Lexical Cohesion for Evaluation of Machine Translation at Document Level

Billy T.M. WONG Cecilia F.K. PUN Chunyu KIT Jonathan J. WEBTER
Department of Chinese, Translation and Linguistics
City University of Hong Kong
83 Tat Chee Avenue, Kowloon, Hong Kong SAR, P R China
{tmwong, fungkpun, ctckit, ctjjw}@cityu.edu.hk

***Abstract*—This paper studies how granularity of machine translation evaluation can be extended from sentence to document level. While most state-of-the-art evaluation metrics focus on the sentence level, we emphasize the importance of document structure, showing that lexical cohesion is a critical feature to highlight the superior quality of human translation to machine translation, which uses cohesive devices to tie salient words between sentences together as a text. An experiment shows that this feature can bring forth a 3-5% improvement in the correlation of automatic evaluation results with human judgments of machine translation outputs at the document level.**

Keywords-machine translation evaluation; evaluation metric; lexical cohesion; text coherence

I. INTRODUCTION

Machine translation (MT) evaluation has undergone a significant evolution over the past decade from human to automatic assessments. Various evaluation metrics were formulated to quantify the quality of MT outputs, based on the assumption that such quality can be estimated by their textual similarity to corresponding professional human translations as references. MT evaluation has been turned into a task of similarity measurement of MT outputs and reference translations. Typical evaluation metrics include BLEU [1] based on n-gram matching, TERp [2] based on edit-distance, METEOR [3] which utilizes morphology and semantic resources, and ATEC [4] which further exploits features like word informativeness and word ordering. Nearly all evaluation metrics in use so far score MT outputs sentence by sentence. The evaluation result for a text is usually a simple average of its sentence scores.

A drawback of this kind of sentence-based evaluation is the neglect of text structure, characterized by the absence of attention to cohesion and coherence within texts. These two linguistic features operate at the inter-sentential level and are realized via the interlinkage of lexical, grammatical and semantic elements across sentences. In the MT evaluation framework from the International Standards of Language Engineering, coherence is defined as "the degree to which the reader can define the role of each individual sentence (or group of sentences) with respect to the text as a whole" [5]. There is no guarantee of this for a text structured by simply putting together well-translated but stand-alone sentences without considering how cohesion and coherence are realized between them. Sentence-based evaluation metrics have no means to distinguish whether a text is cohesive and coherent, and are inevitably prone to falsely over- or underestimate the performance of an MT system.

Accurate MT evaluation on text (document) level is particularly important to MT users, for they mainly care about the overall meaning of a text in question rather than the grammatical correctness of each sentence [6]. Accordingly, the evaluation needs to take into account how individual sentences from MT output are joined together into a text. The connectivity of sentences is apparently a significant factor for assessing the understandability of a text as a whole. This does not mean that the evaluation of sentences versus texts is incompatible with each other, rather it simply acknowledges that both intra- and inter-sentence evaluation are important.

In MT evaluation, both cohesion and coherence are monolingual features in a target text. They can hardly be evaluated in isolation and have to be conjoined with other quality criteria such as adequacy and fluency. A survey of MT postediting [7] suggests that cohesion and coherence serve as higher level quality criteria beyond many others such as syntactic well-formedness. Posteditors tend to correct syntactic errors first before any amendment for improving cohesion and coherence of an MT output. Also as Wilks [8] noted (as cited in [9]), it is rather unlikely for a sufficiently large sample of translations to be coherent and totally wrong at the same time. Cohesion and coherence are thus possible to serve as the criteria when evaluating the overall quality of MT output.

This paper studies the use of a typical type of cohesion, i.e., lexical cohesion, as a potential quality attribute in MT evaluation. We investigate the quantitative variance of lexical cohesion devices in MT output versus human translation, to examine how adequately or inadequately MT systems can handle this feature. We also have devised a manual evaluation method to assess the coherence of MT outputs at the inter-sentence level, as a support to the development of automatic evaluation methods. An experiment to integrate lexical cohesion into a unigram-based MT evaluation metric confirms that this feature can bring about a significant gain in the metric's performance, in terms of the correlation between evaluation results of evaluation metric and human assessment.

