Baltic J. Modern Computing, Vol. 4 (2016), No. 2, pp. 305–317

Predicting and Using Implicit Discourse Elements in Chinese-English Translation

David STEELE, Lucia SPECIA

Department of Computer Science, The University of Sheffield, UK

dbsteele1@sheffield.ac.uk, l.specia@sheffield.ac.uk

Abstract. In machine translation (MT) implicitation can occur when elements such as discourse markers and pronouns are not expected or mandatory in the source language, but need to be realised in the target language for a coherent translation. These 'implicit' elements can be seen as both a barrier to MT and an important source of information. However, identifying where such elements are needed and producing them are non-trivial tasks. In this paper we examine the effect of implicit elements on MT and propose methods to identify and make them explicit. As a starting point, we use human translated and aligned data to decide where to insert place holders for these elements.

We then fully automate this process by devising a prediction model to decide if and where implicit elements should occur and be made explicit. Our experiments compare statistical machine translation models built with and without these explicitation processes. Models built on data marked for discourse elements show substantial improvements over the baseline.

Keywords: Chinese-English machine translation, discourse markers, empty categories, explicitation

1 Introduction

One of the main challenges in machine translation (MT) is to model the multitude of intrinsic differences that occur between the source and target languages. The problem is even more critical when considering distant language pairs such as Chinese[1] and English, which have largely developed in separation from each other and are noted to be markedly different (Wu, 2014). Chinese has a flexible grammar, relatively free word order (Gao, 2008), and is a prolific 'pro-drop' language (Huang, 1989). In addition, the application of cohesive devices (e.g. conjunctions) is one of the most prominent features that distinguishes Chinese and English (Wu, 2014). For instance, implicit links (i.e. the absence of explicit markers) are very common in Chinese and where a relation is not

[1] For this paper 'Chinese' is used to mean 'Mandarin Chinese'.

made explicit it can be inferred from context. However, in MT producing the correct explicit information on the target language when it is not required on the source language poses a significant barrier that often leads to poor quality automated translations.

Examples 1 and 2 (Steele and Specia, 2014) highlight this issue and show that during the translation process[2], even for relatively simple sentences, when elements are not required in the source, but need to be realised in the target then problems can and do occur.

Ex(1) 他因为病了，没来上课。
Because he was sick, he didn't come to class. (Human Translation)
He is ill, absent. (MT)

In the second clause of the Chinese sentence the pronoun 'he' (他) is inferred from the 'he' (他) in the first clause. In addition 'so/therefore' (所以), which commonly co-occurs with 'because'(因为) in the '因为... 所以...' construct, is optional (in this case) and also omitted. Consequently the translation system performs quite poorly.

Example 2 is a modified version of Example 1, with an extra 'so'(所以) and 'he' (他) manually inserted into the second clause of the Chinese sentence.

Ex(2) 他因为病了, 所以他没来上课。
Because he was sick, (so) he didn't come to class. (Human Translation)
Because he was ill, so he did not come to class. (MT)

Grammatically these extra inserted characters are not required in the Chinese , but inserting them has enabled the MT system to produce a better translation. This suggests that recovering such elements can help to produce a smoother translation, although this in turn may still need to be further refined.

In this paper, we examine some of the effects that implicit elements have on MT. We also implement methods for recovering some of the inferred information by inserting explicit place holder tokens into the source data to help inform the automatic alignment and decoding processes. We create an initial benchmark using human translations and oracle alignments (correct word alignments provided by human experts), which we then try to automate using a classifier to predict if and where to insert place holder tokens.

Our primary results show a significant improvement over the baseline models with no place holders for discourse elements (+1 in BLEU) and are close to those obtained with annotations derived from manually produced translations and alignments.

The remainder of this paper is organised as follows: In Section 2 we examine related work. Section 3 explains in detail how we built our benchmark corpus based on datasets translated and word-aligned by humans. Section 4 details our methods used for finding implicit elements and inserting place holder tokens into our data. We also discuss our initial work on building a prediction model. Our experiments, set-up and results are outlined and discussed in Section 5. Finally, Section 6 presents our conclusions and potential directions for future work.

