Dialect MT: A Case Study between Cantonese and Mandarin

Xiaoheng Zhang
Dept. of Chinese &. Bilingual Studies, The Hong Kong Polytechnic University
Hung Hom, Kowloon
Hong Kong
ctxzhang@polyu.edu.hk

Abstract

Machine Translation (MT) need not be confined to inter-language activities. In this paper, we discuss inter-dialect MT in general and Cantonese-Mandarin MT in particular. Mandarin and Cantonese are two most important dialects of Chinese. The former is the national lingua franca and the latter is the most influential dialect in South China, Hong Kong and overseas. The difference in between is such that mutual intelligibility is impossible. This paper presents, from a computational point of view, a comparative study of Mandarin and Cantonese at the three aspects of sound systems, grammar rules and vocabulary contents, followed by a discussion of the design and implementation of a dialect MT system between them.

Introduction

Automatic Machine Translation (MT) between different languages, such as English, Chinese and Japanese, has been an attractive but extremely difficult research area. Over forty years of MT history has seen limited practical translation systems developed or commercialized in spite of the considerable development in computer science and linguistic studies. High quality machine translation between two languages requires deep understanding of the intended meaning of the source language sentences, which in turn involves disambiguation reasoning based on intelligent searches and proper uses of a great amount of relevant knowledge, including common sense (Nirenburg, et. al. 1992). The task is so demanding that some researchers are looking more seriously at machine-aided human translation as an alternative way to achieve automatic machine translation (Martin, 1997a, 1997b).

Translation or interpretation is not necessarily an inter-language activity. In many cases, it happens among dialects within a single language. Similarly, MT can be inter-dialect as well. In fact, automatic translation or interpretation seems much more practical and achievable here since inter-dialect difference is much less serious than inter-language difference. Inter-dialect MT[1] also represents a promising market, especially in China. In the following sections we will discuss inter-dialect MT with special emphasis on the pair of Chinese Cantonese and Chinese Mandarin.

1 Dialects and Chinese Dialects

Dialects of a language are that language's systematic variations, developed when people of a common language are separated geographically and socially. Among this group of dialects, normally one serves as the lingua franca, namely, the common language medium for communication among speakers of different dialects. Inter-dialect differences exist in pronunciation, vocabulary and syntactic rules. However, they are usually insignificant in comparison with the similarities the dialects have. It has been declared that dialects of one language are mutually intelligible (Fromkin and Rodman 1993, p. 276).

Nevertheless, this is not true to the situation in China. There are seven major Chinese dialects: the Northern Dialect (with Mandarin as its standard version), Cantonese, Wu, Min, Hakka, Xiang and Gan (Yuan, 1989), that for the most part are mutually *unintelligible*, and inter-dialect

[1] In this paper, MT refers to both computer-based translation and interpretation.

translation is often found indispensable for successful communication, especially between Cantonese, the most popular and the most influential dialect in South China and overseas, and Mandarin, the lingual franca of China.

2 Linguistic Consideration of Dialect MT

Most differences among the dialects of a language are found in their sound inventory and phonological systems. Words with similar written forms are often pronounced differently in different dialects. For example, the same Chinese word "香港" (Hong Kong) is pronounced *xiang1gang3*[2] in Mandarin, but *hoeng1gong2* in Cantonese. There are also lexical differences although dialects share most of their words. Different dialects may use different words to refer to the same thing. For example, the word "umbrella" is 雨伞 (yu3san3) in Mandarin, and 遮 (ze1) in Cantonese. Differences in syntactic structure are less common but they are linguistically more complicated and computationally more challenging. For example, the positions of some adverbs may vary from dialect to dialect. To express "You go first", we have

Mandarin:

你	先	走	
ni 3	xian1	zou3	(1)
you	first	go	

Cantonese:

你	行	先	
nei5	hang4	sin1	(2)
you	go	first	

Comparative sentences represent another case where syntactic difference is likely to happen. For example the English sentence "A is taller than B" is expressed as

Mandarin:

A	比	B	高	
A	bi3	B	gao1	(3)
A	than	B	tall	

Cantonese:

A	高	过	B	
A	gou1	gwo3	B	(4)
A	tall	more	B	

Sentences with double objects often follow different word orders, too. In a Mandarin sentence with two objects, the one referring to person(s) must be put before the other one. Yet, many dialects allow the order to be reversed, for example:

Mandarin:

我	先	给	他	钱
wo3	xian1	gei3	ta1	qian2
I	first	give	him	money

I will give him some money first.

