Chinese Inter-Dialect Machine Translation

Zhang Xiaoheng and Shi Dingxu
Department of Chinese and Bilingual Studies
Hong Kong Polytechnic University

Abstract: This paper will present, from a computational point of view, a comparative study of Chinese dialects at the three aspects of sound systems, grammar rules and vocabulary contents, with Putonghua, Chaozhouhua and Cantonese as examples. Then we will proceed to discuss the design and implementation of a pinyin-to-pinyin inter-dialect machine translation system, followed by a conclusion and discussion on further research.

Key words: Chinese dialects, machine translation

漢語方言機器翻譯

張小衡 石定栩

(香港理工大學中文及雙語學系)

摘要: 本文以普通話、廣東話和潮州話爲例，從計算語言學的角度出發對漢語方言在語音、詞匯和語法三方面作對比研究。并在此基礎上討論拼音級漢語方言機器翻譯系統的設計與實現。

關鍵詞： 漢語方言　機器翻譯

1 Inter-Language MT and Inter-Dialect MT

Computer-based Machine Translation (MT) between different languages has been an attractive but extremely difficult research area. Over forty years of MT history has seen limited practical translation systems developed or commercialized. Natural languages, such as English, Chinese and Japanese, are very complicated and different from each other. High quality machine translation between two languages requires deep understanding of the intended meaning of the source language sentences. A sentence can be ambiguous at each level of linguistic analysis: phonology, morphology, syntax, semantics and pragmatics. And computer-based disambiguation is an extremely hard job which requires the intelligent search for and proper use of a great amount of relevant knowledge, including common sense [6].

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, where many dialects exist and translation or interpretation among them is often found indispensable for successful communication. In the following sections we will discuss the design and implementation of a Chinese inter-dialect MT system. Our discussion will be done both linguistically and computationally.

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

2 Linguistic Consideration

2.1 Dialects and Chinese Han Dialects

Dialects of a language are systematic variations of that language. Fromkin and Rodman [3] consider dialects mutually intelligible forms of a language, with systematic differences between them. Dialects often develop when people of a common language are separated from each other geographically and socially. Dialects of the same language may differ in their sound inventory, phonological system, morphological build-up, lexicon or even syntax. However, differences among dialects of the same language are usually insignificant in comparison with the similarities they have. Otherwise mutual intelligibility would not exist. Dialects of the same language often share the same writing system. Sometimes there exists a super-dialect that serves as the *lingua franca*, namely, the common dialect, among speakers of different dialects because it enjoys certain social, geographic or economic advantages.

Dialects of *Hanyu,* which is commonly referred to as Chinese, presents a slightly different picture. For one thing, there may not be mutual intelligibility among many dialects of Chinese. For another thing, the lingua franca of all Chinese dialects, namely, *Putonghua* or Mandarin as it is known to the West, is not a natural dialect since it is no one's native tongue. There are seven major dialects in Chinese: the Northern Dialect, Cantonese, Wu, Min, Hakka, Xiang and Gan [9; 3]. These dialects differ from each other mainly in their sound inventory and phonological system, although differences in vocabulary and syntax do exist. Because of the lack of mutual intelligibility among so many dialects, inter-dialect interpretation is frequently required. The following discussion on dialect MT will draw examples from Cantonese, Chaozhouhua and Putonghua. Cantonese is the most popular dialect in Guangdong and Hong Kong. And Chaozhouhua, a sub-dialect of the Min widely used in Fujian and Taiwan, is another influential dialect in Guangdong and Hong Kong[4].

