Document-Level Machine Translation Evaluation with Gist Consistency and Text Cohesion

Zhengxian Gong Min Zhang Guodong Zhou*
School of Computer Science and Technology, Soochow University, Suzhou, China 215006
{zhxgong, minzhang, gdzhou}@suda.edu.cn

Abstract

Current Statistical Machine Translation (SMT) is significantly affected by Machine Translation (MT) evaluation metric. Nowadays the emergence of document-level MT research increases the demand for corresponding evaluation metric. This paper proposes two superior yet low-cost quantitative objective methods to enhance traditional MT metric by modeling document-level phenomena from the perspectives of gist consistency and text cohesion. The experimental results show the proposed metrics can obtain better correlation with human judgments than traditional metrics on evaluating document-level translation quality.

1 Introduction

Since most of current SMT models impose strong independence assumptions on words and sentences, most of these systems only work at sentence level and cannot employ useful relationships among sentences during decoding. However, a text rather than individual words or fragments of sentences is the basic unit of communication (Al-Amri, 2007). Beaugrande and Dressler (1981) define that text is a communicative occurrence which meets seven standards, such as textuality cohesion, coherence. Text is constituted by sentences, but there exist separate principles of text-construction beyond the rules for making sentences (Fowler, 1991).

Document is the carrier of text in modern computer system. Currently more researching work focus on document-level SMT (Tiedemann, 2010; Xiao et al, 2011; Gong et al, 2011; Ture et al., 2012; Hardmeier et al., 2012; Xiong et al, 2013). However, most of these researches show their improvements by using system-level metrics, such as BLEU (Papineni et al., 2002). Whether improvements in performance at system level are really able to reflect the change of text-level translation quality is still to doubt.

Nowadays, the study of real document-level MT metrics has been drawing more and more attention. Based on Discourse Representation Theory (Kamp and Reyle, 1993), Gimenez et al. (2010) propose to use co-reference and discourse relations to build evaluation metrics. The metrics by extending traditional metrics with lexical cohesion devices show some positive experimental results (Wong and Kit, 2012). Bilingual topic model (Blei et al., 2003) is applied to do MT quality estimation(Raphael et al., 2012; Raphael et al, 2013). Guzman et al. (2014) use two discourse-aware similarity measures based on discourse structure to improve existing MT evaluation metrics.

According to the afore-mentioned definition of text, the most important standard of evaluating translation quality for one document should be to what degree the MT output correctly communicates the main idea of origin text. From this regard, this paper first proposes to measure gist consistency of text via topic model. Topic model is a statistical model which assumes each document can be characterized by a particular set of topics. Currently a variety of probabilistic topic models (Landauer et al., 1998; Hofmann, 1999; Blei et al., 2003) have been used to analyze the content of documents and the meaning of words. Our experimental results show the MT evaluation metrics with robust topic model can effectively capture change of translation quality between reference and MT output at document level.

Furthermore, cohesion and coherence are important standards of textuality. Coherence

*Corresponding author.

Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT), pages 33–40,
Lisbon, Portugal, 17 September 2015.

interprets meaning connectedness in the underlying text while cohesion can be formulated quite explicitly on the basis of grammatical and lexical properties (Halliday and Hasan, 1976). This paper describes a simple yet effective cohesion function to measure text cohesion via lexical chain. Our experimental results show that the number of matching lexical chain between reference and MT output can reflect the goodness of translation at document level.

The rest of this paper is organized as follows: Section 2 and 3 respectively describes how to model two kinds of document-level features. Section 4 shows the framework of combing document-level scores with traditional metrics. Section 5 presents the experimental results and Section 6 gives out discussion. Finally, we conclude this paper in Section 7.

2 Gist Consistency Score based on Topic Model

Reeder (2006) proposes to measure MT adequacy at the document level with Latent Semantic Analysis (LSA) (Landauer et al., 1998). However, Reeder only uses a set of complex configuration to show the close correlation between LSA model and human assessments and does not suggest how to use it to design an evaluation metric.

