Crosslingual Annotation and Analysis of Implicit Discourse Connectives for Machine Translation

Frances Yung **Kevin Duh** **Yuji Matsumoto**
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0192 Japan
{pikyufrances-y, kevinduh, matsu}@is.naist.jp

Abstract

Usage of discourse connectives (DCs) differs across languages, thus addition and omission of connectives are common in translation. We investigate how implicit (omitted) DCs in the source text impacts various machine translation (MT) systems, and whether a discourse parser is needed as a preprocessor to explicitate implicit DCs. Based on the manual annotation and alignment of 7266 pairs of discourse relations in a Chinese-English translation corpus, we evaluate whether a preprocessing step that inserts explicit DCs at positions of implicit relations can improve MT.
Results show that, without modifying the translation model, explicitating implicit relations in the input source text has limited effect on MT evaluation scores. In addition, translation spotting analysis shows that it is crucial to identify DCs that should be explicitly translated in order to improve implicit-to-explicit DC translation.
On the other hand, further analysis reveals that the disambiguation as well as explicitation of implicit relations are subject to a certain level of optionality, suggesting the limitation to learn and evaluate this linguistic phenomenon using standard parallel corpora.

1 Introduction

Discourse relations are semantic and pragmatic relations between clauses or sentences. The relations can be explicitly expressed by surface words known as explicit 'discourse connectives' (DCs) or implicitly inferred. The markedness of discourse relations varies across languages. For example, Chinese discourse units are typically clauses separated by commas, so DCs are often implicit. Explicit and implicit DCs account for 45% and 40% of the DCs annotated in the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) respectively, while in the Chinese Discourse Treebank (CDTB), they account for 22% and 76% respectively (Zhou and Xue, 2015).

Comparing with other language pairs, such as Arabic and English, it is found that discourse factors impact machine translation quality more in Chinese-to-English translation, especially when translating discourse relations that are expressed implicitly in one language but explicitly in the other (Li et al., 2014).

When translating from Chinese to English, implicit DCs are explicitated when necessary. For example, a causal relation can be inferred between the 2 clauses of the Chinese sentence below. In the English translation, the 2 clauses should be connected by an explicit DC, such as 'thus'.

- ^{1}[出口快速增长]，(export grows rapidly) ^{2}[成为推动经济增长的重要力量。] (become important strength in promoting the economy to grow.)

An open question in discourse for SMT is how best to handle cases where DCs are implicit in the source (e.g. Chinese) but explicit in the target (e.g. English). In this paper, we investigate how implicit DCs are translated in a translation corpus, and if explicitating implicit DCs in the source can improve MT.

2 Related Work

In translation studies, explicitation of implicit DCs is observed in translations between European languages (Becher, 2011; Zuffery and Cartoni,

Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT), pages 142–152,
Lisbon, Portugal, 17 September 2015. ©2015 Association for Computational Linguistics.

2014). On the other hand, it is also reported that certain English explicit DCs are not translated explicitly in French or German (Meyer and Webber, 2013). We hypothesize that explicitation is more common in Chinese-to-English translation.

To incorporate DC translation in SMT, explicit DCs are annotated in French-English parallel corpus and classifiers are trained to disambiguate DC senses before SMT training (Meyer et al., 2011; Meyer and Popescu-Belis, 2012). Also, translation model based on Rhetorical Structure Theory (Mann and Thompson, 1986) styled discourse parse has been used in Chinese-English SMT (Tu et al., 2013). These works focus on explicit discourse relations.

Chinese sentences can be 'discourse-like', consisting of a sequence of discourse units. Syntactic parsing of Chinese complex sentences (CCS) (Zhou, 2004) covers certain intersentential discourse relations, including both explicit and implicit relations. Tu et al. (2014) presents a CCS-tree-to-string translation model in which translation rules and language model are conditioned by automatic CCS parse. Improved BLEU scores are reported, but it is not clear how much the translation of implicit DCs has been improved.

Sense classification of implicit DCs is a hard task (Lin et al., 2009; Pitler et al., 2009; Park and Cardi, 2012). Echihabi and Marcu (2002) remove DCs in texts to create pseudo implicit DCs training instances. More useful pseudo samples can be generated by classifying ommisable and non-ommisable explicit DCs (Rutherford and Xue, 2015). Concerning the options of explicit and implicit usage, Patterson and Kehler (2013) presents a model that accurately (86.6%) predicts the choice of using an explicit or implicit DC given the discourse sense. However, human performance of the task is only 66%, implying that both choices are acceptable in some cases.

3 Crosslingual manual alignment of DCs

To investigate how DCs are translated from Chinese to English, we manually align DCs in the source to their translations on a parallel corpus. The DCs are further annotated with their nature and senses. This section describes the strategy and findings of our annotation.

