Natural Language Engineering **21** (2): 227–249. © Cambridge University Press 2013
doi:10.1017/S1351324913000260

Improving syntactic rule extraction through deleting spurious links with translation span alignment

JINGBO ZHU, QIANG LI and TONG XIAO

Natural Language Processing Laboratory, Northeastern University, Shenyang 110819, China
e-mail: zhujingbo@mail.neu.edu.cn, liqiangneu@gmail.com, xiaotong@mail.neu.edu.cn

(*Received 18 September 2012; revised 4 August 2013; accepted 6 August 2013; first published online 6 September 2013*)

Abstract

Most statistical machine translation systems typically rely on word alignments to extract translation rules. This approach would suffer from a practical problem that even one spurious word alignment link can prevent some desirable translation rules from being extracted. To address this issue, this paper presents two approaches, referred to as sub-tree alignment and phrase-based forced decoding methods, to automatically learn translation span alignments from parallel data. Then, we improve the translation rule extraction by deleting spurious links and inserting new links based on bilingual translation span correspondences. Some comparison experiments are designed to demonstrate the effectiveness of the proposed approaches.

1 Introduction

Most statistical machine translation (SMT) models typically learn translation rules from parallel data with word alignments. For example, phrase-based models (Koehn *et al.* 2003) extract from each sentence pair phrase translation pairs that are consistent with its word alignment. Syntax-based models utilize word alignments and parse trees on the source/target side to learn syntactic transformation rules (Galley *et al.* 2004; Huang *et al.* 2006; Chiang 2007). Given a huge amount of bilingual training data, word alignments are usually generated by utilizing a word aligner such as the GIZA++ toolkit (Och and Ney 2000); however, this can introduce some alignment errors. For example, observed from our Chinese–English parallel corpus, some frequent Chinese function words such as '了', '地' and '的' are often misaligned because they have no explicit translational equivalent in English. Word alignment mistakes may be caused by word segmentation errors, and these word alignment mistakes in turn cause negative effects on rule extraction and translation quality.

We observed from our experimental results that even one spurious word alignment can rule out some desirable translation rules and, subsequently, affect the translation quality. Figure 1 depicts a real example with a spurious word alignment link between

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

Table 1. *Some desirable syntactic transformation rules are blocked due to a spurious link between* '了' *and* 'the'

Some blocked tree-to-string rules:
r_1: AS(了) → have
r_2: NN(进口) → the imports
r_3: S (NN:x_1 VP:x_2) → x_1x_2
Some blocked tree-to-string rules:
Some blocked tree-to-tree rules:
r_4: AS(了) → VBZ(have)
r_5: NN(进口) → NP(DT(the) NNS(imports))
r_6 : S(NN : x_1VP : x_2) → S(NP : x_1VP : x_2)
r_7 : VP(AD : x_1VP(VV : x_2AS : x_3)) → VP(VBZ : x_3ADVP(RB : x_1VBN : x_2))

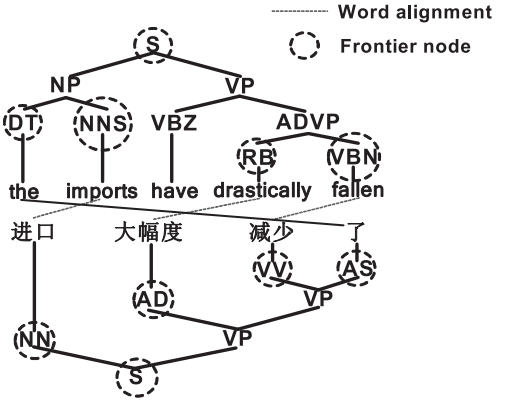


Fig. 1. A real example of the Chinese–English sentence pair with word alignment and both-side parse trees.

a Chinese source word '了' and an English target word '*the*'. Consequently, this spurious link rules out some desirable translation rules shown in Table 1.

To address this challenge, a considerable amount of previous research has been done to improve alignment quality by incorporating statistics and linguistic heuristics or syntactic information into word alignments (Cherry and Lin 2006; DeNero and Klein 2007; May and Knight 2007; Fossum *et al.* 2008; Hermjakob 2009; Liu *et al.* 2010). Since there is poor interaction between word alignment and syntactic rule extraction (DeNero and Klein 2007; Fossum *et al.* 2008), another research direction has been explored to conduct syntactic alignment between both-side parse trees (Tinsley *et al.* 2007; Pauls *et al.* 2010; Sun *et al.* 2010a, 2010b), and implements syntactic rule extraction based on syntactic alignment instead of word alignment. One of the major problems[1] with syntactic alignment is that syntactic divergence between two languages can prevent accurate structure alignments between both-side parse trees and, thus, affect the coverage of the syntactic rule set.

[1] Parse errors can also cause negative effects on syntactic structure alignment.

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

Unlike their efforts, in this work our goal[2] is to remove some spurious word alignment links that would block desirable translation rules from extraction, which in turn improve the extraction of syntactic translation rules. To detect spurious word alignment links, our line of thinking is to utilize the knowledge from span-to-span[3] alignments between each parallel sentence pair, referred to as *translation span alignment* (TSA) in this paper. To achieve this goal, this paper presents two approaches to first automatically building the TSAof each sentence pair in training parallel data, referred to as *sub-tree alignment* and *phrase-based forced decoding* methods, and then improving the extraction of syntactic translation rules by deleting spurious links and inserting new valuable links based on bilingual translation span correspondences.

The proposed approach has several properties, such as (1) TSAs are constructed in an unsupervised learning manner; (2) the proposed approach provides two link operations on word alignment to improve rule extraction, including spurious link deletion and new confident link insertion; (3) the proposed approach is independent of any word-alignment-based rule extraction algorithms used to extract translation rules and easy to implement.

2 Rule extraction with word alignment

In this section, we first briefly discuss translation rule extraction with word alignments. For typical phrase rule extraction (Koehn *et al.* 2003), given a sentence pair (s, t) with a source sentence s and a target sentence t, phrase-based models extract from this sentence pair (s, t) all the phrase translation pairs that are consistent with its word alignment A. We refer to such a phrase pair as an *extractable* phrase pair. Only an extractable phrase pair can form a phrase translation rule.

For syntactic translation rule extraction, some standard algorithms such as the GHKM algorithm (Galley *et al.* 2004) can be used. To ease discussion, we take the tree-to-string model as an example. For each node u in a source parse tree T, tree-to-string syntactic transformation rule extraction begins by identifying a span of the target sentence that satisfies two criteria:

(a) every target word in the span aligns to a source word in the yield of the node u and
(b) no target words in the span align outside the yield of the node u.

If such a target span can be found, this node u is referred to as an *extractable* node. Let us revisit Figure 1. For tree-to-string syntactic rule extraction, nodes NN, AD, VV, AS and S in source Chinese side are extractable nodes under the associated word alignment shown, but the node VP is not an extractable node because two English words '*imports*' and '*drastically*' within its target span align outside its yield.

[2] Our goal is to recover some blocked translation rules instead of directly improving word alignment as done in previous studies, although the removal of spurious links can improve the quality of word alignment.
[3] A span is a sequence of words.

