# Disambiguating “DE” for Chinese-English Machine Translation

**Pi-Chuan Chang**, **Dan Jurafsky**, and **Christopher D. Manning**
Computer Science Department, Stanford University
Stanford, CA 94305
pichuan,jurafsky,manning@stanford.edu

## Abstract

Linking constructions involving 的 (DE) are ubiquitous in Chinese, and can be translated into English in many different ways. This is a major source of machine translation error, even when syntax-sensitive translation models are used. This paper explores how getting more information about the syntactic, semantic, and discourse context of uses of 的 (DE) can facilitate producing an appropriate English translation strategy. We describe a finer-grained classification of 的 (DE) constructions in Chinese NPs, construct a corpus of annotated examples, and then train a log-linear classifier, which contains linguistically inspired features. We use the DE classifier to preprocess MT data by explicitly labeling 的 (DE) constructions, as well as reordering phrases, and show that our approach provides significant BLEU point gains on MT02 (+1.24), MT03 (+0.88) and MT05 (+1.49) on a phrased-based system. The improvement persists when a hierarchical reordering model is applied.

## 1 Introduction

Machine translation (MT) from Chinese to English has been a difficult problem: structural differences between Chinese and English, such as the different orderings of head nouns and relative clauses, cause BLEU scores to be consistently lower than for other difficult language pairs like Arabic-English. Many of these structural differences are related to the ubiquitous Chinese 的(DE) construction, used for a wide range of noun modification constructions (both single word and clausal) and other uses. Part of the solution to dealing with these ordering issues is hierarchical decoding, such as the Hiero system (Chiang, 2005), a method motivated by 的(DE) examples like the one in Figure 1. In this case, the translation goal is to rotate the noun head and the preceding relative clause around 的(DE), so that we can translate to “[one of few countries] 的 [have diplomatic relations with North Korea]”. Hiero can learn this kind of lexicalized synchronous grammar rule.

But use of hierarchical decoders has not solved the DE construction translation problem. We analyzed the errors of three state-of-the-art systems (the 3 DARPA GALE phase 2 teams’ systems), and even though all three use some kind of hierarchical system, we found many remaining errors related to reordering. One is shown here:

| | | | | |
|---|---|---|---|---|
| 当地 | 一所 | 名声不佳 | 的 | 中学 |
| local | a | bad reputation | DE | middle school |

Reference: ‘a local middle school with a bad reputation’
Team 1: ‘a bad reputation of the local secondary school’
Team 2: ‘the local a bad reputation secondary school’
Team 3: ‘a local stigma secondary schools’

None of the teams reordered “bad reputation” and “middle school” around the 的. We argue that this is because it is not sufficient to have a formalism which *supports* phrasal reordering, but it is also necessary to have sufficient linguistic modeling that the system *knows when and how much to rearrange*.

An alternative way of dealing with structural differences is to reorder source language sentences to minimize structural divergence with the target language, (Xia and McCord, 2004; Collins et al., 2005; Wang et al., 2007). For example Wang et al. (2007) introduced a set of rules to decide if a 的(DE) construction should be reordered or not before translating to English:

- For DNPs (consisting of“XP+DEG”):
  - Reorder if XP is PP or LCP;
  - Reorder if XP is a non-pronominal NP
- For CPs (typically formed by “IP+DEC”):
  - Reorder to align with the “that+clause” structure of English.

Although this and previous reordering work has led to significant improvements, errors still remain. Indeed, Wang et al. (2007) found that the precision of their NP rules is only about 54.6% on a small human-judged set.

One possible reason the 的(DE) construction remains unsolved is that previous work has paid insufficient attention to the many ways the 的(DE) construction can be translated and the rich structural cues to the translation. Wang et al. (2007), for example, characterized 的(DE) into only two

*Proceedings of the Fourth Workshop on Statistical Machine Translation* , pages 215–223,
Athens, Greece, 30 March – 31 March 2009. ©2009 Association for Computational Linguistics

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 澳洲 | 是 | 与 | 北韩 | 有 | 邦交 | 的 | 少数 | 国家 | 之一 | 。 |
| Aozhou | shi | yu | Beihan | you | bangjiao | DE | shaoshu | guojia | zhiyi | . |
| Australia | is | with | North Korea | have | diplomatic relations | that | few | countries | one of | . |

‘Australia is one of the few countries that have diplomatic relations with North Korea.’

