Document-Level Machine Translation Evaluation Metrics Enhanced with Simplified Lexical Chain

Zhengxian Gong(✉) and Guodong Zhou

School of Computer Science and Technology,
Soochow University, Suzhou 215006, China
{zhxgong,gdzhou}@suda.edu.cn,
http://nlp.suda.edu.cn

Abstract. Document-level Machine Translation (MT) has been drawing more and more attention due to its potential of resolving sentence-level ambiguities and inconsistencies with the benefit of wide-range context. However, the lack of simple yet effective evaluation metrics largely impedes the development of such document-level MT systems. This paper proposes to improve traditional MT evaluation metrics by simplified lexical chain, modeling document-level phenomena from the perspectives of text cohesion. Experiments show the effectiveness of such method on evaluating document-level translation quality and its potential of integrating with traditional MT evaluation metrics to achieve higher correlation with human judgments.

Keywords: MT · Evaluation metrics · Lexical chain · Document-level MT

1 Introduction

According to modern trends in linguistics, a text rather than individual words or fragments of sentences is preferred as the unit of communication [1]. Therefore, more researchers propose to build document-level Statistical Machine Translation (SMT) systems in recent years [8,11,17–19]. However, most of their experimental results show moderate or even minimal improvement despite of a great deal of efforts. Some authors doubted the failure of current MT evaluation metrics to objectively reflect the changes at document level.

Automatic evaluation metrics have close relation to MT and current SMT especially depends on them since SMT usually requires automatic metrics to tune parameters. Some automatic MT metrics, such as BLEU [15] and METEOR [2], give evaluation by measuring the amount of segment overlap between MT output and reference at sentence level. And document/system-level evaluation scores can be obtained by accumulating (not simply averaging) sentence-level scores. Obviously, such accumulation loses some document-level information, such as the difference of sentences. Thus it is unlikely that a document-level SMT system can obtain significant improvements guided by such inappropriate evaluation metrics.

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 396–403, 2015.
DOI: 10.1007/978-3-319-25207-0_35

A document is made up of sentences, but there exist separate principles of text-construction, beyond the rules for making sentences. Document-level metrics should have the capability of identifying text-level rather than sentence-level change. In recent years, automatic document-level MT metrics have been drawing more and more attention. Gimenez et al. [7] propose a MT evaluation metric based on Discourse Representation Theory [12], which uses co-reference and discourse relations to assess the quality of MT output. However, their metric fails to achieve a higher correlation with human assessments than sentence-level metrics. In order to measure cohesion for translated text, Wong and Kit propose to use lexical cohesion devices and obtain positive experimental results [20] . To better estimate translation adequacy, Rubino et al. propose to do quality estimation for machine translation (without references) by bilingual topic models and show some promising results [16].

A text is a "communicative occurrence which meets seven standards of textuality: cohesion, coherence, intentionality, acceptability, informativity, situationality and intertextuality" [4]. According to this definition, cohesion and coherence are important standards of textuality. Coherence interprets meaning connectedness in the underlying text while cohesion can be formulated quite explicitly on the basis of grammatical and lexical properties. This paper proposes simple yet effective cohesion score to measure text cohesion via lexical chain. Our experimental results show the number of matching lexical chain between reference and MT output has close relevance to the translation quality of full text.

2 Related Work

As the most famous evaluation metric, BLEU is based on n-gram matching. Alternatively, METEOR is based on unigram alignment of references and system translations. METEOR is explicitly designed to improve the correlation with human judgments at sentence level [2]. Document-level BLEU or METEOR score can be generated by aggregating sentences in a document rather than simply averaging scores at sentence level.

Wong and Kit [20] propose to build document-level MT metrics by using lexical cohesion devices. Lexical cohesion devices refer to content words (stopwords are removed) that reiterate once or more times in a document. In their study, the higher ratio of such content words in machine translated text means stronger lexical cohesion.

Text cohesion refers to top-level characteristics of text while document-level BLEU/METEOR score shows the degree to which the detailed information in the original text is conveyed in MT output. Therefore, it is natural to build document-level metrics by extending traditional MT evaluation metrics with document-level feature scores. Thus, Wong and Kit [20] built document-level metrics by extending traditional metrics with lexical cohesion scores as follows:

$$H = \alpha \times CS_{doc} + (1 - \alpha) \times G_{mdoc} \,. \tag{1}$$

where CS_{doc} means lexical cohesion score and G_{mdoc} refers to document-level BLEU/METEOR score, and α is a weight controlling their proportion.

The success of [20] is due to such observation: most SMT systems tend to build less lexical cohesion than human translators. However such observation seems to be controversial. Carpuat and Simard [6] show: MT output tend to have more incorrect repetition than human translation when MT systems especially trained on smaller corpora. Thus, metrics in [20] cannot distinguish such "false" cohesion devices.

