Register-sensitive Translation:
A Case Study of Mandarin and Cantonese

Tak-sum Wong tswong-c@my.cityu.edu.hk
John Lee jsylee@cityu.edu.hk
Department of Linguistics and Translation, City University of Hong Kong,
Hong Kong SAR, P. R. China

Abstract

This paper describes an approach for translation between Chinese dialects that can produce target sentences at different registers. We focus on Mandarin as the source language, and Cantonese as the target. Mutually unintelligible, these two varieties of Chinese exhibit differences at both the lexical and syntactic levels, and the extent of the difference can vary considerably depending on the register of Cantonese. Since only a modest amount of parallel data is available, we adopt a knowledge-based approach and exploit lexical mappings and syntactic transformations from linguistics research. Our system parses a source sentence, uses register-annotated lexical mappings to translate words, and then performs word reordering through syntactic transformations. Evaluation shows that translation models that match the required register of the target sentences yield better translation quality.

1. Introduction

A large number of Chinese dialects are spoken in different regions of China. Many of these dialects are not mutually intelligible (Killingley, 1993; Szeto, 2000); indeed, the differences between the major Chinese varieties have been described as being “at least on the order of the different languages of the Romance family” (Hannas, 1997: 198), or “roughly parallel to English, Dutch, Swedish, and so on among the Germanic group of the Indo-European language family” (Mair, 1991: 3). This paper describes an approach for translating between Chinese dialects, focusing on Mandarin as the source language and Cantonese as the target language.

Mandarin, also known as Pǔtōnghuà, is considered standard Chinese and is the dominant variety in mainland China. Spoken by more than 55 million people, Cantonese, the “most widely known and influential variety of Chinese other than Mandarin” (Matthews and Yip, 2011), is the dominant variety of Chinese spoken in Hong Kong. In Hong Kong, Cantonese is used mainly in speech, while Mandarin is dominant in written contexts. This division of labor is somewhat comparable, for example, to the usage patterns of Swiss German dialects and standard German in Switzerland.

Although mutually unintelligible in their spoken form, Cantonese and Mandarin are genetically related, having both developed from Middle Chinese. They share similar writing systems, as well as many cognates. Most Mandarin lexical items can also be used in Cantonese. In a study on the Leipzig-Jakarta list of 100 basic words (Tadmor et al., 2010: 239–241), 60% of the Mandarin-Cantonese word pairs have identical written forms, most of which have highly regular phonological correspondence; a further 20% have the same core morpheme (Li et al., 2016). However, lexical items in Cantonese can vary considerably depending on the register, *i.e.*, language variation according to context (Halliday and Hasan, 1989; Quirk et al., 1985). Low-register Cantonese, typified by casual, informal speech, is often peppered with lexical items that are not used in Mandarin; in higher-register Cantonese, more lexical items are shared with the standard Chinese lexicon. Put otherwise, “an increase in informality corresponds to an increase in the number of Cantonese lexical items occurring in speech which makes it less like the lexicon of standard Chinese. … On the other hand, an increase in the formality of the social context calls for a corresponding increase in the number of standard Chinese lexical items occurring in the utterance (but of course pronounced in Cantonese)." (Bauer, 1988: 249). These variations among different registers make Cantonese a challenging target language for a case study for machine translation (MT) among Chinese dialects.

Most MT systems do not yet take the notion of register into account. A recent study on English-to-German translation found that both manually and automatically translated texts differ from the original texts in terms of register (Lapshinova-Koltunski and Vela, 2015). Neither does mainstream MT evaluation explicitly consider appropriateness in register, despite recent studies which argue it should (*e.g.*, Vela and Lapshinova-Koltunski, 2015). In this paper, we propose and evaluate a knowledge-based MT system that can translate Mandarin input into Cantonese at different registers.

The rest of the paper is organized as follows. In the next section, we outline previous research on MT for dialects in China and beyond. In section 3, we describe our translation approach. In section 4, we report both automatic and human evaluation, and analyze the main sources of error. Finally, we conclude in Section 5.

2. Previous work

A number of previous studies are related to our research in terms of language, data genre, and approach. Among the earliest attempts on translation between Chinese dialects is the knowledge-based approach taken by Zhang (1998), although no evaluation was reported. Xu and Fung (2012) developed a Cantonese-to-Mandarin MT system that is appended to an automatic speech recognition system for Cantonese, allowing it to output transcription in Mandarin Chinese. The translation capability was implemented with a cross-lingual language model with Inversion Transduction Grammar constraints for syntactic reordering.

