Enhancing Grammatical Cohesion: Generating Transitional Expressions for SMT

Mei Tu **Yu Zhou** **Chengqing Zong**
National Laboratory of Pattern Recognition,
Institute of Automation,
Chinese Academy of Sciences
{mtu,yzhou,cqzong}@nlpr.ia.ac.cn

Abstract

Transitional expressions provide glue that holds ideas together in a text and enhance the logical organization, which together help improve readability of a text. However, in most current statistical machine translation (SMT) systems, the outputs of compound-complex sentences still lack proper transitional expressions. As a result, the translations are often hard to read and understand. To address this issue, we propose two novel models to encourage generating such transitional expressions by introducing the source compound-complex sentence structure (CSS). Our models include a CSS-based translation model, which generates new CSS-based translation rules, and a generative transfer model, which encourages producing transitional expressions during decoding. The two models are integrated into a hierarchical phrase-based translation system to evaluate their effectiveness. The experimental results show that significant improvements are achieved on various test data meanwhile the translations are more cohesive and smooth.

1 Introduction

During the last decade, great progress has been made on statistical machine translation (SMT) models. However, these translations still suffer from poor readability, especially translations of compound-complex sentences. One of the main reasons may be that most existing models concentrate more on producing well-translated local sentence fragments, but largely ignore global cohesion between the fragments. Generally, cohesion, including lexical and grammatical cohesion, contributes much to the understandability and smoothness of a text.

Recently, researchers have begun addressing the lexical cohesion of SMT (Gong et al., 2011; Xiao et al., 2011; Wong and Kit, 2012; Xiong, 2013). These efforts focus mainly on the co-occurrence of lexical items in a similar environment. Grammatical cohesion[1] (Halliday and Hassan, 1976) in SMT has been little mentioned in previous work. Translations without grammatical cohesion is hard to read, mostly due to loss of cohesive and transitional expressions between two sentence fragments. Thus, generating transitional expressions is necessary for achieving grammatical cohesion. However, it is not easy to produce such transitional expressions in SMT. As an example, consider the Chinese-to-English translation in Figure 1.

Source Chinese sentence:

[尽管	*减轻*	*污染*	*的*	*呼声*	*不断*	*，*	*]1*	*[*	*公众*
Although	*reduce*	*pollution*	*of*	*calls*	*continue*	*,*			*public*

日渐	*愤怒*	*，*	*]$_2$*	*[污染*	*还是*	*变得*	*更*	*糟糕*
growing	*angry*	*,*		*pollution*	*still*	*become*	*more*	*worse*

了	*，*	*]$_3$*	*[越发*	*显出*	*环保*	*的*	*紧迫性*	*。]$_4$*
already	*,*		*more*	*show*	*environment protection*	*of*	*urgent*	*.*

Target English golden translation:

Despite *frequent calls for cutting pollution,* ***and*** *growing public anger, the problem has only got worse,* ***which*** *increasingly shows the urgency of environmental protection.*

Figure 1: An example of Chinese-to-English translation. The English translation sentence has three transitional phrases: ***Despite, and, which***.

There are 4 sub-sentences separated by commas in the Chinese sentence. We have tried to translate the Chinese sentence using many well-

[1] Grammatical cohesion can make relations among sentences more explicit. There are various grammatically cohesive devices (reference, substitution ellipsis and conjunction) that tie fragments together in a cohesive way.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 850–860,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

known online translators, but find that it is very difficult to generate the target transitional expressions, especially when there is no explicit connective word in the source sentence, such as generating "***and*** " and "***which***" in Figure 1.

Fortunately, the functional relationships between two neighboring source sub-sentences provide us with a good perspective and the inspiration to generate those transitional phrases. Figure 1 shows that the first and the second Chinese sub-sentences form a ***parallel*** relation. Thus, even though there is no distinct connective word at the beginning of the second source sub-sentence, a good translator is still able to insert or generate an "***and***" as a connection word to make the target translation more cohesive.

Based on the above analysis, this paper focuses on the target grammatical cohesion in SMT to make the translation more understandable, especially for languages with great difference in linguistic structure like Chinese and English. To the best of our knowledge, our work is the first attempt to generate target transitional expressions for SMT grammatical cohesion by introducing the functional relationships of source sentences. In this work, we propose two models. One is a new translation model that is utilized to generate new translation rules combined with the information of source functional relationships. The other is a generative transfer model that encourages producing transitional phrases during decoding. Our experimental results on Chinese-to-English translation demonstrate that the translation readability is greatly improved by introducing the cohesive information.

The remainder of the paper is organized as follows. In Section 2, we describe the functional relationships of Chinese compound-complex sentences. In Section 3, we present our models and show how to integrate the models into an SMT system. Our experimental results are reported in Section 4. A survey of related work is conducted in Section 5, and we conclude our work and outline the future work in Section 6.

2 Chinese Compound-Complex Sentence Structure

To acquire the functional relationships of a Chinese compound-complex sentence, Zhou (2004) proposed a well-annotated scheme to build the Compound-complex Sentence Structure (**CSS**). The structure explicitly shows the minimal semantic spans, called elementary units (***eu***s), and also depicts the hierarchical relations among ***eu***s. There are 11 common types of functional relationships[2] annotated in the Tsinghua Chinese Treebank (Zhou, 2004).