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

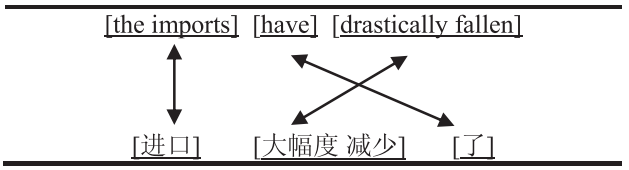


Fig. 2. The TSA of the sentence pair in Figure 1.

In principle, only an extractable tree node can form the root or a leaf of a syntactic translation rule. Since extractable nodes are inferred based on word alignment, spurious word links can possibly rule out some desirable syntactic translation rules and, in turn, causes negative effects on rule quality and translation quality. In the following sections, we will discuss how to learn TSAs to address this challenge.

3 Translation span alignment

3.1 Translation span alignment model

The goal of a typical word alignment model is to identify word-to-word alignments between each parallel sentence pair, whereas the goal of our TSA model is to identify span-to-span alignments between each parallel sentence pair. For each translation span pair, its source (or target) span is a sequence of source (or target) words. Given a source sentence $s = s_1, ..., s_n$, a target sentence $t = t_1, ..., t_m$, and its word alignment A, a translation span pair τ is a pair of source span $(s_i, ..., s_j)$ and target span $(t_p, ..., t_q)$:

$$\tau = (s_i^j \leftrightarrow t_p^q)$$

where τ indicates that the source span $(s_i, ..., s_j)$ and the target span $(t_p, ..., t_q)$ are translational equivalent.

In the TSA model, we do not require that τ is consistent with the word alignment A. In other words, τ is possibly not an extractable phrase pair from the viewpoints of phrase rule extraction algorithms. Due to this reason, it is impossible for us to directly apply typical phrase rule extraction techniques to identify translation span pairs based on word alignment. Figure 2 depicts a TSA of the example sentence pair in Figure 1. In this example TSA, we can see that the translation span pair {[进口]↔[*the imports*]} is not an extractable phrase pair because the English word '*the*' inside the target span is originally aligned to the Chinese word '了' outside the source span. In such a case, it cannot be extracted successfully by a phrase-based system. Next, we will discuss how to automatically build TSAs from parallel sentences.

Following the formulation adopted in statistical word alignment (Brown *et al.*, 1993; Vogel *et al.*, 1996), the TSA between a source sentence s and a target sentence t is modeled as follows:

$$\Pr(t|s) = \sum_{d \in D(s,t)} \Pr_{\theta}(t, d|s) \quad (1)$$

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

where $D(s, t)$ is a set of derivations that each of them can align s with t, and $\Pr_\theta(t, d|s)$ is the *alignment model* that explains the alignment confidence of d; and θ is the associated set of parameters that is learned from the parsed parallel corpus.

In the TSA model, a derivation d can be defined as a sequence of translation span pair applications, i.e., $d = \tau_1 \oplus ... \oplus \tau_k$, in which τ_i indicates the ith translation span pair of (s, t). Afterwards, the best derivation for a given sentence pair (s, t) is obtained by

$$\widehat{d} = \arg\max_{d \in D(s,t)} \Pr_\theta(t, d|s) \tag{2}$$

In principle, the best derivation obtained from (2) implies the best TSA between the given sentence pair (s, t). Since directly calculating $\Pr_\theta(t, d|s)$ is infeasible due to the extremely large number of derivations, we need to decompose this model to obtain a feasible solution. Assume that translation span pairs are conditionally independent given s, then $\Pr_\theta(t, d|s)$ can be decomposed as follows. For conciseness, we will drop the subscript θ from now on:

$$\Pr_\theta(t, d|s) \approx \prod_{\tau \in d} \Pr(\tau|s) \tag{3}$$

Notice that the derivation d yields the given target sentence t. Hereafter, our goal is to find the best derivation that implies the best TSA, referred to as *TSA generation.*

3.2 Sub-tree alignment method

The first intuition for TSA generation is to inject knowledge about syntactic constraints into a word alignment model. A straightforward solution is to adopt some *sub-tree alignment* techniques (Groves *et al.* 2004; Tinsley *et al.* 2007) to build the TSA of each sentence pair with the help of both-side parse trees. Along this line, we can re-model TSA between a given sentence pair (s, t) as a new task of sub-tree alignment between its associated parse tree pair (S, T). In sub-tree alignment methods, the best derivation should yield the given target parse tree T.

The goal of a sub-tree alignment algorithm is to pair up the sub-trees across both-side parse trees, whose lexical leaf nodes covered are semantically translational equivalent. To achieve this goal, a large amount of research has been done to build sub-tree alignments across both-side parse trees by using various methods, such as heuristic rules (Groves *et al.* 2004), score functions based on lexical similarity and co-occurrence (Tinsley *et al.* 2007), tree kernel based score functions and supervised classification techniques (Sun *et al.* 2010a, 2010b).

In this work, we utilized an effective approach proposed by Tinsley *et al.* (2007) to implement sub-tree alignment over its both-side constituent-based parse trees. This approach does not rely on some heuristic rules and any supervised classification techniques. This sub-tree alignment approach defines three well-formed criteria (Tinsley *et al.* 2007) with respect to a tree pair:

(a) a node can only be linked once;

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

Table 2. *TSA generated from sub-tree alignment shown in Figure 3*

(1) [进口 大幅度 减少 了]↔[the imports have drastically fallen]
(2) [进口]↔[the imports]
(3) [大幅度 减少 了]↔[have drastically fallen]
(4) [大幅度]↔[drastically]
(5) [减少]↔[fallen]

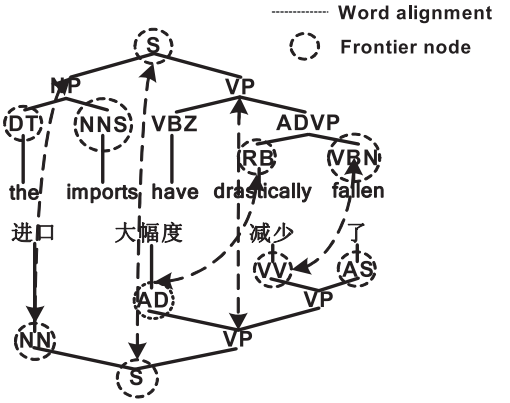


Fig. 3. Sub-tree alignment of the example tree pair in Figure 1.

(b) descendants of a source-linked node may only link to descendants of its target-linked counterpart;
(c) ancestors of a source-linked node may only link to ancestors of its target-linked counterpart.

These criteria can forbid alignments between constituents from crossing each other. Given a tree pair $<S, T>$, the sub-tree alignment approach (Tinsley *et al.* 2007) first takes all the links between the sub-tree pairs as alignment hypotheses, i.e., the Cartesian product of the two sub-tree sets:

$$\psi(S) \times \psi(T) = \{s_1, ..., s_I\} \times \{t_1, ..., t_J\}$$

where $\Psi(S) = s_{1,...}, s_{i,...}, s_I$ and $\Psi(T) = t_{1,...}, t_{j,...}, t_J$ are defined as the sub-tree spaces of the source and target parse trees, respectively. A sub-tree space contains all possible sub-trees of the given parse tree.

