need to reserve four rooms.
  3. I want to inquire, I mean, I want to reserve four rooms.
- rank-C
  1. 您订哪天的房间？
  2. Which room are you sure that I reserved for tomorrow?
  3. When do you need it?

We also employ a traditional error criterion — word error rate (WER) [6] to evaluate our system.

3.3. Experimental Results

Three experiments were conducted in this section. In the first experiment, we trained one TM only using the training set. While in the second experiment, we trained another TM with both the training set and the English-Chinese dictionary. Then these two models are both tested with the same text input. Table III shows the evaluation results of above experiments. From the table, we can see that the dictionary can aid to improve the translation quality.

TABLE III RESULTS OF EXPERIMENT ONE AND TWO

	TrainSet	TrainSet+ECDic
rank-A (%)	67	69
rank-B (%)	28	29
rank-C (%)	5	2
WER (%)	14.4	12.2

In the third experiment, we considered the speech recognition errors and ranked the speech input first according to our performance measure. The evaluations were shown on Table IV.

TABLE IV RECOGNIZING QUALITY

	rank-A (%)	rank-B (%)	rank-C (%)	WER (%)
Recognizer	43.1	41.2	15.7	22.8

Then the model of Experiment two were tested with the speech input. Table V gives out the final results.

TABLE V RESULT OF EXPERIMENT THREE

	rank-A (%)	rank-B (%)	rank-C (%)
Translator	21.6	43.1	35.3

3.4. Error Analyses

There is still a great potential for improving the translation quality because the experimental system is just a preliminary system. Following is the error analyses and the improving advice.

1) The word frequencies in spoken language are quit different. The Chinese vocabulary size of the training set is 804. However, the frequencies of only 24 words are above 100 times. 227 words appear only once and 586 words appear under ten times. This affects the accuracy of the models.

2) The most probable improvement is on the translation quality of rank-B. Hence we especially analyze it from the following three aspects.

- *Searching*: About 33.3% sentences did not search to the optimum in rank-B. However there is no heuristic function in the search algorithm. Thus, adding some heuristic information may mitigate this problem.

- *Grammar*: 69% translations in rank-B have grammar problems. The main problems are as follow:

1) Sentence Patterns. The sentence patterns are expressed in a lexical way in Chinese while they are expressed through the word order in English. For example:

有星期四的票吗？
Is there any ticket available for Thursday?

Thus, it is a big problem for trigram LM to translate an interrogative sentence correctly without any sentence pattern information.

2) Attributive Clauses. The attributive clause is a special phenomenon of English. It never appears in Chinese language. In addition, the relative pronoun in an attributive clause has few meanings. Thus, translating to a subordinate clause is very difficult.

3) Phrases and Idioms. Phrases and idioms are frequently used in the spontaneous speech. Moreover, it is almost impossible to align a phrase or an idiom to its translation word by word. Thus, translating them is a hard thing for our system's TM. If we consider each phrase or idiom as an individual word, this problem may be mitigated. However, it is also a difficult work to define suitable phrases and idioms.

- *Essential points for communication*: We can get most of the important information from the translations in rank-B. However, there are still 13.8% results losing parts of the essential points. A very interesting thing is that nearly 33.3% results in rank-B give more information than what the inputs give. This is because the size of our training corpus is so small that we can not get an accurate translation probability. However, infinitely enlarging the bilingual corpus is an expensive and impossible thing. One possible way to solve this problem is to add more details of the alignment probabilities.

3) There is a clear degradation by speech input. However, the sentence number in rank-B has increased. This is because there are 16.7% sentences whose recognizing qualities belong to rank-C, while their translation qualities are improved to rank-B. That means the statistical approach is a robust method.

4. CONCLUSION

This paper described a statistical approach used in the Chinese-English spoken-language translation. It was tested in a hotel reservation domain and three experiments were reported here. Although there are still some problems in this approach, the effectiveness was confirmed by most of the results. From the analyses, we can see a broad prospect on the statistical spoken-language translation.

5. REFERENCES

[1] E. Sumita, S. Yamada, K. Yamamoto, M. Paul, H. Kashioka, K. Ishikawa, and S. Shirai, "Solutions to problems inherent in spoken-language translation: the ATR-MATRIX approach." *MT Summit VII*, pp. 229-233, Sept. 1999.

[2] E Sumita and H. Iida, "Experiments and prospects of example-based machine translation." In *29th Annual Meeting of the Association for Computational Linguistics*, California, USA, 1991.

[3] I. Garcia-Varea, F. Casacuberta, and H. Ney, "An iterative, DP-based search algorithm for statistical machine translation." In *Proceedings of ICSLP'98*, Sydney, Australia, 1998.

[4] P. F. Brown, J. Cocke, V. J. Della Pietra, S. A. Della Pietra, F. Jelinek, J. D. Lafferty and R. L. Mercer, "A statistical approach to machine translation." *Comput. Linguist*., vol. 16, no. 2, pp. 79-85, 1990.

[5] L. S. Lee, "Structural features of Chinese language – why Chinese spoken language processing is special and where we are." In *Proceedings of ICSLP'98*, Sydney, Australia, 1998.

[6] H. Ney, S. Nießen, F. J. Och, H. Sawaf, C. Tillmann, and S.Vogel, "Algorithms for statistical translation of spoken language." *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, pp.24-36, 2000.

[7] P. F. Brown, V. J. Della Pietra, S. A. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation." *Comput. Linguist*., vol. 19, no. 2, pp. 263-311, 1993.

[8] Y. Y. Wang, A.Waibel, "Fast Decoding For Statistical Machine Translation." In *Proceedings of ICSLP'98*, Sydney, Australia, 1998.