Bilingual Lexical Cohesion Trigger Model for Document-Level Machine Translation

Guosheng Ben[†] **Deyi Xiong**[‡*] **Zhiyang Teng**[†] **Yajuan Lü**[†] **Qun Liu**[§†]

[†]Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
{benguosheng,tengzhiyang,lvyajuan,liuqun}@ict.ac.cn

[‡]School of Computer Science and Technology,Soochow University
{dyxiong}@suda.edu.cn

[§]Centre for Next Generation Localisation, Dublin City University
{qliu}@computing.dcu.ie

Abstract

In this paper, we propose a bilingual lexical cohesion trigger model to capture lexical cohesion for document-level machine translation. We integrate the model into hierarchical phrase-based machine translation and achieve an absolute improvement of 0.85 BLEU points on average over the baseline on NIST Chinese-English test sets.

1 Introduction

Current statistical machine translation (SMT) systems are mostly sentence-based. The major drawback of such a sentence-based translation fashion is the neglect of inter-sentential dependencies. As a linguistic means to establish inter-sentential links, lexical cohesion ties sentences together into a meaningfully interwoven structure through words with the same or related meanings (Wong and Kit, 2012).

This paper studies lexical cohesion devices and incorporate them into document-level machine translation. We propose a **bilingual lexical cohesion trigger model** to capture lexical cohesion for document-level SMT. We consider a lexical cohesion item in the source language and its corresponding counterpart in the target language as a trigger pair, in which we treat the source language lexical cohesion item as the trigger and its target language counterpart as the triggered item. Then we use mutual information to measure the strength of the dependency between the trigger and triggered item.

We integrate this model into a hierarchical phrase-based SMT system. Experiment results show that it is able to achieve substantial improvements over the baseline.

The remainder of this paper proceeds as follows: Section 2 introduces the related work and highlights the differences between previous methods and our model. Section 3 elaborates the proposed bilingual lexical cohesion trigger model, including the details of identifying lexical cohesion devices, measuring dependency strength of bilingual lexical cohesion triggers and integrating the model into SMT. Section 4 presents experiments to validate the effectiveness of our model. Finally, Section 5 concludes with future work.

2 Related Work

As a linguistic means to establish inter-sentential links, cohesion has been explored in the literature of both linguistics and computational linguistics. Cohesion is defined as relations of meaning that exist within the text and divided into grammatical cohesion that refers to the syntactic links between text items and lexical cohesion that is achieved through word choices in a text by Halliday and Hasan (1976). In order to improve the quality of machine translation output, cohesion has served as a high level quality criterion in post-editing (Vasconcellos, 1989). As a part of COMTIS project, grammatical cohesion is integrated into machine translation models to capture inter-sentential links (Cartoni et al., 2011). Wong and Kit (2012) incorporate lexical cohesion to machine translation evaluation metrics to evaluate document-level machine translation quality. Xiong et al. (2013) integrate various target-side lexical cohesion devices into document-level machine translation. Lexical cohesion is also partially explored in the cache-based translation models of Gong et al. (2011) and translation consistency constraints of Xiao et al.

*Corresponding author

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 382–386,
Sofia, Bulgaria, August 4-9 2013.

(2011).

All previous methods on lexical cohesion for document-level machine translation as mentioned above have one thing in common, which is that they do not use any source language information. Our work is mostly related to the mutual information trigger based lexical cohesion model proposed by Xiong et al. (2013). However, we significantly extend their model to a bilingual lexical cohesion trigger model that captures both source and target-side lexical cohesion items to improve target word selection in document-level machine translation.

3 Bilingual Lexical Cohesion Trigger Model

3.1 Identification of Lexical Cohesion Devices

Lexical cohesion can be divided into reiteration and collocation (Wong and Kit, 2012). Reiteration is a form of lexical cohesion which involves the repetition of a lexical item. Collocation is a pair of lexical items that have semantic relations, such as synonym, near-synonym, superordinate, subordinate, antonym, meronym and so on. In the collocation, we focus on the synonym/near-synonym and super-subordinate semantic relations [1]. We define lexical cohesion devices as content words that have lexical cohesion relations, namely the reiteration, synonym/near-synonym and super-subordinate.