Under the annotation scheme of the Tsinghua Chinese Treebank, the Chinese sentence of example in Figure 1 is represented as the tree shown in Figure 2. In this example, each sub-sentence is an ***eu***. ***eu₁*** and ***eu₂*** are combined with a *parallel* relationship, followed by ***eu₃*** with an *adversative* relationship. ***eu₁***, ***eu₂***, and ***eu₃*** form a large semantic span[3], connected with ***eu₄*** by a *consequence* relationship. All of the ***eu***s are organized into various functional relationships and finally form a hierarchical tree.

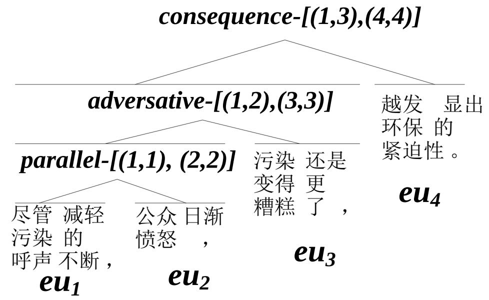

consequence-[(1,3),(4,4)]
adversative-[(1,2),(3,3)]
parallel-[(1,1), (2,2)]
尽管 减轻 污染 的 呼声 不断，
eu1
公众 日渐 愤怒 ，
eu2
污染 还是 变得 更 糟糕 了 ，
eu3
越发 显出 环保 的 紧迫性 。
eu4

Figure 2: The compound-complex sentence structure of the Chinese sentence in Figure 1.

Formally, given a compound-complex sentence structure (CSS), each node in the CSS can be represented as a tuple $R-[(s_1,e_1),...(s_l,e_l),...,(s_L,e_L)]$. R represents the relationship, which has L children. For each child of R , a pair (s_l,e_l) records its *start* and *end* ***eu***s. For example, ***adversative-[(1,2), (3,3)]*** in Figure 2 means that two children are controlled by the relationship ***adversative***, and the left child consists of ***eu₁*** and ***eu₂***, while the right child contains only ***eu₃***.

CSS has much in common with Rhetorical Structure (Mann and Thompson, 1988) in English, which also describe the semantic relation between discourse units. But the Rhetorical Structure involves much richer relations on the document-level, and little corpus is open for Chinese.

In the following, we will describe in detail how to utilize such CSS information for modelling in SMT.

[2] They are *parallel, consequence, progressive, alternative, causal, purpose, hypothesis, condition, adversative, explanation,* and *flowing* relationships.

[3] A semantic span can include one or more ***eu***s.

3 Modelling

Our purpose is to enhance the grammatical cohesion by exploiting the source CSS information. Therefore, theoretically, the conditional probability of a target translation $\boldsymbol{e}_s$ conditioned on the source CSS-based tree $\boldsymbol{f}_\mathrm{t}$ is given by $P(\boldsymbol{e}_s \mid \boldsymbol{f}_t)$, and the final translation $\hat{\boldsymbol{e}}_s$ is obtained with the following formula:

$$\hat{\boldsymbol{e}}_s = \arg\max_{\boldsymbol{e}_s} \{\mathrm{P}(\boldsymbol{e}_s \mid \boldsymbol{f}_t)\} \qquad (1)$$

Following Och and Ney (2002), our model is framed as a log-linear model:

$$P(\boldsymbol{e}_s \mid \boldsymbol{f}_t) = \frac{\exp \sum_k \lambda_k h_k(\boldsymbol{e}_s, \boldsymbol{f}_t)}{\sum_{\boldsymbol{e}'_s} \exp \sum_k \lambda_k h_k(\boldsymbol{e}'_s, \boldsymbol{f}_t)} \qquad (2)$$

where $h(\boldsymbol{e}_s, \boldsymbol{f}_t)$ is a feature with weight λ. Then, the best translation is:

$$\hat{\boldsymbol{e}}_s = \arg\max_{\boldsymbol{e}_s} \exp \sum_k \lambda_k h_k(\boldsymbol{e}_s, \boldsymbol{f}_t) \qquad (3)$$

Our models make use of CSS with two strategies:

1) **CSS-based translation model**: following formula (1), we obtain the cohesion information by modifying the translation rules with their probabilities $P(\boldsymbol{e}_s \mid \boldsymbol{f}_t)$ based on word alignments between the source CSS-tree and the target string;

2) **CSS-based transfer model:** following formula (3), we introduce a transfer score to encourage the decoder to generate transitional words and phrases; the score is utilized as an additional feature $h_k(\boldsymbol{e}_s, \boldsymbol{f}_t)$ in the log-linear model.

3.1 CSS-based Translation Model

For the existing translation models, the entire training process is conducted at the lexical or syntactic level without grammatically cohesive information. As a result, it is difficult to utilize such cohesive information during decoding. Instead, we reserve the cohesive information in the training process by converting the original source sentence into tagged-flattened CSS and then perform word alignment and extract the translation rules from the bilingual flattened source CSS and the target string.