Tinsley's algorithm iteratively selects the best link with maximum score, in which a score function[4] based on the lexical translation probabilities dumped by GIZA++ is used to compute the confidence score of each sub-tree pair. All zero-scored hypotheses are blocked before the algorithm proceeds. The selection procedure iteratively selects the highest scoring link, blocking all hypotheses that contradict the chosen links, until no non-blocked hypotheses remain.

[4] We only present the best results achieved by the score function *skip2_s1_span1* (Tinsley *et al.* 2007).

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

Figure 3 depicts the sub-tree alignment of the example tree pair shown in Figure 1. It is easy to produce TSA based on sub-tree alignment. For each aligned sub-tree pair $(s_i, t_j) \in \Psi(S) \times \Psi(T)$ selected by a sub-tree alignment algorithm, the yields[5] of s_i and t_j can form a translation span pair. Table 2 depicts TSAs generated from sub-tree alignment shown in Figure 3. For example, we can obtain a translation span pair {[进口]↔[*the imports*]} from an aligned sub-tree pair {<NN(进口)>, <NP(DT(*the*), NNS(*imports*))>}.

3.3 Phrase-based forced decoding method

In practice, however, there are some crucial issues with directly utilizing sub-tree alignment techniques for TSA generation. First, sub-tree alignment models still remain limited by word alignment models that underlie it. Second, sub-tree alignment models are essentially inferred from knowledge about constituent boundary-based syntactic constraints and word alignments of the given sentence/tree pair. Third, since sub-tree alignment models are restricted to respect both the source and target parse trees, many desirable node alignments could be ruled out when the syntax of the two languages diverges and parse errors occur, and node alignment quality thus drops.

To address these challenges, we present another effective TSA generation solution in which a *phrase-based forced decoding* technique is adopted to search the best derivation defined in (2). As mentioned in Section 3.1, this best derivation implies the TSA of a given sentence pair (s, t). Forced decoding techniques aim to inject global knowledge learned from the whole training data into a TSA model, and construct a TSA preferred by the translation model. It can alleviate the disconnection problem[6] between the TSA model and the translation model.

In this work, we apply a phrase-based forced decoding technique instead of syntax-based forced decoding techniques because phrase-based models can achieve the state-of-the-art translation quality with a large amount of training data, and are not limited by any constituent boundary-based constraints for decoding. Our preliminary experiments showed that more than 60% of the training sentence pairs cannot be decoded successfully by using a tree-to-string forced decoding technique. Next, we will introduce some detailed implementations of phrase-based forced decoding techniques.

Given a source sentence s, the goal of typical phrase-based decoding (Koehn *et al.* 2003) typically is to find a target string t^* by the following equation

$$t^* = \arg\max_{t} \Pr_{\theta}(t|s) \tag{4}$$

where $\Pr(t \mid s)$ is the probability that t is the translation of the given source sentence s. To model the posterior probability $\Pr(t \mid s)$, most of the state-of-the-art SMT

[5] Lexical leaf nodes of s_i and t_j are used to build the TSA.

[6] Previous work (Fraser and Marcu 2007) has discussed at length the disconnection problem between the word alignment model and the translation model.

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

systems utilize the log–linear model proposed by Och and Ney (2002):

$$\Pr(t|s) = \frac{\exp(\sum_{i=1}^{|H|} \lambda_i \bullet h_i(s,t))}{\sum_{t'} \exp(\sum_{i=1}^{|H|} \lambda_i \bullet h_i(s,t'))} \tag{5}$$

where $H = \{h_i(s,t)|i = 1,...,|H|\}$ is a set of features and λ_i is the feature weight corresponding to the ith feature. Note that $h_i(s,t)$ can be regarded as a function that maps each pair of source string s and target string t into a real number, and λ_i can be regarded as the contribution of $h_i(s,t)$ to $\Pr(t \mid s)$. Ideally, λ_i indicates the pairwise correspondence between the feature $h_i(s,t)$ and the overall score $\Pr(t \mid s)$. A positive value of λ_i indicates a correlation between $h_i(s,t)$ and $\Pr(t \mid s)$, while a negative value indicates an inversion correlation.

Different from typical phrase-based decoding techniques, we require that each candidate derivation d must yield the same target sentence t during the phrase-based forced decoding process. Only such a candidate derivation can survive during the forced decoding process. Formally, given a sentence pair (s, t), the phrase-based forced decoding technique aims to search for the best derivation[7] d^* among all consistent derivations that convert the given source sentence s into the given target sentence t with respect to the current translation model induced from the training data. This can be expressed by

$$d^* = \arg\max_{d \in D(s,t) \wedge \mathrm{TGT}(d)=t} \Pr_{\theta}(\mathrm{TGT}(d)|s) \tag{6}$$

where $D(s,t)$ is the set of candidate derivations that transform s to t, and $\mathrm{TGT}(d)$ is a function that outputs the yield of a derivation d. θ indicates parameters of the translation model learned from the parallel corpus.

Here, we describe how to generate the best derivation d^* for each sentence pair in training corpus C based on the phrase-based forced decoding technique, which performs via the following four steps:

1. Extract translation rules (i.e., grammar induction) needed for a specific phrase-based SMT paradigm M from bilingual training corpus C;
2. Perform minimum error rate training (MERT) on a development data set to obtain a set of optimized feature weights;
3. For each $\{s,t\} \in C$, translate s into accurate t based on M with translation rules learned in step 1 and feature weights optimized in step 2;
4. For each $\{s,t\} \in C$, output the derivation with the maximum model score (defined by (5)) produced in step 3 as d^*.

Notice that most standard features such as language model, bidirectional lexical and phrase-based translation probabilities and reordering models are used to perform MERT in step 2, which is similarly done in typical phrase-based models (Koehn *et al.* 2003). Two reordering models we used include the maximum entropy (ME)-based lexicalized reordering model (Xiong *et al.*, 2006) and the monotone, swap and discontinuous (MSD) reordering model (Tillman 2004; Galley and Manning 2008).

[7] The best derivation is of the highest model score with respect to the translation model.

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

Table 3. *Forced decoding-based TSA generation on the example sentence pair in Figure 1*

c = *进口 大幅度 减少 了*
e = *the imports have drastically fallen*
The best derivation d^* produced by forced decoding:
r_1: 进口 → the imports
r_2: 大幅度 减少 → drastically fallen
r_3: 了 → have
Resulting TSA from d^* as follows:
• [进口]↔[the imports]
• [大幅度 减少]↔[drastically fallen]
• [了]↔[have]

To speed up phrase-based forced decoding in step 3, two optimization techniques can be used as follows.

First, since forced decoding techniques only focus on derivations that yield the same target sentence t, the language model score can be regarded as a constant value when calculating the overall model score of each candidate derivation. In such a case, we can drop the language model feature to speed up forced decoding without adverse effects. In the following comparison experiments, our forced decoding method will consider all translation features except language model feature used by the baseline system.

Second, another simple and effective speed optimization technique we adopted is to decode with some phrase translation rules whose target span matches a span (substring) of the given target sentence t, because as described in step 3, all partial hypotheses that do not match any sequence in t will be discarded during the phrase-based forced decoding process.

The best derivation d^* produced by phrase-based forced decoding can be viewed as a sequence of translation steps (i.e., phrase translation rules), expressed by

$$d* = r_1 \oplus r_2 \oplus \cdots \oplus r_k,$$

where r_i indicates a phrase rule used to form d^*. $\oplus$ is a composition operation that combines rules $\{r_{1,\ldots}, r_k\}$ together to produce the target translation.

