Context-Aware Graph Segmentation for Graph-Based Translation

Liangyou Li and **Andy Way** and **Qun Liu**
ADAPT Centre, School of Computing
Dublin City University, Ireland
{liangyou.li,andy.way,qun.liu}@adaptcentre.ie

Abstract

In this paper, we present an improved graph-based translation model which segments an input graph into node-induced subgraphs by taking source context into consideration. Translations are generated by combining subgraph translations left-to-right using beam search. Experiments on Chinese–English and German–English demonstrate that the context-aware segmentation significantly improves the baseline graph-based model.

1 Introduction

The well-known phrase-based statistical translation model (Koehn et al., 2003) extends the basic translation units from single words to continuous phrases to capture local phenomena. However, one of its significant weaknesses is that it cannot learn generalizations (Quirk et al., 2005; Galley and Manning, 2010). To allow discontinuous phrases (any subset of words of an input sentence), dependency treelets (Menezes and Quirk, 2005; Quirk et al., 2005; Xiong et al., 2007) can be used, which are connected subgraphs on trees. However, continuous phrases which are not connected on trees and thus excluded could in fact be extremely important to system performance (Koehn et al., 2003; Hanneman and Lavie, 2009).

To make use of the merits of both phrase-based models and treelet-based models, Li et al. (2016) proposed a graph-based translation model as in Equation (1):

$$p(\bar{t}_1^I \mid \bar{g}_1^I) = \prod_{i=1}^{I} p(\bar{t}_i|\bar{g}_{a_i}) \times d(\bar{g}_{a_i}, \bar{g}_{a_{i-1}}) \quad (1)$$

where $\bar{t}_i$ is a continuous target phrase which is the translation of a node-induced and connected source subgraph $\bar{g}_{a_i}$.[1] d is a distance-based reordering function which penalizes discontinuous phrases that have relatively long gaps (Galley and Manning, 2010). The model translates an input graph by segmenting it into subgraphs and generates a complete translation by combining subgraph translations left-to-right. However, the model treats different graph segmentations equally.

Therefore, in this paper we propose a context-aware graph segmentation (Section 2): (i) we add contextual information to each translation rule during training (Section 2.2); (ii) during decoding, when a rule is applied, the input context should match with the rule context (Section 2.3). Experiments (Section 3) on Chinese–English (ZH–EN) and German–English (DE–EN) tasks show that our method significantly improves the graph-based model. As observed in our experiments, the context-aware segmentation brings two benefits to our system: (i) it helps to select a better subgraph to translate; and (ii) it selects a better target phrase for a subgraph.

2 Context-Aware Graph Segmentation and Translation

Our model extends the graph-based translation model by considering source context during segmenting input graphs, as in Equation (2):

$$p(\bar{t}_1^I \mid \bar{g}_1^I) = \prod_{i=1}^{I} p(\bar{t}_i \mid \bar{g}_{a_i}, \bar{c}_{a_i}) \times d(\bar{g}_{a_i}, \bar{g}_{a_{i-1}}) \quad (2)$$

where $\bar{c}_{a_i}$ denotes the context of the subgraph $\bar{g}_{a_i}$, which is represented as a set of connections (i.e. edges) between $\bar{g}_{a_i}$ and $[\bar{g}_{a_{i+1}}, \cdots, \bar{g}_{a_I}]$.

[1] All subgraphs in this paper are connected and node-induced.

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 599–604,
Valencia, Spain, April 3-7, 2017.

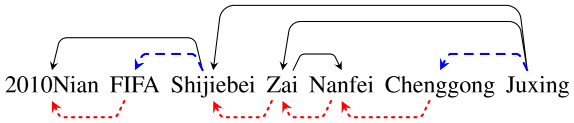


Figure 1: An example graph for a Chinese sentence. Dotted lines are bigram relations. Solid lines are dependency relations. Dashed lines are shared by bigram and dependency relations.