The research described in this paper was partially supported by City University of Hong Kong through the SRG grants 7002267 and 7008003 and by the Research Grants Council (RGC) of HKSAR, China, through the GRF grant 9041597.

II. Cohesion and Coherence

To evaluate a text, besides faithfulness to its purpose, it is also important to study how the text is organized to ensure smooth information flow. This is particularly significant in the case of MT output, in that a well-structured text can greatly facilitate comprehension. A text should exhibit the properties of cohesion and coherence, both of which contribute texture to the text.

Cohesion refers to the "relations of meaning that exist within the text" [10]. It can be realized through cohesive devices, including grammatical and lexical ones. Grammatical cohesion refers to the syntactic links between text items, including reference, substitution and ellipsis, and conjunction. Lexical cohesion is achieved through word choice of two major types: reiteration and collocation. With the aid of these cohesive devices, ties between elements within a text can be built appropriately, so as to ease readers' comprehension [10][11].

Reiteration is the most explicit type of lexical cohesion. It can be recognized as a continuum or a cline of specificity, with repetition of the same lexical item at one end, and, at the other end, using a general noun to point to the same referent; located in between on this cline of specificity is employing a synonym (or near-synonym) and superordinate. Reiteration is the most direct way to create greater texture within a text. The same lexis occurring more than once within a text is unlikely to point to different referents. Collocation is more difficult to define compared with reiteration. It refers to those lexical items that share same or similar semantic relations, including complementarities, antonyms, converses, coordinate terms, holonyms / meronyms, troponym and so on.

Coherence is concerned with sense making in a text. Despite the fact that there is yet no widespread agreement on its definition, factors known to contribute to the realization of coherence include development of a topic, and paragraph unity.

III. Related Works

Cohesion and coherence have long been recognized as important issues in MT and MT evaluation. Early discussion can be traced back to Slype [9]. But progress remains rudimentary in terms of both theoretical and algorithmic aspects. Saggion and Carvalho [12] raised their concern that texts were poorly handled by MT systems as "a disconnected sequence of sentences". They proposed a text structure to preserve the discourse phenomena of a text, treating it as a cohesive and coherent whole, with the aid of a list of coherence relations between propositions. Nevertheless, it is uncertain whether such an idea can be automated. For MT evaluation, in particular, Visser and Fuji [6] suggested to assess the connectivity of consecutive sentences through the use of conjuncts, a subclass of adverbs specific to particular languages. Comelles et al. [13] presented a family of automatic MT evaluation measures based on the Discourse Representation Theory [14]. Their measures generate semantic trees to put together different entities with the same referent according to their contexts and grammatical connections, e.g., between a possessive adjective and a proper noun or between a main verb and an auxiliary verb. However, the parsing process needed for this linguistic-heavy measurement may suffer seriously from grammatical errors, which are unavoidable in MT output. Hence its accuracy and reliability inevitably fluctuate depending on different evaluation data.

The major obstacle to extending the granularity of evaluation from sentence to document level is the absence of a suitable human evaluation approach that can differentiate between the assessments of a document and a set of sentences, consequently resulting in a lack of test data. A parameterized evaluation metric can only be tuned using human evaluation data at the sentence level. Proposals for an evaluation metric at the document level, such as Comelles et al. [13], have to rely on a weak assumption that human evaluation data might reflect document level quality, since MT outputs are usually assessed sentence by sentence in sequence as in a document. Unfortunately, none of these proposals has reported any satisfactory result. Without a sound model of MT quality at document level, there is no way to know whether these evaluation methods can evaluate what they are intended to evaluate. This situation certainly calls for more fundamental work on modeling text quality.