As mentioned above, the best derivation d^* respects the input sentence pair (s, t). It means that for each phrase translation rule r_i used by d^*, its source (or target) side exactly matches a span of the given source (or target) sentence. The source side src(r_i) and the target side tgt(r_i) of each phrase translation rule r_i in d^* form a translation span pair src(r_i) ↔ tgt(r_i) of (s, t). In other words, the TSA of (s, t) is a set of translation span pairs generated from phrase translation rules used by the best derivation d^*. For better understanding, we summarize in Figure 4 the forced decoding-based TSA generation algorithm proposed in this paper. The forced decoding-based TSA generation on the example sentence pair in Figure 1 is shown in Table 3.

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

1. For each $\{s,t\} \in C$, output the best derivation d^* by performing phrase-based forced decoding technique;
2. Build a TSA of each sentence pair $\{s, t\}$ from its best derivation d^*, in which each rule r in d^* is used to form a translation span pair $\{src(r) \leftrightarrow tgt(r)\}$.

Fig. 4. Forced decoding-based TSA generation. src(r) and tgt(r) indicate the source and target sides of rule r.

The phrase-based forced decoding technique can implement a fast search to construct TSAs on the training data. We have empirically shown that our proposed approach requires only about three CPU hours to re-align 1.06M Chinese–English sentence pairs[8] used in our evaluation, and less than thirty minutes for spurious link deletion and new link insertion on this corpus. In practice, there is a risk of forced decoding failure on some training sentence pairs due to beam search-based pruning techniques used for decoding. We observed in our experiments that 10.5% of the training sentence pairs failed in the phrase-based forced decoding, when the beam size is set to 30 in our experimental settings. In such a case, we used their original word alignments for translation rule extraction. Actually increasing the beam size can reduce the decoding failure rate, but would slow down the forced decoding speed.

4 Better rule extraction with TSA

As mentioned in Section 3.3, in the phrase-based forced decoding method, a translation span pair is a phrase translation rule. Since a phrase rule can be extracted from multiple sentence pairs with different word alignments, we cannot assign a consistent word alignment to a translation span pair. In other words, two same phrase rule extracted from two different sentence pairs may have different word alignments. Due to this reason, we are interested in utilizing the knowledge about boundaries of a translation span pair to determine spurious links that could fail to extract some desirable rules.

To better understand the task that we will describe in this section, we first introduce a definition of *inconsistent with a TSA*. Given a sentence pair (s, t) with the word alignment A and the TSA P, we call a link $(s_i, t_j) \in A$ *inconsistent* with P, if s_i andt_j are covered respectively by two different translation span pairs in P, and vice versa.

$(s_i, t_j) \in A$ inconsistent with $P \leftrightarrow$

$$\exists \tau \in P : s_i \in \text{src}(\tau) \wedge t_j \notin \text{tgt}(\tau) \text{ OR } \exists \tau \in P : s_i \notin \text{src}(\tau) \wedge t_j \in \text{tgt}(\tau)$$

where src(τ) and tgt(τ) indicate the source and target span of a translation span pair τ.

By the above definition, a link $(s_i, t_j) \in A$ is a spurious link if it is inconsistent with the given TSA. Table 4 shows that an original link (4→1) is covered by two different translation span pairs ([4,4]↔[3,3]) and ([1,1]↔[1,2]), respectively. In such a case, we think that this link (4→1) is a spurious link according to this TSA, and should

[8] May and Knight (2007) and Fossum *et al.* (2008) reported that their methods require 400 and 18 CPU hours to re-align 330k Chinese–English sentence pairs, respectively.

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

Table 4. *A sentence pair with the original word alignment (WA) and the generated translation span alignment (TSA)*

Source	Target	WA	TSA
1: 进口	1: the	1→2	[1,1]↔[1,2]
2: 大幅度	2: imports	2→4	[2,3]↔[4,5]
3: 减少	3: have	3→5	[4,4]↔[3,3]
4: 了	4: drastically	4→1	
	5: fallen	(null)→3	

be removed before syntactic translation rule extraction. Deleting such a spurious link can avoid blocking some desirable rules from being extracted, resulting in an enlarged coverage of the translation rule set.

Besides deleting spurious links, we are also interested in detecting new valuable links based on the TSA. Given a resulting TSA P, there are four different types of translation span pairs, such as one-to-one, one-to-many, many-to-one and many-to-many cases. For example, the TSA shown in Table 4 contains a one-to-one span pair ([4,4]↔[3,3]), a one-to-many span pair ([1,1]↔[1,2]) and a many-to-many span pair ([2,3]↔[4,5]). As mentioned above, each translation span pair does not have a consistent word alignment. However, we can learn a confident link from a one-to-one translation span pair that is preferred by the translation model in the forced decoding-based TSA generation approach. If such a confident link does not exist in the original word alignment, we consider it as a new valuable link. Adding some new valuable links can improve the precision of translation rule extraction and possibly alleviate a practical problem that unaligned words can result in a huge number of extractable rules due to null alignments involved.

A natural way is to use TSAs to directly delete some spurious links and add some new confident links before translation rule extraction, which in turn improves rule quality and translation quality. In other words, if a desirable translation rule was blocked due to some spurious links, we will output this translation rule. Let us revisit the example in Figure 1 again. The blocked tree-to-string r_5 can be extracted successfully after deleting the spurious link (了, *the*), and a new tree-to-string rule r_3 can be extracted after adding a new confident link (了, *have*) that is inferred from a one-to-one translation span pair [4,4]↔[3,3].

5 Evaluation

In this section, we describe our means of evaluating the resulting TSAs to improve translation rule extraction and translation quality in SMT tasks.

5.1 Baseline systems

We utilized the NiuTrans open-source SMT toolkit (Xiao *et al.* 2012) to build all baseline systems in our comparison experiments. Syntax-based baseline systems were constructed based on tree-to-string and tree-to-tree models, respectively. The feature

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

design is mainly based on a state-of-the-art MT system described in Marcu *et al.* (2006). For example, we use the bidirectional lexical and phrase-based translation probabilities (four features), a syntactic feature – root normalized conditional probability, an n-gram language model, target word penalty, rule penalty, and three binary features – lexicalized rule, low-frequency rule and composed rule. In both baseline systems, for syntactic (tree-to-tree and tree-to-string) translation rule extraction, minimal GHKM (Galley *et al.*, 2004) rules are first extracted from the bilingual corpus whose source and/or target sides are parsed using the Berkeley parser (Petrov *et al.* 2006). To improve the tree-to-tree rule coverage, the parse trees on both language sides are binarized in a head-out manner (Wang *et al.*, 2007). The composed rules are then generated by composing two or three minimal rules. To avoid the overfitting problem and to speed up the system, we restrict the rule composing to the rules of at most depth three, and having at most five terminals and four non-terminals.