Dialect MT is often applied to the task of television subtitle generation. Volk and Harder (2007) implemented an MT system, already in production use, for subtitle machine translation from Swedish to Danish. The system was trained using a statistical MT system, using a parallel corpus with 1 million subtitles. It has been further improved with morphological annotations (Hardmeier and Volk, 2009).

The translation approach taken in this paper is most similar to the knowledge-based system for translating standard German to Swiss German dialects, reported in Scherrer and Rambow (2010) and Scherrer (2011). Their approach uses a word list, compiled by experts, to handle lexical differences; and a set of syntactic transformations, defined by constituent-structure trees, to change sentence structures from German to Swiss German. Most rules achieved 85% accuracy or above. The system customizes the target sentence by selectively applying these rules according to the intended dialect area in Switzerland. Our approach also customizes the target sentence, but according to the register level rather than dialect area.

3. Approach

Statistical machine translation (MT) and neural network approaches have been successfully applied on many language pairs (Koehn et al., 2007), including dialects and other closely related languages (*e.g.*, Volk and Harder, 2007; Delmonte et al., 2009). One critical requirement for these approaches is the availability of a large amount of parallel sentences. In our case, due to the lack of standard written form for Cantonese, and the dominance of Mandarin in the written context, parallel Mandarin-Cantonese sentence pairs do not exist in large quantity. Taking a statistical approach to generate target sentences at different registers would compound the data sparseness issue, since both low- and high-register training data would be needed.

Despite the relative paucity of parallel data, Cantonese has been extensively studied by linguists (Zeng, 1993; Ōuyáng, 1993; *etc.*). It is thus less costly to exploit existing resources such as word lists and syntactic transformations, than to collect bilingual sentence pairs to overcome data sparseness. Hence, we adopt a knowledge-based approach for Mandarin-to-Cantonese translation, similar to that of an MT system for translating standard German into Swiss German dialects (Scherrer and Rambow, 2010; Scherrer, 2011). Our approach consists of three steps. First, it uses the Stanford Chinese parser to perform word segmentation, part-of-speech (POS) tagging and dependency parsing (Levy and Manning, 2003). Second, it uses forward maximal matching to look up Mandarin-to-Cantonese lexical mappings (Section 3.1), conditioned on POS information and register requirement (Section 3.2). Finally, it applies syntactic transformations on the output, with word re-ordering when warranted (Section 3.3).

3.1. Lexical mappings

The lexical mapping contains pairs of equivalent Mandarin and Cantonese words, taken from a parallel corpus of transcribed Cantonese speech and Mandarin Chinese subtitles (Lee, 2011). The speech was transcribed from television programmes broadcast in Hong Kong within the last decade by Television Broadcasts Limited. The Cantonese and Mandarin text were manually word-segmented and aligned. The TV programmes span a variety of genres, including news programmes, current-affairs shows, drama series and talk shows. These programs not only include vocabulary from widely different domains, but also contain Cantonese spoken in different registers. The most formal language is used in news, and the most colloquial in drama series and talk shows. We harvested all word alignments from the corpus to create lexical mappings. We further supplemented these mappings with a Cantonese-Mandarin dictionary that is freely available from the website of Kaifang Cidian (http://kaifangcidian.com).

Overall, our mappings cover 35,196 distinct Mandarin words. Out-of-vocabulary Mandarin words are likely to be infrequently used words; these words, fortunately, tend to be rendered in the same way in Cantonese, and therefore our system leaves them unchanged in the target sentence.

3.2. Annotation on lexical mappings

A Mandarin word may have multiple possible Cantonese translations. This is often because the Mandarin word has multiple meanings, but may also be due to different levels of the register of the Cantonese target word. To guide our system in choosing the most appropriate mapping, we annotate the lexical mappings with the POS of the Mandarin word and the register level in Cantonese. We follow the tagset of the Penn Chinese Treebank (Xia, 2000) in the POS annotation. We label the register level of the Cantonese word, labeling as ‘low’, ‘high’, or ‘both’.

Table 1 shows several examples. The Mandarin word *ràng* 讓 can either mean ‘to give way’, or ‘to let’, both as a verb. In the former case, it has an identical Cantonese counterpart, *yeuhng* 讓 ‘to give way’; in the latter case, however, it must be translated as the Cantonese *dáng* 等 ‘to let’. The Mandarin word *de* 的 is also highly ambiguous. As a relativizer, it is tagged as “DEC” and its Cantonese equivalent is *ge* 嘅. As a sentence-final particle, it is tagged as “SP”, with its high-register translation as *ge* 嘅, but its low-register translation as *ga* 㗎. Table 2 shows an application of these mappings to translate a Mandarin sentence.