3 Evaluation Data

Table 1 shows the evaluation data for this study, including Multiple-Translation Chinese Part 2 (LDC2003T17, MTC2 for short) and Multiple-Translation Chinese Part 4 (LDC2006T04, MTC4 for short). The MTC2 consists of 878 source sentences, translated by 4 human translators (references) as well as 3 MT systems. The MTC4 consists of 919 source sentences, translated by 4 human translators (references) as well as 6 MT systems.

Table 1. Evaluation data

Data Set	MTC2	MTC4
Source language	Chinese	Chinese
Target language	English	English
Systems	3	6
♯ Documents	100	100
♯ Sentences	878	919
♯ References	4	4
♯ Genre	Newswire	Newswire

Besides, each machine translated sentence on the MTC4 and MTC2 was evaluated by 2 to 3 human judges for their adequacy and fluency on a 5-point scale. To avoid the bias in the distributions of different judges' assessments in the evaluation data, we normalize these scores following Blatz et al. [5].

Due to the lack of document-level human assessments on the two evaluation data sets, document-level human assessments are averaged over sentence scores, weighted by sentence length. This method is also adopted by famous Metrics-MaTr (the NIST Metrics for Machine Translation Challenge) and approximated in [7] and [20].

4 Text Cohesion Representing by Simplified Lexical Chain

The major problem in [20] is only to measure the cohesion of MT output and completely ignore the one of references. In another word, we think the cohesion score between MT output and references should be consistent. In this study,

we use simple yet effective lexical chain to measure lexical cohesion at document level. The basic idea is to compute the number of matching lexical chains respectively in reference and MT output.

4.1 Simplified Lexical Chain

Traditional lexical chain is the sequence of semantically related words [14]. During the process of chaining, special thesaurus, such as WordNet and HowNet, is often used to help recognize synonyms and hyper/hypo-nyms. For the general-purpose, this study only focuses on reiterating words including stem-matched words. Another difference is that our lexical chain mainly used to record the difference of position for each content word. To distinguish our lexical chain from traditional lexical chain, we call it as simplified lexical chain.

The establishing procedure of simplified lexical chain is simple. For each document d:

(1) extracting all unique content words occurring at the first sentence, then constructing an array for each word. Here such array corresponds to one lexical chain. And the element in the chain records its location(sentence identity). All these chains are stored into a hash table, ht, shown in Fig.1.
(2) for each successive sentence, if content word w or stem(w) has existed in ht, then inserting its location information into the corresponding chain, else constructing a new chain for w and inserting this chain into ht.
(3) removing chains from ht which only contain one word.

So each content word which appears more than one time at different sentences will have a lexical chain. For example, Fig.1 shows a lexical chain $LC1$ for the word "die" (perhaps with different morphology) which occurs at the 1st, 2nd and 3rd sentence. There are several lexical chains in one document, thus a hash table ht is utilized to organize all these chains. For clarity, ht is called as lexical-chain index. In this hash table, keys are content words and values refer to lexical chains.

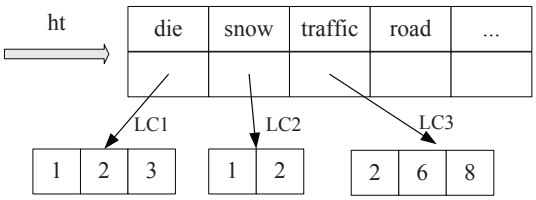

Fig. 1. The structure of the lexical-chain index for one document

4.2 The Characteristics of Lexical-Chain Index

We constructed lexical-chain index for each document on our evaluation data, including 4 human translations (references) and all MT output on different corpus in advance. After that, we carefully studied these chains and achieved the following observations.

The First Observation: lexical chains extracted from reference are more consistent than the ones from MT. Table 2 lists 5 translation versions for one source document. The first 4 columns (E01-E04) correspond to references and the last column refers to one MT output. One word with a list of position represents a lexical chain. The upper part of Table 2 shows some matching lexical chains, which appear in all references. Due to the flexibility of expression, there exist a few un-matching chains shown on the lower part of Table 2. On the whole, the matching number of lexical chain in references greatly exceeds the one in MT output. Moreover, the chains in references are very consistent both in frequency and location while the chains extracted from MT output have large difference.[1]

The Second Observation: it is not always right that lexical cohesion of human translation exceeds the one of MT output.