Raphael et al. (2012; 2013) exploit bilingual topic models to do quality estimation (without references) for machine translation. In this study, since each evaluation document has 4 references, we show a simple way to design document-level metrics with monolingual topic model.

2.1 Topic Model

LDA (Blei et al., 2003) is one of the most common topic models which assumes each document is a mixture of various topics and each word is generated with multinomial distribution conditioned on a topic. We use an off-the-shelf LDA tool[1] to train a topic model with 86070 news (happened in 2004 year) documents coming from the Xinhua portion of the Gigaword corpus (LDC2005T12).

A trained LDA model produces two kinds of distributions: the "document-topic" distribution and the "topic-word" distribution. Suppose there are *K* topics, the *k*-th dimension $P(z = k|d)$ means the probability of topic *k* given document *d*. The whole document-topic distribution over *K* topics for one document *d*, denoted as $P(Z|d)$, can be represented by a K-dimension vector. In this study, when K set to 120, the trained LDA model can be tuned with the minimal perplexity (Blei et al., 2003).

2.2 Measure of Topic Consistency

After constructing a trained topic model, the "document-topic" distribution of MT output and reference on evaluation dataset (see Section 5.1) can be respectively **inferred**. We use Kullback-Leibler divergence to measure topic consistency between MT output and reference with the basic unit of document. Denote the "document-topic" distribution of one reference (d_r) as $P(Z|d_r)$, and the one of its MT output (d_t) as $Q(Z|d_t)$, the KL divergence of Q from P is defined to be:

$$D_{KL}(P||Q) = \sum_{i=1}^{G} P(z_i|d_r) \times \ln \frac{P(z_i|d_r)}{Q(z_i|d_t)} \quad (1)$$

In theory, G should keep same to the value of the trained LDA model ($K = 120$). However our initial experiment results show the hybrid METEOR has a drop on adequacy on evaluation dataset by using a static G.

To address such problem, we output the number of topics whose document-topic probability is great than 0.01 (called as *valid topic*) for each **reference** document and found the range of this number is [7,31]. Obviously the inferred topic model contains plenty of noise topics and we need measure *valid topic* rather than all topics consistency for each document.

Therefore, before computing topic consistency, we first record the IDs of *valid topics* for one reference, then obtain corresponding "document-topic" probability of evaluation document according to these topic IDs. Thus, in this study, G is dynamically set according to the number of valid topics of each reference.

There are 4 references per document in evaluation data. One machine translated document is scored against each reference independently, and the minimal D_{KL} is used. The score of *topic consistency* for each evaluation document, denoted as S_{topic}, is computed by the following formula:

$$S_{topic} = e^{-D_{KL}} \quad (2)$$

[1] http://www.arbylon.net/projects/

3 Cohesion Score based on Simplified Lexical Chain

Text adequacy is the most important standard for the purpose of successful communication. According to the work of Wong and Kit (2012), cohesion is another important element to organize text. They found: SMT systems tend to use less lexical cohesion devices than those of human translators. Here lexical cohesion devices mainly refer to content words reiterating once or more times in a document. They propose to build document-level MT metrics by integrating cohesion score based on lexical cohesion devices.

However, Carpuat and Simard (2012) draw a different conclusion: MT output tend to have more incorrect repetition than human translation when the MT model is especially trained on smaller corpora. Suppose these incorrect repetition as "false" cohesion, metrics in (Wong and Kit, 2012) will fail to distinguish such "false" cohesion devices.

In our opinion, the lack of Wong's work is completely ignoring text cohesion of references, and they only model the cohesion score of MT output. In this study, we assume the correct cohesion of MT output should be consistent with the one of references. Reference is the equivalent of its source text. The MT output might be cohesive only if source text is cohesive, so the assumption is reliable. In this paper, we implement such assumption via a special structure, simplified lexical chain.

3.1 Simplified Lexical Chain

Differing from lexical chain in these work (Morris and Hirst, 1991; Galley and McKeown, 1993; Xiong et al, 2013) which is the sequence of semantically related words based on special thesaurus, our lexical chain refers to reiterating words including stem-matched words. Furthermore, it only records position information for each content word. Our lexical chain is simpler and might gain broader use because it doesn't require special thesaurus, such as WordNet and HowNet. Thus, we call such lexical chain as simplified lexical chain.

