# Learning Better Rule Extraction with Translation Span Alignment

**Jingbo Zhu  Tong Xiao  Chunliang Zhang**
Natural Language Processing Laboratory
Northeastern University, Shenyang, China
{zhujingbo,xiaotong,zhangcl}@mail.neu.edu.cn

## Abstract

This paper presents an unsupervised approach to learning *translation span alignments* from parallel data that improves syntactic rule extraction by deleting spurious word alignment links and adding new valuable links based on bilingual translation span correspondences. Experiments on Chinese-English translation demonstrate improvements over standard methods for tree-to-string and tree-to-tree translation.

## 1 Introduction

Most syntax-based statistical machine translation (SMT) systems typically utilize word alignments and parse trees on the source/target side to learn syntactic transformation rules from parallel data. The approach suffers from a practical problem that even one spurious (word alignment) link can prevent some desirable syntactic translation rules from extraction, which can in turn affect the quality of translation rules and translation performance (May and Knight 2007; Fossum *et al.* 2008). To address this challenge, a considerable amount of previous research has been done to improve alignment quality by incorporating some 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).

Unlike their efforts, this paper presents a simple approach that automatically builds the *translation span alignment* (TSA) of a sentence pair by utilizing a phrase-based forced decoding technique, and then improves syntactic rule extraction by deleting spurious links and adding new valuable links based on bilingual translation span correspondences. The proposed approach has two promising properties.

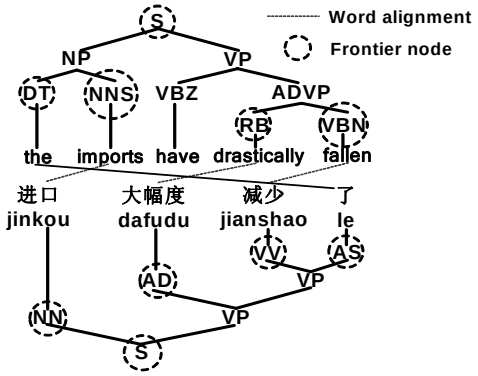


Figure 1. A real example of Chinese-English sentence pair with word alignment and both-side parse trees.

| |
|---|
| 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:$x_2$) → $x_1$ $x_2$ |
| 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:$x_1$ VP:$x_2$) → S(NP:$x_1$ VP:$x_2$) |
| $r_7$: VP(AD:$x_1$ VP(VV:$x_2$ AS:$x_3$)) → VP(VBZ:$x_3$ ADVP(RB:$x_1$ VBN:$x_2$)) |

Table 1. Some useful syntactic rules are blocked due to the spurious link between "了" and "*the*".

Firstly, The TSAs are constructed in an unsupervised learning manner, and optimized by the translation model during the forced decoding process, without using any statistics and linguistic heuristics or syntactic constraints. Secondly, our approach is independent of the word alignment-based algorithm used to extract translation rules, and easy to implement.

## 2 Translation Span Alignment Model

Different from word alignment, TSA is a process of identifying span-to-span alignments between parallel sentences. For each translation span pair,



*Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 280–284,
Jeju, Republic of Korea, 8-14 July 2012. 

1. Extract phrase translation rules $R$ from the parallel corpus with word alignment, and construct a phrase-based translation model $M$.
2. Apply $M$ to implement phrase-based forced decoding on each training sentence pair ($c$, $e$), and output its best derivation $d^*$ that can transform $c$ into $e$.
3. Build a TSA of each sentence pair ($c$, $e$) from its best derivation $d^*$, in which each rule $r$ in $d^*$ is used to form a translation span pair $\{src(r)<=>tgt(r)\}$.

Figure 2. TSA generation algorithm. *src(r)* and *tgt(r)* indicate the source and target side of rule *r*.

its source (or target) span is a sequence of source (or target) words. Given a source sentence $c=c_1...c_n$, a target sentence $e=e_1...e_m$, and its word alignment $A$, a translation span pair $\tau$ is a pair of source span $(c_i...c_j)$ and target span $(e_p...e_q)$

$$\tau = (c_i^j \Leftrightarrow e_p^q)$$

where $\tau$ indicates that the source span $(c_i...c_j)$ and the target span $(e_p...e_q)$ are translational equivalent. We do not require that $\tau$ must be consistent with the associated word alignment $A$ in a TSA model.