[2] The 'MT' here is produced by Google Translate: https://translate.google.co.uk/

2 Related work

In this section we outline some approaches that have been used to deal with the topics of implicitation, explicitation, and empty categories in the context of MT. These are topics that have generated increasing interest for a number of languages in recent years.

In order to contend with the type of language phenomena highlighted in Examples 1 and 2, special empty category tokens have been used in the Penn Treebank (Bies et al., 1995) and its extension, the Chinese Treebank (Xue and Xia, 2000). An empty category is an element that does not have a mapping to a surface word in a parse tree. Essentially, when translating such elements into the target, where they are explicitly required, it is problematic because the implicit information has to be retained, recovered, and realised from what otherwise appears to be non-existent components in the source.

In Meyer and Webber (2013) implicitation of DMs in MT is explored through a detailed corpus analysis. The work highlights how DMs in the source text are not always translated to comparable words in the target language. Disparities in how often this phenomenon occurs in human translated texts (18%) for English, French, and German as opposed to machine translated ones (8%) are observed and the work aims to more widely capture the natural implicitation of DMs in statistical MT (SMT).

More specifically to Chinese, Chung and Gildea (2010) examine the effects that empty categories have on MT with a specific focus on dropped pronouns (little *pro*) and control constructions (big *PRO*) . The work shows that building machine translation systems with explicitly inserted empty elements, either manually or automatically, in the training data improves the overall translation quality. They use and compare three different approaches to recover empty or null elements: pattern matching; parsing; and prediction models. Of the three, the prediction model performed the best. However, they acknowledge that there is a lot of room for improvement in order to better recover empty categories.

In Yang and Xue (2010) the term ‘chasing the ghost’ is used to signify the hunt for empty categories. Identification of empty categories is turned into a tagging task. Essentially, each word in a sentence is given a tag indicating whether or not it follows an empty category. A maximum entropy model is employed for the prediction of the tags. No distinctions are made between the types of tags that are identified. The results show a 63% accuracy rate in recovering empty tags when an automatic parser is used as input.

Luo and Zhao (2011) also try to predict where empty categories may appear in Chinese sentences by using a statistical tree annotator supplemented with additional information. They apply the annotator to a few distinct tasks including: predicting function tags and predicting null elements. The results show favourable comparisons with previously published results using the same data. However, the results for predicting function tags and empty elements in the Chinese were obtained using human annotated data rather than automatically generated data. In addition, some of the empty categories are placed into a single position in the tree, which prevents them from being uniquely recoverable.

Instead of ‘chasing the ghost’ Xiang et al. (2013) outline work that ‘enlists the ghost’. They use a maximum entropy model with additional syntactic features to recover empty categories and then incorporate them into a Chinese-English MT task. The results show

that the recovered empty elements contribute to improvements in both word alignment tasks and the overall quality of their MT system output.

More recently, Steele and Specia (2014) discuss divergences in the usage of DMs for Chinese and English. They illustrate how DMs are vital contextual links, and through a detailed corpus analysis highlight significant divergences in their usage. The findings show how contextual omissions (implicit data) cause problems for MT systems and often lead to incoherent automatic translations.

Steele (2015), builds upon this work with a focus on word alignments for four specific elements: 'if', 'then','because','but'. Automatic alignments are used to ascertain the occurrence of implicit markers, which is found to be quite significant. Experiments show that when artificial tokens are inserted into the data, as a proxy for these markers, and the MT systems are rebuilt, there is a significant improvement over the baseline. However, to achieve the improvement the insertions of the markers were carried out using reference data.

Clearly there is some overlap between the terms 'empty categories' and 'implicit elements', but for this paper we use the latter to refer to, amongst other things, those elements with no corresponding word alignments. Our work detailed in this paper is more general as compared to previous work: is not restricted to big or little *pro* categories, and does not rely on treebank annotations, nor on parsing.

3 A benchmark corpus

Here we describe the pre-processing of a benchmark corpus using human translated and aligned data, which we then use to build and evaluate approaches to make discourse elements explicit.