For the mappings of the 1000 most frequent Mandarin words, we manually annotated the POS and register information. In terms of POS, 32 Mandarin words required POS specification for semantic disambiguation. In terms of register, 174 Mandarin words have different high- and low- register translations into Cantonese.

Mandarin word	**Mandarin POS**	**Cantonese register**	**Cantonese word**
ràng 讓 'to give way'	VV	both	*yeuhng* 讓 'to give way'
ràng 讓 'to let'	VV	both	*dáng* 等 'to let'
de 的	DEC/DEG	both	*ge* 嘅
de 的	SP	high	*ge* 嘅
de 的	SP	low	*ga* 㗎

Table 1. Example lexical mappings from Mandarin to Cantonese, specified by Mandarin POS and Cantonese register.

English	'I (really) have meal first before doing homework!'
Source (Mandarin)	我是先喫了飯再做作業的。 *wǒ shì* ***xiān*** *chī le fàn zài zuò zuòyè* ***de*** . PN VC **AD** VV AS NN AD VV NN **SP** 1SG COP **first** eat PFV meal then do homework **SFP**
High-register target (Cantonese)	我係先食咗飯再做功課嘅。 *ngóh haih* ***sīn*** *sihk jó faahn joi jouh gūngfo* ***ge*** . PN VC **AD** VV AS NN AD VV NN **SP** 1SG COP **first** eat PFV rice then do homework **SFP**
Low-register target (Cantonese)	我係食咗飯先再做功課㗎。 *ngóh haih sihk jó faahn* ***sīn*** *joi jouh gūngfo* ***ga*** . PN VC VV AS NN **AD** AD VV NN **SP** 1SG COP eat PFV rice **first** then do homework **SFP**

Table 2. Application of the lexical mappings on Table 1 and syntactic transformation on Table 3 on an example Mandarin source sentence to generate its high-register and low-register Cantonese target sentence.

3.3. Syntactic transformations

In a comparative analysis of Cantonese and Mandarin, Ōuyáng (1993: 274) noted that although their "grammatical structure is similar in most major respects, the differences are not insignificant". These differences include the use of modal verbs and predicative adjectives; the expression of epistemicity and comparative construction; the word order in double object constructions; and the system of sentence-final particles, which is significantly more complicated in Cantonese. Further, in a quantitative comparison between Mandarin and Cantonese, Wong et al. (2017) showed that Mandarin adverbs are replaced by Cantonese auxiliaries in a number of cases.

Similar to Scherrer (2011), we express syntactic transformations as tree pairs. Rather than constituent trees, however, we used the Stanford dependencies for Chinese (Chang et al., 2009), and also annotated their register level. The system incorporates 10 such transformations, the most frequent of which are shown in Table 3.

Construction	Register	Mandarin	Cantonese
Adverb position	low	先 … <verb> advmod(<verb>, 先)	<verb> …先 discourse:sp(<verb>, 先)
	low	多/少/太/過 <adjective> advmod(<adjective>, 多/少/太/過)	<adjective> 多/少/得滯/過頭 advmod(<adjective>, 多/少/得滯/過頭)

	both	<verb> 不了/不得
		advmod(<verb>, 不了/不得)
			唔 <verb> 得
			advmod(<verb>,唔)
			advmod(<verb>,得)

Ditransitive construction	both	<verb> <noun$_i$> <noun$_d$> obj(<verb>, <noun$_d$>), obj(<verb>, <noun$_i$>)	<verb> <noun$_d$> <noun$_i$> obj(<verb>, <noun$_d$>), obj(<verb>, <noun$_i$>)

Comparative construction	low	比 <noun> <adjective>
		prep(<adjective>, 比)
		pobj(比, <noun>)
			<adjective> 過 <noun>
			prep(<adjective>, 過)
			pobj(過, <noun>)

Adverbs vs. auxiliaries	low	一向都 <verb> advmod(<verb>,一向都)	<verb> 開 aux(<verb>, 開)

Question construction	both	<verb> … 嗎？
		discourse:sp(<verb>,嗎)
			<verb$_i$> 唔 <verb$_j$>？
			conj(<verb$_i$>, <verb$_j$>)
			adv(<verb$_j$>,唔)

Table 3. Syntactic transformations from Mandarin to Cantonese.

