Assessing the Discourse Factors that Influence the Quality of Machine Translation

Junyi Jessy Li
University of Pennsylvania
ljunyi@seas.upenn.edu

Marine Carpuat
National Research Council Canada
marine.carpuat@nrc.gc.ca

Ani Nenkova
University of Pennsylvania
nenkova@seas.upenn.edu

Abstract

We present a study of aspects of discourse structure — specifically discourse devices used to organize information in a sentence — that significantly impact the quality of machine translation. Our analysis is based on manual evaluations of translations of news from Chinese and Arabic to English. We find that there is a particularly strong mismatch in the notion of what constitutes a sentence in Chinese and English, which occurs often and is associated with significant degradation in translation quality. Also related to lower translation quality is the need to employ multiple explicit discourse connectives (*because, but*, etc.), as well as the presence of ambiguous discourse connectives in the English translation. Furthermore, the mismatches between discourse expressions across languages significantly impact translation quality.

1 Introduction

In this study we examine how the use of discourse devices to organize information in a sentence — and the mismatch in their usage across languages — influence machine translation (MT) quality. The goal is to identify discourse processing tasks with high potential for improving translation systems.

Historically MT researchers have focused their attention on the mismatch of linear realization of syntactic arguments (Galley et al., 2004; Collins et al., 2005), lexico-morphological mismatch (Minkov et al., 2007; Habash and Sadat, 2006) and word polysemy (Carpuat and Wu, 2007; Chan et al., 2007). Discourse structure has largely been considered irrelevant to MT, mostly due to the assumption that discourse analysis is needed to interpret multi-sentential text while statistical MT systems are trained to translate a single sentence in one language into a single sentence in another.

However, discourse devices are at play in the organization of information into complex sentences. The mere definition of sentence may differ across languages. Chinese for example is anecdotally known to allow for very long sentences which at times require the use of multiple English sentences to express the same content and preserve grammaticality. Similarly discourse connectives like *because, but, since* and *while* often relate information expressed in simple sentential clauses. There are a number of possible complications in translating these connectives: they may be ambiguous between possible senses, e.g., English *while* is ambiguous between COMPARISON and TEMPORAL; explicit discourse connectives may be translated into implicit discourse relations or translated in morphology rather than lexical items (Meyer and Webber, 2013; Meyer and Poláková, 2013).

In our work, we quantify the relationship between information packaging, discourse devices, and translation quality.

2 Data and experiment settings

We examine the quality of translations to English from Chinese and Arabic using Human-targeted Translation Edit Rates (HTER) (Snover et al., 2006), which roughly captures the minimal number of edits necessary to transform the system output into an acceptable English translation of the source sentence. By comparing MT output with post-edited references, HTER provides more reliable estimates of translation quality than using translated references, especially at the segment level. The data for the analysis is drawn from an extended set of newswire reports in the 2008/2010 NIST Metrics for Machine Translation

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 283–288,
Baltimore, Maryland, USA, June 23-25 2014.

GALE Evaluation set[1]. For Chinese, there are 305 sentences (segments) translated to English by three different translation systems. For Arabic, there are 363 Arabic sentences (segments) translated by two systems.

The presence of discourse devices is analyzed only on the English side: the reference, the system hypothesis and its edited translation. Discourse connectives and their senses are identified using existing tools developed for English. Beyond its practical limitations, analyzing the reference interestingly reflects the choices made by the human translator: whether to choose to use a discourse connective, or to insert one to make an implicit relation on the source side explicit on the target side.

We first conduct analysis of variance (ANOVA) with HTER as dependent variable and the discourse factors as independent variables, and systems as subjects. We examine within-subject significance in each ANOVA model. For discourse factors that are significant at the 95% confidence level or higher according to the ANOVA analysis, we provide detailed breakdown of the system HTER for each value of the discourse factor.