The third baseline system is a phrase-based system with two reordering models such as the ME-based lexicalized reordering model (Xiong *et al.*, 2006) and the MSD reordering model (Tillman 2004; Galley and Manning 2008). In addition to the reordering features, we use most standard features adopted in current state-of-the-art phrase-based systems (Koehn *et al.* 2003), including bidirectional phrase translation probabilities, bidirectional lexical weights, an *n*-gram language model, target word penalty and phrase penalty. In this baseline system, all phrase pairs are limited to having source length of at most five, and the reordering limit is set to eight by default.[9] The phrase-based baseline system is also used to implement the phrase-based forced decoding method with optimization techniques mentioned in Section 3.3.

5.2 Experimental setup

In Chinese-to-English translation tasks, our bilingual data consist of 1.06M sentence pairs selected from the NIST portion of the bilingual data provided within NIST MT 2008 evaluation. The GIZA++ tool[10] was used to perform the bidirectional word alignment between the source and target sentences, and the final alignment was generated using the heuristic '*grow-diag-final-and*' method, referred to as the *baseline* method in the following experiments. A 5-gram language model was trained on the Xinhua portion of English Gigaword corpus in addition to the target side of the bilingual data. We used the NIST 2003 MT evaluation set as our development set (919 sentences) and a large-scale test set (4,529 sentences) of the combination of three NIST evaluation sets of MT04, MT05 and MT06. The translation quality was evaluated in terms of case-insensitive IBM version BLEU metric.

In English-to-Chinese translation tasks, we begin with a training bilingual corpus of English–Chinese bitexts that consists of 1.0M sentence pairs provided by

[9] Our in-house experimental results showed that the NiuTrans phrase-based system performs slightly better than Moses (Koehn *et al.* 2007) on Chinese–English translation tasks.
[10] http://code.google.com/p/giza-pp/

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

Table 5. *Rule sizes and IBM-BLEU4 (%) scores of various methods in tree-to-tree and tree-to-string Chinese-to-English translation on the NIST development set and test set. Here,* * *indicates significantly better on test performance at the* $p = 0.05$ *level*

Method		No. of rules	Dev set	Test set
Tree-to-tree	Baseline	14,603,182	32.09	30.03
	Sub-tree alignment	21,088,087	32.41	30.85*
	Forced decoding	14,925,782	32.58	30.77*
Tree-to-string	Baseline	29,237,206	34.54	32.58
	Sub-tree alignment	47,985,824	34.45	32.30
	Forced decoding	31,157,254	35.08	33.18*

CWMT2011 MT evaluation.[11] Similarly, the baseline word alignment is implemented by using the GIZA++ tool, also referred to as the *baseline* method in the following experiments. A 5-gram language model was trained on the Xinhua portion of Chinese Gigaword corpus in addition to the target side of the bilingual data. We used the CWMT2011 MT evaluation development set (500 sentences) for weight tuning and test set (1359 sentences) for translation performance evaluation. The translation quality was also evaluated in terms of case-insensitive IBM version BLEU metric.

All three baseline systems described above used a CKY-style decoder with beam search and cube pruning (Huang and Chiang 2007) to decode new source sentences. The beam size for beam search was set to 30. All features were linearly combined and their weights were optimized by performing MERT (Och 2003). In the following performance comparison, statistical significance test is conducted using the bootstrap resampling method proposed in Koehn (2004).

5.3 Syntax-based translation

We first investigate the effect of both proposed TSA methods on the translation quality of tree-to-tree models, which consider both-side parse trees for syntactic translation rule extraction. Tables 5 and 6 show that the sub-tree alignment and the forced decoding methods can yield significant improvements on translation quality on both Chinese-to-English and English-to-Chinese test sets. The sub-tree alignment method is to improve syntactic rule extraction with the help of building node alignments across both-side parse trees that are useful indicators for tree-to-tree translation rule extraction (Tinsley *et al.* 2007; Sun *et al.* 2010a, 2010b). In the forced decoding-based method, the resulting TSAs are essentially optimized by the translation model, which is used to detect spurious alignment links. Experimental results show that spurious link deletion based on such TSAs can benefit translation quality of tree-to-tree models.

[11] http://mt.xmu.edu.cn/cwmt2011/

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

Table 6. *Rule sizes and IBM-BLEU4 (%) scores of various methods in tree-to-tree and tree-to-string English-to-Chinese translation on the CWMT development set and test set. Here,* * *indicates significantly better on test performance at the* $p = 0.05$ *level*

Method		No. of rules	Dev set	Test set
Tree-to-tree	Baseline	8,647,996	18.02	17.23
	Sub-tree alignment	12,832,407	18.43	17.74*
	Forced decoding	9,056,934	18.42	17.76*
Tree-to-string	Baseline	18,596,458	19.38	18.25
	Sub-tree alignment	27,397,402	19.34	18.01
	Forced decoding	19,996,670	19.71	18.76*

Table 7. *Rule precision, recall and F1-score of the tree-to-string rule set extracted from 200 sentence pairs of Chinese–English data*

Alignment	Precision %	Recall %	F1 %	Total number (non-unique)
Gold annotation	100	100	100	55823
Baseline (GIZA++)	35.62	38.37	36.94	60123
Sub-tree alignment	28.26	35.69	31.54	70447
Forced decoding	35.18	39.09	37.03	62012

However, when compared with the tree-to-string baseline method, as shown in Tables 5 and 6, the forced decoding method achieves significantly better improvements on translation quality, while the sub-tree alignment method yields unsatisfactory performance in both Chinese-to-English and English-to-Chinese translation tasks. Since the sub-tree alignment method utilizes both-side constituent boundary-based syntactic constraints of each sentence pair, it would result in negative effects on tree-to-string translation models which do not consider target-side parse trees for syntactic translation rule extraction.

To analyze why the sub-tree alignment method yields unsatisfactory performance in tree-to-string translation, we further explore the impact of each TSA method upon rule quality in tree-to-string translation. To achieve this goal, we compare the rule precision, recall and F1-measure of the rule set extracted from our hypothesized word alignments against the rule set extracted from gold word alignments. Two hundreds randomly chosen and manually aligned Chinese–English sentence pairs were used to assess tree-to-string rule quality after incorporating TSA into word alignment.

Table 7 shows the total non-unique number of tree-to-string translation rules extracted from each alignment, the corresponding rule precision, recall and F1-measure. Both TSA methods increase the size of the extracted rule set compared with the baseline method. The forced decoding method achieves better performance on rule recall and F1-measure, and drops a little on rule precision. However, the sub-tree alignment method results in negative effects on the quality of tree-to-string translation rules. As mentioned in Section 3.2, the sub-tree alignment method

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

Table 8. *Rule sizes and IBM-BLEU4 (%) scores of various methods in English-to-Chinese and Chinese-to-English phrase-based translation on development set and test set. Here,* * *indicates significantly better on test performance at the* $p = 0.05$ *level*

	Method	No. of rules	Dev set	Test set
Chinese-to-English	Baseline	47,364,645	36.22	34.10
	Sub-tree alignment	57,340,131	36.17	34.08
	Forced decoding	49,729,974	36.63	34.63*
English-to-Chinese	Baseline	20,803,712	22.96	23.18
	Sub-tree alignment	25,497,209	22.90	23.13
	Forced decoding	22,496,821	23.85	23.74*

Table 9. *The size of training/development/test sets of each minority language–Chinese pair*

Language pair	Training set (no. of sentences)	Development set (no. of sentences)	Test set (no. of sentences)
Kazakh–Chinese	50,000	300	400
Kirgiz–Chinese	50,000	300	400
Mongolian–Chinese	67,288	300	700
Tibetan–Chinese	49,551	300	350
Uyghur–Chinese	50,000	300	400

utilizes bilingual constituent boundary-based constraints to produce the TSA of each sentence pair. It seems that directly injecting bilingual syntactic constraints into word alignment can possibly result in negative effects on tree-to-string translation rule extraction. A similar conclusion was also drawn by Koehn *et al.* (2003) in phrase-based translation tasks. Interestingly, Table 7 shows that the rule precision, recall and F1-score of each automatic method are still low. About two-thirds of desirable tree-to-string translation rules are missed by various methods due to word alignment errors. It means that there is still much room for our future work to improve translation rule extraction for SMT systems.

