Extending Machine Translation Evaluation Metrics with Lexical Cohesion To Document Level

Billy T. M. Wong and **Chunyu Kit**
Department of Chinese, Translation and Linguistics
City University of Hong Kong
83 Tat Chee Avenue, Kowloon, Hong Kong SAR, P. R. China
{tmwong,ctckit}@cityu.edu.hk

Abstract

This paper proposes the utilization of lexical cohesion to facilitate evaluation of machine translation at the document level. As a linguistic means to achieve text coherence, lexical cohesion ties sentences together into a meaningfully interwoven structure through words with the same or related meaning. A comparison between machine and human translation is conducted to illustrate one of their critical distinctions that human translators tend to use more cohesion devices than machine. Various ways to apply this feature to evaluate machine-translated documents are presented, including one without reliance on reference translation. Experimental results show that incorporating this feature into sentence-level evaluation metrics can enhance their correlation with human judgements.

1 Introduction

Machine translation (MT) has benefited a lot from the advancement of automatic evaluation in the past decade. To a certain degree, its progress is also confined to the limitations of evaluation metrics in use. Most efforts devoted to evaluate the quality of MT output so far have still focused on the sentence level without sufficient attention to how a larger text is structured. This is notably reflected in the representative MT evaluation metrics, such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006), that adopt a sentence-by-sentence fashion to score MT outputs. The evaluation result for a document by any of them is usually a simple average of its sentence scores. A drawback of this kind of sentence-based evaluation is the neglect of document structure. There is no guarantee for the coherence of a text if it is produced by simply putting together stand-alone sentences, no matter how well-translated, without adequate inter-sentential connection. As a consequence, MT system optimized this way to any of these metrics can only have a very dim chance of producing translated document that reads as natural as human writing.

The accuracy of MT output at the document level is particularly important to MT users, for they care about the overall meaning of a text in question more than the grammatical correctness of each sentence (Visser and Fuji, 1996). Post-editors particularly need to ensure the quality of a whole document of MT output when revising its sentences. The connectivity of sentences is surely a significant factor contributing to the understandability of a text as a whole.

This paper studies the inter-sentential linguistic features of cohesion and coherence and presents plausible ways to incorporate them into the sentence-based metrics to support MT evaluation at the document level. In the Framework for MT Evaluation in the International Standards of Language Engineering (FEMTI) (King et al., 2003), coherence is defined as “the degree to which the reader can describe the role of each individual sentence (or group of sentences) with respect to the text as a whole”. The measurement of coherence has to rely on cohesion, referring to the “relations of meaning that exist within the text” (Halliday and Hasan, 1976). Cohesion is realized via the interlinkage of grammatical and lexical elements across sentences. *Grammatical*

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1060–1068, Jeju Island, Korea, 12–14 July 2012. ©2012 Association for Computational Linguistics

cohesion refers to the syntactic links between text items, while *lexical* cohesion is achieved through the word choices in a text. This paper focuses on the latter. A quantitative comparison of lexical cohesion devices between MT output and human translation is first conducted, to examine the weakness of current MT systems in handling this feature. Different ways of exploiting lexical cohesion devices for MT evaluation at the document level are then illustrated.

2 Related Works

Cohesion and coherence are both necessary 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 post-editing (Vasconcellos, 1989) suggests that cohesion and coherence serve as higher level quality criteria beyond many others such as syntactic well-formedness. Post-editors tend to correct syntactic errors first before any amendment for improving the cohesion and coherence of an MT output. Also, as Wilks (1978)[1] noted, 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 appropriate to serve as criteria for the overall quality of MT output.

Previous researches in MT predominantly focus on specific types of cohesion devices. For grammatical cohesion, a series of works, including Nakaiwa and Ikehara (1992), Nakaiwa et al. (1995), and Nakaiwa and Shirai (1996), present approaches to resolving Japanese zero pronouns and to integrating them into a Japanese-English transferred-based MT system. Peral et al. (1999) propose an interlingual mechanism for pronominal anaphora generation by exploiting a rich set of lexical, syntactic, morphologic and semantic information. Murata and Nagao (1993) and Murata et al. (2001) develop a rule base to identify the referential properties of Japanese noun phrases, so as to facilitate anaphora resolution for Japanese and article generation for English during translation. A recent COMTIS project (Cartoni et al., 2011) begins to exploit inter-sentential information for statistical MT. A phase of its work is to have grammatical devices, such as verbal tense/aspect/mode, discourse connectives and pronouns, manually annotated in multilingual corpora, in hopes of laying a foundation for the development of automatic labelers for them that can be integrated into an MT model.