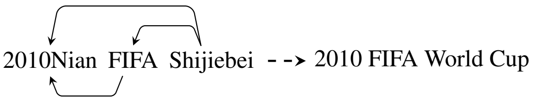
2.1 Building Graphs

The graph used in this paper combines a sequence and a dependency tree as in Li et al. (2016). Each graph contains two kinds of links: **dependency links** from dependency trees which model syntactic and semantic relations between words, and **bigram links** which provide local and sequential information on pairs of continuous words. Figure 1 shows an example graph. Given such graphs, we can make use of both continuous and linguistically informed discontinuous phrases as long as they are connected on graphs. In this paper, we do not distinguish the two kinds of relations, because our preliminary experiments showed no improvement when considering edge types.

2.2 Training

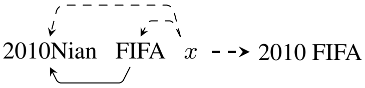
During training, given a word-aligned graph–string pair $\langle g, t, a \rangle$, we extract translation rules $\langle \overline{g}_{a_i}, c_{a_i}, \overline{t}_i \rangle$, each of which consists of a continuous target phrase $\overline{t}_i$, a source subgraph g_{a_i} aligned to $\overline{t}_i$, and a source context c_{a_i}. We first find **initial pairs**. $\langle \tilde{s}_{a_i}, \overline{t}_i \rangle$ is an initial pair, iff it is consistent with the word alignment a (Och and Ney, 2004). $\tilde{s}_{a_j}$ is a set of source words which are aligned to $\overline{t}_i$. Then, the set of rules satisfies the following:

1. If $\langle \tilde{s}_{a_i}, \overline{t}_i \rangle$ is an initial pair and $\tilde{s}_{a_i}$ is covered by a subgraph $\overline{g}_{a_i}$ which is connected, then $\langle \overline{g}_{a_i}, *, \overline{t}_i \rangle$ is a **basic rule**. $c_{a_i} = *$ means that a basic rule is applied without considering context to make sure that at least one translation is produced for any inputs during decoding. Therefore, basic rules are the same as rules in the conventional graph-based model. Rule (3) shows an example of a basic rule:



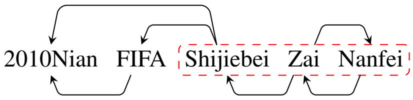
(3)

2. Assume $\langle \overline{g}_{a_i}, *, \overline{t}_i \rangle$ is a basic rule and $\langle \tilde{s}_{a_{i+1}}, \overline{t}_{i+1} \rangle$ is an initial pair where $\overline{t}_{i+1}$ is on the right of and adjacent to $\overline{t}_i$. If there are edges between $\overline{g}_{a_i}$ and $\tilde{s}_{a_{i+1}}$, then $\langle \overline{g}_{a_i}, c_{a_i}, \overline{t}_i \rangle$ is a **segmenting rule**, where c_{a_i} is the set of edges between $\overline{g}_{a_i}$ and $\tilde{s}_{a_{i+1}}$ by treating $\tilde{s}_{a_{i+1}}$ as a single node x. Rule (4) is an example of a segmenting rule:



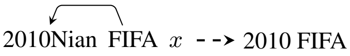
(4)

where dashed links are contextual connections. During decoding, when the context matches, rule (4) translates a subgraph over *2010Nian FIFA* into a target phrase *2010 FIFA*. For example, it can be applied to graph (5) where *Shijiebei Zai Nanfei* (in the dashed rectangle) is treated as x:

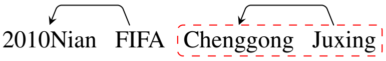


(5)

3. If there are no edges between $\overline{g}_{a_i}$ and $\tilde{s}_{a_{i+1}}$, then c_{a_i} is equal to $\emptyset$ and $\langle \overline{g}_{a_i}, \emptyset, \overline{t}_i \rangle$ is a translation rule, called a **selecting rule** in this paper. During decoding, the untranslated input could be a set of subgraphs which are disjoint with each other. A selecting rule is used to select one of them. For example, rule (6) can be applied to (7) to translate *2010Nian FIFA* to *2010 FIFA*. In this example, the x in rule (6) matches with *Chenggong Juxing* (in the dashed rectangle) in (7).



(6)



(7)

By comparing these three types of rules, we observe that both segmenting rules and selecting rules are based on basic rules. They extend basic rules by adding contextual information to their source subgraphs so that basic rules are split into different groups according to the context. During decoding, the context will help to select target phrases as well.