3.1 The data

The data used to build our benchmark corpus came from sections of the Gale Project provided by the Linguistic Data Consortium (LDC)[3] catalogue[4]. Each section consists of manually translated sentences from news and web broadcasts and contains oracle (i.e. produced by expert linguists) word alignments, as well as additional annotations signalling items such as non translated elements and other metadata.

Our final corpus is made up of a total of 43693 usable parallel aligned sentences[5] consisting of approximately 1.23M English words and 955K Chinese words:

- GALE Chinese-English Word Alignment and Tagging – Broadcast Training Parts 1-4. Total = 19621 usable sentences.
- GALE Chinese-English Word Alignment and Tagging Training – Newswire and Web Parts 1-4. Total = 17966 usable sentences.
- GALE Chinese-English Parallel Aligned Treebank – Training. Total = 6106 usable sentences.

[3] https://catalog.ldc.upenn.edu/

[4] LDC2012T16, LDC2012T20, LDC2012T24, LDC2013T05, LDC2013T23, LDC2014T25, LDC2015T04, LDC2015T18, LDC2015T06

[5] Some sentences had no alignments and so were unusable and consequently removed.

3.2 Building the sentences

Example 3 shows a typical sentence in its original format. The Chinese is character segmented and the English is space delimited. The word alignments reflect the positions (indexes starting with 1) of source and target and contain additional annotations.

Ex(3)
(Sp1) 从 那 时 开 始 这 里 就 成 了 香 港 的 一 个 禁 区 。
<Sp1> Since then , this area has become a prohibited zone in Hong Kong .
19-15(FUN) 17-10(SEM) 7,8-5[DET],6(GIS) 9[COO]-(NTR) 12,13-13,14(SEM) 18-11(SEM) 14[DEP]-12(PDE) 2,5,6-2(FUN) 3,4-3(SEM) 10,11[TEN]-7[TEN],8(GIS) 15,16[MEA]-9(GIF) -4[COO](NTR) -1[MET](MTA) 1[MET]-(MTA)

The separate parts are combined to create a parallel aligned sentence for each line of our corpus. The Chinese segments were segmented into their more common forms typically found in Chinese dictionaries. For instance, '这 里', becomes '这里' ('this area' - in the case of this sentence). This step was performed using the Stanford Chinese Segmenter (Chang, et al., 2008; Chang, et al., 2009; Tseng,et al., 2005). The word alignments were then adjusted to accommodate the changes.

The final stage of the process involves removing the meta-data/additional annotations and zero indexing the word alignments (to match other common word alignment formats). Multiple alignments are split into separate alignment points and then reordered to improve readability. Example 4 is the final version of Example 3 and shows the typical format of the sentences in our corpus.

Ex(4)
从 那时 开始 这里 就 成 了 香港 的 一 个 禁区 。 ||| since then , this area has become a prohibited zone in hong kong . ||| 0-0 1-1 2-0 3-3 3-4 5-5 5-6 6-5 6-6 7-11 7-12 8-10 9-7 10-7 11-8 11-9 12-13

4 Explicitation methods

In this section we first outline the process of recovering the implicit information and inserting tokens into our corpus using a heuristic method based on word alignment information. The main goal of such a method is to produce training data for a fully automated method. We then outline our initial fully automated method to predict implicit elements in the data without resorting to word alignment information.

4.1 A heuristic method to recover implicit elements

Here we outline the method of using data from a parallel corpus to identify and target the missing elements. This method relies on word alignments (oracle or automated) to gain knowledge of where the unaligned elements occur in the corpus. This method is suitable for building training corpora, for gaining insight on where implicitation may occur in the data, and for demonstrating the potential impact of making implicit information explicit.

However, it cannot be used in practice at decoding time, as translations for the test set (and thus word alignment information) will not be available (they will need to be generated).