Figure 1: An example of the DE construction from (Chiang, 2005)

classes. But our investigation shows that there are many strategies for translating Chinese [A 的 B] phrases into English, including the patterns in Table 1, only some involving reversal.

Notice that the presence of reordering is only one part of the rich structure of these examples. Some reorderings are relative clauses, while others involve prepositional phrases, but not all prepositional phrase uses involve reorderings. These examples suggest that capturing finer-grained translation patterns could help achieve higher accuracy both in reordering and in lexical choice.

In this work, we propose to use a statistical classifier trained on various features to predict for a given Chinese 的(DE) construction both whether it will reorder in English and which construction it will translate to in English. We suggest that the necessary classificatory features can be extracted from Chinese, rather than English. The 的(DE) in Chinese has a unified meaning of ‘noun modification’, and the choice of reordering and construction realization are mainly a consequence of facts of English noun modification. Nevertheless, most of the features that determine the choice of a felicitous translation are available in the Chinese source. Noun modification realization has been widely studied in English (e.g., (Rosenbach, 2003)), and many of the important determinative properties (e.g., topicality, animacy, prototypicality) can be detected working in the source language.

We first present some corpus analysis characterizing different DE constructions based on how they get translated into English (Section 2). We then train a classifier to label DEs into the 5 different categories that we define (Section 3). The fine-grained DEs, together with reordering, are then used as input to a statistical MT system (Section 4). We find that classifying DEs into finer-grained tokens helps MT performance, usually at least twice as much as just doing phrasal reordering.

## 2 DE classification

The Chinese character DE serves many different purposes. According to the Chinese Treebank tagging guidelines (Xia, 2000), the character can be tagged as DEC, DEG, DEV, SP, DER, or AS. Similar to (Wang et al., 2007), we only consider the majority case when the phrase with 的(DE) is a noun phrase modifier. The DEs in NPs have a part-of-speech tag of DEC (a complementizer or a nominalizer) or DEG (a genitive marker or an associative marker).

### 2.1 Class Definition

The way we categorize the DEs is based on their behavior when translated into English. This is implicitly done in the work of Wang et al. (2007) where they use rules to decide if a certain DE and the words next to it will need to be reordered. In this work, we categorize DEs into finer-grained categories. For a Chinese noun phrase [A 的 B], we categorize it into one of these five classes:

1. A B
   In this category, A in the Chinese side is translated as a pre-modifier of B. In most of the cases A is an adjective form, like Example 1.1 in Table 1 or the possessive adjective example in Example 1.2. Compound nouns where A becomes a pre-modifier of B also fit in this category (Example 1.3).
2. B *preposition* A
   There are several cases that get translated into the form B *preposition* A. For example, the *of*-genitive in Example 2.1 in Table 1.
   Example 2.2 shows cases where the Chinese A gets translated into a prepositional phrase that expresses location.
   When A becomes a gerund phrase and an object of a preposition, it is also categorized in the B *preposition* A category (Example 2.3).
3. A ’s B
   In this class, the English translation is an explicit *s*-genitive case, as in Example 3.1. This class occurs much less often but is still interesting because of the difference from the *of*-genitive.
4. *relative clause*
   We include the obvious relative clause cases like Example 4.1 where a relative clause is

introduced by a relative pronoun. We also include reduced relative clauses like Example 4.2 in this class.

5. A *preposition* B
   This class is another small one. The English translations that fall into this class usually have some number, percentage or level word in the Chinese A.

Some NPs are translated into a hybrid of these categories, or just don't fit into one of the five categories, for instance, involving an adjectival premodifier and a relative clause. In those cases, they are put into an "other" category.[1]

### 2.2 Data annotation of DE classes

In order to train a classifier and test its performance, we use the Chinese Treebank 6.0 (LDC2007T36) and the English Chinese Translation Treebank 1.0 (LDC2007T02). The word alignment data (LDC2006E93) is also used to align the English and Chinese words between LDC2007T36 and LDC2007T02. The overlapping part of the three datasets are a subset of CTB6 files 1 to 325. After preprocessing those three sets of data, we have 3253 pairs of Chinese sentences and their translations. In those sentences, we use the gold-standard Chinese tree structure to get 3412 Chinese DEs in noun phrases that we want to annotate. Among the 3412 DEs, 530 of them are in the "other" category and are not used in the classifier training and evaluation. The statistics of the five classes are:

1. A B: 693 (24.05%)
2. B *preposition* A: 1381 (47.92%)
3. A 's B: 91 (3.16%)
4. *relative clause*: 669 (23.21%)
5. A *preposition* B: 48 (1.66%)

## 3 Log-linear DE classifier

In order to see how well we can categorize DEs in noun phrases into one of the five classes, we train a log-linear classifier to classify each DE according to features extracted from its surrounding context. Since we want the training and testing conditions to match, when we extract features for the classifier, we don't use gold-standard parses. Instead, we use a parser trained on CTB6 excluding files 1-325. We then use this parser to parse the 3253 Chinese sentences with the DE annotation and extract parse-related features from there.

|  | 5-class Acc. (%) | 2-class Acc. (%) |
|---|---|---|
| baseline | - | 76.0 |
| DEPOS | 54.8 | 71.0 |
| +A-pattern | 67.9 | 83.7 |
| +POS-ngram | 72.1 | 84.9 |
| +Lexical | 74.9 | 86.5 |
| +SemClass | 75.1 | 86.7 |
| +Topicality | 75.4 | 86.9 |

Table 2: 5-class and 2-class classification accuracy. "baseline" is the heuristic rules in (Wang et al., 2007). Others are various features added to the log-linear classifier.

### 3.1 Experimental setting

For the classification experiment, we exclude the "other" class and only use the 2882 examples that fall into the five pre-defined classes. To evaluate the classification performance and understand what features are useful, we compute the accuracy by averaging five 10-fold cross-validations.[2]

As a baseline, we use the rules introduced in Wang et al. (2007) to decide if the DEs require reordering or not. However, since their rules only decide if there is reordering in an NP with DE, their classification result only has two classes. So, in order to compare our classifier's performance with the rules in Wang et al. (2007), we have to map our five-class results into two classes. We mapped our five-class results into two classes. So we mapped *B preposition A* and *relative clause* into the class "*reordered*", and the other three classes into "*not-reordered*".

### 3.2 Feature Engineering

To understand which features are useful for DE classification, we list our feature engineering steps and results in Table 2. In Table 2, the 5-class accuracy is defined by:

$$\frac{\text{(number of correctly labeled DEs)}}{\text{(number of all DEs)}} \times 100$$

The 2-class accuracy is defined similarly, but it is evaluated on the 2-class "*reordered*" and "*not-reordered*" after mapping from the 5 classes.

The DEs we are classifying are within an NP; we refer to them as [A 的 B]$_{\text{NP}}$. A includes all the words in the NP before 的; B includes all the words in the NP after 的. To illustrate, we will use the following NP:

[[韩国 最 大]$_{\text{A}}$ 的 [投资 对象国]$_{\text{B}}$]$_{\text{NP}}$

[1] The "other" category contains many mixed cases that could be difficult Chinese patterns to translate. We will leave this for future work.

[2] We evaluate the classifier performance using cross-validations to get the best setting for the classifier. The proof of efficacy of the DE classifier is MT performance on independent data in Section 4.

| | |
|---|---|
| **1.** | A B |
| **1.1.** | 优越(*excellent*)/的(*DE*)/地理(*geographical*)/条件(*qualification*) → “excellent geographical qualifications” |
| **1.2.** | 我们(*our*)/的(*DE*)/金融(*financial*)/风险(*risks*) → “our financial risks” |
| **1.3.** | 贸易(*trade*)/的(*DE*)/互补性(*complement*) → “trade complement” |
| **2.** | B ***preposition*** A |
| **2.1.** | 投资(*investment*)/环境(*environment*)/的(*DE*)/改善(*improvement*) → “the improvement of the investment environment” |
| **2.2.** | 崇明县(*Chongming county*)/内(*inside*)/的(*DE*)/单位(*organization*) → “organizations inside Chongming county” |
| **2.3.** | 一(*one*)/个(*measure word*)/观察(*observe*)/中国(*China*)/市场(*market*)/的(*DE*)/小小(*small*)/窗口(*window*)<br>→ “a small window for watching over Chinese markets” |
| **3.** | A **’s** B |
| **3.1.** | 国家(*nation*)/的(*DE*)/宏观(*macro*)/管理(*management*) → “the nation ’s macro management” |
| **4.** | ***relative clause*** |
| **4.1.** | 中国(*China*)/不能(*cannot*)/生产(produce)/而(*and*)/又(*but*)/很(*very*)/需要(*need*)/的(*DE*)/药品(*medicine*)<br>→ “medicine that cannot be produced by China but is urgently needed” |
| **4.2.** | 外商(*foreign business*)/投资(*invest*)/企业(*enterprise*)/获得(*acquire*)/的(*DE*)/人民币(*RMB*)/贷款(*loan*)<br>→ “the loans in RMB acquired by foreign-invested enterprises” |
| **5.** | A ***preposition*** B |
| **5.1.** | 四千多万(*more than 40 million*)/美元(*US dollar*)/的(*DE*)/产品(*product*) → more than 40 million US dollars in products |