2.2 Linguistic Consideration on Dialect MT

Most differences among the dialects of a language are found in their sound inventory and phonological systems. Words in similar written forms are normally pronounced differently in different dialects. For example, the same Chinese word 香港 (Hong Kong) is pronounced *xiang1gang3*[2] in Putonghua, *hoeng1gong2* in Cantonese, but *hiang1gang2* in Chaozhouhua. 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 雨傘 (yu3san3) in Putonghua, 雨遮 (hou6zia1) in Chaozhouhua, 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

Putonghua:	你	先	走	
	ni 3	xian1	zou3	(1)
	you	first	go	
Cantonese:	你	行	先	
	nei5	hang4	sin1	(2)
	you	go	first	
Chaozhouhua:	你	先	行	
	le2	sian1	gian5	(3)
	you	first	go	

[2] In this paper, pronunciation of *Putonghua* are expressed in Hanyu Pinyin Scheme [8], *Cantonese* in Yueyu Pinyin Scheme [2], and *Chaozhouhua* in Chaozhouhua Pinyin Scheme [5]. Numbers are used to denote tones of syllables. Both Yueyu Pinyin and Chaozhouhua Pinyin are based on Hanyu Pinyin. That means, across the three pinyin schemes, words with different pinyin symbols are pronounced differently.

where the Putonghua and Chaozhouhua sentences are structurally similar, while their Cantonese counterpart has a different syntax structure.

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

Putonghua:	A	bi3	B	gao1	(4)
	A	than	B	tall	
Cantonese:	A	gou1	gwo3	B	(5)
	A	tall	more	B	
Chaozhouhua:	A	guin5	(gue3)	B	(6)
	A	tall	(more)	B	

where the contents in brackets are optional.

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

Putonghua:	我	先	給	他	錢。
	wo3	xian1	gei3	ta1	qian2
	I	first	give	him	money
	I will give him some money first.				
Cantonese:	我	俾	錢	佢	先。
	ngo3	bei2	cin4	keoi5	sin1
	I	give	money	him	first
Chaozhouhua:	我	先	給	伊	錢
	wa2	sain1	ke3	yi1	zin5
	I	first	give	him	money

Though it happens most frequently, the problem of pronunciation difference is the easiest to solve in dialect MT. It seems wise to use a romanized pinyin scheme for each dialect, for the convenience of computer input with the international standard keyboard. If we put the pinyin words of the relevant dialects in the word entries of the MT dictionary, it will become trivial for the computer to replace the source dialect's words with their target dialect counterparts during the MT procedure.

Similarly, the lexical problem can be solved by adding the dialects' different written words into the dictionary word entries. For example, a tri-dialect dictionary for Putonghua-Cantonese-Chauzhouhua MT may have entries like

word(n, [你,ni3], [你,nei5], [你, le2])
word(v, [走, zou3], [行, hang4], [行, gian5])
word(adv, [先, xian1], [先, sin1], [先, sian1])
word(v, [看, kan4], [睇, tai2], [睇, toin2])

where the word entry flag "word" is followed by four arguments representing part of speech and the corresponding words (in Chinese characters and pinyins) in Putonghua, Cantonese and Chaozhouhua.

Morphologically, there are some useful rules for words formation. For example, in Putonghua, 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 "母/婆" are often used instead. For examples

bull/ox:
Putonghua 公牛(gong1niu1),
Cantonese 牛公(ngau4gung1),
Chaozhouhua 牛牯(ghu5gou2)
male dog:
Putonghua 公狗(gong1gou3),

Cantonese 狗公(gau2gung1),
Chaozhouhua 狗牯(gao2gou2)

while the word "cow" and "female dog" are expressed as

cow:
Putonghua 母牛(mu3niu2),
Cantonese 牛母(ngau4mou5),
Chaozhouhua 牛母(ghu5bho2)
female dog/bitch :
Putonghua 母狗(mu3gou3),
Cantonese 狗母(gau2mou2),
Chaozhouhua 狗母(gao2gbo2)

In Chaozhouhua, doubling a single-syllable noun can produce an adjective, for example: 金(gold), 金金 (golden, shining); 霧(fog), 霧霧 (foggy, blurred); 猴(monkey), 猴猴 (as thin as a monkey); 柴(wood, clump, log), 柴柴 (clumsy, awkward).