Figure 2 depicts the TSA generation algorithm in which a *phrase-based forced decoding* technique is adopted to produce the TSA of each sentence pair. In this work, we do not apply syntax-based forced decoding (e.g., tree-to-string) 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.

Formally, given a sentence pair ($c$, $e$), the phrase-based forced decoding technique aims to search for the *best* derivation $d^*$ among all consistent derivations that convert the given source sentence $c$ into the given target sentence $e$ with respect to the current translation model induced from the training data, which can be expressed by

$$d^* = \arg\max_{d \in D(c,e) \wedge TGT(d)=e} \Pr_\theta(TGT(d) \mid c) \qquad (1)$$

where $D(c,e)$ is the set of candidate derivations that transform $c$ to $e$, and TGT($d$) is a function that outputs the yield of a derivation $d$. $\theta$ indicates parameters of the phrase-based translation model learned from the parallel corpus.

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

$$d^* = r_1 \oplus r_2 \oplus ... \oplus r_k ,$$

| |
|---|
| $c$ = 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 |
| Generating TSA from $d^*$: |
| [进口]<=>[the imports] |
| [大幅度 减少]<=>[drastically fallen] |
| [了]<=>[have] |

Table 2. Forced decoding based TSA generation on the example sentence pair in Fig. 1.

where $r_i$ indicates a phrase rule used to form $d^*$. $\oplus$ is a composition operation that combines rules $\{r_1...r_k\}$ together to produce the target translation.

As mentioned above, the best derivation $d^*$ respects the input sentence pair ($c$, $e$). 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 ($c$,$e$). In other words, the TSA of ($c$,$e$) is a set of translation span pairs generated from phrase translation rules used by the best derivation $d^*$. The forced decoding based TSA generation on the example sentence pair in Figure 1 can be shown in Table 2.

## 3 Better Rule Extraction with TSAs

To better understand the particular task that we will address in this section, we first introduce a definition of *inconsistent with a translation span alignment*. Given a sentence pair ($c$, $e$) with the word alignment $A$ and the translation span alignment $P$, we call a link $(c_i, e_j) \in A$ *inconsistent* with $P$, if $c_i$ and $e_j$ are covered respectively by two different translation span pairs in $P$ and vice versa.

$(c_i, e_j) \in A$ inconsistent with $P \Leftrightarrow$

$$\exists \tau \in P : c_i \in src(\tau) \wedge e_j \notin tgt(\tau)$$

$$OR \ \exists \tau \in P : c_i \notin src(\tau) \wedge e_j \in tgt(\tau)$$

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

By this, we will say that a link $(c_i, e_j) \in A$ is a spurious link if it is inconsistent with the given TSA. Table 3 shows that an original link (4→1) are covered by two different translation span pairs

| 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 | |

Table 3. A sentence pair with the original word alignment (WA) and the translation span alignment (TSA).

([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 be removed for rule extraction.

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 3 contains a one-to-one span pair ([4,4]<=>[3,3]), a one-to-many span pair ([1,1]<=>[1,2]) and a many-many span pair ([2,3]<=>[4,5]). In such a case, 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.

Until now, a natural way is to use TSAs to directly improve word alignment quality by deleting some spurious links and adding some new confident links, 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's revisit the example in Figure 1 again. The blocked tree-to-string $r_3$ can be extracted successfully after deleting the spurious link (了, *the*), and a new tree-to-string rule $r_1$ 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].