Cantonese:

我	俾	钱	佢	先
ngo3	bei2	cin4	keoi5	sin1
I	give	money	him	first

Differences in word pronunciation and word forms can be represented in a bi-dialect dictionary. For example, for Cantonese-Mandarin MT, we can use entries like

word(pron, [你, ni3], [你, nei5]) %you
word(vi,[走, zou3], [行, hang4]) %go
word(n,[行, hang2], [行, hang4]) %row
word(adv, [先, xian1], [先, sin1]) %first
word(n, [雨伞, yu3san3],[遮, ze1]) %ubbrella

where the word entry flag "word" is followed by three arguments: the part of speech and the corresponding words (in Chinese characters and pinyins) in Mandarin and in Cantonese. English comments are marked with "%".

Morphologically, there are some useful rules for word formation. For example, in Mandarin, the prefixes "公" (gong1) and "雄" (xiong2) are for male animals, and "母" (mu3) and "雌"(ci2) female animals. But in most southern China dialects, the suffixes "公/牯" and "鲩/婆" are often used instead. For examples

bull/ox:
Mandarin 公牛 (gong1niu1),
Cantonese 牛公 (ngau4gung1),
cow:
Mandarin 母牛 (mu3niu2),
Cantonese 牛鲩 (ngau4naa2).

And Cantonese "阿" is for calling, e.g., Daddy:

[2] In this paper, pronunciation of *Mandarin* is presented in Hanyu Pinyin Scheme (LICASS, 1996), and *Cantonese* in Yueyu Pinyin Scheme (LSHK, 1997). Numbers are used to denote tones of syllables. Yueyu Pinyin is based on Hanyu Pinyin. That means, across the two pinyin schemes, words with different pinyin symbols are normally pronounced differently.

阿爸 (Cantonese), 爸爸 (Mandarin),
Elder brother:
阿哥 (Cantonese), 哥哥 (Mandarin).

The problem caused by syntactic difference can be tackled with linguistic rules, for example, the rules below can be used for Cantonese-Mandarin MT of the previous example sentences:

Rule 1: NP xian1 VP <--> NP VP sin1
NP first VP <--> NP VP first

Rule 2: bi3 NP ADJP <--> ADJP go3 NP
than more

Rule 3: gei3 (%give) O_{person} O_{thing} <-->
bei2 (%give) O_{thing} O_{person}

Inter-dialect syntactic differences largely exists in word orders, the key task for MT is to decide what part(s) of the source sentence should be moved, and to where. It seems unlikely for words to be moved over long distances, because dialects normally exist in spoken, short sentences.

Another problem to be considered is whether dialect MT should be direct or indirect, i.e., should there be an intermediate language/dialect? It seems indirect MT with the lingua franca as the intermediate representation medium is promising. The advantage is twofold: (a) good for multi-dialect MT; (b) more useful and practical as a lingua franca is a common and the most influential dialect in the family, and maybe the only one with a complete written system.

Still another problem is the forms of the source and target dialects for the MT program. Most MT systems nowadays translate between written languages, others are trying speech-to-speech translation. For dialects MT, translation between written sentences is not that admirable because the dialects of a language virtually share a common written system. On the other hand, speech to speech translation involves speech recognition and speech generation, which is a challenging research area by itself. It is worthwhile to take a middle way: translation at the level of phonetic symbols. There are at least three major reasons: (a) The largest difference among dialects exists in sound systems. (b) Phonetic symbol translation is a prerequisite for speech translation. (c) Some dialect words can only be represented in sound. In our case, pinyins have been selected to represent both input and output sentences, because in China pinyins are the most popular tools to learn dialects and to input Chinese characters to computers. Chinese pinyin schemes, for Mandarin and for ordinary dialects are romanized, i.e., they virtually only use English letters, to the convenience of computer processing. Of course, pinyin-to-pinyin translation is more difficult than translation between written words in Chinese block characters because the former involves linguistics analysis at all the three aspects of sound systems, grammar rules and vocabulary contents in stead of two.

3 The Problem of Ambiguities

Ambiguity is always the most crucial and the most challenging problem for MT. Since inter-dialect differences mostly exist in words, both in pronunciation and in characters, our discussion will concentrate on word disambiguation for Cantonese-Mandarin MT. In the Cantonese vocabulary, there are about seven thousand to eight thousand dialect words (including idioms and fixed phrases), i.e., those words with different character forms from any Mandarin words, or with meanings different from the Mandarin words of similar forms. These dialect words account for about one third of the total Cantonese vocabulary. In spoken Cantonese the frequency of use of Cantonese dialect words is close to 50 percent (Li, et. al., 1995, p236). Because of historical reasons, Hong Kong Cantonese is linguistically more distant from Mandarin than other regions in Mainland China. One can easily spot Cantonese dialect articles in Hong Kong newspapers which are totally unintelligible to Mandarin speakers, while Mandarin articles are easily understood by Cantonese speakers. To translate a Cantonese article into Mandarin, the primary task is to deal with the Cantonese dialect words, especially those that do not have semantically equivalent counterparts in the target dialect. For example, the Mandarin 桔(ju2, orange) has a much larger coverage than the Cantonese 桔(gwat1). In addition to the Cantonese 桔, the Mandarin 桔 also includes the fruits Cantonese refers to as 柑 (gam1) and 橙(caang2). On the other hand, the Cantonese 行 semantically covers the Mandarin 走 (go, walk) and 行 (row). Translation at the sound or pinyin level has to