3.1 Annotation scheme

The parallel corpus comes from 325 newswire articles (2353 sentences) of the the Chinese Treebank and their English translation (Palmer et al., 2005; Bies et al., 2007)[1]. The annotation was carried out by 1 professional Chinese-English translator.

We use translation spotting technique (Meyer et al., 2011) to align the DCs crosslingually, considering both explicit and implicit DCs. Annotation is carried out on the raw texts. Readers are refered to Yung et al. (2015) for details concerning the Chinese side annotation, such as definition of discourse units and annotation policy for parallel connectives. The labels used in the crosslingual annotation are defined as follows:

- **Explicit DC**: An explicit DC is a lexical expression that connects two discourse units with a relation. We do not define a close set of explicit DCs to be annotated. The list is constructed in the course of annotation. We also do not limit the syntactic categories of the DCs. In total, 227 Chinese and 152 English DCs are identified. (See Table 2)

- **Implicit DC**: An implicit DC is an implied relation between two discourse units represented by a lexical expression, e.g. '*and*' for an expansion relation. Since texts are naturally coherent, we assume that two consecutive discourse units are always related by a relation. The list of DCs that is used to annotate implicit relation is the list of 'fine senses'. (see below)

- **Redundant**: The 'redundant' tag is used when it is not grammatically acceptable to insert an implicit DC. Typically, it is annotated on either side of a DC alignment. For example, either half of a pair of parallel Chinese DCs (e.g.'因为'*because*...'所以'*therefore*) is aligned to 'redundant', as it is not grammatical to use both DCs in English.

- **AltLex**: 'AltLex' refers to the 'Alternative lexicalization' of a discourse relation that cannot be isolated from context as an explicit DC, e.g. '*it was followed by*' for a *Temporal* relation. Prepositions that mark discourse

[1] Our annotation is independent of existing monolingual discourse annotation on the Chinese Treebank such as the CDTB(Zhou and Xue, 2015) and Li et al. (2014b)

relations are also labeled ‘AltLex’, such as ‘*through*’ for a *Contingency* relation. This label is defined on English side only.

- **Coarse sense**: We first group the DCs under the 4 top-level discourse senses defined in PDTB, namely *Expansion, Contingency, Comparison* and *Temporal.*

- **Fine sense**: The sense hierarchy of PDTB is always modified in comparable discourse corpora of different languages (Prasad et al., 2014). Instead of defining a list of senses that cover discourse relations of both languages, we group interchangeable explicit DCs under the same category, and the category serves as the ‘fine sense’ label. For example, ‘*besides*’ ,‘*moreover*’ and ‘*in addition*’ are all annotated with the fine sense ‘*in addition*’. Similar to DC identification, the list of fine senses is built in the course of annotation. In total, there are 74 Chinese and 75 English fine senses (See Table 2).

The discourse sense annotation and DC alignment are carried out at one pass by below procedure:

1. Explicit DCs are identified in the source Chinese sentence, and labeled with sense tags.

2. The English translation of the DC is spotted, aligned to the Chinese DC and labeled with sense tags.

3. If the Chinese DC is not translated to an English DC, the annotator first looks for ‘AltLex’. If no ‘AltLex’ can be identified, an implicit DC is inserted. If insertion is not grammatical, the DC is aligned to ‘redundant’.

4. On the Chinese side of the corpus, implicit DCs are inserted between two discourse units if they are not related by an explicit DC[2]. The implicit DC is aligned following the strategy in Step 3.

5. Any explicit DCs on the English side that are not aligned are identified. Further implicit DCs are inserted to the Chinese side for alignment. If insertion of implicit DCs is ungrammatical, they are aligned to ‘redundant’.

Each pair of aligned DCs are thus tagged with 8 labels. Some annotation examples are shown below.

Example 1

中国必须对国有企业进行改革, **[1]**加强本身的竞争力。
China must implement reforms on state-owned enterprises so as to **[1]** improve its own competitiveness. .

	Chinese	English
[1]nature:	implicit	explicit
actual DC:	*nil*	so as to
fine sense:	来	in order to
coarse sense:	*Contingency*	*Contingency*

Example 2

[1] 在投资项目上比上年减少四百四十四件,但 **[2]**投资金额却 **[3]**比上年加一点三亿多美元。
[1] The number of investment projects dropped by 444 as compared with last year, but **[2]** the value of investments **[3]** rose by more than 130 million as compared with last year.