Reiteration is common in texts. Take the following two sentences extracted from a document for example (Halliday and Hasan, 1976).

1. There is a boy climbing the old *elm*.
2. That *elm* is not very safe.

We see that word *elm* in the first sentence is repeated in the second sentence. Such reiteration devices are easy to identify in texts. Synonym/near-synonym is a semantic relationship set. We can use WordNet (Fellbaum, 1998) to identify them. WordNet is a lexical resource that clusters words with the same sense into a semantic group called synset. Synsets in WordNet are organized according to their semantic relations. Let s(*w*) denote a function that defines all synonym words of *w* grouped in the same synset in WordNet. We can use the function to compute all synonyms and near-synonyms for word *w*. In order to represent conveniently, s_0 denotes the set of synonyms in s(*w*). Near-synonym set s_1 is defined as the union of all synsets that are defined by the function s(*w*) where *w*$\in s_0$. It can be formulated as follows.

$$s_1 = \bigcup_{w \in s_0} s(w) \quad (1)$$

$$s_2 = \bigcup_{w \in s_1} s(w) \quad (2)$$

$$s_3 = \bigcup_{w \in s_2} s(w) \quad (3)$$

Similarly s_m can be defined recursively as follows.

$$s_m = \bigcup_{w \in s_{m-1}} s(w) \quad (4)$$

Obviously, We can find synonyms and near-synonyms for word *w* according to formula (4).

Superordinate and subordinate are formed by words with an is-a semantic relation in WordNet. As the super-subordinate relation is also encoded in WordNet, we can define a function that is similar to s(*w*) identify hypernyms and hyponyms.

We use *rep*, *syn* and *hyp* to represent the lexical cohesion device reiteration, synonym/near-synonym and super-subordinate respectively hereafter for convenience.

3.2 Bilingual Lexical Cohesion Trigger Model

In a bilingual text, lexical cohesion is present in the source and target language in a synchronous fashion. We use a trigger model capture such a bilingual lexical cohesion relation. We define xRy (R$\in\{$*rep*, *syn*, *hyp*$\}$) as a trigger pair where x is the trigger in the source language and y the triggered item in the target language. In order to capture these synchronous relations between lexical cohesion items in the source language and their counterparts in the target language, we use word alignments. First, we identify a monolingual lexical cohesion relation in the target language in the form of tRy where t is the trigger, y the triggered item that occurs in a sentence succeeding the sentence of t, and R$\in\{$*rep*, *syn*, *hyp*$\}$. Second, we find word x in the source language that is aligned to t in the target language. We may find multiple words x_1^k in the source language that are aligned to t. We use all of them x_iRt(1$\leq$i$\leq$k) to define bilingual lexical cohesion relations. In this way, we can create bilingual lexical cohesion relations xRy (R$\in\{$*rep*, *syn*, *hyp*$\}$): x being the trigger and y the triggered item.

[1] Other collocations are not used frequently, such as antonyms. So we we do not consider them in our study.

The possibility that y will occur given x is equal to the chance that x triggers y. Therefore we measure the strength of dependency between the trigger and triggered item according to pointwise mutual information (PMI) (Church and Hanks, 1990; Xiong et al., 2011).