On the whole evaluation data, we count the total number of lexical chains extracted from human and MT output respectively. The average number of chains extracted from human translation (2111) is greater than the one of MT output (1999) on MTC4, that means lexical cohesion devices existed in human translation more than the one in MT output, which is consistent to the observation described in [20]. But the number of lexical chain extracted from each MT system on MTC2 (2380) exceeds the one from human translations (2030). So the assumption in [20] is not right on MTC2.

Based on above observations, we know there exist some differences in lexical-chain index even all of them extracted from equivalent references due to the complexity and flexibility of linguistics. But they are more stable and consistent among human translations than those in MT output.

4.3 Text Cohesion Scores Based on the Matching of Lexical Chain

Due to high flexibility of natural language utterances, few lexical chains from MT output can completely match the ones from its references. So we need to design a reasonable function to permit incomplete matching.

For a document, the lexical-chain index in reference and in MT output denoted as ht_{ref} and ht_{mt} . We first extract the word of one chain in ht_{mt}, and find its corresponding lexical chain ID in ht_{ref}. Given a pair of matching lexical chain of ht_{ref} and ht_{mt} is LC_r and LC_t. LC_r contains m elements and

[1] According to the LDC manual, the ranking for the manual translations is $E01 > E02 > E03 > E04$, so the matching lexical chain in E04 has slight difference in position to other references.

Table 2. An Example of lexical-chain indexes extracted from different translation versions for the same source document

References				Machine
E01	E02	E03	E04	Translation
Government:4,5	Government:4,5	Government:4,5	Government:4,5	-
Ople:2,4	Ople:2,4	Ople:2,4	Ople:2,4	-
Ambassador:2,4	Ambassador:2,4	Ambassador:2,4	Ambassador:2,4	Ambassador:1,2
Australian:2,6,7	Australian:2,6,7	Australian:2,6,7	Australian:2,6,7	-
Attack:1,8	Attack:1,8	Attack:1,8	Attack:1,8	Attack:1,8
Terrorist:1,6,8	Terrorist:1,6,8	Terrorist:1,6,8	Terrorist:1,6,8	Terrorist:1,6,8
Threat :1,5,6,8	Threat:1,5,6,8	Threat:1,5,6,8	Threat:1,5,6	Threat:1,6,8
Manila:0-2,6	Manila:0-2,6	Manila:0-2,6	Manila:0,1,6	-
Embassy:0-3,5-7	Embassy:0-3,5-7	Embassy:0-3,5-7	Embassy:0-3,5-7	Embassy:0,1
Reopen:0,1,3	Reopen:0,1,3	Reopen:0,1,3	Reopen:0,1,3	Reopen:0,1
Australium:0,1,3	Australium:0,1,3	Australium:0,1,3	Australium:0,1, 3,6	Australium:0-3, 6-7
Philippine:1,2,4	Philippine:1,2,4,5	Philippine:1,2,4,5	Philippine:2,4,5	Philippine:1,4
Week:1,3	Week:1,3	Week:1,3	Week:1,3	Week:1,3
Shut:1,8	Close:1,8	Shut:1,3,5,8	Close:1,3,5,8	Close:1,3,5,8
-	European:3,7	European:3,7	Eu:3,7	Open:0,1,5,6
-	Union:3,7	Union:3,7	-	Canada:3,6
-	-	So-call:1,5	So-call:1,5	China:1,4
-	Express:4,8	Officer:6,7	Mission:7,8	Chinese:2,5
		Due:1,8		Collaboration: 1,3,5

LC_t contains n elements, but only $m'(m' <= m)$ element both occur in LC_r and LC_t , then the cohesion score of LC_t can be calculated by the following formula:

$$CS_i = m'/m \ . \tag{2}$$

CS_i only refers to one pair of matching chain. If one chain cannot be found in its reference, the chain is invalid ("false"). Suppose ht_{mt} contains K lexical chains, we punish such "false" cohesion by averaging K. Given the number of matching chain is L, the final cohesion score assigned to ht_{mt} is calculated as follows:

$$Doc_{cohesion} = (\sum_{i=1}^{L} CS_i)/K \ . \tag{3}$$

For example, one lexical-chain in column of "E01" in Table 2 is "Ambassador:2,4" while its matching chain in MT shown in Table 2 is "Ambassador:1,2", so the CS value of this chain is 1/2(only 1 item is matching). We use this policy to calculate CS value and finally to obtain cohesion score for the whole document. We choose the best $doc_{cohesion}$ against 4 references.

5 Experiments

Following the formula 1, we build document-level metrics by combining document-level BLEU or METEOR with text cohesion score calculated by the formula 3. Especially, the gradient ascending algorithm described in [13] is utilized to automatically tune the weight α.