# 4 Experiments

## 4.1 Setup

We utilized a state-of-the-art open-source SMT system NiuTrans (Xiao et al. 2012) to implement syntax-based models in the following experiments. We begin with a training parallel corpus of Chinese-English bitexts that consists of 8.8M Chinese words and 10.1M English words in 350K sentence pairs. The GIZA++ tool was used to perform the bi-directional word alignment between the source and the target sentences, referred to as the *baseline* method. For syntactic translation rule extraction, minimal GHKM (Galley *et al.*, 2004) rules are first extracted from the bilingual corpus whose source and target sides are parsed using the Berkeley parser (Petrov *et al.* 2006). The composed rules are then generated by composing two or three minimal rules. A 5-gram language model was trained on the Xinhua portion of English Gigaword corpus. Beam search and cube pruning techniques (Huang and Chiang 2007) were used to prune the search space for all the systems. The base feature set used for all systems is similar to that used in (Marcu *et al.* 2006), including 14 base features in total such as 5-gram language model, bidirectional lexical and phrase-based translation probabilities. All features were log-linearly combined and their weights were optimized by performing minimum error rate training (MERT) (Och 2003). The development data set used for weight training comes from NIST MT03 evaluation set, consisting of 326 sentence pairs of less than 20 words in each Chinese sentence. Two test sets are NIST MT04 (1788 sentence pairs) and MT05 (1082 sentence pairs) evaluation sets. The translation quality is evaluated in terms of the case-insensitive IBM-BLEU4 metric.

| Method | Prec% | Rec% | F1% | Del/Sent | Add/Sent |
|---|---|---|---|---|---|
| Baseline | 83.07 | 75.75 | 79.25 | - | - |
| TSA | 84.01 | 75.46 | 79.51 | 1.5 | 1.1 |

Table 4. Word alignment precision, recall and F1-score of various methods on 200 sentence pairs of Chinese-English data.

## 4.2 Effect on Word Alignment

To investigate the effect of the TSA method on word alignment, we designed an experiment to evaluate alignment quality against gold standard annotations. There are 200 random chosen and manually aligned Chinese-English sentence pairs used to assert the word alignment quality. For word alignment evaluation, we calculated precision, recall and F1-score over gold word alignment.

Table 4 depicts word alignment performance of the baseline and TSA methods. We apply the TSAs to refine the baseline word alignments, involving spurious link deletion and new link insertion operations. Table 4 shows our method can yield improvements on precision and F1-score, only causing a little negative effect on recall.

### 4.3 Translation Quality

| Method | # of Rules | MT03 | MT04 | MT05 |
|---|---|---|---|---|
| Baseline (T2S) | 33,769,071 | 34.10 | 32.55 | 30.15 |
| TSA (T2S) | 32,652,261 | 34.61$^{+}$ (+0.51) | 33.01$^{+}$ (+0.46) | 30.66$^{+}$ (+0.51) |
| | | | | |
| Baseline (T2T) | 24,287,206 | 34.51 | 32.20 | 31.78 |
| TSA (T2T) | 24,119,719 | 34.85 (+0.34) | 32.92$^{*}$ (+0.72) | 32.22$^{+}$ (+0.44) |

Table 5. Rule sizes and IBM-BLEU4 (%) scores of baseline and our method (TSA) in tree-to-string (T2S) and tree-to-tree (T2T) translation on Dev set (MT03) and two test sets (MT04 and MT05). + and * indicate significantly better on performance comparison at $p<.05$ and $p<.01$, respectively.

Table 5 depicts effectiveness of our TSA method on translation quality in tree-to-string and tree-to-tree translation tasks. Table 5 shows that our TSA method can improve both syntax-based translation systems. As mentioned before, the resulting TSAs are essentially optimized by the translation model. Based on such TSAs, experiments show that spurious link deletion and new valuable link insertion can improve translation quality for tree-to-string and tree-to-tree systems.

## 5 Related Work

Previous studies have made great efforts to incorporate statistics and linguistic heuristics or syntactic information into word alignments (Ittycheriah and Roukos 2005; Taskar *et al*. 2005; Moore *et al.* 2006; Cherry and Lin 2006; DeNero and Klein 2007; May and Knight 2007; Fossum *et al.* 2008; Hermjakob 2009; Liu *et al.* 2010). For example, Fossum *et al.* (2008) used a discriminatively trained model to identify and delete incorrect links from original word alignments to improve string-to-tree transformation rule extraction, which incorporates four types of features such as lexical and syntactic features. This paper presents an approach to incorporating translation span alignments into word alignments to delete spurious links and add new valuable links.