The detailed establishing procedure of simplified lexical chain is described in our another work (Gong and Zhou, 2015). The key of this procedure is to assure that each content word occurring at different sentences one more time is assigned an unique lexical chain. Figure 1 shows a lexical chain $LC1$ for the word "die" (perhaps with different morphology) and it records that "die" occurs at the 1st, 2nd and 3rd sentence. One document often contains several lexical chains, 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.

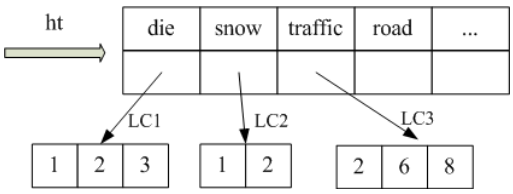

Figure 1: the structure of the lexical-chain index of one document

3.2 Cohesion Score

We constructed lexical-chain index for each document on our evaluation data, including 4 human translations (references) and all MT output on evaluation corpus in advance. Due to high flexibility of natural language utterances, few lexical chains from MT output can completely match the ones from its references. So we design a special function that permits incomplete matching to score text cohesion .

Suppose the lexical-chain index in reference and in MT output as ht_{ref} and ht_{mt} , we can find a pair of matching lexical chain of ht_{ref} and ht_{mt}, denoted as LC_r and LC_t. LC_r contains m elements and LC_t contains n elements, but only m' ($m' <= m$) elements both occur in LC_r and LC_t , then the cohesion score of LC_t can be calculated by the following formula:

$$CS_i = \frac{m'}{m} \tag{3}$$

CS_i only refers to one pair of matching chain. If one chain of MT output 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:

$$Doc_{cs} = \frac{\sum_{i=1}^{L} CS_i}{K} \tag{4}$$

We choose the best Doc_{cs} for one MT output against 4 references.

4 New Metrics by Combining Traditional Metrics with Document-level Scores

4.1 Traditional MT Evaluation Metrics

For fair comparison and possible integration of our proposed document-level features, this section gives a brief introduction on two widely adopted MT evaluation metrics: BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005).

As the most famous evaluation metric, BLEU is based on n-gram matching. Given a system translation, BLEU first collects all n-grams and count how many of them exist in one or more references (sentence by sentence), and then integrate the precisions of n-grams with different lengths into one score as follows:

$$BLEU = BP \cdot \exp(\frac{1}{4}\sum_{n=1}^{4}\log(P_n)) \ . \quad (5)$$

where p_n is the precision of n-gram and BP is a penalty factor, preventing BLEU from favoring short segments due to the lack of direct consideration of recall. It is obvious that, although BLEU takes all n-grams into consideration, the importance of different n-grams is ignored except their lengths.

METEOR is based on unigram alignment of references and MT output. Each unigram in one system translation is at most mapped to one unigram in the references first and then three successive stages of "exact", "porter stem" and "WN synonymy" are used to create alignment in turn. Once the final alignment is produced, unigram precision (P) and recall (R) are calculated and combined into one F_{mean} score:

$$F_{mean} = \frac{PR}{\alpha P + (1-\alpha)R} \ . \quad (6)$$

Finally, the METEOR score is obtained as follows:

$$score = (1 - pen)F_{mean} \ . \quad (7)$$

Where pen is a penalty factor. METEOR is explicitly designed to improve the correlation with human judgments of MT quality at the sentence level and the performance of METEOR outperforms BLEU at sentence level.

Based on the formula 5 or 7, document-level BLEU/METEOR score can be generated by aggregating sentences in a document rather than simply averaging scores at sentence level.