The PMI for the trigger pair xRy where x is the trigger, y the triggered item that occurs in a target sentence succeeding the target sentence that aligns to the source sentence of x, and R∈{*rep, syn, hyp*} is calculated as follows.

$$PMI(xRy) = log(\frac{p(x,y,R)}{p(x,R)p(y,R)}) \quad (5)$$

The joint probability $p(x,y,R)$ is:

$$p(x,y,R) = \frac{C(x,y,R)}{\sum_{x,y} C(x,y,R)} \quad (6)$$

where $C(x,y,R)$ is the number of aligned bilingual documents where both x and y occur with the relation R in different sentences, and $\sum_{x,y} C(x,y,R)$ is the number of bilingual documents where this relation R occurs. The marginal probabilities of $p(x,R)$ and $p(y,R)$ can be calculated as follows.

$$p(x,R) = \sum_{y} C(x,y,R) \quad (7)$$

$$p(y,R) = \sum_{x} C(x,y,R) \quad (8)$$

Given a target sentence y_1^m, our bilingual lexical cohesion trigger model is defined as follows.

$$MI_R(y_1^m) = \prod_{y_i} exp(PMI(\cdot Ry_i)) \quad (9)$$

where y_i are content words in the sentence y_1^m and PMI($\cdot$Ry_i)is the maximum PMI value among all trigger words x_1^q from source sentences that have been recently translated, where trigger words x_1^q have an R relation with word y_i.

$$PMI(\cdot Ry_i) = max_{1 \le j \le q} PMI(x_j Ry_i) \quad (10)$$

Three models $MI_{rep}(y_1^m)$, $MI_{syn}(y_1^m)$, $MI_{hyp}(y_1^m)$ for the reiteration device, the synonym/near-synonym device and the super-subordinate device can be formulated as above. They are integrated into the log-linear model of SMT as three different features.

3.3 Decoding

We incorporate our bilingual lexical cohesion trigger model into a hierarchical phrase-based system (Chiang, 2007). We add three features as follows.

- $MI_{rep}(y_1^m)$
- $MI_{syn}(y_1^m)$
- $MI_{hyp}(y_1^m)$

In order to quickly calculate the score of each feature, we calculate PMI for each trigger pair before decoding. We translate document one by one. During translation, we maintain a cache to store source language sentences of recently translated target sentences and three sets S_{rep}, S_{syn}, S_{hyp} to store source language words that have the relation of {*rep, syn, hyp*} with content words generated in target language. During decoding, we update scores according to formula (9). When one sentence is translated, we store the corresponding source sentence into the cache. When the whole document is translated, we clear the cache for the next document.

4 Experiments

4.1 Setup

Our experiments were conducted on the NIST Chinese-English translation tasks with large-scale training data. The bilingual training data contains 3.8M sentence pairs with 96.9M Chinese words and 109.5M English words from LDC[2]. The monolingual data for training data English language model includes the Xinhua portion of the Gigaword corpus. The development set is the NIST MT Evaluation test set of 2005 (MT05), which contains 100 documents. We used the sets of MT06 and MT08 as test sets. The numbers of documents in MT06, MT08 are 79 and 109 respectively. For the bilingual lexical cohesion trigger model, we collected data with document boundaries explicitly provided. The corpora are selected from our bilingual training data and the whole Hong Kong parallel text corpus[3], which contains 103,236 documents with 2.80M sentences.

[2]The corpora include LDC2002E18, LDC2003E07, LDC2003E14,LDC2004E12,LDC2004T07,LDC2004T08(Only Hong Kong News), LDC2005T06 and LDC2005T10.

[3]They are LDC2003E14, LDC2004T07, LDC2005T06, LDC2005T10 and LDC2004T08 (Hong Kong Hansards/Laws/News).

We obtain the word alignments by running GIZA++ (Och and Ney, 2003) in both directions and applying "grow-diag-final-and" refinement (Koehn et al., 2003). We apply SRI Language Modeling Toolkit (Stolcke, 2002) to train a 4-gram language model with Kneser-Ney smoothing. Case-insensitive NIST BLEU (Papineni et al., 2002) was used to measure translation performance. We used minimum error rate training MERT (Och, 2003) for tuning the feature weights.

4.2 Distribution of Lexical Cohesion Devices in the Target Language

Cohesion Device	Percentage(%)
rep	30.85
syn	17.58
hyp	18.04

Table 1: Distributions of lexical cohesion devices in the target language.