Algorithm 1 illustrates a simple process for rule extraction. Given a word-aligned graph–string pair, we first extract all initial pairs (Line 1). Then, we find basic rules from these pairs (Lines 3–4). Basic

Algorithm 1: An algorithm for extracting translation rules from a graph–string pair.

```
   Data: Word-aligned graph–string pair $\langle g, t, a \rangle$
   Result: A set of translation rules $R$
1  find a set of initial pairs $P$;
2  for each $p = \langle \overline{s}_{a_i}, \overline{t}_i \rangle$ in $P$ do
3      if $s_i^j$ is connected then
               // basic rules
4              add $\langle \overline{g}_{a_i}), *, \overline{t}_i \rangle$ to $R$ ;
               // segmenting and selecting
                   rules
5              for $q = \langle \overline{s}_{a_{i+1}}, \overline{t}_{i+1} \rangle$ in $P$ do
6                  $c$ is the set of edges between $\overline{g}_{a_i}$
                     and $\overline{s}_{a_{i+1}}$ ;
7                  add $\langle \overline{g}_{a_i}, c, \overline{t}_i \rangle$ to $R$ ;
8              end
9      end
10 end
```

rules are then used to generate segmenting and selecting rules by extending them with contextual connections (Lines 5–8).

2.3 Model and Decoding

Following Li et al. (2016), we define our model in the well-known log-linear framework (Och and Ney, 2002). In our experiments, we use the following standard features: two translation probabilities $p(g, c|t)$ and $p(t|g, c)$, two lexical translation probabilities $p_{lex}(g, c|t)$ and $p_{lex}(t|g, c)$, a language model $p(t)$, a rule penalty, a word penalty, and a distortion function as defined in Galley and Manning (2010). In addition, we add one more feature into our system: a basic-rule penalty to distinguish basic rules from segmenting and selecting rules.

Our decoder is very similar to the one in the conventional graph-based model, which generates hypotheses left-to-right using beam search. A hypothesis can be extended on the right by translating an uncovered source subgraph. The translation process ends when all source words have been translated.

However, when extending a hypothesis, our decoder considers the context of the translated subgraph, i.e. edges connecting it with the remaining untranslated source words. Figure 2 shows a derivation which translates an input graph in Chinese to an English string. In this example, both rules $\mathbf{r_1}$ and $\mathbf{r_2}$ are segmenting rules.