As introduced in Section 2, a CSS consists of nodes, and a node can be represented as a tuple $R-[(s_1,e_1),...(s_l,e_l),...,(s_L,e_L)]$. In this representation, the relationship R is the most important factor because different relationships directly reflect different cohesive expressions. In addition, the children's positions always play a strong role in choosing cohesive expressions because transitional expressions vary for children with different positions. For example, when translating the last child of a *parallel* relation, we always use word "*and*" as the transitional expression seen in Figure 3, but we will not use it for the first child of a *parallel* relation. Therefore, in the training process we just keep the information of relationships and children's positions when converting

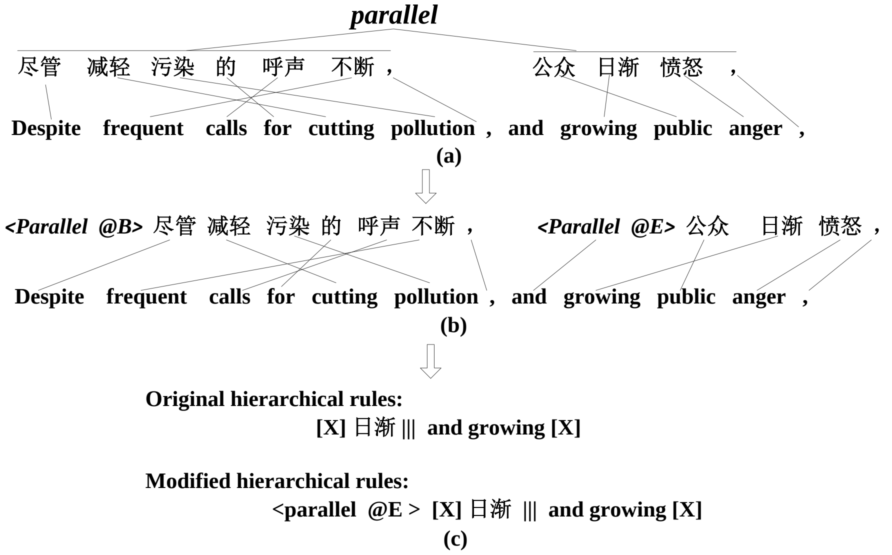

parallel
尽管 减轻 污染 的 呼声 不断，
公众 日渐 愤怒 ，
Despite frequent calls for cutting pollution , and growing public anger ,
(a)
<Parallel @B> 尽管 减轻 污染 的 呼声 不断，
<Parallel @E> 公众 日渐 愤怒 ，
Despite frequent calls for cutting pollution , and growing public anger ,
(b)
Original hierarchical rules:
[X] 日渐 ||| and growing [X]
Modified hierarchical rules:
<parallel @E > [X] 日渐 ||| and growing [X]
(c)

Figure 3: An example of modifying translation rules. @B means the current structure information comes from the first child, and @E means from the last child.

the source CSS to a tagged-flattened string.

Considering that the absolute position (index of the ***eu***, such as *1*, *2*, *3*) is somehow sparse in the corpus, we employ the relative position instead. *B* (*Beginning*) represents the first child of a relationship, *E* (*End*) means the last child of a relationship, and *M* (*Middle*) represents all the middle children.

Under this agreement, the original Chinese CSS-based tree will be converted to a new tagged-flattened string. Note the converting example from Figure 3(a) to Figure 3(b): node *parallel-[(1,1), (2,2)]* (see Figure 2) is converted to a flat string. Its first child is represented as <*parallel, @B*> with the semantic span, while the last child is <*parallel, @E*> with the corresponding semantic span.

We then perform word alignment on the modified bilingual sentences, and extract the new translation rules based on the new alignment, as shown in Figure 3(b) to Figure 3(c). Now the newly extracted rule "<*parallel, @E* > *[X] 日渐* ||| *and growing [X]* " is tagged with cohesive information. Thus, if the similar relationship *parallel* occurs in the test source sentence, this type of rule is more likely to be chosen to generate the cohesive word "*and*" during decoding because it is more discriminating than the original rules (*[X] 日渐* ||| *and growing [X]*). The conditional probabilities of the new translation rules are calculated following (Chiang, 2005).

3.2 CSS-based Transfer model

In general, according to formula (3), the translation quality based on the log-linear model is related tightly with the features chosen. Most translation systems adopt the features from a translation model, a language model, and sometimes a reordering model. To give a bonus to generating cohesive expressions during decoding, we have designed a special additional feature. The additional feature is represented as a probability calculated by a transfer model.

Given the source CSS information, we want our transfer model to predict the most possible cohesive expressions. For example, given two semantic spans with a *parallel* relationship and many translation candidates, our transfer model is expected to assign higher scores to those with transitional expressions such as "*and*" or "*as well as*".

Let $\boldsymbol{w} = w_0, w_1, ...w_n$ represent the transitional expressions observed in the target string. Our transfer model can be represented as a conditional probability:

$$P(\boldsymbol{w} \mid CSS) \qquad (4)$$

By deriving each node of the CSS, we can obtain a factored formula:

$$P(\boldsymbol{w} \mid CSS) = \prod_{i,j} P(\boldsymbol{w}_{ij} \mid R_i, RP_j) \qquad (5)$$

where $\boldsymbol{w}_{ij}$ is the transitional expression produced by the j^{th} child of the i^{th} node of the CSS. R_i is the relationship type of the i^{th} node. For the j^{th} child in the i^{th} node, RP_j is its relative position (*B, M* or *E*) introduced in Section 3.1.