5.4 Other translation comparisons

In this section, we first investigate the effectiveness of both proposed TSA methods on the translation quality of a typical phrase-based model in English-to-Chinese and Chinese-to-English translation tasks, as shown in Tables 8 and 9. Both TSA methods can increase the size of the phrase translation table. For both translation tasks, the forced decoding method yields significant improvements on translation performance, while the sub-tree alignment method results in slightly lower performance than the baseline method. It seems that directly injecting bilingual syntactic constraints into word alignment can result in negative effects on phrase-based translation.

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

Table 10. *Rule sizes and IBM-BLEU4 (%) scores of various methods in minority language-to-Chinese phrase-based translation on development set and test set. Here,* * *indicates significantly better on test performance at the* $p = 0.05$ *level*

Method		No. of rules	Dev set	Test set
Kazakh-to-Chinese	Baseline	2,447,104	28.47	28.00
	Forced decoding	2,483,132	28.77	29.10*
Kirgiz-to-Chinese	Baseline	2,435,528	39.71	37.60
	Forced decoding	2,469,289	40.29	38.11*
Mongolian-to-Chinese	Baseline	2,323,569	17.22	14.18
	Forced decoding	2,399,867	17.61	14.73*
Tibetan-to-Chinese	Baseline	1,198,202	37.66	36.13
	Forced decoding	1,398,074	38.64	37.05*
Uyghur-to-Chinese	Baseline	2,802,299	41.68	40.10
	Forced decoding	2,855,387	42.01	40.90*

Table 11. *Rule sizes and IBM-BLEU4 (%) scores of various methods in English-to-Chinese and Chinese-to-English hierarchical phrase-based translation on development set and test set. Here,* * *indicates significantly better on test performance at the* $p = 0.05$ *level*

	Method	No. of rules	Dev set	Test set
Chinese-to-English	Baseline	143,759,540	36.50	34.35
	Sub-tree alignment	141,856,343	36.45	34.31
	Forced decoding	156,447,366	36.95	34.89*
English-to-Chinese	Baseline	76,272,925	23.20	23.13
	Sub-tree alignment	74,770,179	23.23	23.08
	Forced decoding	77,575,686	23.59	23.71*

We also further investigate the effect of the proposed TSA method on the translation between minority languages and Chinese. The training/development/test sets come from CWMT2011 MT evaluation as shown in Table 9, involving five minority languages, such as Kazakh, Kirgiz, Mongolian, Tibetan and Uyghur. Because there are not any publicly-available syntactic parsers for these minority languages, we choose to demonstrate the effectiveness of the forced decoding method for phrase-based translation, compared with the baseline method. As shown in Table 10, the forced decoding method not only increases the size of the extracted phrase translation table but also improves the translation performance for each language pair, compared with the baseline method. Accordingly, our experimental results have shown that the forced decoding-based TSA model can benefit rule quality by removing spurious links, which thus in turn improves translation performance.

Finally, we estimate the effectiveness of both proposed TSA methods on the translation quality of a hierarchical phrase-based model (Chiang 2007) in English-to-Chinese and Chinese-to-English translation tasks, as shown in Table 11. For the hierarchical phrase-based model, we follow the general framework of SCFG where

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

Table 12. *Word alignment precision, recall and F1-score of various methods on 200 sentence pairs of Chinese–English data*

Method	Precision %	Recall %	F1 %	Del/sent	Add/sent
Baseline	83.07	75.75	79.25	–	–
Sub-tree alignment	80.75	68.73	74.26	6.6	7.1
Forced decoding	84.01	75.46	79.51	1.5	1.1

a grammar rule has three parts – a source-side, a target-side and alignments between source and target non-terminals. To learn SCFG rules from word-aligned sentences, we choose the algorithm proposed in Chiang (2007) and estimate the associated feature values as in the phrase-based system. This hierarchical phrase-based baseline system is built using the NiuTrans open-source SMT toolkit (Xiao *et al.* 2012). Similar to performance comparisons of phrase-based translation models shown in Table 8, for both Chinese-to-English and English-to-Chinese translation tasks, the forced decoding method yields significant better improvements on translation performance, while the sub-tree alignment method results in slightly lower translation performance than the baseline hierarchical phrase-based method.

5.5 Effects on word alignment

Our goal is to detect some spurious word alignment links for improving the translation rule extraction, instead of directly improving the overall quality of word alignments. However, it is also interesting to investigate the effect of the proposed TSA method on word alignments. Here, we examine the effectiveness of the phrase-based forced decoding method for TSA generation because this method results in significantly better improvements on the translation quality of all systems in previously described experiments. We designed some experiments to evaluate our alignment quality against gold-standard annotations. To achieve this goal, there are 200 randomly chosen and manually aligned Chinese–English sentence pairs used to assessing the word alignment quality after incorporating TSA into word alignment. For word alignment evaluation, we calculated precision, recall and F1-score over word alignment.

Table 12 depicts the word alignment performance of both proposed methods and the number of spurious links and new links detected. Notice that the baseline is the GIZA++ alignment method. We apply the TSAs produced by the sub-tree alignment-based and the forced decoding-based methods to refine the baseline word alignments, respectively, involving spurious link deletion and new link insertion operations.

Table 12 shows that the sub-tree alignment-based method causes negative effects on word alignment precision, recall and F1-score performances, while the forced decoding-based method can yield improvements on precision and F1-score, only causing a little negative effect on recall performance. Table 12 shows that there are average 1.5 spurious links deleted and 1.1 new valuable links added by the forced

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

decoding method, and average 6.6 spurious links deleted and 7.1 new links added by the sub-tree alignment method.

To get more intuition on how to apply TSA methods to revise original word alignments, we present some illustrations of word alignments conducted by the baseline method and the proposed forced decoding method in Table 13. In case 1, the forced decoding method removed two original links {17-20 19-21} based on two generated translation span pairs [,]↔[,] | {17-31} and [暑假]↔[during the summer vacation] | {19-39 19-40 19-41 19-42}, because the two deleted links violated the two translation span pairs. The one-to-one generated translation span pair [,]↔[,] | {17-31} indicates the new link {17-31}. Notice that the proposed TSA model only considers one-to-one translation span pairs for new link generation.