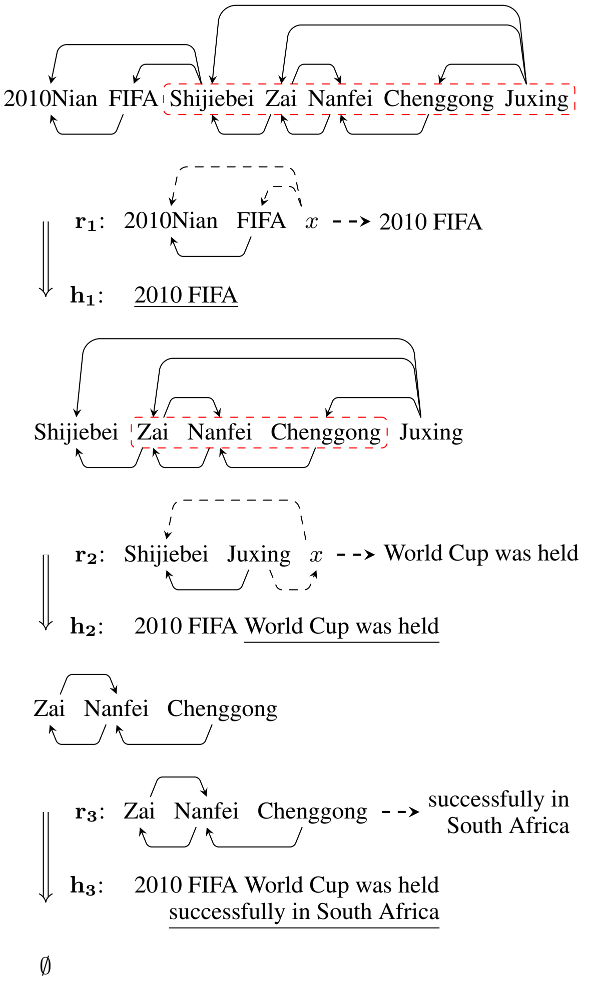

2010Nian FIFA Shijiebei Zai Nanfei Chenggong Juxing
$\mathbf{r_1}$: 2010Nian FIFA x 2010 FIFA
$\mathbf{h_1}$: 2010 FIFA
Shijiebei Zai Nanfei Chenggong Juxing
$\mathbf{r_2}$: Shijiebei Juxing x World Cup was held
$\mathbf{h_2}$: 2010 FIFA World Cup was held
Zai Nanfei Chenggong
$\mathbf{r_3}$: Zai Nanfei Chenggong successfully in South Africa
$\mathbf{h_3}$: 2010 FIFA World Cup was held successfully in South Africa
∅

Figure 2: Example of translating an input graph. Each rule $\mathbf{r_i}$ generates a new hypothesis $\mathbf{h_i}$ by appending translations on the right. Edges connected to x denote contextual information. Nodes in dashed rectangles are treated as x during decoding for matching contexts.

3 Experiments

We conduct experiments on ZH–EN and DE–EN corpora.

3.1 Data and Settings

The ZH–EN training corpus contains 1.5M+ sentences from LDC. NIST 2002 is taken as a development set to tune weights. NIST 2004 (MT04) and NIST 2005 (MT05) are two test sets to evaluate systems. The DE–EN training corpus (2M+ sentence pairs) is from WMT 2014, including Europarl V7 and News Commentary. News-Test 2011 is taken as a development set while News-Test 2012 (WMT12) and News-Test 2013 (WMT13) are our test sets.

System	ZH–EN		DE–EN	
	MT04	MT05	WMT12	WMT13
PBMT	33.2	31.8	19.5	21.9
TBMT	33.8^{*}	31.7	19.6	22.1^{*}
GBMT	34.7^{*+}	32.4^{*+}	19.8^{*+}	22.4^{*+}
GBMT_{ctx}	$\mathbf{35.4}^{*+}$	$\mathbf{33.7}^{*+}$	$\mathbf{20.1}^{*+}$	$\mathbf{22.8}^{*+}$

Table 1: BLEU scores of all systems. Bold figures mean GBMT_{ctx} is significantly better than GBMT at $p \leq 0.01$. $*$ means a system is significantly better than PBMT at $p \leq 0.01$. $+$ means a system is significantly better than TBMT at $p \leq 0.01$.

Following Li et al. (2016), Chinese and German sentences are parsed into projective dependency trees which are then converted to graphs by adding bigram edges. Word alignment is performed by GIZA++ (Och and Ney, 2003) with the heuristic function *grow-diag-final-and*. We use SRILM (Stolcke, 2002) to train a 5-gram language model on the Xinhua portion of the English Gigaword corpus 5th edition with modified Kneser-Ney discounting (Chen and Goodman, 1996). Batch MIRA (Cherry and Foster, 2012) is used to tune feature weights. We report BLEU (Papineni et al., 2002) scores averaged on three runs of MIRA (Clark et al., 2011).

We compare our system $\mathbf{GBMT}_{\mathbf{ctx}}$ with several other systems. A system **PBMT** is built using the phrase-based model in Moses (Koehn et al., 2007). **GBMT** is the graph-based translation system described in Li et al. (2016). To examine the influence of bigram links, GBMT is also used to translate dependency trees where treelets (Menezes and Quirk, 2005; Quirk et al., 2005; Xiong et al., 2007) are the basic translation units. Accordingly, we name the system **TBMT**. All systems are implemented in Moses.

3.2 Results and Discussion

Table 1 shows BLEU scores of all systems. We found that GBMT_{ctx} is better than PBMT across all test sets. Specifically, the improvements are +2.0/+0.7 BLEU on average on ZH–EN and DE–EN, respectively. This improvement is reasonable as our system allows discontinuous phrases which can reduce data sparsity and handle long-distance relations (Galley and Manning, 2010). In addition, the system TBMT does not show consistent improvements over PBMT while both GBMT and GBMT_{ctx} achieve better BLEU scores than TBMT on both ZH–EN (+1.8 BLEU, in terms of GBMT_{ctx}) and DE–EN (+0.6 BLEU, in terms of GBMT_{ctx}). This suggests that continuous phrases connected by bigram links are essential to system performance since they help to improve phrase coverage (Hanneman and Lavie, 2009).