The process of training this transfer model and smoothing is similar to the process of training a language model. We obtain the factored transfer probability as follows,

$$P(\boldsymbol{w}_{ij} \mid R_i, RP_j)$$
$$= P(w_0 \mid R_i, RP_j) \prod_{k=1}^{n} P(w_k \mid w_0{}^{k-1}, R_i, RP_j) \qquad (6)$$

where

$$\boldsymbol{w}_{ij} = w_0{}^n = w_0, ...w_n \qquad (7)$$

Following (Bilmes and Kirchhoff, 2003), the conditional probabilities $P(w_k \mid w_0{}^{k-1}, R_i, RP_j)$ in formula (6) are estimated in the same way as a factored language model, which has the advantage of easily incorporating various linguistic information.

Considering that $\boldsymbol{w}_{ij}$ commonly appears at the beginning of the target translation of a source semantic span such as "*which …*", namely, the **left-frontier phrases,** we focus only on the left-frontier phrases when training this model. Note that if there exists a target word before a left frontier, and this word is aligned to *NULL*, we will expand the left frontier to this word. The expansion process will be repeated until there is no such word. For example, if we take the CSS and the alignment in Figure 3(a) for training, the left frontier of the second child will be expanded from "*growing*" to "*and*". In addition, taking the tri-gram left-frontier phrase for example, we can obtain a training sample such as $\boldsymbol{w}_{ij}$ = *and growing public*, *R=parallel*, *RP = E*.

By learning such probabilities for different transitional expressions conditioned on different relationships, we are able to capture the inner connection between the source CSS and the projected target cohesive phrases. Thus, during decoding, if we add the probability generated by the transfer model of $P(\boldsymbol{w} \mid CSS)$ as a feature in formula (3), it will certainly contribute to selecting more cohesive candidates.

3.3 Elementary-Unit Cohesion Constraint

As mentioned in Section 3.2, in the transfer model, the transitional phrases are expected to occur at the left frontier of a projected span on target side. In fact, this depends on the assumption that the projected translations of any two disjoint source semantic spans are also disjoint to keep their own semantic integrity. We call this assumption the integrity assumption. This assumption is intuitive and supported by statistics. After analyzing 1,007 golden aligned Chinese-English sentence-pairs, we find that approximately 90% of the pairs comply with the assumption. However, in real automatically aligned noisy data, the ratio of complying pairs reduces to 71%[4]. Two projected translations that violate the integrity assumption may mutually overlap, which causes our confusion on where to extract the transitional phrases. In this case, extracted transitional phrases are likely to be wrong.

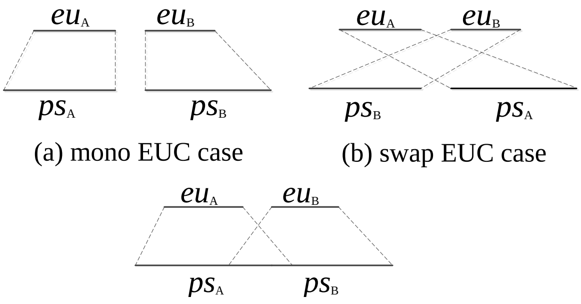
To increase the chance of extracting correct transitional phrases, the alignment results must be modified to reduce the impact of incorrect alignment. We propose a dynamic cleaning method to ensure that the most expressive transitional phrases fall in the accessible extraction range before training the transfer model.

3.3.1 EUC and non-EUC

As we have defined in Section 2, the minimal semantic span is called elementary unit (***eu***). If the source ***eu*** and its projected target span comply with the integrity assumption, we say that such an ***eu*** and its projected span have **Elementary-Unit-Cohesion (EUC).** We define EUC formally as follows.

Given two elementary units eu_A and eu_B, and their projected target spans ps_A and ps_B bound by the word alignment, the alignment complies with EUC only if there is no overlap between ps_A and ps_B. Otherwise, the alignment is called ***non-EUC***. The common ***EUC*** and ***non-EUC*** cases are illustrated in Figure 4.

EUC is the basic case for the integrity assumption. For the best cases, the elementary units comply with EUC, and thus the semantic spans combined by elementary units are certainly subject to the integrity assumption.



(a) mono EUC case

(b) swap EUC case

(c) non-EUC case

Figure.4 The schematic diagram of *EUC* cases and *non-EUC* case.

3.3.2 A Dynamic Cleaning Method

An intuitive method to clean the alignment results is to drop off the noisy word-to-word links that cause non-EUC. Considering that the dropping process is a post-editing method for the original alignment obtained by a state-of-the-art aligner such as GIZA++, we do not expect over-deleting. Therefore, we tend to take a relatively conservative strategy to minimize the deleting operation.