Table 1: Examples for the 5 DE classes

to show examples of each feature. The parse structure of the NP is listed in Figure 2.

```
(NP
  (NP (NR 韩国))
  (CP
    (IP
      (VP
        (ADVP (AD 最))
        (VP (VA 大))))
    (DEC 的))
  (NP (NN 投资) (NN 对象国))))))
```

Figure 2: The parse tree of the Chinese NP.

### DEPOS: part-of-speech tag of DE

Since the part-of-speech tag of DE indicates its syntactic function, it is the first obvious feature to add. The NP in Figure 2 will have the feature “DEC”. This basic feature will be referred to as DEPOS. Note that since we are only classifying DEs in NPs, ideally the part-of-speech tag of DE will either be DEC or DEG as described in Section 2. However, since we are using automatic parses instead of gold-standard ones, the DEPOS feature might have other values than just DEC and DEG. From Table 2, we can see that with this simple feature, the 5-class accuracy is low but at least better than simply guessing the majority class (47.92%). The 2-class accuracy is still lower than using the heuristic rules in (Wang et al., 2007), which is reasonable because their rules encode more information than just the POS tags of DEs.

### A-pattern: Chinese syntactic patterns appearing before 的

Secondly, we want to incorporate the rules in (Wang et al., 2007) as features in the log-linear classifier. We added features for certain indicative patterns in the parse tree (listed in Table 3).

| |
|---|
| 1. **A is ADJP**:<br>true if A+DE is a DNP which is in the form of “ADJP+DEG”. |
| 2. **A is QP**:<br>true if A+DE is a DNP which is in the form of “QP+DEG”. |
| 3. **A is pronoun**:<br>true if A+DE is a DNP which is in the form of “NP+DEG”, and the NP is a pronoun. |
| 4. **A ends with VA**:<br>true if A+DE is a CP which is in the form of “IP+DEC”, and the IP ends with a VP that’s either just a VA or a VP preceded by a ADVP. |

Table 3: A-pattern features

Features 1–3 are inspired by the rules in (Wang et al., 2007), and the fourth rule is based on the observation that even though the predicative adjective VA acts as a verb, it actually corresponds to adjectives in English as described in (Xia, 2000).[3] We call these four features A-pattern. Our example NP in Figure 2 will have the fourth feature “A ends with VA” in Table 3, but not the other three features. In Table 2 we can see that after adding A-pattern, the 2-class accuracy is already much higher than the baseline. We attribute this to the fourth rule and also to the fact that the classifier can learn weights for each feature.[4]

[3] Quote from (Xia, 2000): “VA roughly corresponds to adjectives in English and stative verbs in the literature on Chinese grammar.”

[4] We also tried extending a rule-based 2-class classifier with the fourth rule. The accuracy is 83.48%, only slightly lower than using the same features in a log-linear classifier.

### POS-ngram: unigrams and bigrams of POS tags

The POS-ngram feature adds all unigrams and bigrams in A and B. Since A and B have different influences on the choice of DE class, we distinguish their ngrams into two sets of features. We also include the bigram pair across DE which gets another feature name for itself. The example NP in Figure 2 will have these features (we use b to indicate boundaries):

- POS unigrams in A: "NR", "AD", "VA"
- POS bigrams in A: "b-NR", "NR-AD", "AD-VA", "VA-b"
- cross-DE POS bigram: "VA-NN"
- POS unigram in B: "NN"
- POS bigrams in B: "b-NN", "NN-NN", "NN-b"

The part-of-speech ngram features add 4.24% accuracy to the 5-class classifier.