	Chinese	English
[1]nature:	implicit	implicit
actual DC:	*nil*	*nil*
fine sense:	其实	in fact
coarse sense:	*Expansion*	*Expansion*
[2]nature:	explicit	explicit
actual DC:	但	but
fine sense:	但是	but
coarse sense:	*Comparison*	*Comparison*
[3]nature:	explicit	redundant
actual DC:	却	*nil*
fine sense:	却	*nil*
coarse sense:	*Comparison*	*nil*

3.2 How many DCs are identified?

In total, 7266 pairs of discourse relations are aligned. Table 1 shows the distribution of coarse DC senses (*Comparison* (COM), *Contingency* (CON), *Expansion* (EXP) and *Temporal* (TEM)).

Similar to the findings in PDTB and CDTB, there are more implicit DCs than explicit DCs on the Chinese side but they are of similar proportion in English. *Comparison*, *Contingency*, and *Expansion* relations are more often expressed by implicit DCs than explicit DCs in Chinese. On the other hand, *Contingency* and *Expansion* relations are more often expressed by implicit DCs than explicit DCs in English.

Similar tendency is found in the PDTB. In CDTB, among the 9 coarse senses, *Causation*, *Entailment*, *Expansion* and *Conjunction* relations are more often implicit than explicit.

Table 2 shows the number of unique DCs and

[2] We treat each component of a paired DC independently: when only half of a paired DC occurs explicitly, the other half is inserted as an implicit DC.

Chi.	Explicit	Implicit		Total
COM	248 (36%)	446 (64%)		694 (9.9%)
CON	379 (20%)	1551(80%)		1930 (27.5%)
EXP	683 (18%)	3022(82%)		3705 (52.8%)
TEM	522 (76%)	165 (24%)		687 (9.8%)
Total	1832(26%)	5184(74%)		7016

Eng.	Explicit	Implicit	AltLex	Total
COM	287 (51%)	274 (48%)	6 (1%)	567 (9.3%)
CON	308 (25%)	584 (47%)	338(27%)	1230 (20.3%)
EXP	1545(42%)	1927(52%)	218 (6%)	3690 (60.8%)
TEM	408 (70%)	108 (19%)	63 (11%)	579 (9.5%)
Total	2548(42%)	2893(48%)	625(10%)	6066

Table 1: Proportion of various DCs per coarse sense. On top of above, there are 250 Chinese and 1200 English 'redundant' cases

fine senses that are identified in the annotation process. A smaller variety of DCs are used in the English translation than the Chinese source. The number of fine senses recognized in implicit DCs is smaller than that of explicit DCs, implying that some fine senses are only expressed explicitly.

Exp.	COM	CON	EXP	TEM	Total
Chi.	30(11)	63(18)	72(26)	62(19)	227(74)
Eng.	20(11)	41(13)	55(23)	40(14)	156(61)
Imp.	COM	CON	EXP	TEM	Total
Chi.	−(9)	−(15)	−(17)	−(13)	−(54)
Eng.	−(7)	−(11)	−(12)	−(9)	−(39)

Table 2: Number of unique DCs and DC fine senses (in brackets)[3]

Eng. / Chi.	Explicit	Implicit	Redun.	TTL
Explicit	1332	1193	23	2548
Implicit	81	2812	0	2893
Redund.	198	775	227	1200
AltLex	221	404	0	625
TTL	1832	5184	250	7266

Table 3: Number of alignments between discourse relations of different nature

Table 3 shows the number of alignments between discourse relations of different nature. Among the 5184 implicit DCs in Chinese, about 70% are not explicitly translated in English (2812 aligned to implicit DCs and 775 to 'redundant'). The rest 30% are translated to explicit DCs or other explicit lexicalization in English. We further examine the crosslingual alignment of discourse senses in Section 5.2.

Statistics of the annotated parallel corpus shows the divergence in DC usage between Chinese and English. It suggests that certain implicit Chinese DCs are explicitated in the English translation. To correctly model the translation of implicit relations, do we need a discourse parser that classifies an implicit source DC to its fine sense or coarse sense? Or will SMT robustly handle implicit-to-explicit DC translation without any discourse preprocessing? We seek to answer these questions in the next section.

4 Explicitating implicit DCs for MT based on manual annotation

With an automatic discoure parser, a discourse-tree-to-string translation model can be built. Nonetheless, state-of-the-art accuracy of implicit discourse sense classification is still low for downstream application (Rutherford and Xue, 2014). In this work, we design oracle experiments to evaluate the MT of implicit DCs assuming that the gold discourse sense is given.

4.1 Method

In our annotation scheme, implicit DCs senses are defined by DCs that are identified during explicit DC annotation. In other words, the implicit DCs are represented by explicit DC that acturally occur in Chinese discourse. We hypothize that explicitating implicit DCs in the source based on manual annotation will improve implicit-to-explicit DC translations and thus the overall MT result.