IV. Lexical Cohesion in Machine Translation

The quality of MT output at document level is largely attributed to its coherence, whilst its measurement has to rely on cohesion, particularly lexical cohesion. Lexical cohesion contributes a lot to the identification of topic. Topic-oriented lexical items tend to be repeated throughout a text, via both reiteration and collocation, to refer to the same referent, to signify its importance and to help expand and elaborate on it. This indicator, therefore, should be assigned a heavier weight when used as one of the criteria in MT evaluation. In human translations (HT), cohesion is normally attained through strategic use of both grammatical and lexical cohesion devices, so as to achieve necessary coherence. In MT output, however, cohesion is weakly attained only through reiteration of lexical entities, which is more likely to be the result of lexical mapping from the repeated item in source texts. Consistency is normally maintained in MT output, particularly in the lexical items denoting topics, although such consistency applies to both correct and incorrect translations.

Despite the importance of reiterated topic-oriented lexical items in target texts, the appropriate use of reiteration is also important, for excessive use of reiteration may decrease the readability of a text. The difference between the use of cohesive devices in MT and HT reveals that human translators strategically change the ways of expression to achieve coherence and also avoid unnecessary overuse of the same lexical item. To a certain extent, this may be an underlying reason for the unnaturalness of some MT output even though it may contain counterpart lexical items on the same topic as in source texts. Another issue that hinders the comprehensiveness of MT output is grammaticality. An MT output is inevitably prone to grammatical errors, and poor grammaticality certainly affects the cohesion, coherence and meaning of a text.

We have carried out a study to compare reiteration of lexical items in MT and HT, and examine their contribution to coherence, in order to evaluate its potential as a useful feature in MT evaluation. The MetricsMATR08 dataset [15] was

selected for our study. It consists of 25 documents with a total of 249 segments. Each segment has eight different versions of MT outputs and four versions of human reference translations. The MT outputs are assessed by humans according to adequacy of translation.

Table I presents a case study from an excerpt of this dataset, consisting of a human translation and three samples of MT output for the same source text. The top 5 topic-oriented lexical items with the highest frequencies of occurrence denote that the text is about a political movement. Although these lexical items are on the same topic, one can see the differences between the HT and MT output. HT is most similar to MT03 with only slight difference in the number of occurrence, and is less similar to MT01 and MT02, which contain processes (i.e. *meet*, *participate*) besides participants. As a reference, the human evaluation results of these three MT outputs are ranked as MT03 > MT01 > MT02 in the dataset, which is exactly in agreement with the similarity between MT and HT in terms of the pattern of lexical repetition. Although this human evaluation is carried out at the sentence level and cannot fully represent the translation quality at the document level, it shows the possible relationship of lexical cohesion to the quality of MT outputs. However, in this stage, further investigation is needed before any conclusion be made.

The difference between MT and HT in terms of the use of lexical cohesion devices in MetricsMATR data is presented in Table II. A further categorization breaks down content words into lexical cohesion devices and those that are not. The former include those content words that reiterate at least once in a document, in the form of repetition, synonym or others (including superordinate and collocation). In general there are more content words in HT than in MT. The numbers of ordinary content words (i.e., not lexical cohesion devices) are basically the same in MT and HT, at around 2400. The surplus content words in HT over MT are all lexical cohesion devices. Among the subtypes of lexical cohesion, repetition is more widely used than other alternatives in both MT and HT.

TABLE I. TOP FIVE LEXICAL REPETITIONS OF MACHINE TRANSLATIONS AND HUMAN TRANSLATION IN A SELECTED DOCUMENT

HT		**MT**					
		01		**02**		**03**	
Hamas	6	Hamas	8	government	5	Hamas	6
Palestinian	5	leader	5	Fatah movement	3	government	5
leader	5	government	5	Hamas	3	Fatah movement	5
government	5	meet	4	leader	3	Palestinain	4
Fatah movement	5	participate	3	meet	3	leader	4

TABLE II. STATISTICS OF LEXICAL COHESION DEVICES BETWEEN MACHINE TRANSLATION AND HUMAN TRANSLATION

Average frequencies per each version of MT/HT	**MT**	**HT**
Content words	4428	4646
- Not lexical cohesion device	2403	2391
- Lexical cohesion devices	2025	2255
- Repetitions	1297	1445
- Synonyms	318	350
- Others	410	460

V. HUMAN EVALUATION OF COHERECE

The above study reveals that lexical cohesion can be a significant indicator of the quality difference between MT and HT. It is thus reasonable to hypothesize the quantification of lexical cohesion may reflect the quality of MT outputs. This certainly has to be verified with a human MT evaluation method at the document level, rather than at the sentence level (e.g., the adequacy assessment in MetricsMATR08 data).