Similarly, in case 2, the forced decoding method deleted two baseline links {2-7 17-12} and added a new link {17-24} according to two generated translation span pairs [采取了]↔[conducted] | {1-5 2-5} and [深受]↔[received] | {17-24}. In case 3, the forced decoding method deleted a baseline link {10-23} and added a new link {10-25} based on a one-to-one translation span pair [集中]↔[gather] | {10-25}. In case 4, a baseline link {13-2} was removed and a new link {13-16} was inserted based on a translation span pair [,]↔[and] | {13-26}. In case 5, our method deleted a baseline link {13-5} based on a one-to-many translation span pair [驱车]↔[traveled by vehicles] | {13-19 13-20 13-21}.

6 Related work

Previous studies have made great efforts to incorporate statistics and linguistic heuristics or syntactic information into word alignments, so as to improve alignment quality and translation quality (Ittycheriah and Roukos 2005; Liu *et al.* 2005, 2010; Taskar *et al.* 2005; Cherry and Lin 2006; Moore *et al.* 2006; DeNero and Klein 2007; May and Knight 2007; Fossum *et al.* 2008; Hermjakob 2009). One of representative work is to use discriminative models for word alignment. For example, Fossum *et al.* (2008) used a discriminatively trained model to identify and delete incorrect links from original word alignments in order to improve string-to-tree transformation rule extraction. This discriminative model incorporates four types of features such as lexical and syntactic features. Since most state-of-the-art word alignment systems use HMM or IBM model 4-based methods, which tend to misalign words separated with a long distance, some previous studies (Genzel 2010; Khalilov and Sima'an 2011) used source-side reordering as a preprocessing step to improve word alignment. Besides, the word alignment combination that combines links from multiple word alignments was also adopted to improve alignment quality in some previous studies (Deng and Zhou 2009). In this work, we propose an approach that incorporates TSAs into word alignments to delete spurious links and add new valuable links, which can in turn improve word alignment quality and translation quality.

Since there is the poor interaction between word alignments and syntactic transformation rule extraction (DeNero and Klein 2007; Fossum *et al.* 2008), some previous work directly models the syntactic correspondence in the training data for syntactic rule extraction (Imamura 2001; Groves *et al.* 2004; Tinsley *et al.*

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

Table 13. *Five examples of word alignments conducted by the baseline method and the proposed forced decoding-based TSA method. The differences between word alignment results of both methods are marked in bold*

Case 1	Chinese	为 改变 这 一 状况, $date $number 名 教师 组成 宣讲 小组,[17] 利用 暑假 [19] 深入 农村 牧区, 给 群众 摆事实 讲道理.
	English	in order to change this situation, in $date, the yili prefecture cpc committee took the lead in organizing and[20] sending[21] some $number teachers, divided into various propaganda teams,[31] deep into the rural and pastoral areas during[39] the[40] summer[41] vacation[42] to present the facts and reason things out for the masses .
	Baseline	0-0 0-1 0-2 1-3 2-4 3-4 4-5 5-6 6-7 6-8 6-9 7-11 8-11 9-17 9-18 10-19 11-22 11-23 13-24 15-29 16-30 **17-20 19-21** 20-32 21-35 22-37 22-38 24-51 25-53 26-42 27-42 28-54
	Forced decoding	0-0 0-1 0-2 1-3 2-4 3-4 4-5 5-6 6-7 6-8 6-9 7-11 8-11 9-17 9-18 10-19 11-22 11-23 13-24 15-29 16-30 **17-31** 20-32 21-35 22-37 22-38 24-51 25-53 26-42 27-42 28-54
Case 2	Chinese	由于 采取 了 [2] 群众 喜闻乐见 的 宣讲 方式, 加上 紧 扣 当地 实际, 宣讲 活动 深受 [17] 欢迎 .
	English	as the propaganda activity was conducted[5] in a[7] way welcomed by the masses[12] and with the local realities fully taken into account, it received[24] an enthusiastic response from the masses .
	Baseline	0-0 1-5 **2-7** 3-30 4-25 4-26 4-27 4-28 6-2 6-3 7-8 8-13 9-14 10-20 11-23 12-16 13-17 14-22 **17-12** 18-9 19-31
	Forced decoding	0-0 1-5 3-30 4-25 4-26 4-27 4-28 6-2 6-3 7-8 8-13 9-14 10-20 11-23 12-16 13-17 14-22 **17-24** 18-9 19-31
Case 3	Chinese	开始 下来 时, 认为 集中 村民 很难, 即使 集中 [10] 起来 也 怕 中途 人 走 掉.
	English	when i first came here, i thought that it would be difficult for us to gather the villagers together and even if we[23] could gather[25] them together, some of them would leave the site before a propaganda lecture ended .
	Baseline	0-16 1-16 2-0 3-5 4-7 4-8 6-18 6-19 7-9 7-10 7-11 7-12 7-13 8-20 9-21 9-22 **10-23** 11-25 13-24 14-1 14-2 14-3 14-4 15-26 16-33 17-38 17-39 17-40 18-41
	Forced decoding	0-16 1-16 2-0 3-5 4-7 4-8 6-18 6-19 7-9 7-10 7-11 7-12 7-13 8-20 9-21 9-22 **10-25** 11-25 13-24 14-1 14-2 14-3 14-4 15-26 16-33 17-38 17-39 17-40 18-41
Case 4	Chinese	但 后来 宣讲 团 每 到 一 地, 都 受到 群众 欢迎,[13] 有时 听讲 的 人 越来越 多, 自己 也 受到 很大 教育.
	English	ater on,[2] when i saw our propaganda team welcomed by the masses wherever it went and[16] a large number of people coming to listen to our propaganda lectures, i myself was greatly educated .
	Baseline	0-0 1-0 2-6 2-7 2-8 4-13 5-15 6-17 10-10 11-12 12-9 **13-2** 14-3 15-4 15-5 17-21 20-29 21-31 22-31 23-32 24-33 25-34 26-35
	Forced decoding	0-0 1-0 2-6 2-7 2-8 4-13 5-15 6-17 10-10 11-12 12-9 **13-16** 14-3 15-4 15-5 17-21 20-29 21-31 22-31 23-32 24-33 25-34 26-35

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

Table 13. *Continued*

Case 5	Chinese	在 乌苏市 甘家湖 牧场, 宣讲 人员 自 带 干粮 连续 七 天 驱车 [13] $number 公里, 进 村 入户 宣讲.
	English	in the ganjiahu pasture of wusu[5] city, the propaganda team members carried with them their own food and traveled[19] by[20] vehicles[21] for several hundred kilometers for seven consecutive days, going into the households to give propaganda lectures .
	Baseline	0-0 1-2 2-2 3-2 5-9 5-10 6-11 9-15 10-28 11-27 12-29 **13-5** 14-24 15-25 16-30 17-32 19-34 20-37 20-38 21-39
	Forced decoding	0-0 1-2 2-2 3-2 5-9 5-10 6-11 9-15 10-28 11-27 12-29 14-24 15-25 16-30 17-32 19-34 20-37 20-38 21-39

2007; Pauls *et al.* 2010; Sun *et al.* 2010a, 2010b). Some previous methods infer syntactic correspondences between the source and target languages through word alignments and constituent boundary-based syntactic constraints. Such a syntactic alignment method is sensitive to word alignment behavior. To combat this, Pauls *et al.* (2010) presented an unsupervised ITG alignment model that directly aligns syntactic structures for string-to-tree rule extraction. One major problem with syntactic structure alignment is that syntactic divergence between languages can prevent accurate syntactic alignments between the source and target languages.