In this paper we do not compare the performance of individual systems, but instead seek to understand if a discourse phenomena is problematic across systems.[2]

3 Sentence length and HTER

The presence of complex discourse structure is likely to be associated with longer sentences. It stands to reason that long sentences will be harder to process automatically and this reasoning has motivated the first approaches to text simplification (Chandrasekar et al., 1996). So before turning to the analysis of discourse phenomena, we examine the correlation between translation quality and sentence length. A strong correlation between the two would call for revival of interest in text simplification where syntactically complex sentences are transformed into several shorter sentences as a preprocessing step.

We find however that no strong relationship exists between the two, as shown by the correlation coefficients between HTER values and the number of words in each segment in Table 1.

[1]Data used in this work includes more documents and the human edits not present in the official release.

[2]For the readers with keen interest in system comparison, we note that according to ANOVA none of the differences in system performance on this data is statistically significant.

Lan.	Sys1	Sys2	Sys3
ZH	0.097 (0.099)	0.117 (0.152)	0.144 (0.173)
AR	0.071(0.148)	-0.089 (-0.029)	-

Table 1: Pearson (Spearman) correlation coefficient between segment length and HTER values.

Next we examine if sentence–discourse divergence between languages and the presence of (ambiguous) discourse connectives would be more indicative of the expected translation quality.

4 When a sentence becomes discourse

Some languages allow more information to be packed into a single sentence than is possible in another language, making single-sentence translations cumbersome and often ungrammatical. Chinese is known for sentences of this kind; for example, the usage of punctuation is very different in Chinese in the sense that a comma can sometimes function as a full stop in English, motivating a series of disambiguation tasks (Jin et al., 2004; Xue and Yang, 2011; Xu and Li, 2013). Special handling of long Chinese sentences were also shown to improve machine translation (Jin and Liu, 2010; Yin et al., 2007).

To investigate the prevalence of sentences in the source language (Chinese and Arabic in our case) that do not confirm to the notion of sentence in the target language (English for the purposes of this study), we separate the translation segments in the source language into two classes: a source sentence is considered 1-1 if the reference translation consists of exactly one sentence, and 1-many if the reference contains more than one sentence.

For Chinese, 26.2% of the source segments are 1-many. These sentences tend to be much longer than average (36.6% of all words in all reference translations are part of such segments). For Arabic, the numbers are 15.2% and 26.3%, respectively. Below is an example of a 1-many Chinese segment, along with the human reference and its translation by one of the systems:

[**source**] 俄警方宣称，Erinys有一重要竞争对手RISC，利特维年科生前最后见面的人卢戈沃伊与友人都是从事这些行业。

[**ref**] Russian police claim that Erinys has an important competitor RISC. The last people Litvinenko saw while he was alive, Lugovoi and his friends, were all engaged in these industries.

[**sys**] Russian police have claimed that a major competitor, Litvinenko his last meeting with friends are engaged in these industries.

We conducted ANOVA on HTER, separately for each language, with type of segment (1-1 or

AOV	Arabic	Chinese
$Pr(> F)$	0.209	0.0045*

	1-1	1-many
System	HTER	HTER
ZH-Sys1	16.22	19.03*
ZH-Sys2	19.54	21.02
ZH-Sys3	20.64	23.86*

Table 2: ANOVA for both languages; average HTER for the three Chinese to English systems, stratified on type of segment (1-1 and 1-many). An (*) denotes significance at $p < 0.05$.

1-many) as the independent variable and systems treated as subjects. The test revealed that there is a significant difference in translation quality between 1-1 and 1-many segments for Chinese but not for Arabic. For the Chinese to English systems we further ran a Wilcoxon rank sum test to identify the statistical significance in performance for individual systems. For two of the three systems the difference is significant, as shown in Table 2.

We have now established that 1-many segments in Chinese to English translation are highly prevalent and their translations are of consistently lower quality compared to 1-1 segments. This finding suggests a cross language discourse analysis task of identifying Chinese sentences that cannot be translated into single English sentences. This task may be related to existing efforts in comma disambiguation in Chinese (Jin et al., 2004; Xue and Yang, 2011; Xu and Li, 2013) but the relationship between the two problems needs to be clarified in follow up work. Once 1-many segments are identified, source-side text simplification techniques may be developed (Siddharthan, 2006) to improve translation quality.