The human evaluation method in this work is revised from Miller and Vanni [5], following the Rhetorical Structure Theory (RST) [16], a theory of text organization specifying coherence relations in an authentic text. Coherence is demonstrated if any spans (i.e., functionally significant text parts) in a text can be joined together by some designated RST relations. In our evaluation method, a loose definition of RST is applied, with sentence as the unit of analysis. For a document of MT output, raters are instructed to read its sentences and attempt to assign each sentence an RST relation in light of other sentences. The sentence is then given the score

- 1 if an RST relation can be determined,
- 0 if no relation is observed, and
- 0.5 if it can only be partially understood and assigned a possible relation with weak confidence.

The coherence score of a document is the average score of all its sentences. Notice that it does not matter which RST relation is assigned to a sentence and whether this relation is assigned correctly. It only matters if logical relations exist and can be determined across sentences. The RST theory provides a classification of existing logical relations for raters as a reference.

We attempt to compare the document coherence of MT outputs with the above method in terms of sentence adequacy (i.e., how faithfully meaning is translated) in the MetricsMATR08 data. Figure 1 illustrates the results of the comparison, where the evaluation results of MT output are sorted according to the adequacy scores at the document level rated in a 7-point scale (the higher the better). The two evaluation types correlate in general, especially when adequacy scores are above 6, but the coherence scores highly fluctuate for adequacy scores below 5. This suggests that coherence cannot be entirely obtained from sentence meaning. Without a basic understanding of a sentence it is difficult to relate the sentence to a text.

VI. EXPERIMENT OF LEXICAL COHESION EVALUATION

An experiment was performed to examine the extent to which the coherence of a document can be estimated by the use of lexical cohesion devices, and to evaluate whether lexical cohesion is a useful feature for automatic MT evaluation. The performance of an MT evaluation metric is represented as the magnitude of correlation between its evaluation scores and human assessment results on the same set of MT outputs, so as

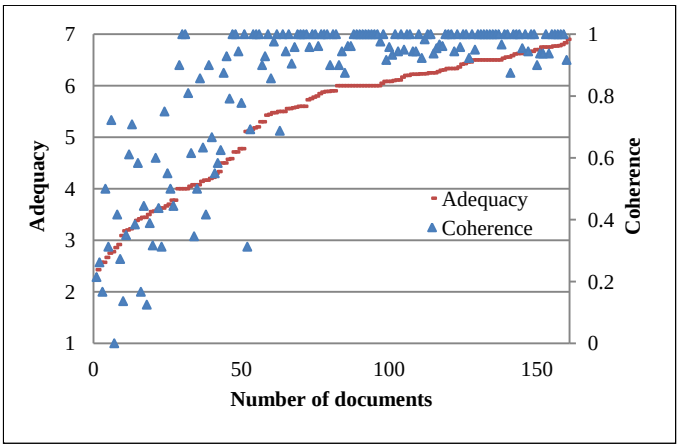

Figure 1. Evaluation scores of adequacy and coherence

to show how consistent the evaluation metric is to predict human judgment of MT quality.

A basic MT evaluation metric counts the number of words in an MT output that occur in corresponding human translation as well. This is used as a baseline in our experiment. In detail, it first measures the unigram matches $m(c,t)$ between an MT output c and its reference translation t, then calculates the precision p and recall r, and their harmonic F-measure f, as

$$f(c,t) = 2pr/(p+r), \quad (1)$$

$$p(c,t) = m(c,t)/l(c), \qquad r(c,t) = m(c,t)/l(t)$$

where $l(c)$ and $l(t)$ are the lengths of c and t respectively. This baseline metric is sensitive to word choice only, and disregards all other features such as word order or syntax. All words in the dataset are reduced to their stems with Porter stemmer, so as to group together the morphological variants of a word into one.