To mark implicit elements, we first POS tag[6] the corpus. In this process sentence a) is transformed into sentence b), for example.

a) 自然 资源 相对 缺乏，
||| natural resources are relatively scarce .
||| 0-0 1-1 2-3 3-4 4-5

b) _ 自然 #NN _ 资源 #NN _ 相对 #AD _ 缺乏 #VV _，#PU
||| natural_JJ resources_NNS are_VBP relatively_RB scarce_JJ ._.
||| 0-0 1-1 2-3 3-4 4-5

The next step retrieves the positions of all the words on the English side that have no corresponding alignment on the Chinese side. In the case of sentence a) the word 'are' with index position 2 has not been aligned to any Chinese counterpart. This element is then tagged in one of a number of ways:

i) are_VBP[7] (both the word and its POS type)
ii) _VBP (a more general POS category token)
iii) <are> (just the word)
iv) <TOK> (a hold all general place holder token)

Once the element is tagged it is inserted into the Chinese segment. In order to do this, each side of the element is examined to find the nearest aligned English neighbour. The tagged element is then placed, as a token, next to the Chinese counterpart of said neighbour. If both neighbours are equally close, the left neighbour is given preference.

In this case 'resources' (position 1) and 'relatively' (position 3) are aligned to '资源' (resources) and '相对' (relatively), respectively, on the Chinese side. The token is hence inserted immediately after the alignment point for its left neighbour ('资源', resources). Everything to the right of the insertion then moves over. Sentence c) shows the final result after the insertion of the token using the markup in i):

c) 自然 资源 are_VBP 相对 缺乏，||| natural resources are relatively scarce .

A quick check, done in the same way as Examples 1 and 2 (Section 1), showed that inserting the word 'are' (with the POS tag removed) into the sentence improves the automated translation[8] (Example 5 is the original, and Example 6 is the modified sentence).

Ex(5) 自然资源相对缺乏，

[6] For all POS tagging tasks (Chinese and English) we use the Stanford Log-Linear Part-Of-Speech Tagger (Toutanova et al., 2003).
[7] _VBP = Verb, non-3rd person singular present.
[8] Google Translate (https://translate.google.co.uk/)

natural resources are relatively scarce . (Human Translation)
Relative lack of natural resources, (MT)

Ex(6) 自然资源 are 相对缺乏，(token replaced with the relevant word)
natural resources are relatively scarce . (Human Translation)
Natural resources are relatively scarce, (MT)

Choosing insertions Depending on test criteria there are a number of options to consider when making the insertions. Firstly, a choice has to be made as to what POS types to make insertions for (hence the POS tagging step). We can choose to make insertions for every element with no alignment or we can, for example, exclude certain elements, such as punctuation.

For this paper we chose to include a specific subset of elements based on the discussion in Section 1. Table 1 shows a representative list of the POS tagged elements, with their corresponding Penn Treebank descriptions, and POS category frequency counts, that we used in our experiments (Section 5). The first three rows relate to our primary focus on DMs and pronouns, whereas the final row includes two elements that are often linked with the pronouns (e.g. 'it's' in Ex 7, Section 5.3). The final decision on which elements to include was made based upon frequency counts.

Table 1. The words and POS elements used in our experiments (Section 5).

POS description	*Word coupled with POS tag*	*Frequency*
Coordinating conjunctions	and_CC, or_CC, but_CC	13373
Personal pronouns	it_PRP, you_PRP, they_PRP, he_PRP, she_PRP	9672
Subordinating conjunction	if_IN, because_IN	1037
Verb singular present	's_VBZ (3rd person), are_VBP (non-3rd person)	915, 1711

With different corpora and languages it may be necessary to experiment with the number and type of tags to include. Some tags may be aligned more often in one language, but less often in another. In addition, the method or software used to process the word alignments may also give different results. For our training split of the dataset, using the oracle word alignments, 'and_CC' was inserted 12350 times across 41693 sentences, whilst 'or_CC' was only inserted 554 times. Insertions were made for the POS groups in Table 1 in over 26000 of the sentences.