deal with another kind of ambiguity: the homophones of a word in the source dialect may not have their counterpart synonyms in the target dialect pronounced as homophones as well. For example, the words 香蕉(banana) and 相交 (intersection) are both pronounced *xiang1jiao1* in Mandarin, but in Cantonese they are pronounced *hoeng1ziu1* and *soeng1gaau1* respectively, though their written characters remain unchanged.

To tackle these ambiguities, we employs the techniques of hierarchical phrase analysis (Zhang and Lu, 1997) and word collocation processing (Sinclair, 1991), both rule-based and corpus-based. Briefly speaking, the hierarchical phrase analysis method firstly tries to solve a word ambiguity in the context of the smallest phrase containing the ambiguous word(s), then the next layer of embedding phrase is used if needed, and so on. As a result, the problem will be solved within the minimally sufficient context. To further facilitate the work, large amount of commonly used phrases and phrase schemes are being collected into the dictionary. Further more, interaction between the users and the MT system should be allowed for difficult disambiguation (Martin, 1997a).

4 System Design and Implementation

A rudimentary design of a Cantonese-Mandarin dialect MT system has been made, as shown in Figure 1. The system takes Cantonese Pinyin sentences as input and generates Mandarin sentences in Hanyu Pinyin and in Chinese characters. The translation is roughly done in three steps: syntax conversion, word disambiguation and source-target words substitution. The knowledge bases include linguistic rules, a word collocation list and a bi-dialect MT dictionary.

A simplified example will make the basic ideas clearer. Suppose the example word entries and transformational rules in Section 2 are included in the MT system's knowledge base. Example sentence (2) in Cantonese, i.e.,

nei5	hang4	sin1	
你	行	先	(2)
you	go	first	

is given as input for the system to translate into Mandarin. Because the input sentence contains the time adverb "sian1" (first), according to grammar rules, it is syntactically different from its counterpart in Mandarin. According to the flowchart, the Cantonese pinyin sentence is converted into a Mandarin structure. Rule 1 in the knowledge base is applied, producing

nei5	sin1	hang4
你	先	行
you	first	go

Then the dictionary is accessed. The Cantonese word 行(hang4) corresponds to two Mandarin words, i.e., 走(vi. go, walk) and 行(n. row). According to Rule 1, the verb Mandarin word is selected. And the individual Cantonese words in the sentence are substituted with their Mandarin counterparts, a target Mandarin sentence

ni 3	xian1	zou3
你	先	走
you	first	go

like sentence (1) is then correctly produced.

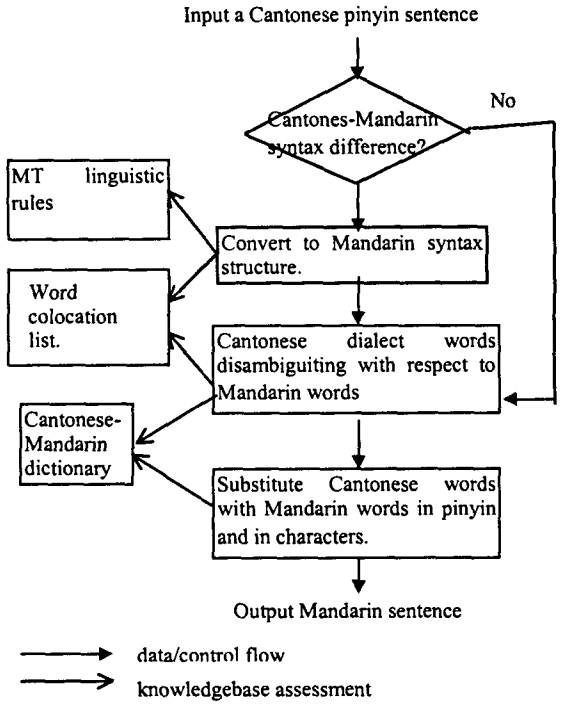

Figure 1: A Design for Cantonese-Mandarin MT

Similarly, with transformational rule 1-3, a more complicated Cantonese sentence like

高 过 我 既 人 俾 钱 佢 先。

gou1gwo3 wo3 ge3 yan4 bei2 cin4 keoi5 sin1

tall more me PART person give money him first

can be correctly translated into Mandarin:

比 我 高 的 人 先 给 他 钱。
bi3 wo3 gao1 de ren2 xian1 gei3 ta1 qian2
than me tall PART persons first give him money
Those who are taller than me will give him some money first.