We define lexical cohesion LC in a document according to the following ratio,

$$LC = d/w \quad (2)$$

where d and w are the numbers of lexical cohesion devices and content words (all tokens with stopwords removed) respectively. In this experiment, lexical cohesion devices refer to content words that reiterate once or more in a document, in the form of repetition, synonym, near-synonym, superordinate and collocation. To classify the semantic relationship of words, a lexical resource WordNet [17] is used where words of the same senses, i.e., synonyms, are clustered into a semantic group, namely a *synset*. Synsets are interlinked in WordNet according to their semantic relationships. Superordinates and collocations are words in proximate semantic relationship, such as bicycle and vehicle (hypernyms), bicycle and wheel (meronyms), bicycle and car (coordinate terms), and so on. They are defined as synset pairs with a distance of 1. A measure of semantic distance [18] is also applied to identify near-synonyms, i.e., words that are synonyms in a broad sense but not grouped in the same synset. This measure quantifies the semantic similarity of word pairs as a real number in between 0 and 1 (the higher the more similar). A threshold is set at 0.96 for words to be considered near-synonyms of each other, based on the empirical observation in a previous study [19]. In sum, our measure of lexical cohesion accounts for the proportion of content words in a document that are used as lexical cohesion devices. It is a monolingual feature that does not require any reference translation for scoring purpose.

However, lexical cohesion alone is a weak, though valid, indicator of translation quality, as it is a high-order feature at discourse level, which has to be realized on top of other fundamental features. It is shown in the previous section that coherence is largely related to sentence meaning. We therefore combine the lexical cohesion measure with the baseline evaluation metric, in hope of covering both the sentence and document level quality of MT outputs. A hybrid evaluation metric H is formulated as follows to combine LC and the baseline metric f,

$$H = LC \cdot \alpha + f \cdot (1-\alpha) \quad (3)$$

where α is a weight to set the proportion of LC and f. It is empirically set as 0.2, which is found to attain an optimal balance in this experiment.

Tables III and IV report the performance of this evaluation metric with lexical cohesion, in terms of the rates of Pearson correlation coefficient between its evaluation result and the human assessments of adequacy and coherence of MT output in the MetricsMATR dataset at the document level. Besides the hybrid evaluation metric, we also test the use of lexical cohesion alone and the baseline metric for comparison. Two evaluation settings, single and multiple reference translations, are included in this experiment. Reference translations are used as the resource for the baseline metric providing choices of legitimate translation variants. In general, the baseline metric works better in the setting of multiple references. The correlation results demonstrate an observable improvement of the baseline measure when lexical cohesion is taken into account, i.e., a 3-5% increase in the correlations of the hybrid evaluation metric, from 0.66/0.72 to 0.68/0.74 for adequacy assessment and from 0.63/0.68 to 0.66/0.70 for coherence assessment in the single/multiple reference groups. The use of lexical cohesion alone also shows a positive correlation, but the magnitude is much weaker than its combination with the baseline metric. This confirms our idea that MT evaluation at the document level is not an independent practice, but is one to be performed on top of the sentence level in order to achieve a holistic account of MT quality. It is also worth noting that the human assessment of adequacy, a kind of evaluation at the sentence level, can be also better modeled if lexical cohesion is combined into the baseline metric, as the case in coherence assessment.

VII. Conclusion

We have attempted to address the problem that most MT evaluation metrics in use disregard the connectivity of sentences in a document. Such connectivity depends on how cohesive and coherent is the text, both important features to aid our comprehension. A typical type of cohesion, i.e., lexical cohesion, is brought into our focus. We have shown that this feature is one of the critical factors causing the difference between MT and HT. To combine lexical cohesion into an evaluation metric, a human assessment of coherence is first carried out on MT outputs, to see how the inter-relationships across sentences are determined and rated by humans. This is

TABLE V. CORRELATION OF EVALUATION SCORES WITH HUMAN ASSESSMENT (ADEQUACY)

Evaluation Measures	**Single Reference**	**Multiple References**
Lexical Cohesion	0.25	0.25
Hybrid (Lexical Cohesion + Baseline)	**0.68**	**0.74**
Baseline	0.66	0.72