### Lexical: lexical features

In addition to part-of-speech features, we also tried to use features from the words themselves. But since using full word identity resulted in a sparsity issue,[5] we take the one-character suffix of each word and extract suffix unigram and bigram features from them. The argument for using suffixes is that it often captures the larger category of the word (Tseng et al., 2005). For example, 中国 (China) and 韩国 (Korea) share the same suffix 国, which means "country". These suffix ngram features will result in these features for the NP in Figure 2:

- suffix unigrams: "国", "最", "大", "的", "资", "国"
- suffix bigrams: "b-国", "国-最", "最-大", "大-的", "的-资", "资-国", "国-b"

Other than the suffix ngram, we also add three other lexical features: first, if the word before DE is a noun, we add a feature that is the conjunction of POS and suffix unigram. Secondly, an "NR only" feature will fire when A only consists of one or more NRs. Thirdly, we normalize different forms of "percentage" representation, and add a feature if they exist. This includes words that start with "百分之" or ends with the percentage sign "%". The first two features are inspired by the fact that a noun and its type can help decide "B prep A" versus "A B". Here we use the suffix of the noun and the NR (proper noun) tag to help capture its animacy, which is useful in choosing between the *s*-genitive (*the boy's mother*) and the *of*-genitive (*the mother of the boy*) in English (Rosenbach, 2003). The third feature is added because many of the cases in the "A *preposition* B" class have a percentage number in A. We call these sets of features Lexical. Together they provide 2.73% accuracy improvement over the previous setting.

### SemClass: semantic class of words

We also use a Chinese thesaurus, CiLin, to look up the semantic classes of the words in [A 的 B] and use them as features. CiLin is a Chinese thesaurus published in 1984 (Mei et al., 1984). CiLin is organized in a conceptual hierarchy with five levels. We use the level-1 tags which includes 12 categories.[6] This feature fires when a word we look up has one level-1 tag in CiLin. This kind of feature is referred to as SemClass in Table 2. For the example in Figure 2, two words have a single level-1 tag: "最"(most) has a level-1 tag K[7] and "投资"(investment) has a level-1 tag H[8]. "韩国" and "对象国" are not listed in CiLin, and "大" has multiple entries. Therefore, the SemClass features are: (*i*) before DE: "K"; (*ii*) after DE: "H"

### Topicality: re-occurrence of nouns

The last feature we add is a Topicality feature, which is also useful for disambiguating *s*-genitive and *of*-genitive. We approximate the feature by caching the nouns in the previous two sentences, and fire a topicality feature when the noun appears in the cache. Take this NP in MT06 as an example:

"南韩 与 北韩 的 奥运 代表队"

For this NP, all words before DE and after DE appeared in the previous sentence. Therefore the topicality features "cache-before-DE" and "cache-after-DE" both fire.

After all the feature engineering above, the best accuracy on the 5-class classifier we have is 75.4%, which maps into a 2-class accuracy of 86.9%. Comparing the 2-class accuracy to the (Wang et al., 2007) baseline, we have a 10.9% absolute improvement. The 5-class accuracy and confusion matrix is listed in Table 4.

"A *preposition* B" is a small category and is the most confusing. "A 's B" also has lower accuracy, and is mostly confused with "B *preposition* A".

[5] The accuracy is worse when we tried using the word identity instead of the suffix.

[6] We also tried adding more levels but it did not help.

[7] K is the category 助语 (auxiliary) in CiLin.

[8] H is the category 活动 (activities) in CiLin.

| real → | A 's B | AB | A *prep.* B | B *prep.* A | *rel. clause* |
|---|---|---|---|---|---|
| A 's B | 168 | 36 | 0 | 110 | 0 |
| AB | 48 | 2473 | 73 | 227 | 216 |
| A *prep.* B | 0 | 18 | 46 | 23 | 11 |
| B *prep.* A | 239 | 691 | 95 | 5915 | 852 |
| *rel. clause* | 0 | 247 | 26 | 630 | 2266 |
| Total | 455 | 3465 | 240 | 6905 | 3345 |
| Accuracy(%) | 36.92 | 71.37 | 19.17 | 85.66 | 67.74 |

Table 4: The confusion matrix for 5-class DE classification

This could be due to the fact that there are some cases where the translation is correct both ways, but also could be because the features we added have not captured the difference well enough.