In spite of the difference in word forms, word translation is trivial for computers so long as there is a semantic one-to-one relation among the relevant dialects, i.e., for each word in a source dialect, there is a corresponding word of similar meaning in the target dialect. Unfortunately, a handful of exceptions do exists. For example, in Chaozhouhua, the word 鼻(pin7) can refer to the Putonghua word 鼻子(bi4zi, nose) and 鼻涕(bi2ti4, snivel). To deal with this kinds of ambiguity, analysis at syntax or higher levels is required. Statistical word collocation data gained from corpora can also help [7].

The problem caused by syntactic difference can be tackled with parsing and transformational rules, For example, the rules below can be used for Putonghua-Cantonese MT of the previous example sentences:

Rule 1: NP xian1 VP --> NP VP sin1
Rule 2: bi3 NP ADJP --> ADJP go3 NP
Rule 3 gei3 O_{person} O_{thing}-->bei2 O_{thing} O_{person}

The first rule only involves word order, while the other two rules involves word order as well as word substitution. Inter-dialect syntactic difference 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 of the limitation of human's short term memory. Dialects normally exist in spoke forms. Chafe [1] declared that in spontaneous conversation, there are basic sentence components called *idea units* of about six English words (or two seconds) long.

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. However, pure indirect MT may not be the most desirable. A close study will soon reveal that some source sentences are grammatically similar to the target sentences, while different from their counterparts in the lingua franca. Such cases happen more often when the source dialect is linguistically closer to the target dialect than to the intermediate dialect. For example, to express "I walk faster than you",

Cantonese:	ngo5	hang4	dak1	fai3	gwo3	nei5
	I	walk	PART.	fast	more	you
Chaozhouhua:	wa2	gian5	(deg4)	meng2	gue3	le2
	I	walk	(PART.)	fast	more	you
Putonghua:	wo3	zou3	de	bi3	ni3	kuai4
	I	walk	PART.	than	you	fast

In a Cantonese-Chaozhouhua dialect MT system with Putonghua as intermediate representation, such source dialect sentences should be allowed to be directly converted into the target dialect. In other words, some bypasses should be allowed for special cases like this.

Still another problem to be considered 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 not be written as Chinese characters. In our case, pinyins have been selected to represent both input and output sentences, because in China pinyins are the most popularly used tools to learn dialects and to input Chinese character to computers. Of course, pinyin-to-pinyin translation is more difficult than translation between written words because the former involves linguistics analysis at all the three aspects of sound systems, grammar rules and vocabulary contents in stead of two.

3 System Design and Implementation

Based on the previous linguistic consideration. A design of a Chinese dialect MT system has been made, as shown in Figure 1.