We use the annotated corpus as the *test set* for the MT experiments. The source input is prepro-

[3] DCs and fine senses that have multiple course senses are counted as different DCs/senses. If counted only once, the total numbers of unique DCs and DC fine senses (in brackets) are: explicit-Chinese: 200(70); explicit-English: 139(56); implicit-Chinese: (52); implicit-English: (38)

cessed based on the manual DC annotations. We compare a number of variations of the preprocess:

- **Implicit fine sense (FIN)**: We insert the annotated lexicalized fine sense to the source text. For example, referring to Example 2 in Section 3.1, '其实 ('in fact') ' is inserted at position **[1]** in the source sentence.
- **Implicit coarse sense (COA)**: Classification up to the coarse discourse sense could be helpful enough to translate the implicit DCs. We insert the most frequent fine sense of the annotated coarse sense to the source text[4]. Referring to the same example, '而且' ('and') is inserted at position **[1]** because it is the most frequent fine sense under the coarse sense *Expansion*.
- **Most explicitated DCs (TOP)**: According to findings in translation studies, explicitation of DCs is DC-dependent (Zuffery and Cartoni, 2014). We thus preprocess the input source text by explicitating only the N most frequently explicitated implicit DCs (implicit in source but explicit in target) according to the manual annotation[5]. Referring to the same example, no DC is inserted at position **[1]** because the annotated fine sense '其实' ('in fact') is not within the top 4.
- **Same DC for all implicit relations (SAM)**: To evaluate the effect of inserting explicit DCs to the source text independent of the discourse sense, we homogenously insert the most frequently explicited DC, '而且' ('and'), to all positions where an implicit DC is annotated in the source text. Therefore, '而且' is inserted to position **[1]** of both Example 1 and Example 2 under this setting.

We compare the 4 kinds of preprocessing (FIN, COA, TOP, SAM) to see what kind of explication of implicit DCs could improve MT. For each of the 4 kinds of preprocessing, we also experimented with an additional variant 'implicit-to-explicit only' (i2e), which restrictively explicitate only those DCs that are actually aligned to explicit target DCs. This is to evaluate the importance of identifying which implicit DC has to be explicitly translated. Referring to Example 2, no DC is inserted to position **[1]** since it is not an 'implicit-to-implicit' alignment. These various versions of source texts are decoded by SMT systems.

4.2 MT Settings

We train baseline MT systems with 2.5 million sentences of bitexts through the LDC[6], including newswire, broadcast news and law genres. To see if there is any bias of DC translation to certain framework, we build 3 types of SMT systems with default settings: a phrase-based model and a hierarchical model using MOSES (Koehn et al., 2007), and a tree-to-string model using TRAVATAR (Neubig, 2013). All models use a 5-gram language model trained on the English Gigaword (Parker et al., 2011) and are tuned by MERT (Och, 2003). We use GIZA^{++} (Och and Ney, 2003) for automatic word alignment and the Stanford Parser (Levy and Manning, 2003) to parse the source text for tree-to-string MT training. Tuning and testing with the newswire portions of OpenMT08 and OpenMT06 respectively, the phrase-based, Hiero and tree-to-string systems yield BLEU scores of 26.7, 26.1 and 20.4 respectively, evaluating against 4 reference translations.

We use these SMT models to translate the source text in which implicit DCs are explicitated by the methods described in Section 4.1. 1178 sentences and 1175 sentences of the manually annotated parallel corpus are used as the tuning and test sets respectively. The systems are tuned with the tuning set preprocessed by the **FIN** method.

Note that the SMT training data is not discourse annotated and thus the translation models are not trained with any discourse markups. Nonetheless, the source side of the training data contains abundant examples of both implicit and explicit DCs and we believe that the translation model will contain translation rules for both natures. The question is whether explicitating implicit DC senses in the source input will the improve final performance.

[4] The top frequent DCs per coarse sense for *Expansion, Comparison, Contingency* and *Temporal* relations are '而且' ('and'), '但' ('but'), '然后' ('then'), and '从而' ('thus') respectively.

[5] We use the 4 most often explicitated fine senses, which are '而且' ('and'), '而' ('whearas'), '和' ('and'), '并' ('also').

[6] LDC2004T08, LDC2005E47, LDC2005T06, LDC2007T23, LDC2008T08, LDC2008T18, LDC2012T16, LDC2012T20, LDC2014T04, LDC2014T11, LDC2014T15

4.3 Result

Figure 4 shows the BLEU and METEOR scores of the SMT outputs resulting from various preprocessed test sets. Explicitation of implicit DCs in the source input generally results in evaluation scores comparable to that of the unprocessed input. Similar results are produced by the 3 SMT frameworks. Only the **SAM** preprocess results in higher evaluation scores using Hiero SMT.