4.2 The Combining Framework

Gist consistency and text cohesion refer to top-level characteristics of text while traditional MT evaluation metrics, such as document-level BLEU, show the degree to which the n-grams also occur in the MT output. Inspired by the work of Wong and Kit (2012), we construct document-level metric by extending traditional metric with aforementioned two kinds of document-level scores as

$$H = \alpha \times S_{m_{doc}} + \beta \times G_{m_{doc}} \quad (8)$$

where $G_{m_{doc}}$ refers to document-level BLEU or METEOR score (one score per document), $S_{m_{doc}}$ to gist consistency score(S_{topic}) or text cohesion score(Doc_{cs}) proposed in this paper. α and β are weights which are tuned on MTC2 evaluation dataset (see Section 5.1) by a gradient ascending algorithm with the optimum goal of maximum correlation value (Liu and Gildea, 2007).

5 Experiments

5.1 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.

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 the scores following Blatz et al. (2003).

It is worth noting that, due to the lack of document-level human assessments on the two evaluation dataset, document-level human assessments are averaged over sentence scores, weighted by sentence length. This method is also adopted by famous MetricsMaTr (the NIST Metrics for Machine Translation Challenge) and approximated in Gimenez et al. (2010) and Wong and Kit (2012).

LDC corpus	LDC2003T17	LDC2006T04
Source language	Chinese	Chinese
Target language	English	English
Number of Systems	3	6
Number of Documents	100	100
Number of Sentences	878	919
Number of References	4	4
Genre	Newswire	Newswire

Table 1: Evaluation Data

5.2 The Performance of Extending Metrics

In this study, Pearson and Kendall coefficients are both used to formulate correlation following the way of MetricsMaTr. It noted, Pearson ranges from -1 to 1 with 1 for total positive correlation, 0 for no correlation and -1 for total negative correlation, while Kendall ranges from 0 to 1 with 0 for no agreement and 1 for complete agreement.

The document-level BLEU and METEOR scores (one score per document) are first obtained via the NIST BLEU script (version 13) and the METEOR toolkit 1.4. The correlation between traditional metrics and human judgements is shown in Table 2.

After introducing gist consistency score into traditional MT metrics, the Kendall correlation between the hybrid BLEU (HBLEU(s_{topic})) and human judgements rise from 42.56% to 48.66% on adequacy on MTC4, and with a similar increase on MTC2. The Kendall correlation of the hybrid METEOR (HMETEOR(s_{topic})) scores also obtain a significant rise (0.8%-1.4%) both on MTC4 and MTC2.

After introducing cohesion score into traditional metrics, the Kendall correlation between the hybrid BLEU (HBLEU) and human judgements 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 Wong's work, our hybrid METEOR (HMETEOR) scores also obtain a moderate rise (0.64%-0.67%) both on MTC4 and MTC2.

It seems gist consistency outperforms text cohesion on evaluating document-level MT output. It is worth noting the α and β is 1.47 and 0.51 on methods of combing gist consistency score with METEOR. The α and β is 1.82 and 0.02 on methods of combing text cohesion score with METEOR. It seems that cohesion score only plays a minor role on improving METEOR in this study. We think the approximated document-level human judgments may be the major reason (see section 5.1).

6 Discussion

6.1 The Impacts of Associating Gist Consistency with Text Cohesion

In this paper, Gist consistency is obtained based on LDA topic model that uses representative term for major topics existed in one document, and the training procedure of LDA actually relies on term repetition. Text cohesion is obtained based on simplified lexical chain which also depends on iterating words. In a sense, both of these measures are based on same kind of information (although measured differently). It would be interesting to see whether BLEU or METEOR with their combination can increase performance or not.

According to the results shown in Table 3, both document-level BLEU and METEOR enhanced with the combination of gist consistency and text cohesion is subordinate to its corresponding metrics only with gist consistency. BLEU with such combination is still superior to its enhanced metrics only with text cohesion while METEOR with such combination has a slight drop compared with its enhanced metrics only with text cohesion.

Metrics	MTC2	MTC4
HBLEU(combination)	0.0736	0.4850
HMETEOR(combination)	0.2083	0.5211

Table 3: The Kendall correlation between human judgments and the proposed metrics with the combination of gist consistency and text cohesion

METEOR uses WordNet to help evaluation, so METEOR can utilize synonym information. In this paper, LDA model utilize an additional large training corpora (see section 2), thus it may contain synonym information in some topics. Furthermore, we only focus on major topics of one document, which may help METEOR highlight some important words in the scope of documents.