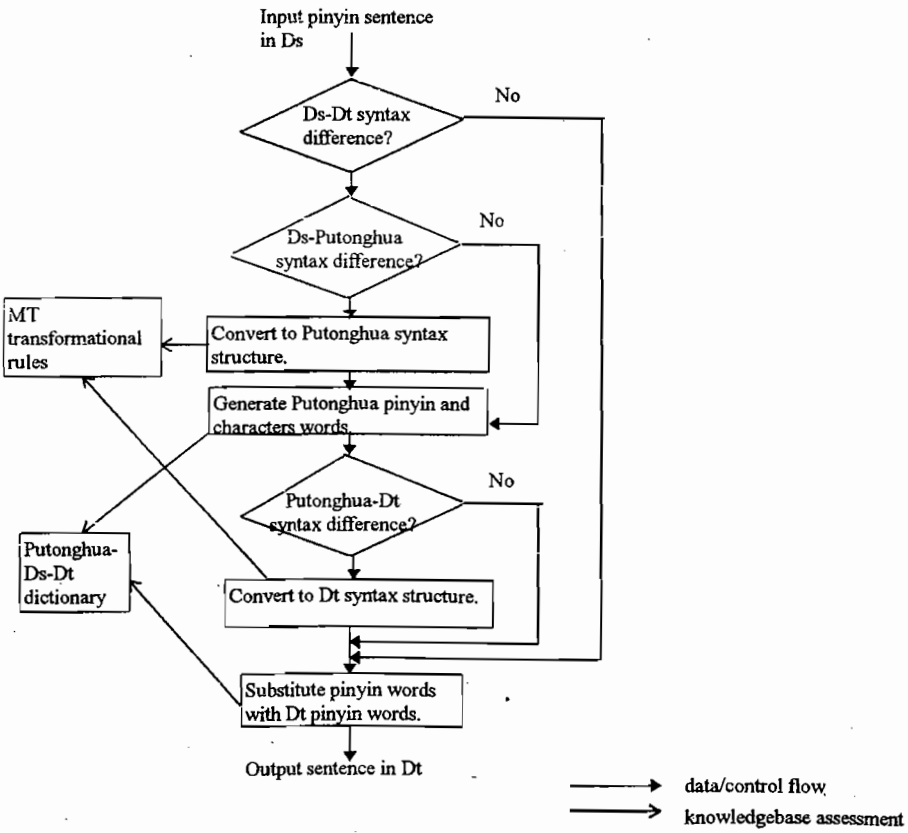

Figure 1: Design of a Chinese dialect MT system

An input sentence in source dialect (Ds) is translated into a target dialect (Dt) with the lingua franca Putonghua as an intermediate representation. Bypasses are provided for special cases where a source sentence and its corresponding target sentence are grammatically similar. Sentences of both source and target dialects are in romanized pinyin forms, while Putonghua can have a Chinese character form as well, enabling the system to generate output in three versions: Dt sentences in pinyin, Putonghua sentences in pinyin and Putonghua sentences in Chinese characters. The translation is roughly done in three steps: syntax conversion, word substitution and pinyin substitution, according to the three categories of inter-dialect differences mentioned in last section. The knowledgebase include linguistic transformational rules and a MT dictionary, as the figure shows.

An example will help to make the basic ideas clearer. Suppose the example word entries and transformational rules in last section are included in the MT system's knowledgebase. Example sentence (3) in Chaozhouhua, i.e.,

le2 sian1 gian5
you first go

is given as Ds input for the system to translate into the Dt of Cantonese. Because the Ds sentence contains the time adverb "sian1"(first), according to grammar rules, it is syntactically different from its counterpart in Dt, but is syntactically similar to the sentence in Putonghua. According to the flowchart, words in the input sentence are substituted with Putonghua pinyin and characters words. Then the Putonghua pinyin sentence is converted into a Cantonese structure. Transformational Rule 1 in the knowledgebase is applied, producing

ni3 zou3 sian1
you go first

Finally, the dictionary is accessed and the individual Putonghua words in the sentence are substituted with their Cantonese counterparts, a target Cantonese sentence

nei5 hang4 sin1
you go first

like sentence (2) is then correctly produced. Similarly, with transformational rule 1-3, a complicate Putonghua sentence like

比	我	高	的	人	先	給	他	錢
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.

can be correctly translated into Cantonese:

高過	我	既	人	俾	錢	佢	先
gou1gwo3	wo3	ge3	yan4	bei2	cin4	keoi5	sin1
tall more	me	PART	person	give	money	him	first

We have set out to implement an inter-dialect MT prototype, called CPC, for translation between Cantonese and Chauzhouhua using Putonghua as the intermediate dialect. Input and output sentences are in pinyins. Putonghua sentences can also be in Chinese characters. The programming languages used are Prolog and C. At its current state, the program can process a number of typical sentence patterns.