For lexical cohesion, it has been only partially and indirectly addressed in terms of translation consistency in MT output. Different approaches to maintaining consistency in target word choices are proposed (Itagaki et al., 2007; Gong et al., 2011; Xiao et al., 2011). Carpuat (2009) also observes a general tendency in human translation that a given sense is usually lexicalized in a consistent manner throughout the whole translation.

Nevertheless there are only a few evaluation methods explicitly targeting on the quality of a document. Miller and Vanni (2001) devise a human evaluation approach to measure the comprehensibility of a text as a whole, based on the Rhetorical Structure Theory (Mann and Thompson, 1988), a theory of text organization specifying coherence relations in an authentic text. Snover et al. (2006) proposes HTER to assess post-editing effort through human annotation. Its automatic versions TER and TERp (Snover et al., 2009), however, remain sentence-based metrics. Comelles et al. (2010) present a family of automatic MT evaluation measures, based on the Discourse Representation Theory (Kamp and Reyle, 1993), that generate semantic trees to put together different text entities for the same referent according to their contexts and grammatical connections. Apart from MT evaluation, automated essay scoring programs such as E-rater (Burstein, 2003) also employ a rich set of discourse features for assessment. However, the parsing process needed for these linguistic-heavy approaches may suffer seriously from grammatical errors, which are unavoidable in MT output. Hence their accuracy and reliability inevitably fluctuate in accord with different evaluation data.

Lexical cohesion has far been neglected in both MT and MT evaluation, even though it is the single most important form of cohesion devices, accounting for nearly half of the cohesion devices in English (Halliday and Hasan, 1976). It is also a significant feature contributing to translation equivalence of texts by preserving their texture (Lotfipour-Saedi, 1997). The lexical cohesion devices in a text can be

[1] As cited in van Slype (1979).

represented as lexical chains conjoining related entities. There are many methods of computing lexical chains for various purposes, e.g., Morris and Hirst (1991), Barzilay and Elhadad (1997), Chan (2004), Li et al. (2007), among many others. Contrary to grammatical cohesion highly depending on syntactic well-formedness of a text, lexical cohesion is less affected by grammatical errors. Its computation has to rely on a thesaurus, which is usually available for almost every language. In this research, a number of formulations of lexical cohesion, with or without reliance on external language resource, will be explored for the purpose of MT evaluation.

3 Lexical Cohesion in Machine and Human Translation

This section presents a comparative study of MT and human translation (HT) in terms of the use of lexical cohesion devices. It is an intuition that more cohesion devices are used by humans than machines in translation, as part of the superior quality of HT. Two different datasets are used to ensure the reliability and generality of the comparison. The results confirm the incapability of MT in handling this feature and the necessity of using lexical cohesion in MT evaluation.

3.1 Data

The MetricsMATR 2008 development set (Przybocki et al., 2009) and the Multiple-Translation Chinese (MTC) part 4 (Ma, 2006) are used for this study. They consist of MT outputs of different source languages in company with reference translations. The data of MetricsMATR is selected from the NIST Open MT 2006 evaluation, while MTC4 is from the TIDES 2003 MT evaluation. Both datasets include human assessments of MT output, from which the part of adequacy assessment is selected for this study. Table 1 provides overall statistics of the datasets.

	MetricsMATR	MTC4
Number of systems	8	6
Number of documents	25	100
Number of segments	249	919
Number of references	4	4
Source language	Arabic	Chinese
Genre	Newswire	Newswire

Table 1: Information about the datasets in use

3.2 Identification of Lexical Cohesion Devices

Lexical cohesion is achieved through word choices of two major types: reiteration and collocation. Reiteration can be realized in a continuum or a cline of specificity, with repetition of the same lexical item at one end and the use of a general noun to point to the same referent at the other. In between the two ends is to use a synonym (or near-synonym) and superordinate. Collocation refers to those lexical items that share the same or similar semantic relations, including complementarity, antonym, converse, coordinate term, meronym, troponym, and so on.