We are in the progress of implementing an inter-dialect MT prototype, called CPC, for translation between Cantonese and Putonghua (i.e., Mandarin), both Cantonese-to-Putonghua and Putonghua-to-Cantonese. Input and output sentences are in pinyins or Chinese characters. The programming languages used are Prolog and Java. We are doing Cantonese-to-Putonghua first, based on the design. At its current state, we have built a Cantonese-Mandarin bi-dialect dictionary of about 3000 words and phrases based on some well established books (e.g., Zeng, 1984; Mai and Tang, 1997), (When completed, there will be around 10,000 word entries) and a handful of rules. A Cantonese-Mandarin dialect corpus is also being built. The program can process sentences of a number of typical patterns. The funded project has two immediate purposes: to facilitate language communication and to help Hong Kong students write standard Mandarin Chinese.

Conclusion

Compared with inter-language MT, inter-dialect MT is much more manageable, both linguistically and technically. Though generally ignored, the development of inter-dialect MT systems is both rewarding and more feasible. The present paper discusses the design and implementation of dialect MT systems at pinyin and character levels, with special attention on the Chinese Mandarin and Cantonese. When supported by the modern technology for multimedia communication of the Internet and the WWW, dialect MT systems will produce even greater benefits (Zhang and Lau, 1996).

Nonetheless, the research reported in this paper can only be regarded as an initial exploratory step into a new exciting research area. There is large room for further research and discussion, especially in word disambiguation and syntax analysis. And we should also notice that the grammars of ordinary dialects are normally less well described than those of lingua francas.

Acknowledgements

The research is funded by Hong Kong Polytechnic University, under the project account number of 0353 131 A3 720.

References

Fromkin V. and Rodman R. (1993) *An Introduction to Language* (5th edition). Harcourt Brace Jovanovich College Publishers, Orlando, Florida, USA., p. 276.

Li X., Huang J., Shi Q., Mai Y. and Chen D. (1995) *Guangzhou Fangyan Janjiu (Research in Cantonese Dialect).* Guangdong People's Press, Guangzhou, China, p. 236.

LICASS (Language Institute, the Chinese Academy of Social Sciences) (1996) *Xiandai Hanyu Cidian (Contemporary Chinese Dictionary).* Commercial Press, Beijing, China.

LSHK (1997) *Yueyu Pinyin Zibiao (The Chinese Character List with Cantonese Pinyin).* Linguistic Society of Hong Kong, Hong Kong.

Mai Y. and Tang B. (1997) *Shiyong Guangzhouhua Fenlei Cidian (A Practical Semantically-Classified Dictionary of Cantonese).* Guandong People's Press, Guangzhou, China.

Martin K. (1997a) *The proper place of men and machines in language translation.* Machine Translation, 1-2/12, pp. 3-23.

Martin K. (1997b) *It's still the proper place.* Machine Translation, 1-2/12, pp. 35-38.

Nirenburg S., Carbonell J., Tomita M. and Goodman K. (1992) *Machine Translation: A Knowledge-Based Approach.* Morgan Kaufmann Publishers, San Mateo, California, USA.

Sinclair J. (1991) *Corpus, Concordance and Collocation.* Collins, London, UK.

Yuan J. (1989) *Hanyu Fangyan Gaiyao (Introduction to Chinese Dialects).* Wenzi Gaige Press, Beijing, China.

Zeng Z. F. (1984) *Guangzhouhua-Putonghua Kouyuci Duiyi Shouce (A Translation Manual of Cantonese-Mandarin Spoken Words and Phrases).* Joint Publishing, Hong Kong.

Zhang X. and Lau C. F. (1996) *Chinese inter-dialect machine translation on the Web.* In "Collaboration via the Virtual Orient Express: Proceedings of the Asia-Pacific World Wide Web Conference" S. Mak, F. Castro & J. Bacon-Shone, ed., Hong Kong University, pp. 419--429.

Zhang X. and Lu F. (1997) *Intelligent Chinese pinyin-character conversion based on phrase analysis and dynamic semantic collocation.* In "Language Engineering", L. Chen and Q. Yuan, ed., Tsinghua University Press, Beijing, China, pp. 389-395.