4. Conclusion and Discussion

Comparing 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 feasible. The present paper has discussed the design and implementation of dialect MT both linguistically and computationally. Though the example dialects used in this paper are confined to the Chinese lingua franca Putonghua, and the authors' native tongues: Cantonese and Chaozhouhua, the basic ideas illustrated should be applicable to other Chinese and non-Chinese dialect MT as

well. When supported by the modern technology for multimedia communication provided by the Internet and the Web, dialect MT systems will produce even greater benefits[10]. Nonetheless, the research reported in this papers can only be regarded as an initial exploratory step into a new exciting research area. There is large room for further research and discussion.

Our methodology for inter-dialect MT heavily relies on pinyins, both input and output of the system are in pinyin. The reasons include (a). Many dialects do not have their own complete written systems. Some dialect words have no written forms at all. Instead the dialects of a common language normally share a common writing system. (b). Because a common writing system is shared, it doesn't make much sense if both input and output are in written character forms. (c). Pinyin MT is essential to speech MT. Nevertheless, pure pinyin input and output also has its negative side effects. For example, 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 Putonghua, but in Chaozhouhua they are pronounced *hiang1zio1* and *siang1gao1* respectively, though the written characters remain unchanged.

In addition, as in inter-language MT, we have the problem that some (though not many) source dialect words do not have semantically identical counterparts in the target dialect, i.e., there is no strict one-to-one relation between the source vocabulary and the target vocabulary. For example, the Putonghua 橘(*ju2*, orange) has a much larger coverage than the Cantonese 橘(*gwat1*). In addition to the Cantonese 橘, the Putonghua 橘 also includes the fruits Cantonese refers to as 柑(*gam1*) and 橙(*caang2*). Such problems sometimes requires substantial linguistic analysis. Unfortunately, the grammar of many dialects is not well described yet. There is no free lunch. Though more manageable as a whole, inter-dialect MT is not easier than inter-language MT in every aspects. Interaction between the users and the MT system should be allowed for difficult disambiguation.

Theoretically, MT can be carried out between any dialects in a language. Yet, considering that a lingua franca, such as Putonghua in China, is a shared special dialect in the language family, especially in its written form, it seems more important and urgent to concern about the dialect-to-lingua-franca MT first.

References

[1] Chafe, W. Integration and involvement in speaking, writing, and oral literature. In D. Tannen ed., Spoken and Written Language. Norwood: Ablex, 1982, pp35-53.

[2] Cheung, Kwan-hin Reply to reviews of Cantonese Pinyin Scheme, Chinese Language Review, 43:34-37, 1994.

[3] Fromkin, V. and Rodman, R. An Introduction to Language, (5th edition), Harcourt Brace Jovanovich College Publishers, Orlando, FL, 1993.

[4] Li, Xinkiu. Guangdong de Fangyan (Dialects in Guangdong Province), Guangdong People's Press, Guangzhou, China, 1994.

[5] Li, Xinkui. Putonghua-Chaoshanfangyan Changyong Zidian (Putonghua-Chaozhouhua Popular Characters Dictionary), Guangdong People's Press, Guangzhou, China, 1993.

[6] Nirenburg, S., Carbonell, J., Tomita, M. and Goodman, K. Machine Translation: A Knowledge-Based Approach, Morgan Kaufmann Publishers, San Mateo, California, 1992.

[7] Sinclair, J. Corpus, Concordance and Collocation. Collins, London, 1991.

[8] ------, Xinhua zidian (Xinhua Chinese Character Dictionary), Commercial Press, Hong Kong, 1989.

[9] Yuan, Jiahua. Hanyu Fangyan Gaiyao (Introduction to Chinese Dialects). Wenzi Gaige Press, Beijing, 1989.

[10] Zhang, Xiaoheng and Lau, Chun-fat. Chinese inter-dialect machine translation on the Web. In Mak, S. Castro, F. and Bacon-Shone, J.(eds.), Collaboration via The Virtual Orient Express: Proceedings of the Asia-Pacific World Wide Web Conference , Hong Kong and Beijing, August 1996, pp. 419--429.