In this study, lexical cohesion devices are defined as content words (i.e., tokens after stopword having been removed) that reiterate once or more times in a document, including synonym, near-synonym and superordinate, besides those repetition and collocation. Repetition refers to the same words or stems in a document. Stems are identified with the aid of Porter stemmer (1980).

To classify the semantic relationships of words, WordNet (Fellbaum, 1998) is used as a lexical resource, which clusters words of the same sense (i.e., synonyms) into a semantic group, namely a synset. Synsets are interlinked in WordNet according to their semantic relationships. Superordinate and collocation are formed by words in a proximate semantic relationship, such as *bicycle* and *vehicle* (hypernym), *bicycle* and *wheel* (meronym), *bicycle* and *car* (coordinate term), and so on. They are defined as synset pairs with a distance of 1 in WordNet. The measure of semantic distance (Wu and Palmer, 1994) is also applied to identify near-synonyms, i.e., words that are synonyms in a broad sense but not grouped in the same synset. It quantifies the semantic similarity of word pairs as a real number in between 0 and 1 (the higher the more similar) as

$$sim(c_1, c_2) = \frac{2\ d(lcs(c_1, c_2))}{d(c_1) + d(c_2)}$$

where c_1 and c_2 are the concepts (synsets) that the two words in question belong to, d is the distance in terms of the shortest path from a concept to the

Word type	MetricsMATR			MTC4		
	MT	HT	Difference (%)	MT	HT	Difference (%)
Content word	4428	4636	208 (4.7)	16162	16982	830 (5.1)
- Not lexical cohesion device	2403	2381	-22 (-1.0)	8657	8814	157 (1.8)
- Lexical cohesion device	2025	2255	230 (11.4)	7505	8168	663 (8.9)
- Repetition	1297	1445	148 (11.4)	4888	5509	621 (12.7)
- Synonym and near-synonym	318	350	32 (10.1)	1323	1311	-12 (-0.9)
- Superordinate and collocation	410	460	50 (12.4)	1294	1348	54 (4.2)

Table 2: Statistics of lexical cohesion devices in machine versus human translation (average frequencies per version of MT/HT)

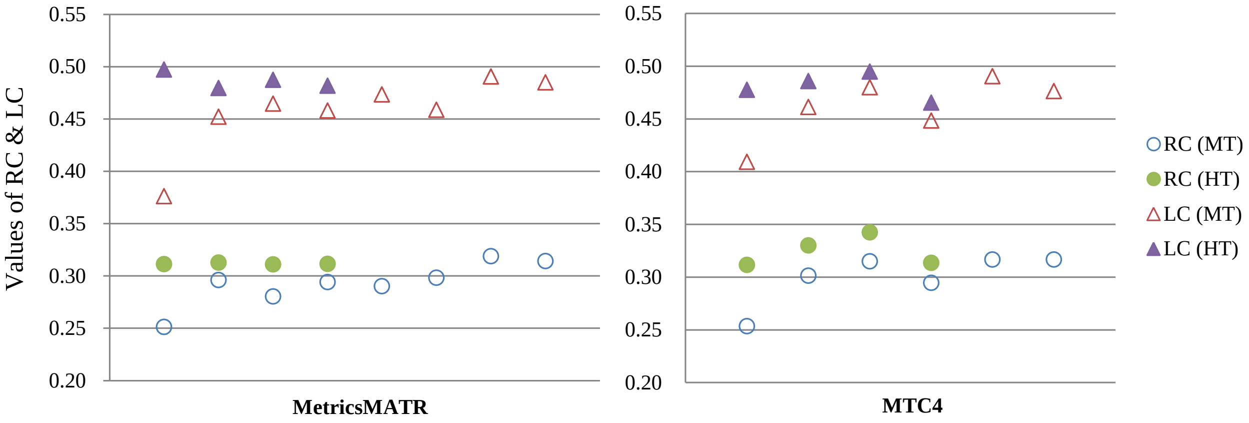

Figure 1: Use of lexical cohesion devices in machine versus human translation

global root node in WordNet, and lcs is the least common subsumer (i.e., the most specific ancestor concept) of c_1 and c_2. A threshold is set to 0.96 for words to be considered near-synonyms of each other, based on the empirical observation in a previous study (Wong, 2010).