## 4 Machine Translation Experiments

### 4.1 Experimental Setting

For our MT experiments, we used a re-implementation of Moses (Koehn et al., 2003), a state-of-the-art phrase-based system. The alignment is done by the Berkeley word aligner (Liang et al., 2006) and then we symmetrized the word alignment using the grow-diag heuristic. For features, we incorporate Moses' standard eight features as well as the lexicalized reordering model. Parameter tuning is done with Minimum Error Rate Training (MERT) (Och, 2003). The tuning set for MERT is the NIST MT06 data set, which includes 1664 sentences. We evaluate the result with MT02 (878 sentences), MT03 (919 sentences), and MT05 (1082 sentences).

Our MT training corpus contains 1,560,071 sentence pairs from various parallel corpora from LDC.[9] There are 12,259,997 words on the English side. Chinese word segmentation is done by the Stanford Chinese segmenter (Chang et al., 2008). After segmentation, there are 11,061,792 words on the Chinese side. We use a 5-gram language model trained on the Xinhua and AFP sections of the Gigaword corpus (LDC2007T40) and also the English side of all the LDC parallel data permissible under the NIST08 rules. Documents of Gigaword released during the epochs of MT02, MT03, MT05, and MT06 were removed.

To run the DE classifier, we also need to parse the Chinese texts. We use the Stanford Chinese parser (Levy and Manning, 2003) to parse the Chinese side of the MT training data and the tuning and test sets.

### 4.2 Baseline Experiments

We have two different settings as baseline experiments. The first is without reordering or DE annotation on the Chinese side; we simply align the parallel texts, extract phrases and tune parameters. This experiment is referred to as BASELINE. Also, we reorder the training data, the tuning and the test sets with the NP rules in (Wang et al., 2007) and compare our results with this second baseline (WANG-NP).

The NP reordering preprocessing (WANG-NP) showed consistent improvement in Table 5 on all test sets, with BLEU point gains ranging from 0.15 to 0.40. This confirms that having reordering around DEs in NP helps Chinese-English MT.

### 4.3 Experiments with 5-class DE annotation

We use the best setting of the DE classifier described in Section 3 to annotate DEs in NPs in the MT training data as well as the NIST tuning and test sets.[10] If a DE is in an NP, we use the annotation of 的$_{\text{AB}}$, 的$_{\text{AsB}}$, 的$_{\text{BprepA}}$, 的$_{\text{relc}}$, or 的$_{\text{AprepB}}$ to replace the original DE character. Once we have the DEs labeled, we preprocess the Chinese sentences by reordering them.[11] Note that not all DEs in the Chinese data are in NPs, therefore not all DEs are annotated with the extra labels. Table 6 lists the statistics of the DE classes in the MT training data.

| class of 的(DE) | counts | percentage |
|---|---|---|
| 的$_{\text{AB}}$ | 112,099 | 23.55% |
| 的$_{\text{AprepB}}$ | 2,426 | 0.51% |
| 的$_{\text{AsB}}$ | 3,430 | 0.72% |
| 的$_{\text{BprepA}}$ | 248,862 | 52.28% |
| 的$_{\text{relc}}$ | 95,134 | 19.99% |
| 的 (unlabeled) | 14,056 | 2.95% |
| total number of 的 | 476,007 | 100% |

Table 6: The number of different DE classes labeled for the MT training data.

After this preprocessing, we restart the whole MT pipeline – align the preprocessed data, extract phrases, run MERT and evaluate. This setting is referred to as DE-Annotated in Table 5.

### 4.4 Hierarchical Phrase Reordering Model

To demonstrate that the technique presented here is effective even with a hierarchical decoder, we

[9] LDC2003E07, LDC2003E14, LDC2005E83, LDC2005T06, LDC2006E26, LDC2006E85, LDC2006E85, LDC2005T34, and LDC2005T34

[10] The DE classifier used to annotate the MT experiment was trained on all the available data described in Section 2.2.