Given a sentence-pair (***f***, ***e***), suppose that $\boldsymbol{f} = \{f_0, ..., f_i, ..., f_I\}$ is divided into M elementary units $U = \{u_0, ..., u_m, ..., u_M\}$, and ***e*** has N words, that is, $\boldsymbol{e} = \{e_0, ..., e_n, ..., e_N\}$. If A is the word alignment of (***f***, ***e***), then the goal is to construct the maximum subset $A^* \subseteq A$ under the condition that A^* is the word alignment with the constraint of EU. The search process can be described as the pseudo code in Figure 5.

In Figure 5, we scan each target word and each source ***eu*** to assign each word to a unique ***eu*** under the EUC constraint with the lowest cost. Function $\text{cost}(n, m)$ in line 6 computes the counts of deleted links that force the n^{th} target word to align only to words in the range of the m^{th} ***eu***. For example, if the n^{th} target word is aligned to the i^{th}, $(i+1)^{th}$, and $(i+2)^{th}$ word in source side, while the i^{th} word belongs to u_{m_1} and the $(i+1)^{th}$ and $(i+2)^{th}$ words belong to u_{m_2}, then $\text{cost}(n, u_{m_1}) = 2$, and $\text{cost}(n, u_{m_2}) = 1$. In line 6, *Score*[*n*][*m*] saves a list of scores, each score computed by adding the current cost(*n*, *m*) with the history score of each list of *Score*[*n-1*].

[4] The aligning tool is GIZA++ with 5 iterations of Model 1, 5 iterations of HMM, and 10 iterations of Model 4. The GIZA++ code can be downloaded from https://code.google.com/p/giza-pp/

Before the next iteration, the bad branches are pruned, as seen in line 5. We adopt the following two ways to prune:

(1) EUC constraint: if the current link violates EUC alignment, delete it.

(2) Keep the hypothesis with a fixed maximum size to avoid too large a searching space.

```
//Pseudo code for dynamic cleaning
1: $Score\ [N+1][M] = \{[0]\}_{N\times M}$    /* initialize
                                   cumulative cost score chart*/
2: $Path\ [M]=[[]]$                /*initialize tracking path*/
3: for $n = 1 \rightarrow N$ :{          /* scan target words*/
4:   for $m = 0 \rightarrow M-1$ :{       /*scan source U set */
5:     PrunePath();
           /* prune invalid path and high-cost path*/
6:     $Score[n][m]$=GetScore($Score[n-1]$, cost($n$, $m$))
       /*compute current cumulative cost score by previ-
                  ous score and current cost*/
7:     SaveCurrentPath($Path[m]$);
                      /*add current index to Path*/
8:   }//end $m$
9: }//end $n$
10: OptimalPath = $\arg\max_{Path[m]}\{Score[N][m]\}$;
```

Figure 5. The pseudo code of dynamic cleaning method.

4 Experiments

4.1 Experimental Setup

To obtain the CSSs of Chinese sentences, we use the Chinese parser proposed in (Tu et al., 2013a). Their parser first segments the compound-complex sentence into a series of elementary units, and then builds structure of the hierarchical relationships among these elementary units. Their parser was reported to achieve an F-score for elementary unit segmentation of approximately 0.89. The *progressive*, *causal*, and *condition* terms of functional relationships can be recognized with precisions of 0.86, 0.8, and 0.75, respectively, while others, such as *purpose*, *parallel*, and *flowing*, achieve only 0.5, 0.59 and 0.62, respectively.

The translation experiments have been conducted in the Chinese-to-English direction. The bilingual training data for translation model and CSS-based transfer model is FBIS corpus with approximately 7.1 million Chinese words and 9.2 million English words. We obtain the word alignment with the grow-diag-final-and strategy with GIZA++. Before training the CSS-based transfer model, the alignment for transfer model is modified by our dynamic cleaning method. During the cleaning process, the maximum size of hypothesis is limited to 5. A 5-gram language model is trained with SRILM[5] on the combination of the Xinhua portion of the English Gigaword corpus combined with the English part of FBIS. For tuning and testing, we use NIST03 evaluation data as the development set. NIST04/05/06, CWMT08-Development[6] and CWMT08-Evaluation data are used for testing under the measure metric of BLEU-4 (Papineni et al. 2002) with the shortest length penalty.

Table 1 shows how the CSS is distributed in all testing sets. According to the statistics in Table 1, we see that CSS is really widely distributed in the NIST and CWMT corpora, which implies that the translation quality may benefit substantially from the CSS information, if it is well considered in SMT.

	Total	CSS	Ratio(%)
NIST04	1,788	1,307	73.1
NIST05	1,082	849	78.5
NIST06	1,000	745	74.5
CWMT08-Dev.	1,006	818	81.3
CWMT08-Eval.	1,006	818	81.3

Table 1. The numbers of sentences and the CSS ratios of all sentences. CWMT08-Dev. is short for CWMT08 Development data and CWMT08-Eval. is CWMT08 Evaluation data.

4.2 Extracted Transitional Expressions

Eleven types of Chinese functional relationships and their English left-frontier phrases (tri-gram) learned by our transfer model are given in Table 2.

The results in Table 2 show that some left-frontier phrases reflect the source functional relationship well, especially for those with better precision of relationship recognition, such as *progressive, causal* and *condition*. Conversely, lower precision of relationship recognition may weaken the learning ability of the transfer model. For example, noisy left-frontier phrases are easily generated under relationships such as *parallel* and *purpose*.