In this section we want to study how these lexical cohesion devices distribute in the training data before conducting our experiments on the bilingual lexical cohesion model. Here we study the distribution of lexical cohesion in the target language (English). Table 1 shows the distribution of percentages that are counted based on the content words in the training data. From Table 1, we can see that the reiteration cohesion device is nearly a third of all content words (30.85%), synonym/near-synonym and super-subordinate devices account for 17.58% and 18.04%. Obviously, lexical cohesion devices are frequently used in real-world texts. Therefore capturing lexical cohesion devices is very useful for document-level machine translation.

4.3 Results

System	MT06	MT08	Avg
Base	30.43	23.32	26.88
rep	31.24	23.70	27.47
syn	30.92	23.71	27.32
hyp	30.97	23.48	27.23
rep+syn+hyp	**31.47**	**23.98**	**27.73**

Table 2: BLEU scores with various lexical cohesion devices on the test sets MT06 and MT08. "Base" is the traditonal hierarchical system, "Avg" is the average BLEU score on the two test sets.

Results are shown in Table 2. From the table, we can see that integrating a single lexical cohesion device into SMT, the model gains an improvement of up to 0.81 BLEU points on the MT06 test set. Combining all three features *rep+syn+hyp* together, the model gains an improvement of up to 1.04 BLEU points on MT06 test set, and an average improvement of 0.85 BLEU points on the two test sets of MT06 and MT08. These stable improvements strongly suggest that our bilingual lexical cohesion trigger model is able to substantially improve the translation quality.

5 Conclusions

In this paper we have presented a bilingual lexical cohesion trigger model to incorporate three classes of lexical cohesion devices, namely the reiteration, synonym/near-synonym and super-subordinate devices into a hierarchical phrase-based system. Our experimental results show that our model achieves a substantial improvement over the baseline. This displays the advantage of exploiting bilingual lexical cohesion.

Grammatical and lexical cohesion have often been studied together in discourse analysis. In the future, we plan to extend our model to capture both grammatical and lexical cohesion in document-level machine translation.

Acknowledgments

This work was supported by 863 State Key Project (No.2011AA01A207) and National Key Technology R&D Program(No.2012BAH39B03). Qun Liu was also partially supported by Science Foundation Ireland (Grant No.07/CE/I1142) as part of the CNGL at Dublin City University. We would like to thank the anonymous reviewers for their insightful comments.

References

Bruno Cartoni, Andrea Gesmundo, James Henderson, Cristina Grisot, Paola Merlo, Thomas Meyer, Jacques Moeschler, Sandrine Zufferey, Andrei Popescu-Belis, et al. 2011. Improving mt coherence through text-level processing of input texts: the comtis project. http://webcast. in2p3. fr/videos-the_comtis_project.

David Chiang. 2007. Hierarchical phrase-based translation. *computational linguistics*, 33(2):201–228.

Kenneth Ward Church and Patrick Hanks. 1990. Word

association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Christine Fellbaum. 1998. Wordnet: An electronic lexical database.

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

M.A.K Halliday and Ruqayia Hasan. 1976. Cohesion in english. *English language series*, 9.

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. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Proceedings of the international conference on spoken language processing*, volume 2, pages 901–904.

Muriel Vasconcellos. 1989. Cohesion and coherence in the presentation of machine translation products. *Georgetown University Round Table on Languages and Linguistics*, pages 89–105.

Billy T. M. Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea, July. Association for Computational Linguistics.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level consistency verification in machine translation. In *Machine Translation Summit*, volume 13, pages 131–138.

Deyi Xiong, Min Zhang, and Haizhou Li. 2011. Enhancing language models in statistical machine translation with backward n-grams and mutual information triggers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1288–1297, Portland, Oregon, USA, June. Association for Computational Linguistics.

Deyi Xiong, Guosheng Ben, Min Zhang, Yajuan Lv, and Qun Liu. 2013. Modeling lexical cohesion for document-level machine translation. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, Beijing,China.