[11] Reordering is applied on DNP and CP for reasons described in Wang et al. (2007). We reorder only when the 的 is labeled as 的$_{\text{BprepA}}$ or 的$_{\text{relc}}$.

| BLEU | | | | |
|---|---|---|---|---|
| | MT06(tune) | MT02 | MT03 | MT05 |
| BASELINE | 32.39 | 32.51 | 32.75 | 31.42 |
| WANG-NP | 32.75(+0.36) | 32.66(+0.15) | 33.15(+0.40) | 31.68(+0.26) |
| DE-Annotated | **33.39**(+1.00) | **33.75**(+1.24) | **33.63**(+0.88) | **32.91**(+1.49) |
| BASELINE+Hier | 32.96 | 33.10 | 32.93 | 32.23 |
| DE-Annotated+Hier | **33.96**(+1.00) | **34.33**(+1.23) | **33.88**(+0.95) | **33.01**(+0.77) |
| **Translation Error Rate (TER)** | | | | |
| | MT06(tune) | MT02 | MT03 | MT05 |
| BASELINE | 61.10 | 63.11 | 62.09 | 64.06 |
| WANG-NP | 59.78(−1.32) | 62.58(−0.53) | 61.36(−0.73) | 62.35(−1.71) |
| DE-Annotated | **58.21**(−2.89) | **61.17**(−1.94) | **60.27**(−1.82) | **60.78**(−3.28) |

Table 5: MT experiments of different settings on various NIST MT evaluation datasets. We used both the BLEU and TER metrics for evaluation. All differences between DE-Annotated and BASELINE are significant at the level of 0.05 with the approximate randomization test in (Riezler and Maxwell, 2005)

conduct additional experiments with a hierarchical phrase reordering model introduced by Galley and Manning (2008). The hierarchical phrase reordering model can handle the key examples often used to motivated syntax-based systems; therefore we think it is valuable to see if the DE annotation can still improve on top of that. In Table 5, BASELINE+Hier gives consistent BLEU improvement over BASELINE. Using DE annotation on top of the hierarchical phrase reordering models (DE-Annotated+Hier) provides extra gain over BASELINE+Hier. This shows the DE annotation can help a hierarchical system. We think similar improvements are likely to occur with other hierarchical systems.

# 5 Analysis

## 5.1 Statistics on the Preprocessed Data

Since our approach DE-Annotated and one of the baselines (WANG-NP) are both preprocessing Chinese sentences, knowing what percentage of the sentences are altered will be one useful indicator of how different the systems are from the baseline. In our test sets, MT02 has 591 out of 878 sentences (67.3%) that have DEs under NPs; for MT03 it is 619 out of 919 sentences (67.4%); for MT05 it is 746 out of 1082 sentences (68.9%). This shows that our preprocessing affects the majority of the sentences and thus it is not surprising that preprocessing based on the DE construction can make a significant difference.

## 5.2 Example: how DE annotation affects translation

Our approach DE-Annotated reorders the Chinese sentence, which is similar to the approach proposed by Wang et al. (2007) (WANG-NP). However, our focus is on the annotation on DEs and how this can improve translation quality. Table 7 shows an example that contains a DE construction that translates into a relative clause in English.[12] The automatic parse tree of the sentence is listed in Figure 3. The reordered sentences of WANG-NP and DE-Annotated appear on the top and bottom in Figure 4. For this example, both systems decide to reorder, but DE-Annotated had the extra information that this 的 is a 的$_{\text{relc}}$. In Figure 4 we can see that in WANG-NP, “的” is being translated as “for”, and the translation afterwards is not grammatically correct. On the other hand, the bottom of Figure 4 shows that with the DE-Annotated preprocessing, now “的$_{\text{relc}}$” is translated into “which was” and well connected with the later translation. This shows that disambiguating 的 helps in choosing a better English translation.

```
(IP
 (NP (NN 比亚吉))
 (VP
   (ADVP (AD 曾))
   (VP (VV 协助)
     (IP
       (VP (VV 草拟)
         (NP
           (QP (CD 一)
             (CLP (M 份)))
           (CP
             (IP
               (VP (VV 遭)
                 (NP
                   (NP (NN 工会)
                     (CC 和)
                     (NN 左翼) (NN 分子))
                   (ADJP (JJ 强烈))
                   (NP (NN 反对)))))
             (DEC 的))
           (NP (NN 就业) (NN 改革) (NN 方案)))))))
 (PU 。))
```

Figure 3: The parse tree of the Chinese sentence in Table 7.