To our surprise, disambiguating the implicit discourse sense up to the fine sense does not yeild better translation comparing with disambiguation up to the coarse sense. In turn, homogenously inserting '而且' ('and') without sense disambiguation yeilds even better result. Similar scores are produced by explicitating only the most frequently explicitated implicit DCs. The 'implicit-to-explicit only' restriction generally produces higher scores, suggesting that it is crucial to identify which DCs should be explicitated in translation and which should not.

Results of the oracle MT experiment show that MT performance is hardly improved by explicitating implicit DCs even based on manual annotation. It will be more difficult to improve MT based on predicted implicit discourse senses.

	PBMT		Hiero		T2S	
	B	M	B	M	B	M
original	**15.6**	**24.5**	15.6	24.4	**12.6**	**22.7**
FIN	15.5	24.4	15.3	24.4	12.3	22.6
FIN+i2e	**15.6**	24.4	15.6	24.4	12.4	22.6
COA	15.4	**24.5**	15.4	24.4	12.4	**22.7**
COA+i2e	15.5	24.4	15.5	24.4	12.5	22.6
TOP	**15.6**	**24.5**	15.6	24.5	12.5	22.6
TOP+i2e	**15.6**	24.4	15.6	24.4	12.5	**22.7**
SAM	15.4	**24.5**	**15.7**	**24.6**	12.4	**22.7**
SAM+i2e	15.5	24.4	15.5	24.4	12.4	**22.7**

Table 4: BLEU (B) and METEOR (M) scores of MT outputs resulting from various DC insertions. Highest scores of each SMT system are bolded

5 Analysis

The negative MT results could be due to the following possibilities: (1) Improvement of DC translation is not captured by automatic evaluation scores. (2) The sense of the implicit DCs that requires explicitation is unevenly distributed, such that disambiguating the sense has limited effect. (3) The context in which a discourse relation is expressed explicitly in the source largely differs from the context in which it is expressed implicity. As a result, translation rules of actual explicit DCs cannot correctly translate artificially expliciated DCs.

We analyze these possibilities in this section.

5.1 Is the translation of implicit-to-explicit DCs improved?

Since DCs contribute to a small portion of word counts in the MT output, the difference in DC translation is not sensitive to global n-gram-based evaluation metrics. Translation of DCs can be actually improved while BLEU scores remain similar (Meyer et al., 2012).

We manually analyze 100 sentences of the baseline Hiero output, the reference translation, as well as the Hiero MT outputs produced by the preprocesses TOP and TOP with 'i2e' restriction. It is done by spotting how each implicit source DC is translated - to which explicit DC or not translated as explicit DC. Table 5 shows the proportion of different DC alignments produced by different MT systems and the reference translation.

(1)	**implicit-to-explicit rate**			
Ref.	19%			
Original	23%			
TOP	73%			
TOP+i2e	33%			
(2)	**correct**		**incorrect**	
Original	22%		78%	
TOP	23%		77%	
TOP+i2e	48%		52%	
(3)	**insert=explicit**		**nil=explicit**	
TOP	90%		10%	
TOP+i2e	44%		56%	
(4)	**correct**	**incorrect**	**correct**	**incorrect**
TOP	25%	75%	6%	94%
TOP+i2e	97%	3%	9%	91%

Table 5: Comparison of implicit DC translations in different preprocessing schemes

Part (1) of Table 5 compares the rate in which implicit source DCs are explicitated in the translation outputs. As expected, more implicit DCs are translated explicitly in the output of the preprocessed source text than that of the original source text. However, the original output already explicitates more implicit DCs than the reference does.

Part (2) of the table shows how much of the target DCs aligned to (originally) implicit source DCs are correct translation. The explicit target DC is considered **correct** if it matches with the explicit DC in the reference translation, and **incorrect** if the explicit DC is different from the reference DC or the relation is not translated as an explicit DC in the reference. It is seen that the preprocess (23%) hardly improves the accuracy comparing with the original output (22%), unless we only explicitate source DCs that are known to be explicitly translated (48%).

Part (3) of the table shows how often explicitating source DCs actually produces explicit DC translations. '**insert=explicit**' means the target explicit DC is aligned to a source explicit DC inserted by preprocess. '**nil=explicit**' means the target explicit DC is not aligned to any source DCs (inserted or not). It is observed that implicit DCs are sometimes explicitly translated by the MT systems even without source explicitation, yet the translation accuracy is low, comparing with translation from explicitated source DCs, as shown in Part (4) of the table.

Result of this analysis supports our hypothesis that the improvement in implicit-to-explicit DC translation is not captured by MT evaluation metrics. Although the MT outputs under comparison have similar scores, implicit-to-explicit DC translation is improved under the TOP+i2e setting, but not under the other settings. In addition, the result suggests that certain implicit-to-explicit DC translation is captured by SMT even without source explicitation preprocessiing.