Rule Type	# Rules	
	ZH–EN	DE–EN
Basic Rule	84.7M+	115.7M+
Segmenting Rule	128.4M+	167.3M+
Selecting Rule	30.2M+	35.7M+
Total	243.5M+	318.9M+

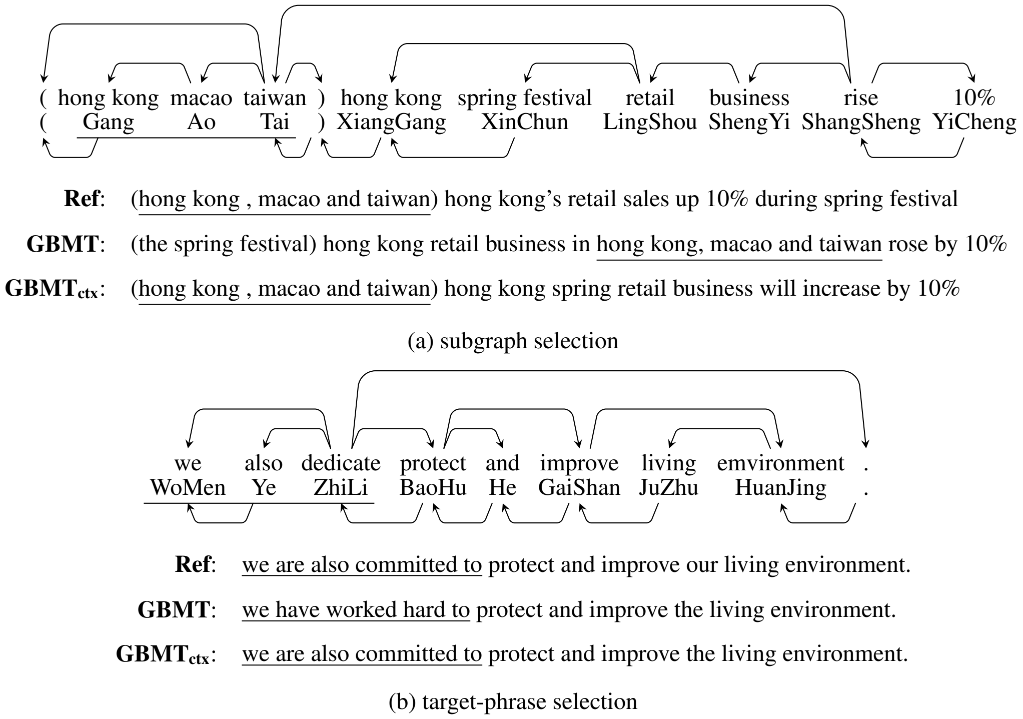
Table 2: The number of rules in GBMT_{ctx} according to their type

We also found that GBMT_{ctx} is significantly better than GBMT on both ZH–EN (+1.0 BLEU) and DE–EN (+0.4 BLEU), which indicates that explicitly modeling a segmentation using context is helpful. The main reason for the improvement is that context helps to select proper subgraphs and target phrases. Figure 3 shows example translations. We found that in Figure 3a, after translating a parenthesis, GBMT_{ctx} correctly selects a subgraph *Gang Ao Tai* and generates a target phrase *hong kong, macao and taiwan*. In Figure 3b, both GBMT and GBMT_{ctx} choose to translate the subgraph *WoMen Ye ZhiLi*. However, given the context of the subgraph, GBMT_{ctx} selects a correct target phrase *we are also committed to* for it.

3.3 Influence of Different Types of Rules

Recall that, compared with GBMT, GBMT_{ctx} contains three types of rules: basic rules, segmenting rules, and selecting rules. While basic rules exist in both systems, segmenting and selecting rules make GBMT_{ctx} context-aware. Table 2 shows the number of rules in GBMT_{ctx} according to their types. We found that on both language pairs 35%–36% of rules are basic rules. While the proportion of segmenting rules is $\sim$53%, selecting rules only account for 11%–12%. This is because segmenting rules contain richer contextual information than selecting rules.