4. Evaluation

For evaluation, we used two test sets that correspond to high- and low-register Cantonese. The "High Register" test set consists of speeches in the Hong Kong Legislative Council 12th October, 2017, and their official translation into Mandarin.[1] The set contains 176 sentences, with 29 as the median sentence length. The "Low Register" test set is extracted from a Cantonese movie produced by students as a term project at a university in Hong Kong, with their Mandarin subtitles. This set contains 391 sentences, with 6 as the median sentence length. We first report results with automatic evaluation measures (Section 4.1), and then discuss results from a human evaluation on both translation quality and register appropriateness (Section 4.2).

4.1. Automatic evaluation

As shown in Table 4, we evaluated two translation models. The "High Register" model excludes lexical mappings and syntactic transformations that are labelled as 'low', while the "Low Register" model excludes mappings and transformations that are labelled as 'high' (Section 3.2). We evaluated translation quality using the translation edit rate (Snover et al., 2006) (TER). TER is similar to the Word Error Rate, but also allows "shift" as an edit in addition to insertion, deletion, and substitution.

On both datasets, both the "High Register" and "Low Register" models outperform the "no change" baseline, *i.e.*, output the Mandarin source sentence as target sentence. The TER of this baseline is lower for the "High Register" dataset, confirming that high-register Cantonese more closely resembles Mandarin.

[1] Retrieved from http://webcast.legco.gov.hk/public/zh-hk .

On the low-register set, the "Low Register" model gave better performance than the "High register" model; and *vice versa* on the high-register set. This suggests that our lexical mappings succeed in capturing Cantonese usage at different registers. Overall performance is higher on the high-register set, reflecting the difficulty in predicting sentence-final particles, which are missing in the Mandarin source sentences and must be inserted more frequently in the low-register set. Elsewhere, disfluencies, false-starts, slangs, and English code-switching (Chan, 1998) also contributed some of the other errors.

Model	Low Register test set	High Register test set
Baseline (No change)	59.13%	49.40%
High Register	53.34%	**45.32%**
Low Register	**52.56%**	45.50%

Table 4. TER for Mandarin-to-Cantonese translation on two register levels. "High-register test set" include news, legco; "low-register test set" includes movies, drama

4.2. Human evaluation

To perform a more detailed analysis on translation quality and register appropriateness, we randomly selected 90 Mandarin sentences for human evaluation; 40% of these were drawn from the High Register test set, and 60% from the Low Register test set.

Translation quality. We translated these 90 sentences with the translation model with matching register (Section 4.1). We then asked two native speakers of Cantonese to judge the quality of the system output, rating each as "Wrong", "Fair", or "Good". "Wrong" means the MT output contains at least one obvious mistake; "Fair" means the output is technically correct but can be made more fluent. On average, 62% of the sentences were rated as "Good", 12% rated as "Fair", and 26% as "Wrong'. In some cases, inappropriate Cantonese words are chosen because of semantic ambiguity for the Mandarin word. In others, dependency parsing errors affected the detection of some of the syntactic transformations.

Register appropriateness. In a second evaluation, we generated low- and high-register Cantonese output from the same 90 pair of Mandarin sentences, and asked two native speakers of Cantonese to choose which one has lower register. On average, 86% of the human judgment corresponded with the system's.

5. Conclusion

We have presented an approach for translation between Chinese dialects, and implemented it for Mandarin-to-Cantonese translation. The MT system uses lexical mapping from Mandarin to Cantonese, coupled with syntactic transformations defined with dependency trees. A significant novelty is that the lexical mappings are register-sensitive. Automatic evaluation shows that translation models that match the required register of the target sentences yield better translation quality, and significantly outperformed a baseline. Further, in human evaluations, 62% of the sentences were rated as "good", and 86% of the system output matches human judgment on its level of register. In future work, we plan to apply this approach to other Chinese dialects, and to allow more fine-grained specification of Cantonese register.

Acknowledgement

This work is partially supported by Knowledge Transfer Initiatives in Business, Law and Liberal Arts and Social Sciences Disciplines (project #6354028), and by a Strategic Research Grant (project #7004494) at City University of Hong Kong.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Bauer, R. S. (1988). Written Cantonese of Hong Kong. *Cahiers de linguistique – Asie orientale*, 17(2):245−293.

Chan, B. H.-S. (1998). How does Cantonese-English code-mixing work? In *Language in Hong Kong at Century's End*, M. C. Pennington (ed.), pages 191−216, Hong Kong: Hong Kong University Press.

Chang, P.-C., Tseng, H., Jurafsky, D., and Manning, C. D. (2009). Discriminative Reordering with Chinese Grammatical Relations Features. In *Proceedings of SSST*.