5 Explicit discourse relations

Explicit discourse relations such as COMPARISON, CONTINGENCY or TEMPORAL are signaled by an explicit connective, i.e., *however* or *because*. The Penn Discourse Treebank (PDTB) (Prasad et al., 2008) provides annotations for the arguments and relation senses of one hundred pre-selected discourse connectives over the news portion of the Penn Treebank corpus (Marcus et al., 1993). Based on the PDTB, accurate systems for explicit discourse relation identification have been developed (Pitler and Nenkova, 2009; Lin et al., 2014). The accuracy of these systems is 94% or higher, close to human performance on the task. Here we

AOV	Arabic	Chinese
$Pr(> F)$	0.39	0.0058*

		No Conn	> 1 Conn
all	% data (ZH)	53.77	15.08
1-many	% data (ZH)	13.77	5.25
		HTER mean	HTER mean
all	ZH-Sys1	16.11	19.84^{+}
	ZH-Sys2	19.96	22.39
	ZH-Sys3	20.70	25.00*
1-many	ZH-Sys1	16.94	22.75^{+}
	ZH-Sys2	20.47	23.25
	ZH-Sys3	22.30	29.68*

Table 3: Number of connectives: ANOVA for both languages; proportion of data in each factor level and average HTER for the three Chinese-English systems, of the entire dataset and of 1-many translations. An (*) or (+) sign denotes significance at 95% and 90% confidence levels, respectively.

study the influence of explicit discourse relations on machine translation quality and their interaction with 1-1 and 1-many segments.

5.1 Number of connectives

We identify discourse connectives and their senses (TEMPORAL, COMPARISON, CONTINGENCY or EXPANSION) in each reference segment using the system in Pitler and Nenkova (2009)[3]. We compare the translation quality obtained on segments with reference translation containing no discourse connective, exactly one discourse connective and more than one discourse connective.

The ANOVA indicates that the number of connectives is not a significant factor for Arabic translation, but significantly impacts Chinese translation quality. A closer inspection using Wilcoxon rank sum tests reveals that the difference in translation quality is statistically significant only between the groups of segments with no connective vs. those with more than one connective. Additionally, we ran Wilcoxon rank sum test over 1-1 and 1-many segments individually and find that the presence of discourse connectives is associated with worse quality only in the latter case. Effects above are illustrated in Table 3.

5.2 Ambiguity of connectives

A number of discourse connectives are ambiguous with respect to the discourse relation they convey. For example, *while* can signal either COMPARI-

[3]http://www.cis.upenn.edu/∼epitler/discourse.html; We used the Stanford Parser (Klein and Manning, 2003).

AOV	Arabic	Chinese
$Pr(> F)$	0.57	0.00014*

	has-amb-conn	no-amb-conn
System	HTER mean	HTER mean
ZH-Sys1	21.57	16.34*
ZH-Sys2	21.44	19.72
ZH-Sys3	27.47	20.69*

Table 4: ANOVA for both languages; average HTER for the three Chinese systems for segments with (11.80% of all data) and without an ambiguous connective in the reference translation. An (*) denotes significance at $p < 0.05$.

SON or TEMPORAL relations and *since* can signal either CONTINGENCY or TEMPORAL. In translation this becomes a problem when the ambiguity is present in one language but not in the other. In such cases the sense in source ought to be disambiguated before translation. Here we compare the translation quality of segments which contain ambiguous discourse connectives in the reference translation to those that do not. This analysis gives lower bound on the translation quality degradation associated with discourse phenomena as it does not capture problems arising from connective ambiguity on the source side.

We base our classification of discourse connectives into ambiguous or not according to the distribution of their senses in the PDTB. We call a connective ambiguous if its most frequent sense among COMPARISON, CONTINGENCY, EXPANSION, TEMPORAL accounts for less than 80% of occurrence of that connective in the PDTB. Nineteen connectives meet this criterion of ambiguity.[4]

In the ANOVA tests for each language, we compared the quality of segments which contained an ambiguous connective in the reference with those that do not, with systems treated as subjects. For Arabic the presence of ambiguous connective did not yield a statistically significant difference. The difference however was highly significant for Chinese, as shown in Table 4.

