Phrase-Based Statistical Machine Translation for MANOS System

Bo XU, Z.B Chen., W WEI W PAN and Z.D.YANG

National Lab of Pattern Recognition,Institute of Automation, CAS

P.O.Box 2728, Beijing China,

xubo@nlpr.ia.ac.cn

Abstract

MANOS (Multilingual Application Network for Olympic Services) project. aims to provide intelligent multilingual information services in 2008 Olympic Games. By narrowing down the general language technology, this paper gives an overview of our new work on Phrase-Based Statistical Machine Translation (PBT) under the framework of the MANOS. Starting with the construction of large scale Chinese-English corpus (sentence aligned) and introduction four methods to extract phrases, The promising results from PBT systems lead us to confidences for constructing a high-quality translation system and harmoniously integrate it into MANOS platform.

Keywords: Statistical Machine Translation (MT), Speech and Language Processing, Multilingual information processing,

1 Introduction

Beijing has won the right to stage the Games of the XXIX Olympiad in year 2008. With the promise of highlighting the theme of "green, high-tech and people's Olympiad", a "High-tech Olympics Action Plan" was released to accelerate its scientific research efforts in a move to provide the world the "best Olympiad featuring the latest technologies". As one of the key projects in “Digital Olympiad”, Multilingual Intelligent Information Service Network System- MANOS is to realize convenient information service related to the Olympics at any time, in any place and with facilities of various kinds by porting state-of-the-art available speech and language processing technologies.

This paper gives a survey of MANOS in Section 2. Section 3 describes the key technology in MANOS---Phrase-Based SMT. Finally; we aim to harmoniously port SMT system into MANOS in Section 4.

2 Outline of MANOS system

MANOS system has been sketched as an Olympic-oriented intelligent multilingual information network service center in which some advanced multilingual speech and language processing measures, such as speech interactive, machine translation, information retrieval & management, dialog management, and so on, could be integrated. Figure 1 shows a concept structure of MANOS system. Bearing in mind the usability of state-of-the-art spoken language processing

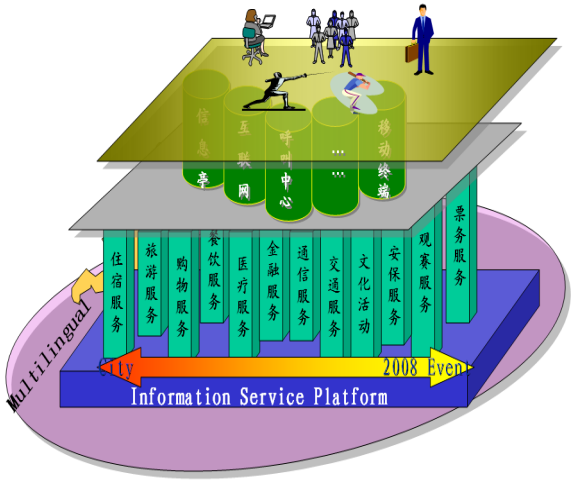

Figure 1 A concept Framework of Olympic-oriented Multilingual Network Information Service System

technologies, MANOS system is located on the Olympic-related information inquiry service based on three different channels, such as wire or wireless telephone, voice-enabled multilingual information kiosk, and voice-enabled wearable terminal (PDA). The scope of information services covers the traffic,

hotel reservation, ticket booking, medical, shopping, weather forecast, Games program, travel sites and so on. *Multilingual service, as one of add-value service, are designed to cross various service if the vocabulary and scenery is feasible for current technology.*

As the entrance of multilingual information service, speech interaction, machine translation and speech translation (ST) all could find their positions for different purposes. Especially in our designed Conceptual Olympic Terminal speech translation plays as possible practical application in whole MANOS system, which include three components:

1) Automatic speech recognition (ASR);
2) Machine Translation (MT);
3) Text to speech (TTS) technology.

With long-time accumulation on speech interactive technology, the project team have made significant progress achieved in ASR and TTS technology, and realized successful demos and commercial applications. However, the application of Machine Translation has some troubles, as it must bear the imperfection of speech recognition and the relative scarcity of the training corpus. Therefore, we will pay more attention on the construction of Machine Translation system in the following section.

