Extending Phrase-Based Translation with Dependencies by Using Graphs

Liangyou Li and **Andy Way** and **Qun Liu**
ADAPT Centre, School of Computing
Dublin City University, Ireland
{liangyouli,away,qliu}@computing.dcu.ie

Abstract

In this paper, we propose a graph-based translation model which takes advantage of discontinuous phrases. The model segments a graph which combines bigram and dependency relations into subgraphs and produces translations by combining translations of these subgraphs. Experiments on Chinese–English and German–English tasks show that our system is significantly better than the phrase-based model. By explicitly modeling the graph segmentation, our system gains further improvement.

1 Introduction

One significant weakness of conventional phrase-based (PB) models (Koehn et al., 2003) is that it only uses continuous phrases and thus cannot learn generalizations, such as French *ne. . . pas* to English *not* (Galley and Manning, 2010). Although using tree structures is believed to be a promising way to solve this problem by learning either translation patterns (Chiang, 2005; Galley et al., 2004; Liu et al., 2006) or treelets (Menezes and Quirk, 2005; Quirk et al., 2005; Xiong et al., 2007), handling non-syntactic phrases is still a big challenge.

In this paper, we propose a graph-based translation model which translates a graph into a string by segmenting the graph into subgraphs. Each subgraph is connected and may cover discontinuous phrases. Experiments show that our model is significantly better than the PB model. Explicitly modeling the graph segmentation further improves our system.

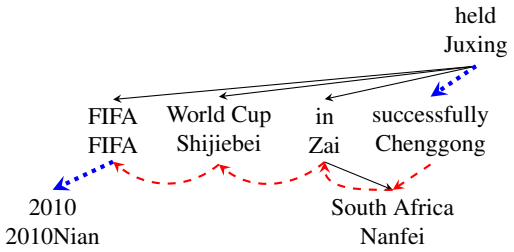

Figure 1: An example of constructing a graph for a Chinese sentence. Each node includes a Chinese word and its English meaning. Dashed red lines are bigram relations. Dark lines are dependency relations. Dotted blue lines are shared by bigram and dependency relations.

2 Graph-Based Translation

Our graph-based translation model extends PB translation by translating an input graph rather than a sequence to a target string, as in Equation (1):

$$p(t \mid G(s)) = \prod_{i=1}^{I} P(\bar{t}_i|G(\tilde{s}_{a_i}))d(\tilde{s}_{a_i}, \tilde{s}_{a_{i-1}}) \quad (1)$$

where $\tilde{s}$ denotes a source phrase which may be discontinuous and $G(\tilde{s})$ indicates a connected graph covering $\tilde{s}$. d is a distortion function.[1]

2.1 Building Graphs

As a more powerful and natural structure for sentence representation, a graph can model various word-relations together in a unified way. In this paper, we use graphs to combine two commonly

[1] In this paper, we use a distortion function, defined in Galley and Manning (2010), to penalize discontinuous phrases that have relatively long gaps.

Proceedings of the 2nd Workshop on Semantics-Driven Machine Translation, pages 10–12,
San Diego, California, June 16, 2016.

used relations: bigram relations and dependency relations. In this way, we can make use of both continuous and linguistic-informed discontinuous phrases as long as they are connected subgraphs. Figure 1 shows an example of a graph.

2.2 Training and Decoding

Different from the PB model, the basic translation units in our model are subgraphs. During training, we extract subgraph–phrase pairs instead of phrase pairs on parallel graph-string sentences associated with word alignments.

Our graph-based decoder is based on beam search and generates hypotheses (partial translations) from left to right. Each hypothesis can be extended by translating an uncovered source subgraph. The translation process ends when no untranslated words remain.

2.3 Graph Segmentation Model

We define a set of sparse features to explicitly model a graph segmentation. Given previous subgraphs, for each node in a current subgraph, we extract the following features:

$$\begin{Bmatrix} n.w \\ n.c \end{Bmatrix} \times \begin{Bmatrix} n'.w \\ n'.c \end{Bmatrix} \times \begin{Bmatrix} C \\ P \\ H \end{Bmatrix} \times \begin{Bmatrix} in \\ out \end{Bmatrix}$$

where $n.w$ and $n.c$ are the word and class of a current node n, and n' is a node connected to n. C, P, and H denote that n' is in the current subgraph or the last previous subgraph or other previous subgraphs, respectively. *in* and *out* denote that an edge is an in-coming edge or out-going edge of n.

In this paper we lexicalize only on the top-100 frequent words (Cherry, 2013). In addition, we group source words into 50 classes by using *mkcls*.

3 Experiments and Results

Our Chinese–English (ZH–EN) training corpus contains 1.5M+ sentence pairs from LDC. Our German–English (DE–EN) training corpus (2M+ sentence pairs) is from WMT 2014. **GBMT** is our graph-based translation system and **GSM** adds the graph segmentation model into GBMT. **DTU** extends the PB model by allowing source discontinuous phrases (Galley and Manning, 2010). All systems are implemented in Moses (Koehn et al., 2007).

System	ZH–EN		DE–EN	
	NIST04	NIST05	WMT12	WMT13
PBMT	32.8	31.4	19.6	21.9
DTU	33.4*	31.5	19.8*	22.3*
GBMT	33.7^{*+}	31.7	19.8*	22.4*
GSM	33.8^{*+}	32.0^{*+}	20.3^{*+}	22.9^{*+}

Table 1: BLEU (Papineni et al., 2002) scores for all systems on two datasets. Each score is the average score over three MIRA (Cherry and Foster, 2012) runs (Clark et al., 2011). $*$ means a system is significantly better than PBMT at $p \leq 0.01$. $+$ means a system is significantly better than DTU at $p \leq 0.01$.

System	# Rules	
	ZH–EN	DE–EN
DTU	224M+	352M+
GBMT	99M+	153M+

Table 2: The number of rules in DTU and GBMT.

Table 1 shows our main results. Our system GBMT is better than PBMT as measured by all three metrics across all testsets. This improvement is reasonable as GBMT allows discontinuous phrases which can reduce data sparsity and handle long-distance relations (Galley and Manning, 2010).

Since phrases from syntactic structures are fewer in number but more reliable (Koehn et al., 2003), our system GBMT achieves slightly better performance than DTU but uses significantly fewer rules, as shown in Table 2. After integrating the graph segmentation model to help subgraph selection, our system (GSM) achieves significantly better BLEU than DTU on both language pairs.

4 Conclusion

In this paper, we present a graph-based translation model which extends the phrase-based model by allowing discontinuous phrases.

Acknowledgments

This research has received funding from the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement n^{o} 317471. 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.

References

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.

Colin Cherry. 2013. Improved Reordering for Phrase-Based Translation using Sparse Features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 22–31, Atlanta, Georgia, June.

David Chiang. 2005. A Hierarchical Phrase-based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270, Ann Arbor, Michigan, 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.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a Translation Rule? In *HLT-NAACL 2004: Main Proceedings*, page 273280, Boston, Massachusetts, USA, May.

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.

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string Alignment Template for Statistical Machine Translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 609–616, Sydney, Australia, July.

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.

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.

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, June.