The finding that discourse connective ambiguity is associated with change in translation quality for Chinese but not for Arabic is rather interesting. It appears that the language pair in translation impacts the expected gains from discourse analysis on translation.

[4]The ambiguous connectives are: as, as if, as long as, as though, finally, if and when, in the end, in turn, lest, meanwhile, much as, neither...nor, now that, rather, since, ultimately, when, when and if, while

AOV	Event	Arabic	Chinese
$Pr(> F)$	Contingency	0.61	0.028*
	Comp.:Temp.	0.047*	0.0041*

Chinese	HTER	HTER
	Contingency	¬ Contingency
Sys1	20.15	16.72
Sys2	21.69	19.80
Sys3	25.87	21.16^{+}
	Comp.∧Temp.	¬(Comp.∧Temp.)
Sys1	23.58	16.64*
Sys2	26.16	19.63*
Sys3	27.20	21.21^{+}

Table 5: ANOVA for both languages; average HTER for Chinese sentences containing a CONTINGENCY relation (6.89% of all data) or both COMPARISON and TEMPORAL (4.59% of all data). An (*) or (+) sign denotes significance at 95% and 90% confidence levels, respectively.

5.3 Relation senses

Here we study whether discourse relations of specific senses pose more difficulties on translations than others and whether there are interactions between senses. In the ANOVA analysis we used a binary factor for each of the four possible senses. For example, we compare the translation quality of segments that contain COMPARISON relations in the reference translation with those that do not.

The relation sense makes a significant difference in translation quality for Chinese but not for Arabic. For Chinese specifically sentences that express CONTINGENCY relations have worse quality translations than sentences that do not express CONTINGENCY. One explanation for this tendency may be that CONTINGENCY in Chinese contains more ambiguity with other relations such as TEMPORAL, as tense is expressed lexically in Chinese (no morphological tense marking on verbs). Finally, the interaction between COMPARISON and TEMPORAL is significant for both languages.

Table 5 shows the effect of relation sense on HTER values for Chinese.

6 Human edits of discourse connectives

A relation expressed implicitly without a connective in one language may need to be explicit in another. Moreover, the expressions themselves are used differently; for example, the paired connective "虽然...但是" (despite...but) in Chinese should not be translated into two redundant connectives in English. It is also possible that the source language contains an explicit discourse

connective which is not translated in the target language, as has been quantitatively studied recently by Meyer and Webber (2013). An example from our dataset is shown below:

[source] 还有些人可到大学的游戏专业深造，而后被聘请为大游戏厂商的技术顾问等。

[ref] Still some others can receive further professional game training in universities and later(*Temporal*) be employed as technical consultants by large game manufacturers, etc.

[sys] Some people may go to the university games professional education, which is appointed as the big game manufacturers such as technical advisers.

[edited] Some people may go to university to receive professional game education, and later(*Temporal*) be appointed by the big game manufacturers as technical advisers.

The system fails to translate the discourse connective "而后" (later), leading to a probable misinterpretation between receiving education and being appointed as technical advisors.

Due to the lack of reliable tools and resources, we approximate mismatches between discourse expressions in the source and MT output using discourse-related edits. We identify explicit discourse connectives and their senses in the system translation and the human edited version of that translation. Then we consider the following mutually exclusive possibilities: *(i)* there are no discourse connectives in either the system output or the edit; *(ii)* the system output and its edited version contain exactly the same discourse connectives with the same senses; *(iii)* there is a discourse connective present in the system output but not in the edit or vice versa. In the ANOVA we use a factor with three levels corresponding to the three cases described above. The factor is significant for both Chinese and Arabic. In both languages, the mismatch case *(iii)* involves significantly higher HTER than either case *(i)* or *(ii)*. The human edit rate in the mismatch class is on average four points greater than that in the other classes.