3. Phrase-Based Statistical Machine Translation

Machine Translation accuracy has increased, due to better techniques and the availability of larger parallel training set. Especially, the use of statistical techniques in machine translation has led to dramatic improvements in the quality of research systems in the recent years. (Verbmobil and NIST/TIDES MT evaluations).

In China, there are more and more research groups committed themselves to the study on MT, especially Statistical MT in recent year. In the first workshop on Conference on Statistical Machine translation held in July 2005, the project team showed the leading MT system based on statistical methods besides discussing the mainstream of MT' development in the future. The promising result shows that the quality of SMT system has outperformed classical approaches based on interpretation, transfer, and generation in certain field which has long research achievement.

3.1 Construction of large-scale bilingual corpus(Chinese-English)

Statistical MT systems are trained on bilingual (human translated) documents. Most researchers believe that deep secrets of translation lie buried in these large data sets, waiting to be uncovered by automatic analysis.

However, it is time-consuming work to "clean" natural bilingual data into sentence-aligned corpus which is necessity in Statistical MT training procedures. Here, we use some technology for substantial cleaning, such as: text de-formatting, encoding detection/ conversion, and so on.

Through ChineseLDC(Linguistic Data Consortium for Chinese Language) data sharing mechanism and ten years bilingual corpus accumulation in group ,we have more than 800,000 sentences pairs could be adopted for our research purposes. The abundant data resources are the basis of Statistical MT system. The following experiments are based on

3.2 Phrase-Based Statistical Machine Translation

In order to overcome the intrinsic weakness of purely word based translation model, our translation system takes the advanced statistical phrase-based translation model as baseline, which is similar to (Och, 1999) alignment template model. Instead of using word class and alignment template, however, we use directly a phrase translation table. This allows us to build a more compact, more transparent and faster decoder,

3.2.1 The methods of phrase-extraction

Because the quality of translations is largely dependent on the quality of phrase translation pairs extracted from bilingual corpora. Our system integrated with four methods to extract bilingual phrase pairs:

1. Integrated segmentation and phrase alignment (ISA) (Zhang, 2003)
2. Extracting phrase pairs from HMM word alignment model (Vogel et al., 1996)
3. phrases from bi-direction Word-Based Alignment (Och et al., 1999).
4. Phrase-extraction using Inversion Transduction Grammar (Wu, 1997)

Table 1 gives the different scale of extracted phrases

(Maxlength: 3 words), Training corpus: 130 thousand sentence pairs.

	ISA	HMM	WBA	ITG
Size	100,000	130,000	250,000	300,000

Table 1 the scale of extracted phrase

3.2.2 Decoder

The decoding process works in two stages: First, the phrase translations should be generated for the input text, this is done before the searching begin. Second, the search process takes place through which phrase translation model, language model, distortion model and length model will be applied. Both steps will now be described in more detail.

1. Generation of Translation Options

A phrase translation table can be achieved through a bilingual corpus by the methods introduced in Section 3.2.1. They are stored with some information about the source phrase, the target phrase and phrase translation probability

2. Searching algorithm

The phrase-based decoder we developed employs a beam search algorithm, similar to the one in (Koehn. P, 2003) However, considering the big difference between Chinese and English, we insert the English Functional Words (F-Word) as complementary when translating Chinese sentence into English. So after every new hypothesis expanded, F-Word can be applied, that is to say, a NULL is added after the source phrase translated. Because perhaps not all words of the input sentence are necessary to be translated, we select the final hypothesis of the best translation in the last several stacks according to their scores when tracing back.