5.2 Which senses are more common in implicit-to-explicit aligments?

On average, 18.5 Chinese and 15.25 English fine senses are identified under each of the 4 coarse senses. Nonetheless, the oracle MT experiment suggests that classifying the implicit discourse senses more precisely does not improve MT more. A possible explanation is that the senses of implicit-to-explicit DCs only limit to a small set of senses that are already captured by coarse sense classification.

Among the 7266 aligned relations, there are 1193 implicit-explicit alignments (refer to Table 3). Table 6 shows the sense distribution of these pairs. While the sense distribution on the Chinese side is comparable to the overall sense distribution (refer to Table 1), over 80% of which are translated by explicit DCs that signal an *Expansion* sense. In fact, 88% of the implicit source DCs are aligned to the explicit target DC '*and*'.

Chi.	Chinese		English	
COM	131	11.0%	90	7.5%
CON	300	25.1%	109	9.1%
EXP	715	59.9%	958	80.3%
TEM	47	3.9%	36	3.0%
Total	1193		1193	

Table 6: Sense distribution of imp.-exp. DC

Table 7 lists the top 10 frequent implicit-explicit alignments. It shows that '*and*' is used to explicitate a range of discourse relations. On the other hand, although '*and*' ambiguously signal various senses, non-*Expansion* senses only occur marginally in PTDB, as shown in Table8. The distinct discrepancy suggests that DC usage differs between spontaneous writing and translation.

source implicit fine sense	target explicit DC	count	(coverage)
而且 'and'	and	203	(17%)
而 'whearas'	and	117	(15%)
和 'and'	and	139	(12%)
并 'also'	and	81	(11%)
从而 'thus'	and	61	(7%)
所以 'therefore'	and	46	(5%)
来 'in order to'	and	26	(4%)
因此 'therefore'	and	23	(3%)
然后 'and then'	and	18	(2%)
即 'which is'	and	18	(2%)

Table 7: Top 10 frequent imp.-exp. alignments

sense of explicit *'and'*		count	(coverage)
Conjunction	(expansion)	2543	(85%)
result	(contingency)	38	(1%)
Conjunction and *result*		138	(5%)
others		281	(9%)
sense of implicit *'and'*		count	(coverage)
Conjunction	(expansion)	891	(70%)
List	(expansion)	346	(27%)
others		35	(3%)

Table 8: Sense distribution of DC *'and'* in PDTB.

Analysis of the implicit-explicit alignments explains why more precise sense disambiguation of the source relations does not improve MT. It is because the reference translation uses *'and'* as the *'wild card'* to translate most implicit DCs 'explicitly', but without explicitating the discourse sense. This finding is similar to the analysis based on word-aligned Chinese-English translation corpus, which also reports that *'and'* is the most frequently added DC to the reference translation (Li et al., 2014a). Therefore, to improve implicit-to-explicit DC translation, an additional task should be defined to identify whether a source implicit DC is kept implicit, explicitly translated to an ambigous DC such as *'and'*, or explicitly translated to other unambiguous DCs.

Generally, it is pragmatically correct to use *'and'* to translate an implicit discourse relation, or to keep the relation implicit as in the source. Nonetheless, repetatively using this stragegy will result in excessively long sentences, as in the example below. In this case, insertion of explicit DCs to the target text is desirable, instead of duplicating the source writing style.

Source

^{1}[天津港保税区投入运行五年来，]^{2}[已建成了中国第一货物分拨中心，]^{3}[具备了口岸关的功能，]^{4}[开通了天津港保税区经西安、兰州到新疆阿拉山口口岸的铁路专用线；]^{5}[建立了一批集仓储、运输、销售于一体的大型物流配给中心，]^{6}[开办了铁路和国际集装箱多式联运，]^{7}[月接卸集装箱能力达六千标准箱；]^{8}[形成了七千门程控电话的装机能力，]^{9}[供电能力达二点五万千伏、日供水能力一万吨。]

Reference

^{1}[Since being put into operation five years ago,] ^{2}[the Tianjin Port Bonded Area has completed the construction of China's first goods distribution center,] ^{3}[functioned like a customs port,] ^{4}[opened up the special use the railway line from the Tianjin Port Bonded Area passing Xi'an and Lanzhou to arrive at Xinjiang's Allah Mountain pass customs port,]^{5}[established a number of large-scale materials circulation distribution and supply centers integrating storage, transportation and sales,] ^{6}[opened multiple railway and international container joint-operations] ^{7}[with a monthly loading and unloading capacity reaching 6,000 standard containers.] ^{8}[It has built up an installation capacity of 7,000 sets of program-controlled telephones,] ^{9}[with a power supply capacity of 25,000 kilovolts, and a daily water supply capacity of 10,000 tons.]