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.* 2007; Sun *et al.* 2010a, 2010b; Pauls *et al.* 2010). Some previous methods infer syntactic correspondences between the source and the 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 transformation 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 differs from (May and Knight 2007) 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 translation span alignments that are used to improve the extraction of translation rules. Translation span alignments are optimized by the translation model. Secondly, their models are only applicable for syntax-based systems while our method can be applied to both phrase-based and syntax-based translation tasks.

## 6 Conclusion

This paper presents an unsupervised approach to improving syntactic transformation rule extraction by deleting spurious links and adding new valuable links with the help of bilingual translation span alignments that are built by using a phrase-based forced decoding technique. In our future work, it is worth studying how to combine the best of our approach and discriminative word alignment models to improve rule extraction for SMT models.

## Acknowledgments

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

## References


Colin Cherry and Dekang Lin. 2006. Soft syntactic constraints for word alignment through discriminative training. In *Proc. of ACL*.

John DeNero and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *Proc. of ACL*.

Victoria Fossum, Kevin Knight and Steven Abney. 2008. Using syntax to improve word alignment precision for syntax-based machine translation. In *Proc. of the Third Workshop on Statistical Machine Translation*, pages 44-52.

Michel Galley, Mark Hopkins, Kevin Knight and Daniel Marcu. 2004. What's in a translation rule? In *Proc. of HLT-NAACL 2004*, pp273-280.

Declan Groves, Mary Hearne and Andy Way. 2004. Robust sub-sentential alignment of phrase-structure trees. In *Proc. of COLING,* pp1072-1078.

Ulf Hermjakob. 2009. Improved word alignment with statistics and linguistic heuristics. In *Proc. of EMNLP*, pp229-237

Liang Huang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *Proc. of ACL*, pp144-151.

Kenji Imamura. 2001. Hierarchical Phrase Alignment Harmonized with Parsing. In *Proc. of NLPRS*, pp377-384.

Abraham Ittycheriah and Salim Roukos. 2005. A maximum entropy word aligner for Arabic-English machine translation. In *Proc. of HLT/EMNLP*.

Yang Liu, Qun Liu and Shouxun Lin. 2010. Discriminative word alignment by linear modeling. *Computational Linguistics*, 36(3):303-339

Daniel Marcu, Wei Wang, Abdessamad Echihabi and Kevin Knight. 2006. SPMT: Statistical machine translation with syntactified target language phrases. In *Proc. of EMNLP*, pp44-52.

Jonathan May and Kevin Knight. 2007. Syntactic realignment models for machine translation. In *Proc. of EMNLP-CoNLL*.

Robert C. Moore, Wen-tau Yih and Andreas Bode. 2006. Improved discriminative bilingual word alignment. In *Proc. of ACL*

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL*.

Adam Pauls, Dan Klein, David Chiang and Kevin Knight. 2010. Unsupervised syntactic alignment with inversion transduction grammars. In *Proc. of NAACL*, pp118-126

Slav Petrov, Leon Barrett, Roman Thibaux and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proc. of ACL*, pp433-440.

Jun Sun, Min Zhang and Chew Lim Tan. 2010a. Exploring Syntactic Structural Features for Sub-Tree Alignment Using Bilingual Tree Kernels. In *Proc. of ACL*, pp306-315.

Jun Sun, Min Zhang and Chew Lim Tan. 2010b. Discriminative Induction of Sub-Tree Alignment using Limited Labeled Data. In *Proc. of COLING*, pp1047-1055.

Ben Taskar, Simon Lacoste-Julien and Dan Klein. 2005. A discriminative matching approach to word alignment. In *Proc. of HLT/EMNLP*

John Tinsley, Ventsislav Zhechev, Mary Hearne and Andy Way. 2007. Robust language pair-independent sub-tree alignment. In *Proc. of MT Summit XI*.

Tong Xiao, Jingbo Zhu, Hao Zhang and Qiang Li. 2012. NiuTrans: An Open Source Toolkit for Phrase-based and Syntax-based Machine Translation. In *Proceedings of ACL*, demonstration session