Obviously, the mismatch in implicit/explicit expression of discourse relation is related to the first problem we studied, i.e., if the source segment is translated into one or multiple sentences in English, since discourse relations between adjacent sentences are more often implicit (than intra-sentence ones). For this reason we performed a Wilcoxon rank sum test for the translation quality of segments with discourse mismatch conditioned on whether the segment was 1-1 or 1-many. For both languages a significant difference was found for 1-1 sentences but not 1-many. Table 6 shows the proportion of data in each of the conditioned classes and the average HTER for sentences from the mismatch case *(iii)* where a discourse connective was edited and the others (no such edits). Translation quality degrades significantly for all systems for the mismatch case, over all data as well as 1-1 segments.

% data	Mismatch	Mismatch (1-1)	¬Mismatch (1-1)
Arabic	21.27	15.47	69.34
Chinese	29.51	17.05	56.82

AOV	Arabic	Chinese
$Pr(>F)$	4.0×10^{-6}*	4.1×10^{-11}*

	HTER ¬ Mismatch	HTER Mismatch
AR-Sys1	11.23	15.92*
AR-Sys2	11.64	15.74*
ZH-Sys1	15.57	20.72*
ZH-Sys2	19.02	22.34*
ZH-Sys3	11.64	15.74*
	¬ Mismatch 1-1	Mismatch 1-1
AR-Sys1	10.86	16.24*
AR-Sys2	11.58	16.65*
ZH-Sys1	15.47	19.13*
ZH-Sys2	18.68	22.52*
ZH-Sys3	19.57	26.07*

Table 6: Data portions, ANOVA for both languages and average HTER for segments where there is a discourse mismatch between system and edited translations. An (*) denotes significance at $p < 0.05$.

7 Conclusion

We showed that translation from Chinese to English is made more difficult by various discourse events such as the use of discourse connectives, the ambiguity of the connectives and the type of relations they signal. None of these discourse factors has a significant impact on translation quality from Arabic to English. Translation quality from both languages is adversely affected by translations of discourse relations expressed implicitly in one language but explicitly in the other or by paired connectives. Our experiments indicate that discourse usage may affect machine translation between some language pairs but not others, and for particular relations such as CONTINGENCY. Finally, we established the need to identify sentences in the source language that would be translated into multiple sentences in English. Especially in translating from Chinese to English, there is a large number of such sentences which are currently translated much worse than other sentences.

References

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 33–40.

Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th Conference on Computational Linguistics (COLING)*, pages 1041–1044.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 531–540.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 273–280.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Short Papers*, pages 49–52.

Yaohong Jin and Zhiying Liu. 2010. Improving Chinese-English patent machine translation using sentence segmentation. In *International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, pages 1–6.

Meixun Jin, Mi-Young Kim, Dongil Kim, and Jong-Hyeok Lee. 2004. Segmentation of Chinese long sentences using commas. In *Proceedings of the Third SIGHAN Workshop on Chinese Language Processing (SIGHAN)*, pages 1–8.

Dan Klein and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems*, volume 15.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184, 4.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics - Special issue on using large corpora*, 19(2):313–330.

Thomas Meyer and Lucie Poláková. 2013. Machine translation with many manually labeled discourse connectives. In *Proceedings of the Workshop on Discourse in Machine Translation (DiscoMT)*, pages 43–50.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation (DiscoMT)*, pages 19–26.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 128–135.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference: Short Papers*, pages 13–16.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Shengqin Xu and Peifeng Li. 2013. Recognizing Chinese elementary discourse unit on comma. In *International Conference on Asian Language Processing (IALP)*, pages 3–6.

Nianwen Xue and Yaqin Yang. 2011. Chinese sentence segmentation as comma classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT): Short Papers*, pages 631–635.

Dapeng Yin, F. Ren, Peilin Jiang, and S. Kuroiwa. 2007. Chinese complex long sentences processing method for Chinese-Japanese machine translation. In *International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, pages 170–175.