Table 3 shows BLEU scores of GBMT_{ctx} when different types of rules are used. Note that when only basic rules are allowed, our system degrades to the conventional GBMT system. The results in Table 3 suggest that both segmenting and selecting rules consistently improve GBMT on both language pairs. However, segmenting rules are more useful than selecting rules. This is reasonable since


(hong kong macao taiwan) hong kong spring festival retail business rise 10%
(Gang Ao Tai) XiangGang XinChun LingShou ShengYi ShangSheng YiCheng

Ref: (hong kong , macao and taiwan) hong kong's retail sales up 10% during spring festival

GBMT: (the spring festival) hong kong retail business in hong kong, macao and taiwan rose by 10%

GBMT$_{\text{ctx}}$: (hong kong , macao and taiwan) hong kong spring retail business will increase by 10%

(a) subgraph selection

we also dedicate protect and improve living emvironment .
WoMen Ye ZhiLi BaoHu He GaiShan JuZhu HuanJing .

Ref: we are also committed to protect and improve our living environment.

GBMT: we have worked hard to protect and improve the living environment.

GBMT$_{\text{ctx}}$: we are also committed to protect and improve the living environment.

(b) target-phrase selection

Figure 3: Example translations of GBMT and GBMT$_{\text{ctx}}$

System	ZH–EN		DE–EN	
	MT04	MT05	WMT12	WMT13
Basic Rule	34.7	32.4	19.8	22.4
+Seg. Rule	**34.9**	**33.0**	**20.2**	**23.0**
+Sel. Rule	34.8	32.5	**20.0**	**22.7**
All	**35.4**	**33.7**	**20.1**	**22.8**

Table 3: BLEU scores of GBMT$_{\text{ctx}}$ when different types of rules are used, including Basic Rule, Segmenting (Seg.) Rule, and Selecting (Sel.) Rule. Bold figures mean a system is significantly better than the one only using basic rules at $p \leq 0.01$.

the number of segmenting rules is much larger than the number of selecting rules. We further observed that, while our system achieves the best performance when all rules are used on ZH–EN, the combination of basic rules and segmenting rules on DE–EN results in the best system. This is probably because reordering (including long-distance reordering) is performed less often in DE–EN than in ZH–EN (Li et al., 2016) which makes selecting rules less preferable on DE–EN.

4 Conclusion

In this paper, we present a graph-based model which takes subgraphs as the basic translation units and considers source context during segmenting graphs into subgraphs. Experiments on Chinese–English and German–English show that our model is significantly better than the conventional graph-based model which equally treats different graph segmentations.

In this paper, source context is used as hard constraints during decoding. In future, we would like to try soft constraints. In addition, it would also be interesting to extend this model using a synchronous graph grammar.

Acknowledgments

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n^{o} 645452 (QT21). The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. The authors thank all anonymous reviewers for their insightful comments and suggestions.

References

Stanley F. Chen and Joshua Goodman. 1996. An Empirical Study of Smoothing Techniques for Language Modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Santa Cruz, California, June.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montreal, Canada, June.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, pages 176–181, Portland, Oregon, June.

Michel Galley and Christopher D. Manning. 2010. Accurate Non-hierarchical Phrase-Based Translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 966–974, Los Angeles, California, June.

Greg Hanneman and Alon Lavie. 2009. Decoding with Syntactic and Non-syntactic Phrases in a Syntax-based Machine Translation System. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*, pages 1–9, Boulder, Colorado, June.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 48–54, Edmonton, Canada, July.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, Prague, Czech Republic, June.

Liangyou Li, Andy Way, and Qun Liu. 2016. Graph-Based Translation Via Graph Segmentation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 97–107, Berlin, Germany, August.

Arul Menezes and Chris Quirk. 2005. Dependency Treelet Translation: The Convergence of Statistical and Example-Based Machine-translation? In *Proceedings of the Workshop on Example-based Machine Translation at MT Summit X*, September.

Franz Josef Och and Hermann Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 295–302, Philadelphia, PA, USA, July.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, March.

Franz Josef Och and Hermann Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(4):417–449, December.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, Philadelphia, Pennsylvania, July.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 271–279, Ann Arbor, Michigan, June.

Andreas Stolcke. 2002. SRILM An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference Spoken Language Processing*, pages 901–904, Denver, CO.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2007. A Dependency Treelet String Correspondence Model for Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 40–47, Prague, Czech Republic, June.