Thus far, the focus has been on insertions being made using oracle word alignments. However, we also experimented with automated word alignments, where we created an equivalent corpus using the same insertion rules, but with inserts made based on alignments extracted by Fast-Align (Dyer et al., 2013). Counts for insertions made on the same corpus, but using Fast-Align alignments, vary considerably. For example, 'and_CC' has 5015 insertions (previously 12350) whilst 'or_CC' has 95 (previously 554). Table 2 shows the difference in frequency of insertions made for 'and_CC', 'or_CC', and 'the_DT' using the oracle alignments and automated alignments, respectively.

Table 2. Highlighting the differences between insertions of words based on oracle and automated alignments.

Word	*Oracle alignments*	*Automated alignments*
and_CC	12350	5015
or_CC	554	95
the_DT	1160	33751

The word 'the_DT' is included in Table 2 as an additional observation (to be explored in future work) because it does not have a direct equivalent in Chinese. In the GALE corpus 'the' is often merged with the noun it is restricting or modifying. For instance, '城市' ('city') is actually aligned to 'the city'. This method is formally applied to the function words: 'the', 'a', 'an', 'this', 'that' (Li et al., 2009).

Conversely, Fast-Align makes no such distinctions. As a result, when making insertions using the oracle alignments, insertions for 'the_DT' were made 1160 times. When performing insertions on the same data, but using Fast-Align alignments, insertions for 'the_DT' were made on 33751 occasions (roughly 29 times as many). This highlights yet another difficulty for automatic word alignment tools.

4.2 A method to predict implicit elements

Section 4.1 showed that by using heuristics based on word alignments we can locate specific unaligned elements in a sentence. These methods cannot be used for unseen test data at translation time. Our ultimate aim is to use our data with insertions made using this method to train a classifier that predicts whether or not an insertion should occur after a word in a given Chinese sentence, without resorting to any information on the English side.

For our initial model we use CRFsuite (Okazaki, 2007) and our training set of 41693 sentences (annotated automatically with insertions) to build a prediction model, treating the problem as a sequence labelling task. The test set is made up of 1000 sentences (annotated in the same way for evaluation purposes) that do not appear in the training set. Individual sentences are converted into a sequence of tokens (one per line) and each is attached to its POS category and a label signalling whether it precedes an insert or not. As an example the first word from the sentence discussed in Section 4.1 would be placed into a file like so: _ 自然 #NN #NN NON .

A template file is then used to describe each word with a number of features. For our initial experiments, the following simple features for each word in the sentence were extracted:

- the preceding two individual words and the following two individual words
- a bigram including the word itself and the word immediately to the left
- a bigram including the word itself and the word immediately to the right
- POS tags for the preceding two words and the following two words
- POS bigrams for the preceding two words through to the following two words
- POS trigrams for the preceding two words (and the word itself) through to the following two words (e.g. POS[-2] | POS[-1] | POS[0] = #DT | #LB | #NN).

The performance of the model is measured using precision and recall. For our test set that contains 598 insertions, our model labelled 123 (21%) elements as 'PRE' (preceding an insert), with a precision of 84%.

As these are early tests the results are promising, but our future work will need to address two main issues: Firstly, we are currently only predicting 21% of the implicit elements and it is anticipated that experimenting with feature extraction will yield better results. Secondly, we are currently only tagging whether a word precedes an implicit element or not (i.e. no distinction between words). We are not currently recovering any other specific, fine grained information.

However, experimentation has shown that even only having a single catch all place holder token (e.g. <TOK>) inserted into the data can still positively affect the alignment and decoding processes (Table 3, Section 5).

5 Experiments with SMT

In this section we present experiments using our corpora annotated as per Section 4 to build SMT systems. The overall aim is to compare SMT systems built and tested with raw parallel data against SMT systems where the source side of the corpus is annotated with place holder information. The corpus annotations are derived from either oracle or automated word alignments. Predicted annotations (Predicted_Inserts) in the source of the test set are produced through a fully automated process, using a classifier trained on oracle alignments. This section also provides a number of examples highlighting how some translations have changed either for better or worse.