Delmonte, R., Bristot, A., Tonelli, S., and Pianta, E. (2009). English/Veneto resource poor machine translation with STILVEN. In *Proceedings of ISMTCL*, Bulag 33:82−29, Besancon, France.

Halliday, M. and Hasan, R. (1989). Language, Context and Text: Aspects of Language in a Social-semiotic Perspective. Oxford University Press, Oxford, UK.

Hannas, W. C. (1997) *Asia's orthographic dilemma*. Honolulu: University of Hawaii Press.

Hardmeier, C. and Volk, M. (2009). Using Linguistic Annotations in Statistical Machine Translation of Film Subtitles. In *Proceedings of NODALIDA*.

Huang, G., Gorin, A., Gauvain, J.-L., and Lamel, L. (2016). Machine Translation Based Data Augmentation for Cantonese Keyword Spotting. In *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP)*.

Killingley, S. Y. (1993). *Cantonese*. Lincom Europa: Newcastle.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL*.

Lapshinova-Koltunski, E. and Vela, M. (2015). Measuring 'Registerness' in Human and Machine Translation: A Text Classification Approach. In *Proceedings of 2nd Workshop on Discourse in Machine Translation*.

Lee, J. (2011). Toward a Parallel Corpus of Spoken Cantonese and Written Chinese. In *Proceedings of IJCNLP*.

Lee, S. and Chan, K. K. W. (2017). From Cantonese L2 data to theoretical analysis: An Analysis of Num-Cl-N construction in Cantonese. In *Proceedings of 22nd International Conference on Yue Dialects*.

Levy, R. and Manning, C. D. (2003). Is it harder to parse Chinese, or the Chinese Treebank?. In *Proceedings of ACL*.

Li, D. C. S., Wong, C. S. P., Leung, W. M. and Wong, S. T. S. (2016). Facilitation of transference: The case of monosyllabic salience in Hong Kong Cantonese. *Linguistics*, 54(1):1−58.

Mair, V. H. (1991). What is a Chinese “dialect/topolect”? Reflections on some key Sino-English linguistic terms. *Sino-Platonic Papers*, 29: 1–31.

Matthews, S. and Yip, V. (2011). *Cantonese: A comprehensive grammar*. New York: Routledge.

Ōuyáng, J. (1993). *Pŭtōnghuà Guăngzhōuhuà de bĭjiào yŭ xuéxí* (The comparison and learning of Mandarin and Cantonese). Peking: China Social Science Press.

Prokopidis, P., Karra, V., Papagianopoulou, A., and Piperidis, S. (2008). Condensing Sentences for Subtitle Generation. In *Proceedings of Linguistic Resources and Evaluation Conference* (LREC).

Ramsey, S. R. (1987). *The Languages of China*. Princeton: Princeton University Press.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). A Comprehensive Grammar of the English Language. Longman, London, UK.

Scherrer, Y. (2011). Syntactic transformations for Swiss German dialects. In *Proceedings of EMNLP*.

Scherrer, Y. and Rambow, O. (2010). Natural Language Processing for the Swiss German Dialect Area. In *Proceedings of KONVENS*.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas* (AMTA).

Szeto, C. (2000). Testing Intelligibility among Sinitic Dialects. In *Proceedings of the Conference of the Australian Linguistic Society*.

Tadmor, U., Haspelmath, M. and Taylor, B. (2010). Borrowability and the notion of basic vocabulary. *Diachronica*, 27(2):226–246.

Vela, M. and Lapshinova-Koltunski, E. (2015). Register-based Machine Translation with Text Classification Techniques. In Proceedings of *MT Summit XV: MT Researchers' Track*.

Volk, M. and Harder, S. (2007). Evaluating MT with translations or translators. What is the difference? In *Proceedings of MT Summit XI*.

Wong, T.-S., Gerdes, K., Lee, J. and Leung, H. (2017). Quantitative Comparative Syntax on the Cantonese-Mandarin Parallel Dependency Treebank. In *Proceedings of International Conference on Dependency Linguistics*.

Xia, F. (2000). The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0). *IRCS Technical Reports Serie*s. 38.

Xu, P. and Fung, P. (2012). Cross-Lingual Language Modeling with Syntactic Reordering for Low-Resource Speech Recognition. In *Proceedings of EMNLP-CoNLL*.

Zeng, Z. (1986). *Colloquial Cantonese and Putonghua Equivalents*. Hong Kong: Joint Publisher.

Zhang, X. (1998). Dialect MT: A Case Study between Cantonese and Mandarin. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*.