[5] http://www.speech.sri.com/projects/srilm/

[6] The China Workshop on Machine Translation

Relation	**Left-frontier phrases (tri-gram)**
parallel	as well as; at the same; …
progressive	but will also; in addition to;…
causal	therefore , the; for this reason; as a result; because it is; so it is;…
condition	as long as; only when the…
hypothesis	if we do; if it is; if the us; …
alternative	regardless of whether;…
purpose	it is necessary; further promote the ;…
explanation	that is ,; the first is; first is the;…
adversative	however , the ; but it is; …
flowing	this is a; which is an; …
consequence	so that the; to ensure that…

Table 2. Chinese functional relations and their corresponding English left-frontier phrases learned by our transfer model. The noun phrases starting with a definite / indefinite word are filtered because they are unlikely to be the transitional phrases.

4.3 Results on SMT with Different Strategies

For this work, we use an in-house decoder to build the SMT baseline; it combines the hierarchical phrase-based translation model (Chiang, 2005; Chiang, 2007) with the BTG (Wu, 1996) reordering model (Xiong et al., 2006; Zens and Ney, 2006; He et al., 2010).

To test the effectiveness of the proposed models, we have compared the translation quality of different integration strategies. First, we adopted only the tagged-flattened rules in the hierarchical translation system. Next, we added the log probability generated by the transfer model as a feature into the baseline features. The baseline features include bi-directional phrase translation probabilities, bi-directional lexical translation probabilities, the BTG re-ordering features, and the language model feature. The tri-gram left-frontier phrase was adopted in the experiment. Then the probability generated by the transfer model with EUC constraint is added. Finally, we incorporated the tagged-flattened rules and the additional transfer model feature together.

Table 3 shows the results of these different integrated strategies. In Table 3, almost all BLEU scores are improved, no matter what strategy is used. In particular, the best performance marked in bold is as high as 1.24, 0.94, and 0.82 BLEU points, respectively, over the baseline system on NIST04, CWMT08 Development, and CWMT08 Evaluation data. The strategy of "**TFS+ Flattened Rule**" is the most stable. Meanwhile the "**Flattened Rule**" achieves better performance than "**TFS**". The merits of "**Flattened Rule**" are two-fold: 1) In training process, the new word alignment upon modified sentence pairs can align transitional expressions to flattened CSS tags; 2) In decoding process, the CSS-based rules are more discriminating than the original rules, which is more flexible than "TFS". From the table, we cannot conclude that the EUC constraint will certainly promote translation quality, but the transfer model performs better with the constraint on most testing sets.

4.4 Analysis of Different Effects of Different N-grams

As mentioned in Section 4.3, we have noted the effectiveness of tri-gram transfer model, which means $n = 2$ in formula (7). In fact, the lengths of common transitional expressions vary from one word to several words. To evaluate the effects of different n-grams for our proposed transfer model, we compared the uni-/bi-/tri-gram transfer models in SMT, and illustrate the results in Fig-

	NIST04	NIST05	NIST06	CWMT08's Dev.	CWMT08's Eval.
Baseline	33.42	31.99	33.88	26.14	23.88
+Flattened Rule	34.54**	32.32	__34.58**__	26.79**	__24.70__**
+TFS (without EUC)	33.93**	32.04	34.40*	26.44	24.58**
+TFS	33.84**	__32.63__*	34.15	__27.08__**	24.65**
+TFS+ Flattened Rule	__34.66**__	32.54	34.52**	26.87**	24.49**

+ **Flattened Rule**: only use the tagged-flattened translation rules

+ **TFS:** only use the transfer model score as an additional feature (based on 3-gramtransitional phrase)

+ **TFS** + **Flattened Rule**: both are used

*: value with * means that it is significantly better than the baseline with p<0.05

**: value with ** means that it is significantly better than the baseline with p<0.01

Table 3. BLEU scores of the testing sets with different integrating strategies

ure 6. In this experiment, the CSS-based translation rules and the CSS-based transfer model are both incorporated. Considering time and computing resources, in the rest of our paper, our analysis is conducted on NIST05 and NIST06.

We choose $n = 0, 1, 2$ in this experiment for that the common English transitional expressions are primarily conjunctions, most of which are less than 4 words. Results in Figure 6 show that the uni-gram and tri-gram transitional expressions seem more fitting for our transfer model. One possible reason is that uni-gram or tri-gram conjunctions are more utilized in an English text. In a conjunction expression list proposed by (Williams, 1983) which summarizes the different kinds of conjunctions based on the work of Halliday and Hassan (1976), we obtain the statistical results on uni-/bi-/tri-gram expressions, which are about 52.1%/16.9%/23.9% respectively.

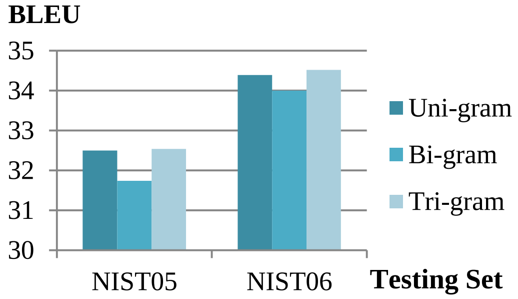


Figure 6. Different translation qualities along with different n-grams for transfer model.