5.1 Settings and methodology

Our SMT systems are built using the corpus described in Section 3. CDEC (Dyer et al., 2010) is used for rule extraction and decoding following the hierarchical phrase-based approach (Chiang, 2007) for Chinese-English translation. We use BLEU (Papineni et al., 2002) as the metric to evaluate the systems. For consistency, default parameters are used during different builds with the only change being the source of the word alignments.

We perform the same experiments twice, with two different splits of the corpus. For each experiment, the corpus is first randomly shuffled. The development and test sets (dev and tst in the table) are then created using the first 2000 sentences (1000 for each) in the shuffled corpus, while the training set is made up of the remaining 41693 sentences. For an oracle build, all sets (dev, tst, and training) include the human created alignments, whereas for the full automated build, Fast-Alignment (FA) alignments are used. Once the alignment points are added to the sets, each individual sentence has the format shown in Example 4 (Section 3.2).

Each experiment consists of five builds:

- **Oracle**: an SMT system built using oracle alignments (no insertions).
- **Baseline_FA**: a baseline SMT system using Fast-Align (FA) (no insertions).
- **Oracle+Inserts**: an oracle SMT system with insertions made using heuristics based on oracle alignments.

- **FA+Inserts**: an automated SMT system with insertions made using heuristics based on Fast-Align alignments.
- **Predicted_Inserts**: an SMT system with insertions made by a classifier using oracle alignments for training the classifier.

Two scores are produced for each of the five builds per experiment, one for the development set and one for the test set.

5.2 Results

Table 3 shows BLEU scores for the different experiments with the two different splits of the data. In all cases our experiments have shown that having insertions has a strong positive effect on the scores.

Table 3. Examples of the benchmark, baseline, and insertion scores.

(Experiment A)		(Experiment B)	
Build Type	*BLEU*	*Build Type*	*BLEU*
Oracle (dev)	17.81	Oracle (dev)	18.54
Oracle (tst)	18.34	Oracle(B) (tst)	18.59
Baseline_FA (dev)	16.59	Baseline_FA (dev)	17.16
Baseline_FA (tst)	16.76	Baseline_FA (tst)	17.02
Oracle+Inserts (dev)	18.62	Oracle+Inserts (dev)	19.37
Oracle+Inserts (tst)	19.11	Oracle+Inserts(tst)	19.38
FA+Inserts (dev)	17.80	FA+Inserts (dev)	18.08
FA+Inserts (tst)	18.00	FA+Inserts (tst)	18.40
Predicted_Inserts (dev)	16.95	Predicted_Inserts (dev)	17.42
Predicted_Inserts (tst)	17.20	Predicted_Inserts (tst)	17.41

As expected, out of all systems, the Oracle builds perform the best. However, the builds using inserts and Fast-Align (FA+Inserts) show a compelling improvement of up to 1.38 BLEU points over the baseline (Baseline_FA) on the test sets. Similarly, the Oracle builds with inserts (Oracle+Inserts) show a convincing improvement over the plain Oracle builds. More noteworthy is the fact that modest but credible improvements of up to 0.44 are made with our fully automated builds (Predicted_Inserts) over the baseline (Baseline_FA).

5.3 Going beyond BLEU scores

BLEU by itself does not give information about what improvements have been made and why, so here we provide some examples, taken from our translations, to show the changes. Upon manual inspection of our test translations we noted that a large negative factor was the abundance of out of vocabulary (OOV) words, a possible side effect of the limited sized corpus we used.

Each of the following examples are taken from translations produced by our SMT systems built using inserts based on Fast-Align (FA+Inserts) and have four distinct parts: i) the original sentence (source and target); ii) the source with (automatic) insertions (if any) that appear in the sentence; iii) our (FA+Inserts) system translation (with inserts in the source data); iv) baseline translation (no inserts in the source data). In each case the ideal output is for item iii) to be a good coherent sentence that closely maps to the target sentence in item i) and is smoother than the baseline translation shown in item iv).

Ex(7)

i 因为 便宜。|||| because it 's cheap .
ii 因为 it_PRP 's_VBZ 便宜。
iii because it 's cheaper .
iv because cheaper .