In each expansion English words are generated, additional Chinese words are covered (marked by *), and the probability cost so far is adjusted. In the example the input sentence is 王 先生 想去 打 篮球 。

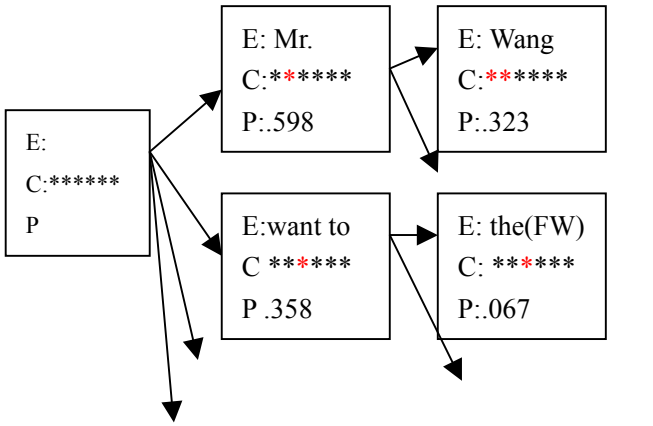

Figure 3: Hypotheses Expansion in the beam decoder:

3.3 Results from PBT system

Table 2 shows the Phrase-based translation performance in several cases. Two test sets are used: 1500 test sentences is domain-specific test corpus which is extracted from travel domain, 863-03 test corpus is benchmark test adopted in 863 test in year 2003.

Table 2: Some results based on Phrase-based translation

Training and Test corpus	NIST	BLEU
50Ktraining, 1500 test	6.7	0.27
130K training, 1500 test	7.43	0.33
50K training, 863-03Test	6.0	0.22
130K training, 863-03Test	6.82	0.28

By Comparison of different experiments, we have following experiences:

1) Phrase-Based Statistical MT systems has some output performed in BLEU score than rule-based approaches: the BLEU score was 0.2833 in comparison with 0.2439 to 0.2657 for the other translation approaches.

2) When we add more corpus to training which are more general domain corpus ,the output model achieves more high performance in domain-specific 1500 test corpus. Which shows the good model is more important than just conduct the domain-specific model in statistical machine translation.

4 Porting Available Language Technologies in MANOS System

Based on the advanced MT technology, we have developed some successful demos in certain domain, and applied in commercial applications. The goal is to construct a high-quality, simultaneous MT system which can be seamlessly integrated with other parts in MANOS.

Based on the gradually clear goal of MANOS project, especially gradually clear target of services kinds, services scale and services coverage, we are making effort towards the realistic application of available speech and language technologies. In addition to speech translation described above, the related techniques such as multilingual content management and back-off human-edited machine translation for multilingual content synchronize have been preliminary developed for the construction of MANOS multilingual information processing platform. To make multilingual value-added applied to various possible services, consistent standard, interface and open architecture in order to harmoniously integrate so many modules into the whole MANOS system are needed.

5 Conclusion

Aimed at the realization of intelligent multilingual information services in 2008 Beijing Olympic Games, we investigated the work on Statistical MT which could be possibly ported into the MANOS project, one of 2008 Beijing digital Olympiad action plan. The paper gives an overview of the whole Translation system, including both training and decoding part. All progress made so far through the internal and international cooperation encourages us to construct a high-quality multilingual Information network service in the 2008 Olympic Games in Beijing using more advanced phrase-based translation methods.

6 Acknowledgements

This work is supported by High Technology Research & Development (863) Program in China (2001AA114015, 2001AA114070) and Olympic Program of Beijing Committee of Science and Technology (H030130050430).

References

Dekai Wu. 1997. *Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora.* Computational Linguistics, 23(3):377-404

Koehn. P, Och, F. J., and Marcu .D, 2003 *Statistical Phrase-Based Translation.* In Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics. 2003.

Och, F. J., Tillmann, C., and Ney, H. 1999. *Improved alignment models for statistical machine translation.* In Proc. of the Joint Conf. of Empirical Methods in Natural Language Processing and Very Large Corpora, pages 20–28.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996 *HMMbased Word Alignment in Statistical Translation*. In COLING'96: The 16th Int. Conf. on Computational Linguistics,pp. 836–841, Copenhagen..

Ying Zhang, Stephan Vogel and Alex Waibel. 2003 *Integrated Phrase Segmentation and Alignment Model for Statistical* Machine Translation. Submitted to Proc. of International Confrerence on Natural Language Processing and Knowledge Engineering (NLP-KE) , Beijing, China