4.5 Experiments on Big Training Data

To further evaluate the effectiveness of the proposed models, we also conducted an experiment on a larger set of bilingual training data from the LDC corpus[7] for translation model and transfer model. The training corpus contains 2.1M sentence pairs with approximately 27.7M Chinese words and 31.9M English words. All the other settings were the same as the SMT experiments of sub-section 4.3. The final BLEU scores on NIST05 and NIST06 are given in Table 4.

The results in Table 4 further verify the effectiveness of our proposed models. The best performance with bold marking scored as high as 0.83 and 0.64 BLEU points, respectively over the baseline system on NIST05 and NIST06 evaluation data.

	NIST05	NIST06
Baseline	35.20	35.52
+Flattened Rule	**36.03****	36.10*
+TFS	35.56*	36.04*
+TFS +Flattened Rule	36.02**	36.16**

+ **Flattened Rule**: only use the tagged-flattened translation rules
+ **TFS:** only use the transfer model score as an additional feature (3-gram transitional phrase)
+ **TFS** + **Flattened Rule**: both are used
*: value with * means that it is significantly better than the baseline with $p<0.05$
**: value with ** means that it is significantly better than the baseline with $p<0.01$

Table 4. BLEU scores on the large-scale training data.

4.6 Translation Examples

Two SMT examples of Chinese-to-English are given in Table 5. We observe that compared to the baseline, our approach has obvious advantages on translating the implicit relations, due to generating translational expressions on target side. Moreover, with the transitional expressions, cohesion of the entire translation improves. Notably, the transitional expressions in this work like "*including, there are, the core of which*" are not linguistic conjunctions. We would like to call them "generalized" conjunctions, because they tie semantic fragments together, analogously to linguistic conjunctions.

5 Related Work

Improving cohesion for complex sentences or discourse translation has attracted much attention in recent years. Such research efforts can be roughly divided into two groups: 1) research on lexical cohesion, which mainly contributes to the selection of generated target words; 2) efforts to improve the grammatical cohesion, such as disambiguation of references and connectives.

In lexical cohesion work, (Gong et al., 2011; Xiao et al., 2011; Wong and Kit, 2012) built discourse-based models to ensure lexical cohesion or consistency. In (Xiong et al., 2013a), three different features were designed to capture the lexical cohesion for document-level machine translation. (Xiong et al., 2013b) incorporated lexical-chain-based models (Morris and Hirst, 1991) into machine translation. They generated the target lexical chains based on the source

[7] LDC category number: LDC2000T50, DC2002E18, LDC2003E07, LDC2004T07, LDC2005T06, LDC2002L27, LDC2005T10 and LDC2005T34.

Source	过去三年中，已有三对染色体完成排序， 包括第二十对、第二十一对和第二十二 对 。
Reference	In the past three years, the sequencing of three chromosomes has been completed, including chromosomes 20 , 21 , and 22 .
Baseline	In the past three years , now has three terms of the completion of the chromosomes , 20 , 21 and 22 .
Improved	In the past three years , **there are** three chromosomes to accomplish , **including** 20 , 21 and 22 .
Source	上述主张构成了一个中国原则的基本涵义，核心是维护中国的主权和领土完整。
Reference	The above-mentioned propositions constitute the basic connotation of this one-china principle with safeguarding china ' s sovereignty and territorial integrity as its core .
Baseline	The above-mentioned propositions constitute the basic meaning of the one-china principle is the core of safeguard china ' s sovereignty and territorial integrity .
Improved	The above-mentioned propositions constitute the basic meaning of the one-china principle , **the core of which** is to safeguard china ' s sovereignty and territorial integrity .

Table 5. Examples of baseline and the improved system outputs.

chains via maximum entropy classifiers, and used the target chains to work on the word selection.

Limited work has been conducted on grammatical cohesion. (Marcu et al., 2000) designed a discourse structure transfer module, but it focused on converting the semantic structure rather than actual translation. (Tu et al., 2013b) provided a Rhetorical-Structure-Theory-based tree-to-string translation method for complex sentences with explicit relations inspired by (Marcu et al., 2000), but their models worked only for explicit functional relations, and they were concerned mainly with the translation integrity of semantic span rather than cohesion. (Meyer and Popescu-Belis, 2012) used sense-labeled discourse connectives for machine translation from English to French. They added the labels assigned to connectives as an additional input to an SMT system, but their experimental results show that the improvements under the evaluation metric of BLEU were not significant. (Nagard and Koehn, 2010) addresses the problems of reference or anaphora resolution inspired by work of Mitkov et al. (1995).

To the best of our knowledge, our work is the first attempt to exploit the source functional relationship to generate the target transitional expressions for grammatical cohesion, and we have successfully incorporated the proposed models into an SMT system with significant improvement of BLEU metrics.

6 Conclusion

In this paper, we focus on capturing cohesion information to enhance the grammatical cohesion of machine translation. By taking the source CSS into consideration, we build bridges to connect the source functional relationships in CSS to target transitional expressions; such a process is very similar to human translating.