[12] In this example, all four references agreed on the relative clause translation. Sometimes DE constructions have multiple appropriate translations, which is one of the reasons why certain classes are more confusable in Table 4.

| Chinese | 比亚吉 曾 协助 草拟 [一 份 遭 工会 和 左翼 分子 强烈 反对]$_A$ 的 [就业 改革 方案]$_B$ 。 |
|---|---|
| Ref 1 | biagi had assisted in drafting [an employment reform plan]$_B$ [that was strongly opposed by the labor union and the leftists]$_A$ . |
| Ref 2 | biagi had helped in drafting [a labor reform proposal]$_B$ [that provoked strong protests from labor unions and the leftists]$_A$ . |
| Ref 3 | biagi once helped drafting [an employment reform scheme]$_B$ [that was been strongly opposed by the trade unions and the left - wing]$_A$ . |
| Ref 4 | biagi used to assisted to draft [an employment reform plan]$_B$ [which is violently opposed by the trade union and leftest]$_A$ . |

Table 7: A Chinese example from MT02 that contains a DE construction that translates into a relative clause in English. The $[]_A$ $[]_B$ is hand-labeled to indicate the approximate translation alignment between the Chinese sentence and English references.

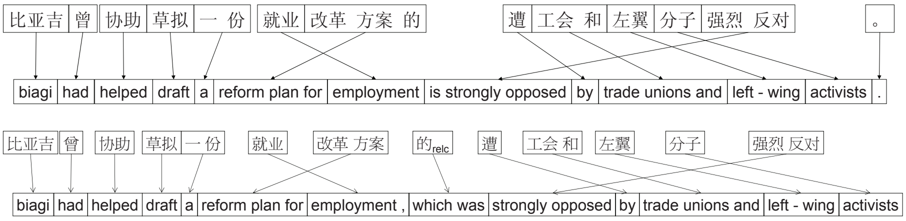

Figure 4: The top translation is from WANG-NP of the Chinese sentence in Table 7. The bottom one is from DE-Annotated. In this example, both systems reordered the NP, but DE-Annotated has an annotation on the 的.

## 6 Conclusion

In this paper, we presented a classification of Chinese 的(DE) constructions in NPs according to how they are translated into English. We applied this DE classifier to the Chinese sentences of MT data, and we also reordered the constructions that required reordering to better match their English translations. The MT experiments showed our preprocessing gave significant BLEU and TER score gains over the baselines. Based on our classification and MT experiments, we found that not only do we have better rules for deciding what to reorder, but the syntactic, semantic, and discourse information that we capture in the Chinese sentence allows us to give hints to the MT system which allows better translations to be chosen.

## Acknowledgments

The authors would like to thank Michel Galley and Daniel Cer for useful discussions and technical help, and Spence Green for his comments on an earlier draft of the paper. This work is funded by a Stanford Graduate Fellowship to the first author and gift funding from Google for the project "Translating Chinese Correctly".

## References

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232, Columbus, Ohio, June. Association for Computational Linguistics.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL*, pages 263–270, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *ACL '05: Proceedings of ACL*, pages 531–540, Morristown, NJ, USA. Association for Computational Linguistics.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of EMNLP*, pages 847–855, Honolulu, Hawaii, October. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of NAACL-HLT*.

Roger Levy and Christopher Manning. 2003. Is it harder to parse Chinese, or the Chinese treebank? In *Proceedings of ACL*, pages 439–446, Morristown, NJ, USA. Association for Computational Linguistics.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of HLT-NAACL*, pages 104–111, New York City, USA, June. Association for Computational Linguistics.

Jia-ju Mei, Yi-Ming Zheng, Yun-Qi Gao, and Hung-Xiang Yin. 1984. *TongYiCi CiLin*. Shanghai: the Commercial Press.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *ACL*.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Anette Rosenbach. 2003. Aspects of iconicity and economy in the choice between the s-genitive and the of-genitive in English. *Topics in English Linguistics*, 43:379–412.

Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning. 2005. Morphological features help pos tagging of unknown words across language varieties. In *Proc. of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of EMNLP-CoNLL*, pages 737–745, Prague, Czech Republic, June. Association for Computational Linguistics.

Fei Xia and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of Coling 2004*, pages 508–514, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Fei Xia. 2000. The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0).