5.3 Contexts of explicit/implicit DC usage

Lastly, we compare the contexts in which a particular sense is expressed explicitly or implicitly in the source. If the contexts are distinctly different, it suggests that artificially explicitated source implicit DCs cannot be captured by a translation model trained only with naturally occuring explicit DCs.

In addition, we compare the contexts in which a source implicit DC is translated into an explicit DC or by other means (by implicit DC or alternative lexicalization). If the contexts are similar, it suggests that the translation strategy could be an option independent of the context.

Following Rutherford and Xue (2015), we define the context of a discourse relation as the unigram distribution of words in the 2 arguments connected by the relation. The context of a particular discourse usage is thus the sum of the unigram distributions of all discourse relations associated with that usage. We also use the Jensen-Shannon Divergence (JSD) to evaluate the similarity of the contextual distributions (Rutherford and Xue, 2015; Hutchinson, 2005; Lee, 2001). This metric compares 2 distributions with the average. If both distributions are close to the average, it means they are close to each other as well. The metric value ranges from 0 (identical) to $\ln 2$.

Table 9 shows the difference between the context of each source sense against the context of other senses, when the discourse relation is expressed implicitly (Column [1]) and explicitly (Column [2]). The difference suggests that implicit and explict DCs are used in different contexts, supporting our hypothesis. In particular, the difference between the context of each sense against others is smaller in implicit usage, thus making implicit relations harder to disambiguate.

Comparing with the difference in context between implicit and explicit usage (Column [3]), the context of source implicit relations that are explicitated in the target is similar to the context of source implicit relations that are kept implicit (Column [4]). This suggests that to explicitate the implicit DC or not in translation is independent of the local context to certain extent.

The example below shows the optionality of DC translation. It is taken from the test data of OpenMT 06. The implicit relations between the 3 discourse units in the source are translated by different DC usage in the target. For example, the re-

	$JSD(q,r)$			
	[1]	[2]	[3]	[4]
	1 sense	1 sense	exp	imp-imp
source	vs all	vs all	vs	vs
fine sense	imp	exp	imp	imp-exp
而且 ‘and’	.025	.149	.142	.059
而 ‘whearas’	.052	.111	.124	.076
和 ‘and’	.066	.166	.186	.106
并 ‘also’	.064	.052	.068	.110
从而 ‘thus’	.052	.182	.189	.094
所以 ‘therefore	.051	.238	.239	.142
来 ‘in order to	.053	.126	.124	.178
因此 ‘therefore’	.039	.164	.164	.119
然后 ‘and then’	.154	.286	.316	.218
即 ‘which is’	.131	.321	.393	.205

Table 9: Jensen-Shannon Divergence (JSD) of various discourse usage of the top imp-exp DCs

lation between Unit 1 and Unit 2 is translated to a *Temporal* DC *‘as’* in Reference 1, while translated to a *Contingency* DC *‘so that’* in Reference 3. In Reference 2, 4, it is kept implicit. This suggests that multiple reference are necessary for evaluation of DC translation.

Source:

^{1}[这厚重的历史回声,通过电视台”连线”大陆和香港],^{2}[南京市民与香港同胞”天涯共此时” ,] ^{3}[共同庆祝香港回归祖国十周年。]

Reference 1:

^{1}[This rich echo of history connected the mainland and Hong Kong via television,] ^{2}[***as*** the citizens of Nanjing and Hong Kong compatriots ”shared the same occasion from the far corners of the earth”]^3[***and*** celebrated together the tenth anniversary of Hong Kong’s reversion to the motherland.]

Reference 2

^{1}[This echo of profound historical significance ”connected” the Mainland and Hong Kong through television;]^2[citizens of Nanjing and their fellow countrymen in Hong Kong ”shared this moment with the entire world” together]^3[***celebrating*** the 10th anniversary of Hong Kong’s handover to the motherland]

Reference 3

^{1}[The sophisticated echo of history ”connected” the mainland and Hong Kong through a TV channel,] ^{2}[***so that*** Nanjing citizens and Hong Kong compatriots ”shared the moments across the land”]^3[***to*** celebrate together the 10th anniversary of Hong Kong’s return to the motherland.]

Reference 4

^{1}[The heavy historical echo ”connected” the Mainland with Hong Kong through television station.]^2[Residents of Nanjing shared the moment with Hong Kong compatriots from afar]^3[***to*** celebrate the 10th Anniversary of the return of Hong Kong to its motherland together.]

6 Conclusion

Motivated by the difference in DC usage between Chinese and English, we investigate the translation of implicit to explicit DCs given the gold crosslingual DC senses. We present a scheme to annotate and align DCs crosslingually and annotate 7266 relations in a Chinese-English translation corpus.

To simulate the incorporation of implicit DC information to MT, we explicitate the implicit DCs in the input source text based on annotation, and decode the preprocessed input by baseline, non-discourse-aware SMT models. Results show that artificially explicitating source implicit DCs in the input text alone does not improve the MT performance significantly.