3.3 Results

The difference between MT and HT (reference translation) in terms of the frequencies of lexical cohesion devices in MetricsMATR and MTC4 datasets is presented in Table 2. The frequencies are averaged by the number of MT/HT versions. A further categorization breaks down content words into lexical cohesion devices and those that are not. The count of each type of lexical cohesion device is also provided. In general the two datasets provide highly similar statistics. There are 4.7–5.1% more content words in HT than in MT. The numbers of ordinary content words (i.e., not lexical cohesion devices) are close in MT and HT. The difference of content words in HT and MT is mostly due to that of lexical cohesion devices, which are mostly repetition. 8.9–11.4% more lexical cohesion devices are found in HT than in MT in the datasets.

A further analysis is carried out to investigate into the use of lexical cohesion devices in each version of MT and HT in terms of the following two ratios,

LC = lexical cohesion devices / content words,
RC = repetition / content words.

A higher LC or RC ratio means that a greater proportion of content words are used as lexical cohesion devices.

Figure 1 illustrates the RC and LC ratios in the two datasets. The ratios of different MT systems are presented in an ascending order in each graph from left to right, according to their human assessment results. The distributions of these values show a strong similarity between the two datasets. First, most of the RC and LC ratios are within an observable range, i.e., 0.25–0.35 for the former and 0.40–0.50 for the latter, except a particularly low LC for

MT 1

1 Chine scrambled research on 16 key *technical*

2 These techniques are from within headline everyones boosting science and *technology* and *achieving* goals and contend of delivered on time bound through *achieving* breakthroughs in essential *technology* and complimentarity resources . national

BLEU: 0.224 (1-gram:7, 2-gram:0, 3-gram:2, 4-gram:1)
LC: 0.107 (number of lexical cohesion devices: 5)
Human assessment: 2.67

MT 2

1 China is accelerating research 16 *main technologies*

2 These *technologies* are within the *important* realm to promote sciences and *technology* and *achieve* national goals and must be completed in a timely manner through *achieving main* discoveries in *technology* and integration of resources .

BLEU: 0.213 (1-gram:5, 2-gram:3, 3-gram:2, 4-gram:1)
LC: 0.231 (number of lexical cohesion devices: 9)
Human assessment: 4.33

Reference

1 China Accelerates Research on 16 Main Technologies

2 These technologies represent a significant part in the development of science and technology and the achievement of national goals. They must be accomplished within a fixed period of time by realizing breakthroughs in essential technologies and integration of resources.

Table 3: An example of MT outputs of different quality (underlined: matched n-grams; italic: lexical cohesion devices)

one MT system. Second, the ratios in those different HT versions are very stable in comparison with those of MT. Especially, all four HT versions in the MetricsMATR dataset share the same RC ratio 0.31. This shows a typical level of the use of lexical cohesion device. Third, the ratios in MT are lower than or at most equal to those in HT, suggesting their correlation with translation quality: the closer their RC and LC ratios to those in HT, the better the MT. These results verify our assumption that lexical cohesion can serve as an effective proxy of the level of translation quality.

4 MT Evaluation at Document Level

As a feature at the discourse level, lexical cohesion is a good complement to current evaluation metrics focusing on features at the sentence level. Table 3 illustrates an example selected from the MetricsMATR dataset, consisting two versions of MT output for a short document of two segments only. The n-grams matched with the reference are underlined, while the lexical cohesion devices are italicized. The two MT outputs have a similar number of matched n-grams and hence receive similar BLEU scores. These scores, however, do not reflect their real difference in quality: the second version is better, according to human assessment of adequacy. Instead, their LC ratios seem to represent such a variation more accurately. The theme of the second output is also highlighted through the lexical chains, including *main/important*, *technology/technologies* and *achieve/achieving*, which create a tight texture between the two sentences, a crucial factor of text quality.