Table 3. The Correlation of different metrics with human assessments at document level

DataSet	MTC2		MTC4	
Metrics	Pearson	Kendall	Pearson	Kendall
BLEU	0.0994	0.0449	0.5862	0.4256
METEOR	0.3069	0.2037	0.7390	0.5180
$Doc_{cohesion}$	0.1284	0.0609	0.6891	0.4601
HBLEU	0.1240	0.0698	0.6551	0.4800
HMETEOR	0.3107	0.2103	0.7467	0.5244

The results of our proposed metrics are shown in Table 3. To our surprise, the solely use of lexical cohesion already outperforms document-level BLEU, but it still subordinates to METEOR. The hybrid BLEU (HBLEU) scores rise from 42.56% to 48.00% on Kendall score on MTC4 and with a similar increase on MTC2. Furthermore, differing with the results in [20], our hybrid METEOR (HMETEOR) scores also obtain a moderate rise (0.64%- 0.67%) both on MTC4 and MTC2.

6 Conclusion

This paper describes how to modeling text cohesion for MT output based on simplified lexical chains. We successfully build reasonable document-level evaluation metrics by extending traditional MT evaluation metrics with text cohesion score based on lexical chains.

Since important words will be repeated in one text, lexical chains can not only model text cohesion but also highlight key words. So our proposed metrics can obtain very significant improvements for BLEU and also give moderate improvements for METEOR.

In the future work, we will explore how to estimate more document-level features, such as co-reference matching, and hope our study can bring more inspirations to document-level SMT.

Acknowledgments. This research is supported by the National Natural Science Foundation of China under grant No.61305088 and No.61401295.

References

1. Al-Amri, K.H.: Text-linguistics for Students of Translation. King Saud University (2007)
2. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72 (2005)
3. Barzilay, R., Lapata M.: Modeling local coherence: an entity-based approach. In: Proceedings of ACL, pp. 141–148 (2008)
4. Beaugrande, R.D., Dressler, W.U.: Introduction to Text Linguistics. Longman, London (1981)
5. Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., Ueffing, N.: Confidence Estimation for Machine Translation. Technical report, Natural Language Engineering Workshop Final Report (2003)
6. Carpuat, M., Simard, M.: The trouble with SMT consistency. In: Proceedings of the 7th Workshop on Statistical Machine Translation, pp. 442–449 (2012)
7. Gimenez, J., Marquez, L., Comelles, E., Castellon, I., Arranz, V.: Document-level automatic MT evaluation based on discourse representations. In: Proceedings of WMT and MetricsMATR, pp. 333–338 (2010)
8. Gong, Z.X., Zhang, M., Zhou, D.: Cache-based document-level statistical machine translation. In: Proceedings of EMNLP, pp. 909–919 (2011)
9. Guzman, F., Joty, S., M'arquez, L.: Using discourse structure improves machine translation evaluation. In: Proceedings of ACL, pp. 687–698 (2014)
10. Halliday, M.A.K., Hasan, R.: Cohesion in English. Longman, London (1976)
11. Hardmeier, C., Nivre, J., Tiedemann, J.: Document-wide decoding for phrase-based statistical machine translation. In: Proceedings of EMNLP, pp. 1179–1190 (2012)
12. Kamp, H., Reyle, U.: From Discourse to Logic. Introduction to Model Theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. Kluwer Academic Publishers, Dordrecht (1993)
13. Liu, D., Gildea, D.: Source-language features and maximum correlation training for machine translation evaluation. In: Proceedings of NAACL, pp. 41–48 (2007)
14. Morris, J., Hirst, G.: Lexical Cohesion Computed by Thesauri Relations as an Indicator of the Structure of Text. Computational Linguistics **17**(1), 21–48 (1991)
15. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for au-tomatic evaluation of machine translation. In: Proceedings of ACL, pp. 311–318 (2002)
16. Rubino, R., Jos'e, G.C.S., Foster, J., Specia, L.: Topic models for translation quality estimation for gisting purposes. In: Proceedings of the XIV Machine Translation Summit, pp. 295–302 (2013)
17. Tiedemann, J.: Context adaptation in statistical machine translation using models with exponentially decaying cache. In: Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, pp. 8–15 (2010)
18. Xiao, T., Zhu, J.B., Yao, S.J., Zhang, H.: Document-level consistency verification in machine translation. In: Proceedings of MT Summit XIII, pp. 131–138 (2011)
19. Xiong, D.Y., Ding, Y., Zhang, M., Tan, C.L.: Lexical chain based cohesion models for document-level statistical machine translation. In: Proceedings of EMNLP, Seattle, Washington, USA, pp. 1563–1573 (2013)
20. Wong, B.T.M., Kit, C.: Extending machine translation evaluation metrics with lexical cohesion to document level. In: Proceedings of EMNLP, pp. 1060–1068 (2012)