Ex(8)

i 就说 这个 人 长得 像 猴子。|||| say, this person looks like a monkey.
ii No direct insertions made in this sentence
iii that is to say this person looks like a monkey .
iv that is to say , who looks like a monkey .

Ex(9)

i 这 次 会谈 主要 讨论 三 国 在 经贸 文化 等 领域 的 合作，没有 涉及 历史 问题。|||| the meeting focused on the three nations ' cooperation in economy , trade and culture , and did not touch on any history problems .
ii 这 次 会谈 主要 讨论 三 国 在 经贸 and_CC 文化 等 领域 的 合作，and_CC 没有 涉及 历史 问题。
iii this meeting primarily discuss cooperation in the fields of economics and trade, culture, in the three countries , and there is no problem involved in history.
iv talks this time will primarily discuss cooperation in areas such as economics and trade , culture , the three countries have on the issue of history .

Ex(10)[9]

i 后来 又 说 学生 会 人 太少，没 精力。|||| later he said the student association had no energy due to a shortage of hands .
ii 后来 he_PRP 又 说 学生 会 人 太少，没 精力。
iii later , he also said that the student association people . no , energy .
iv later , people will also said that students 太少, i did n't energy .

Ex(11)

[9] The original Chinese sentence in example 8 does not contain the phrase 'shortage of hands' but rather: 人 (people) 太少 (too few)... An MT system will therefore struggle to produce the actual phrase 'shortage of hands'.

i 中朝 友谊 已经 成为 双方 共同 的 宝贵 财富。lll the friendship between china and north korea has become a precious treasure for the two sides .
ii 中朝 and_CC 友谊 已经 成为 双方 共同 的 宝贵 财富。
iii north korea and friendship has become the peoples of both sides together .
iv the friendship between china and north korea has become a precious wealth of both sides together .

In examples 7-10 the sentences translated using data containing inserts are generally much better than the baseline translations. Having inserts appears to affect the overall alignment and decoding process (e.g. weights), so even those sentences without inserts within the actual sentence boundary itself (example 8) often still show improvements.

Occasionally, having inserts did not help. In example 11, the baseline translation is clearly better. The insert appears to cause degradation, which could be attributed to conflict with the character pair '中朝' ('zhōng cháo'). By itself '中朝' already has the meaning 'China and North Korea', but the way it is written here is akin to 'sino-DPRK (Democratic People's Republic of Korea)'. That is, the common forms of each country (中国 - China, 北朝鲜 - North Korea) have been truncated and used in a specific (less common) way, which already carries the 'and' information within. Essentially, our insertion of 'and_CC', outside of this tight character pair, introduces extra complexity that is clearly difficult for the MT system to deal with.

6 Conclusions and future work

In this paper, we first presented an approach for locating and tagging implicit elements in a parallel aligned corpus. We applied this information to an insertion task that placed proxy tokens for implicit elements into the source data. The data was then used to train SMT systems that were stronger than our baselines.

The source data with the newly inserted elements was also used to train a binary classifier that ultimately was able to predict where implicit elements should occur on unseen data. The data was again used to train SMT systems and the results showed improvements over the baseline.

We faced a barrier with OOV words, which could perhaps be resolved by using a larger dataset. In addition, we observed a strong variance in how items such as function words are treated by oracle and automated alignments. Alignment software lacks the judgement factor of human translation and the gulf in the variance is something that needs to be addressed, or in the least, explored.

Future work will target improvements on our prediction method. We only experimented with a relatively simple set of features. We believe that improving the CRF template and using a wider array of pertinent features will significantly enhance the prediction model, particularly in terms of recall. This, in turn, should lead to further improvements in the quality of translations produced by our SMT systems.

References

Bies, A., Ferguson, M., Katz, K., and MacIntyre, R. (1995). *Bracketing Guidelines for Treebank II Style*. Penn Treebank Project.

Chang P.C., Gally, M., and Manning, C. (2008) *Optimizing Chinese Word Segmentation for Machine Translation Performance*. In ACL 2008 Third Workshop on Statistical Machine Translation.