To perform MT evaluation at the document level, the LC and RC ratios can be used alone or integrated into a sentence-level metric. The former way has an advantage that it does not have to rely on any reference translation. LC mainly requires a thesaurus for computing semantic relation, while RC only needs a morphological processor such as stemmer, both of which are available for most languages. Its drawback, however, lies in the risk of relying on a single discourse feature. Although lexical cohesion gives a strong indication of text coherence, it is not indispensable, because a text can be coherent without any surface cohesive clue. Furthermore, the quality of a document is also reflected in that of its sentences. A coherent translation may be mistranslated, and on the other hand, a text containing lots of sentence-level errors would make it difficult to determine its document-level quality. A previous study comparing MT evaluation at the sentence versus document level (Wong et al., 2011) reports a poor consistency in the evaluation results at these two levels when the sentence-level scores of MT output are low. In regard of these, how to integrate these two levels of MT evaluation is particularly worth studying.

5 Experiments

We examine, through experiments, the effectiveness of using LC and RC ratios alone and integrating them into other evaluation metrics for MT evaluation at the document and system levels. Three evaluation metrics, namely BLEU, TER and METEOR,[2] are selected for testing. They represent three distinctive types of evaluation metrics: n-gram, edit-distance, and unigram with external language resources, respectively. These metrics are evaluated in terms of their correlation with human assessments, using Pearson's *r* correlation coefficient. The MetricsMATR and MTC4 datasets and their adequacy assessments are used as evaluation data. Note that the adequacy assessment is in fact an evaluation method for the sentence level. We have to rely on an assumption that this evaluation data may emulate document-level quality, since its MT outputs were assessed sentence by sentence in sequence as in a document. All experiments are performed under a setting of multiple reference translations.

The integration of the two ratios into an evaluation metric follows a simple weighted average approach. A hybrid metric H is formulated as

$$H = \alpha \, m_{doc} + (1 - \alpha) \, m_{seg}$$

where m_{doc} refers to the document-level feature in use (i.e., LC or RC), m_{seg} to a sentence-level metric, and α to a weight controlling their proportion. The MetricsMATR dataset is used as training data to optimize the values of α for different metrics, while the MTC4 is used as evaluation data. Table 4 shows the optimized weights for the metrics for evaluation at the document level.

Metrics	RC	LC
BLEU	0.28	0.29
TER	0.40	0.38
METEOR	0.19	0.18

Table 4: Optimized weights for the integration of discourse feature into sentence-level metrics

Table 5 presents the correlation rates of evaluation metrics obtained in our experiments under different settings, with their 95% conference intervals (CI) provided. The LC and RC ratios are found to have strong correlations with human assessments at the system level even when used alone, highly comparable to BLEU and TER. At the document level, however, they are not as good as the others. They show their advantages when integrated into other metrics, especially BLEU and TER. LC raises the correlation of BLEU from 0.447 to 0.472 and from 0.861 to 0.905 at the document and system levels, respectively. It improves TER even more significantly, in that the correlation rates are boosted up from -0.326 to -0.390 at the document level, and even from -0.601 to -0.763 at the system level. Since there are only six systems in the MTC4 data, such a dramatic change may not be as meaningful as the smooth improvement at the document level. METEOR is a special case in this experiment. Its correlation cannot be improved by integrating LC or RC, and is even slightly dropped at the document level. The cause for this is yet to be identified. Nevertheless, these results confirm the close relationship of an MT system's capability to appropriately generate lexical cohesion devices with the quality of its output.

Table 6 presents the Pearson correlations between evaluation results at the document level using different evaluation metrics in the MTC4 data. It illustrates the homogeneity/heterogeneity of different metrics and helps explain the performance change

[2] METEOR 1.0 with default parameters optimized over the adequacy assessments.