In this study, the performance of METEOR with text cohesion has a slight improvement since our lexical chain ignores synonym for the general purpose. However, using different target words to translate the same source word in different context is common. In the future work, we will

Metrics	MTC2		MTC4	
	Pearson	Kendall	Pearson	Kendall
BLEU	0.0994	0.0449	0.5862	0.4256
METEOR	0.3069	0.2037	0.7401	0.5180
HBLEU(s_{topic})	0.1350	0.0741	0.6601	0.4866
HMETEOR(s_{topic})	0.3149	0.2177	0.7481	0.5260
HBLEU(Doc_{cs})	0.1240	0.0698	0.6551	0.4800
HMETEOR(Doc_{cs})	0.3107	0.2103	0.7467	0.5244

Table 2: The correlation between the proposed metrics combining with gist consistency/text cohesion with human judgments

build lexical chain by introducing synonyms.

Furthermore, it noted that one additional weight of formula 8 needs to be tuned with the gradient ascending algorithm, and it might be the another reason for degrading the performance.

6.2 The Characteristic of Text Cohesion based on Simplified Lexical Chain

We output the lexical chains on two evaluation dataset shown in Table 4. On MTC4, the average number of chains extracted from references (2111) is really more than the one of evaluated documents (1999), which is consistent to the observation in Wong's work. But such observation is not true on MTC2. Table 4 also shows each MT system on MTC2 produces more lexical chains (2380) than the average number of its reference (2030).

Genres	Item	Data	
		MTC4	MTC2
Reference	1	2125	2124
	2	2194	2079
	3	2087	2018
	4	2036	1897
	Avg	2111	2030
MT System	1	2488	2333
	2	2066	2469
	3	2029	2337
	4	2001	-
	5	2152	-
	6	1259	-
	Avg	1999	2380

Table 4: The number of lexical chains extracted from human translation and MT output on MTC4 and MTC2 (MTC2 only involves 3 MT systems)

Furthermore, compared with the column of $\sharp chain$ and $\sharp match_{chain}$ shown in Table 5, we observed there are plenty of invalid lexical chains existed in MT output.

Data	System	$\sharp$ chain	$\sharp match_{chain}$
MT System	1	2333	1180
	2	2469	1222
	3	2337	1262
	Avg:	2380	1221

Table 5: The number of lexical chains($\sharp$ chain) extracted from MT output and the number of lexical chain($\sharp match_{chain}$) refers to the chain which have corresponding lexical chain in its references on MTC2

7 Conclusion

We describes two kinds of document-level measures and successfully use them to construct document-level evaluation metrics.

Hybrid metrics based on topic model can produce significant positive impacts when given a robust trained topic model. 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 improvement for BLEU and also give might improvement for METEOR. Furthermore, hybrid metrics based on text cohesion has less limitation than topic-based method since it doesn't need additional training data, and it can be easily integrated into existing traditional metrics.

In the future, we will explore how to model 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

Al-Amri K.H. 2007. *Text-linguistics for students of translation*. King Saud University.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Banerjee Satanjeev and Lavie Alon. 2005. *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, pages: 65-72.

Blatz John, Fitzgerald Erin, Foster George, Gandrabur Simona, Goutte Cyril, Kulesza Alex, Sanchis Alberto and Ueffing Nicola. 2003. *Confidence estimation for machine translation*. In Technical Report Natural Language Engi-neering Workshop Final Report, pages: 97-100.

Blei David M, Ng Andrew Y and Jordan Michael. 2003. *Latent Dirichlet allocation* . Journal of Machine Learning Research, pages: 993-1022.

Carpuat M. and Simard M.. 2012. *The Trouble with SMT Consistency*. Proceedings of the 7th Workshop on Statistical Machine Translation, pages: 442-449.