Chang, P.C., Tseng, H., Jurafsky, D., and Manning, C.D. (2009). *Discriminative Reordering with Chinese Grammatical Relations Features*. Proc. Third Workshop on Syntax and Structure in Statistical Translation.

Chiang, D. (2007). *Hierarchical phrase-based translation*. Proc. ACL, 33(2), pp. 201–228.

Chung, T. and Gildea, D. (2010) *Effects of Empty Categories on Machine Translation*. Proc. Conference on Empirical Methods in Natural Language processing, pp. 636-645.

Dyer, C., Lopez, A., Ganitkevitch, J., Weese, J., Ture, F., Blunsom, P., Setiawan, H., Eidelman, V., and Resnik, P. (2010). *CDEC: A decoder, Alignment, and Learning Framework for Finite-state and Context-free Translation Models*. Proc. ACL, System Demonstrations, Uppsala, Sweden, pp. 7–12.

Dyer, C., Chahuneau, V., and Smith N.A. (2013) *A Simple, Fast and Effective Reparameterization of IBM Model 2*. Proc. NAACL, Atlanta, June 09-15.

Gao, Q. (2008). *Word Order in Mandarin: Reading and Speaking*. Proc. 20th North American Conference on Chinese Linguistics (NACCL-20) Conf., Ohio, USA.

Huang, J. (1989) *Pro-drop in Chinese a Generalized Control Approach*. In: Jaeggli, O and Safir, K. (editors) The NUll Subject Parameter, pp. 185-214.

Li, X., Ge, N., and Strassel, S. (2009) *Tagging Guidelines for Chinese-English Word Alignment - Version 1.0*. LDC

Luo, X. and Zhao, B. (2011) *A Statistical Tree Annotator and Its Applications*. Proc. 49th annual meeting ACL, pages 1230-1238, Portland, Organ, June 19-24.

Meyer T. and Webber B. (2013). *Implicitation of Discourse Connectives in (Machine) Translation*. Proc. 1st DiscoMT Workshop at ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics), Sofia, Bulgaria, pp. 19-26.

Okazaki, N. (2007). *CRFsuite: a fast implementation of Conditional Random Fields (CRFs),* `http://www.chokkan.org/software/crfsuite/`.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). *BLEU: a Method for Automatic Evaluation of Machine Translation*. Proc. 40th ACL, Philadelphia, PA, pp. 311-318.

Steele, D. and Specia, L. (2014) *Divergences in the Usage of Discourse Markers in English and Mandarin Chinese*. In: (TSD) Lecture Notes in Computer Science, 8655:189–200, Springer Berlin Heidelberg. pp. 189-200.

Steele, D. (2015) *Improving the Translation of Discourse Markers for Chinese into English*. Proc. Proceedings of NAACL-HLT (ACL) 2015 Student Research Workshop (SRW),Denver, Colorado, June 1st, pp. 311-318,

Toutanova, K., Klein, D., Manning, C., and Singer, Y. (2003). *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network*. Proc. NAACL-HLT, pp. 252-259.

Tseng, H., Chang P.C., Andrew, G., Jurafsky, D., and Manning, C. (2005) *A Conditional Random Field Word Segmenter*. In Fourth SIGHAN Workshop on Chinese Language Processing.

Wu, J. (2014) *Shifts of Cohesive Devices in English-Chinese Translation*. Proc. Theory and practice in Language Studies, Vol. 4, No. 8, Finland, pp. 1659-1664,

Xiang, B., Luo, X., and Zho, B. (2013) *Enlisting the Ghost: Modeling Empty Categories for Machine Translation*. Proc. 51st annual meeting of ACL, Bulgaria, pp. 822-831.

Xue, N. and Xia, F. (2000) *The Bracketing Guidelines for The Penn Chinese Treebank 3.0* IRCS-00-08, IRCS, University of Pennsylvania

Yang, Y. and Xue, N. (2010) *Chasing the Ghost: Recovering Empty Categories in the Chinese Treebank*. Proc. 23rd International Conference on Computational Linguistics, Beijing, China, pp. 1382-1390.

Received May 3, 2016 , accepted May 15, 2016