TABLE VI. CORRELATION OF EVALUATION SCORES WITH HUMAN ASSESSMENT (COHERENCE)

Evaluation Measures	**Single Reference**	**Multiple References**
Lexical Cohesion	0.18	0.18
Hybrid (Lexical Cohesion + Baseline)	**0.66**	**0.70**
Baseline	0.63	0.68

followed by the establishment of empirical relationship between coherence and lexical cohesion, by correlating human judgments of coherence to the use of lexical cohesion devices in documents of MT output. It is shown that lexical cohesion alone, however, is not a feature strong enough to predict coherence. Its successful application requires combination with an evaluation metric working at the sentence level. This result broadens our limited understanding of the necessary parameters of translation and their inter-operation. This work is thus an essential, yet necessary, step towards a better MT evaluation measure.

REFERENCES

[1] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu. "BLEU: A method for automatic evaluation of machine translation," Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002), pp. 311-318. Philadelphia, PA, 2002.

[2] M. Snover, N. Madnani, B.J. Dorr and R. Schwartz. "Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric," Proceedings of the EACL-2009 Workshop on Statistical Machine Translation (WMT09), pp. 259-268. Athens, Greece, 2009.

[3] S. Banerjee and A. Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, 2005.

[4] B. Wong and C. Kit. "ATEC: Automatic evaluation of machine translation via word choice and word order," Machine Translation, vol. 23(2), pp. 141-155, 2009.

[5] K. J. Miller and M. Vanni. "Scaling the ISLE taxonomy: Development of metrics for the multi-dimensional characterisation of machine translation quality," Proceedings of MT Summit VIII, pp. 229-234. Santiago de Compostela, Spain, 2001.

[6] E. M. Visser and M. Fuji. "Using sentence connectors for evaluating MT output," Proceedings of the 16th International Conference on Computational Linguistics (COLING-1996), pp. 1066-1069. Center for Sprogteknologi, Copenhagen, 1996.

[7] M. Vasconcellos. "Cohesion and coherence in the presentation of machine translation products," In J. E. Alatis, ed., Georgetown University Round Table on Languages and Linguistics 1989, pp. 89-105, Washington, D.C., Georgetown University Press, 1989.

[8] Y. Wilks. "The value of the monolingual component in MT evaluation and its role in the battelle," Report on SYSTRAN. Luxembourg CEC Memorandum, 1978.

[9] V. G. Slype. "Critical study of methods for evaluating the quality of machine translation," Prepared for the Commission of European Communities Directorate General Scientific and Technical Information and Information Management, Report BR 19142, 1979.

[10] M. A. K. Halliday and R. Hasan. Cohesion in English, London: Longman, 1976.

[11] J. Martin. English Text: System and structure, Philadelphia: John Benjamins, 1992.

[12] H. Saggion and A. Carvalho. "Text structure aiming at machine translation," XII Artificial Intelligence National Meeting - RNIA'95, Cuernavaca, Morelos, Mexico, 1995.

[13] E. Comelles, J. Giménez, L. Màrquez, I. Castellón and V. Arranz. "Document-level automatic MT evaluation based on discourse representations," Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, 2010.

[14] H. Kamp and U. Reyle. From Discourse to Logic: An Introduction to Model Theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory, Dordrecht: Kluwer, 1993.

[15] M. Przybocki, K. Peterson and S. Bronsart. "2008 NIST metrics for machine translation (MetricsMATR08) development data," Linguistic Data Consortium, Philadelphia, 2009.

[16] W. Mann and S. Thompson. "Rhetorical structure theory: Toward a functional theory of text organization," Text, vol. 8(3), pp. 243-281, 1988.

[17] C. Fellbaum. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press, 1998.

[18] Z. Wu and M. Palmer. "Verb semantics and lexical selection," Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-1994). Las Cruces, New Mexico, 1994.

[19] B. T.M. Wong. "Semantic evaluation of machine translation," Proceedings of the 7th International Conference on Language Resource and Evaluation (LREC 2010). Valletta, Malta, 2010.