De Beaugrande R. and Dressler W.U. 1981. *Introduction to text linguistics* , London. New York : Longman.

Fowler Roger. 1991. *Language in the News: Discourse and Ideology in the press*, London: Routledge.

Galley Michel and McKeown Kathleen. 1993. *Improving word sense disambiguation in lexical chaining*. In Proceedings of the 18th international joint conference on Artificial intelligence, IJCAI03, pages: 1486-1488.

Gimenez Jesus, Marquez Lluis, Comelles Elisabet,Castellon Irene and Arranz Victoria. 1993. *Document-level automatic MT evaluation based on discourse representations*. In Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages: 333-338.

Gong Z.X., Zhang M. and Zhou G.D. 2011. *Cache-based document-level statistical machine translation*. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages: 909-919.

Gong Z.X. and Zhou G.D. 2015. *Document-level Machine Translation Evaluation Metrics Enhanced with Simplified Lexical Chain*. In Proceedings of the 4th Conference on Natural Language Processing & Chinese Computing(To be published).

Guzmn Francisco, Joty Shafiq and Mrquez Lluís. 2014. *Using Discourse Structure Improves Machine Translation Evaluation*. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages: 687-698.

Halliday M.A.K and Hasan Ruqayia. 1976. *Cohesion in English*, London: Longman.

Hardmeier Christian, Nivre Joakim and Tiedemann Jörg. 2012. *Document-wide decoding for phrase-based statistical machine translation*. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages: 1179-1190.

Hofmann Thomas. 1999. *Probabilistic Latent Semantic Indexing*. In Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval, pages: 50-57.

Kamp H. and Reyle U. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht C Boston C London: Kluwer Academic Publishers.

Landauer T. K., Foltz P. and Laham D. 1998. *Introduction to Latent Semantic Analysis*. Discourse Processes 25.

Liu D., Gildea D. 2007. *Source-Language Features and Maximum Correlation Training for Ma-chine Translation Evaluation*. In Proceedings of NAACL, pages:41-48.

Morris Jane and Hirst Graeme. 1991. *Lexical cohesion computed by thesaural relations as an indicator of the structure of text*. Computational linguistics, 17(1):21-48.

Papineni Kishore, Roukos Salim, Ward Todd and Zhu WeiJing. 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.

Raphael Rubino, Jennifer Foster, Joachim Wagner, et al. 2012. *DCU-Symantec Submission for the WMT 2012 Quality Estimation Task*. Proceedings of the 7th Workshop on Statistical Machine Translation, pages: 138-144.

Raphael Rubino, Jos'e G. C. de Souza, Jennifer Foster, Lucia Specia. 2013. *Topic Models for Translation Quality Estimation for Gisting Purposes*. Proceedings of the XIV Machine Translation Summit, pages: 295-302.

Reeder, F. 2006. *Measuring MT Adequacy Using Latent Semantic Analysis*. In Proceedings of the 7th Conference of the Association for Machine Translation of the Americas. Cambridge, Massachusetts, pages: 176-184.

Tiedemann Jörg. 2010. *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 (DANLP), pages: 8-15.

Ture Ferhan, Oard Douglas W and Resnik Philip. 2010. *Encouraging consistent translation choices*. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages: 417-426.

Xiao Tong, Zhu Jingbo , Yao Shujie and Zhang Hao. 2011. *Document-Level Consistency Verification in Machine Translation*. In Proceedings of MT Summit XIII,pages: 131-138.

Xiao Xinyan,Xiong Deyi,Zhang Min, Liu Qun and Lin Shouxun. 2012. *A Topic Similarity Model for Hierarchical Phrase-based Translation*. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages: 750-758.

Xiong Deyi, Ding Yang, Zhang Min and Tan Chew Lim. 2013. *Lexical Lexical Chain Based Cohesion Models for Document-Level Statistical Machine Translation*. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages: 1563-1573. Seattle, Washington, USA.

Van Rijsbergen C. 1979. *Information Retrieval*. Butterworths,London, UK.

Wong B.T.M. and Kit C. 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.