Further analysis by translation spotting suggests that discourse usage as well as sense disambiguation can be subject to a certain level of optionality. In our annotated corpus, explicitation of implicit source DCs in translation is suppressed, either by traslation not using an explicit DC, or by translation using an ambiguous, sense-neutral explicit DC.

Nonetheless, our analysis is based on written-text in the news domain, while the discrepancy of Chinese-English DC usage is different in conversation dialogues and other domains (Steele and Specia, 2014). The suppression in explicitation of implicit DC could be due to the fact that subjective interpretation is avoided in news report. The future direction of our work is thus to exploit data from other domains, and to identify implicit DC relations that require explicitation in translation. The annotation used in this work is openly released on `http://cl.naist.jp/nldata/zhendisco`.

References

Viktor Becher. 2011. When and why do translators add connectives? a corpus-based study. *Target*, 23(1).

Ann Bies, Martha Palmer, Justin Mott, and Colin Warner. 2007. English chinese translation treebank v1.0 (ldc2007t02).

Ben Hutchinson. 2005. Modelling the similarity of discourse connectives. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Lillian Lee. 2001. On the effectiveness of the skew divergence for statistical language analysis. *Artificial Intelligence and Statistics*.

Roger Levy and Christopher Manning. 2003. Is it harder to parse chinese, or the chinese treebank. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. Assessing the discourse factors that influence the quality of machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014a. Cross-lingual discourse relation analysis: A corpus study and a semi-supervised classifacation system. *Proceedings of the International Conference on Computational Linguistics*.

Yancui Li, Wenhi Feng, Jing Sun, Fang Kong, and Guodong Zhou. 2014b. Building chinese discourse corpus with connective-driven dependency tree structure. *Proceedings of the Conference on Empirical Methods on Natural Language Processing*.

Ziheng Lin, Minyen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. *Proceedings of the Conference on Empirical Methods on Natural Language Processing*.

William C Mann and Sandra A Thompson. 1986. Rhetorical structure theory: Description and construction of text structures. Technical report, DTIC Document.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Thomas Meyer and Andrei Popescu-Belis. 2012. Using sense-labeled discourse connectives for statistical machine translation. *Proceedings of the Workshop on Hybrid Approaches to Machine Translation*.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. *Proceedings of the Discourse in Machine Translation Workshop*.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. 2011. Multilingual annotation and disambiguation of discourse connectives for machine translation. *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Thomas Meyer, Andrei Popescu-Belis, and Najeh Hajlaoui. 2012. Machine translation of labeled discourse connectives. *Proceedings of the Biennial Conference of the Association for Machine Translation in the Americas*.

Graham Neubig. 2013. Travatar: A forest-to-string machine translation engine based on tree transducers. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Demonstration Track)*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Martha Palmer, Fu-Dong Chiou, Nianwen Xue, and Tsan-Kuang Lee. 2005. Chinese treebank 5.0 (ldc2005t01).

Joonsuk Park and Claire Cardi. 2012. Improving implicit discourse relation recognition through feature set optimization. *Proceedings of Annual Meeting on Discourse and Dialogue*.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition ldc2011t07. Linguistic Data Consortium.

Gary Patterson and Andrew Kehler. 2013. Predicting the presence of discourse connectives. *Proceedings of the Conference on Empirical Methods on Natural Language Processing*.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*.

Rashmi Prasad, Nikhit Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. *Proceedings of the Language Resource and Evaluation Conference*.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the penn discourse treebank, comparable corpora, and complementary annotation. *Computational Linguistics*.

Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.

Attapol Rutherford and Nianwen Xue. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. *Proceedings of the North American Chapter of the Association of Computational Linguistics*.

David Steele and Lucia Specia. 2014. Divergences in the usage of discourse markers in english and mandarin chinese. *Text, Speech and Dialogue*.

Mei Tu, Yu Zhou, and Chengqing Zong. 2013. A novel translation framework based on rhetori- cal structure theory. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Mei Tu, Yu Zhou, and Chengqing Zong. 2014. Enhancing grammatical cohesion: Generating transitional expressions for smt. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Frances Yung, Kevin Duh, and Yuji Matsumoto. 2015. Sequential annotation and chunking of chinese discourse structure. *The SIGHAN Workshop on Chinese Language Processing*.

Yuping Zhou and Nianwen Xue. 2015. The chinese discourse treebank: a chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2):397–431.

Qiang Zhou. 2004. Annotation scheme for chinese treebank. *Journal of Chinese Information Processing*.

Sandrine Zuffery and Bruno Cartoni. 2014. A multifactorial analysis of explicitation in translation. *Target*, 26(3).