Our contributions can be summarized as: 1) the new translation rules are more discriminative and sensitive to cohesive information by converting the source string into a CSS-based tagged-flattened string; 2) the new additional features embedded in the log-linear model can encourage the decoder to produce transitional expressions. The experimental results show that significant improvements have been achieved on various test data, meanwhile the translations are more cohesive and smooth, which together demonstrate the effectiveness of our proposed models.

In the future, we will extend our methods to other translation models, such as the syntax-based model, to study how to further improve the performance of SMT systems. Besides, more language pairs with various linguistic structures will be taken into consideration.

Acknowledgement

We would like to thank Jiajun Zhang for providing the BTG-based hierarchical decoder. The research work has been partially funded by the Natural Science Foundation of China under Grant No. 61333018, the Hi-Tech Research and Development Program (“863” Program) of China under Grant No. 2012AA011101, and also the Key Project of Knowledge Innovation Program of Chinese Academy of Sciences under Grant No. KGZD-EW-501 as well.

References

Jeff A. Bilmes and Katrin Kirchhoff. *Factored language models and generalized parallel backoff*. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers-Volume 2: 4-6.

David Chiang. 2005. *A hierarchical phrase-based model for statistical machine translation*. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pages 263–270.

David Chiang. 2007. *Hierarchical phrase-based translation*. Computational Linguistics, pages 33(2):201–228.

Zhengxian Gong, Min Zhang, and Guodong Zhou. *Cache-based document-level statistical machine translation*, 2011, Edinburgh, Scotland, UK. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 909–919.

Liane Guillou. 2013. *Analysing lexical consistency in translation*. In Proceedings of the Workshop on Discourse in Machine Translation, pages 10–18, Sofia

Michael A.K. Halliday, Hasan R. *Cohesion in English*. 1976. London: Longman.

Zhongjun He, Yao Meng, and Hao Yu. 2010b. *Maximum Entropy Based Phrase Reordering for Hierarchical Phrase-based Translation*. In Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP), pages 555–563.

Annie Louis and Ani Nenkova. 2012. *A coherence model based on syntactic patterns*. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1157–1168, Jeju Island, Korea, July.

William C Mann and Sandra A Thompson. 1988. *Rhetorical structure theory: Toward a functional theory of text organization*. Text, 8(3):243–281.

Ruslan Mitkov, Sung-Kwon Choi, and Randall Sharp. 1995. *Anaphora resolution in Machine Translation*. In Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation.

Thomas Meyer and Andrei Popescu-Belis. *Using sense-labeled discourse connectives for statistical machine translation*, 2012, In Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), pages:129-138.

Jane Morris and Graeme Hirst. 1991. *Lexical cohesion computed by thesaural relations as an indicator of the structure of text*. Comput. Linguist., 17(1):21–48, March.

Ronan L Nagard and Philipp Koehn. 2010, *Aiding pronoun translation with co-reference resolution*, In proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pages 252-261.

Franz J Och and Hermann Ney. 2002. *Discriminative training and maximum entropy models for statistical machine translation*. In Proc. of ACL, pages 295–302.

Kishore Papineni, Salim Roukos, Todd Ward, et al. 2002, *BLEU: a method for automatic evaluation of machine translation*. In proceedings of the 40th annual meeting on association for computational linguistics. pages: 311-318.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. *The Penn Discourse Treebank 2.0*. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

Williams Ray. *Teaching the Recognition of Cohesive Ties in Reading a Foreign*, 1983. Reading in a foreign language, 1(1), pages: 35-52.

Radu Soricut and Daniel Marcu. 2003. *Sentence level discourse parsing using syntactic and lexical information*. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pages 149–156.

Mei Tu, Yu Zhou, and Chengqing Zong. 2013a, *A Novel Translation Framework Based on Rhetorical Structure Theory*. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, short paper, Sofia, Bulgaria, pages 370–374.

Mei Tu, Yu Zhou, Chengqing Zong. 2013b, *Automatically Parsing Chinese Discourse Based on Maximum Entropy*. In The 2nd Conference on Natural Language Processing & Chinese Computing.

Ashish Vaswani, Liang Huang and David Chiang, Huang L, Chiang D. 2012, *Smaller alignment models for better translations: unsupervised word alignment with the l 0-norm*. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1,pages 311-319.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. *Document-level consistency verification in machine translation*. September 2011, Xiamen, China. In Proceedings of the 2011 MT summit XIII, pages 131–138.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. *Maximum entropy based phrase reordering model for statistical machine translation*. In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics, pages 521–528.

Deyi Xiong, Guosheng Ben, Min Zhang, Yajuan Lv, and Qun Liu. 2013 (a). *Modeling lexical cohesion for document-level machine translation*. In Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI-13), Beijing, China, August.

Deyi Xiong, Ding Yang, Min Zhang and Chew Lim Tan, 2013 (b). *Lexical Chain Based Cohesion Models for Document-Level Statistical Machine Translation.* In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages: 1563-1573.

Richard Zens and Hermann Ney. 2006. *Discriminative reordering models for statistical machine translation*. In Proceedings of theWorkshop on Statistical Machine Translation, pages 55–63.

Qiang Zhou, 2004, *Annotation Scheme for Chinese Treebank*, Journal of Chinese Information Processing, 18(4): 1-8.