May and Knight (2007) presented a syntactic re-alignment model for syntax-based MT that uses syntactic constraints to re-align a parallel corpus with word alignments. The motivation behind their methods is similar to ours, our work being different in two major respects. First, the approach proposed by May and Knight (2007) first utilizes the EM algorithm to obtain Viterbi derivation trees from derivation forests of each (tree, string) pair, and then produces Viterbi alignments based on obtained derivation trees. Our forced decoding-based approach searches for the best derivation to produce TSAs that are used to improve the extraction of translation rules. In other words, TSAs are optimized by the translation model. Second, their models are only applicable to syntax-based systems while our model can be applied to both phrase-based and syntax-based MT tasks.

7 Discussion and conclusion

This paper presents a TSA-based approach to improving rule extraction by deleting spurious links and adding new valuable links with the help of bilingual TSAs. The TSAs are constructed in an unsupervised learning manner, and essentially optimized by the translation model during the forced decoding process, without using any statistics and linguistic heuristics or syntactic constraints. Spurious link deletion can avoid some desirable rules from being blocked and thus enlarge the coverage of the translation rule set. Adding some new valuable links can also improve the precision of translation rule extraction and possibly alleviate a practical problem that unaligned words can result in a huge number of extractable rules due to null alignments

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

involved. While our approach directly utilizes TSAs to refine word alignments for rule extraction, there is another option worth exploring where TSA can be used to identify blocked desirable translation rules for extraction, rather than used for word alignment refinement. This option is similar to word alignment combination, except that TSAs are only used to release an originally blocked desirable rule for extraction in this option. In addition, it is also worth studying how to combine the best of our approach with other discriminative word alignment models to improve rule extraction for SMT systems, e.g., by incorporating part-of-speech information for word alignment refinement.

Acknowledgements

This work was supported in part by the National Science Foundation of China (61073140 and 61272376), Specialized Research Fund for the Doctoral Program of Higher Education (20100042110031) and the Fundamental Research Funds for the Central Universities (N100204002).

References

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics* **19**(2): 263–311.

Cherry, C., and Lin, D. 2006. Soft syntactic constraints for word alignment through discriminative training. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, Sydney, Australia, pp. 105–112.

Chiang, D. 2007. Hierarchical phrase-based translation. *Computational Linguistics* **33**(2), 201–28.

DeNero, J., and Klein, D. 2007. Tailoring word alignments to syntactic machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, pp. 17–24.

Deng, Y., and Zhou, B. 2009. Optimizing word alignment combination for phrase table training. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL–IJCNLP)*, short paper, Suntec, Singapore, pp. 229–232.

Fossum, V., Knight, K., and Abney, S. 2008. Using syntax to improve word alignment precision for syntax-based machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, Ohio, USA, pp. 44–52.

Fraser, A., and Marcu, D. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics* **33**(3): 293–303.

Galley, M., Hopkins, M., Knight, K., and Marcu, D. 2004. What's in a translation rule? In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Boston, Massachusetts, USA, pp. 273–80.

Galley, M., and Manning, C. D. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Honolulu, Hawaii, USA, pp. 848–56.

Genzel, D. 2010. Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, Beijing, China, pp. 376–84.

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

Groves, D., Hearne, M., and Way, A. 2004. Robust sub-sentential alignment of phrase-structure trees. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, Geneva, Switzerland, pp. 1072–8.

Hermjakob, U. 2009. Improved word alignment with statistics and linguistic heuristics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, pp. 229–37.

Huang, L., and Chiang, D. 2007. Forest rescoring: faster decoding with integrated language models. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, pp. 144–51.

Huang, L., Knight, K., and Joshi, A. 2006. Statistical syntax-directed translation with extended domain of locality. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas (AMTA)*, Cambridge, Massachusetts, USA, pp. 66–73.

Imamura, K. 2001. Hierarchical phrase alignment harmonized with parsing. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS)*, Tokyo, Japan, pp. 377–84.

Ittycheriah, A., and Roukos, S. 2005. A maximum entropy word aligner for Arabic–English machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, pp. 89–96.

Khalilov, M., and Sima'an, K. 2011. Statistical translation after source reordering: oracles, context-aware models, and empirical analysis. *Natural Language Engineering* **18**(4): 491–519.

Koehn, P. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Boston, Massachusetts, USA, pp. 388–95.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, pp. 177–80.

Koehn, P., Och, F. J., and Marcu, D. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Edmonton, pp. 48–54.

Liu, Y., Liu, Q., and Lin, S. 2005. Log-linear models for word alignment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, Michigan, USA, pp. 459–66.

Liu, Y., Liu, Q., and Lin, S. 2010. Discriminative word alignment by linear modeling. *Computational Linguistics* **36**(3): 303–39.

Marcu, D., Wang, W., Echihabi, A., and Knight, K. 2006. SPMT: statistical machine translation with syntactified target language phrases. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Sydney, Australia, pp. 44–52.

May, J., and Knight, K. 2007. Syntactic re-alignment models for machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP–CoNLL)*, Prague, pp. 360–68.

Moore, R. C., Yih, W.-t., and Bode, A. 2006. Improved discriminative bilingual word alignment. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, Sydney, Australia, pp. 513–20.

Och, F. J. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, pp. 160–67.

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press

Och, F. J., and Ney, H. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, Hongkong, pp. 440–47.

Och, F., and Ney, H. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, Pennsylvania, USA, p. 295–302.

Pauls, A., Klein, D., Chiang, D., and Knight, K. 2010. Unsupervised syntactic alignment with inversion transduction grammars. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT–NAACL)*, Los Angeles, California, USA, pp. 118–26.

Petrov, S., Barrett, L., Thibaux, R., and Klein, D. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, Sydney, Australia, pp. 433–40.

Sun, J., Zhang, M., and Tan, C. L. 2010a. Exploring syntactic structural features for sub-tree alignment using bilingual tree kernels. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, Uppsala, Sweden, pp. 306–15.

Sun, J., Zhang, M., and Tan, C. L. 2010b. Discriminative induction of sub-tree alignment using limited labeled data. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, Beijing, China, pp. 1047–55.

Taskar, B., Lacoste-Julien, S., and Klein, D. 2005. A discriminative matching approach to word alignment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, pp. 73–80.

Tillman, C. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Boston, Massachusetts, USA, pp. 101–4.

Tinsley, J., Zhechev, V., Hearne, M., and Way, A. 2007. Robust language pair-independent sub-tree alignment. In *Proceedings of MT Summit XI*, Copenhagen, Denmark, pp. 467–74.

Vogel, S., Ney, H., and Tillmann, C. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, Copenhagen, Denmark, pp. 836–41.

Wang, W., Knight, K., and Marcu, D. 2007. Binarizing syntax trees to improve syntax-based machine translation accuracy. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP–CoNLL)*, Prague, pp. 746–54.

Xiao, T., Zhu, J., Zhang, H., and Li, Q. 2012. NiuTrans: an open source toolkit for phrase-based and syntax-based machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, Jeju, Republic of Korea, pp. 19–24.

Xiong, D., Liu, Q., and Lin, S. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, Sydney, Australia, pp. 521–8.

https://doi.org/10.1017/S1351324913000260 Published online by Cambridge University Press