	Document		System	
Metrics	Correlation	95% CI	Correlation	95% CI
RC	0.243	(0.167, 0.316)	0.873	(0.211, 0.985)
LC	0.267	(0.192, 0.339)	0.818	(0.020, 0.979)
BLEU	0.447	(0.381, 0.508)	0.861	(0.165, 0.984)
BLEU+RC	0.463	(0.398, 0.523)	0.890	(0.283, 0.987)
BLEU+LC	0.472	(0.408, 0.531)	0.905	(0.352, 0.989)
TER	-0.326	(-0.253, -0.395)	-0.601	(-0.411, -0.949)
TER+RC	-0.370	(-0.299, -0.437)	-0.740	(-0.179, -0.969)
TER+LC	-0.390	(-0.320, -0.455)	-0.763	(-0.127, -0.972)
METEOR	0.557	(0.500, 0.609)	0.961	(0.679, 0.995)
METEOR+RC	0.555	(0.498, 0.608)	0.960	(0.672, 0.995)
METEOR+LC	0.556	(0.499, 0.609)	0.962	(0.687, 0.995)

Table 5: Correlation of different metrics with adequacy assessment in MTC4 data

BLEU	1				
TER	-0.699	1			
METEOR	0.834	-0.510	1		
RC	0.287	-0.204	0.405	1	
LC	0.263	-0.097	0.437	0.736	1
	BLEU	TER	METEOR	RC	LC

Table 6: Correlation between the evaluation results of different metrics

by combining sentence- and document-level metrics. The table shows that the two ratios LC and RC highly correlate with each other, as if they are two variants of quantifying lexical cohesion devices. The three sentence-level metrics, BLEU, TER and METEOR, also show strong correlations with each other, especially between BLEU and METEOR. The correlations are generally weaker between sentence- and document-level metrics, for instance, 0.263 between BLEU and LC and only -0.097 between TER and LC, showing that they are quite heterogeneous in nature. This accounts for the significant performance gain from their combination: their difference allows them to complement each other. It is also worth noting that between METEOR and LC the correlation of 0.437 is mildly strong, explaining the negative result of their integration. On the one hand, lexical cohesion is word choice oriented, which is only sensitive to the reiteration and semantic relatedness of words in MT output. On the other hand, METEOR is strong in unigram matching, with multiple strategies to maximize the matching rate between MT output and reference translation. In this sense they are homogeneous to a certain extent, explaining the null effect of their combination.

6 Discussion and Conclusion

In this study we have attempted to address the problem that most existing MT evaluation metrics disregard the connectivity of sentences in a document. By focusing on a typical type of cohesion, i.e., lexical cohesion, we have shown that its use frequency is a significant factor to differentiate HT from MT and MT outputs of different quality from each other. The high correlation rate of its use with translation adequacy also suggests that the more lexical cohesion devices in use, the better the quality of MT output. Accordingly we have used two ratios, LC and RC, to capture such correlativity. Our experimental results have confirmed the effectiveness of this feature in accounting for the document-level quality of MT output. The performance of two evaluation metrics, BLEU and TER, is highly improved through incorporating this document-level feature, in terms of the change of their correlation with human assessments.

This finding is positive and sheds light on a region of MT research that is still severely under-explored. Our approach to extending the granularity of MT evaluation from sentence to document through lexical cohesion is highly applicable to different languages. It has a relatively weak demand for language resource in comparison with the processing of other discourse features like grammatical cohesion. It is also much unaffected by grammatical problems or errors commonly seen in natural languages and, in particular, MT outputs.

Our future work will continue to explore the relationship of lexical cohesion to translation quality, so as to identify, apart from its use frequency, other significant aspects for MT evaluation at the document level. A frequent use of cohesion devices in a text is not necessarily appropriate, because an excess of them may decrease the quality and readability of a text. Human writers can strategically change the ways of expression to achieve appropriate coherence and also avoid overuse of the same lexical item. To a certain extent, this is one of the causes for the unnaturalness of MT output: it may contain a large number of lexical cohesion devices which are simply direct translation of those in a source text that do not fit in the target context. How to use lexical cohesion devices appropriately instead of frequently is thus an important issue to tackle before we can adopt them in MT and MT evaluation by a suitable means.

Acknowledgments

The research described in this paper was substantially supported by the Research Grants Council (RGC) of Hong Kong SAR, P. R. China through GRF grant 144410.

References

Satanjeev Banerjee and Alon Lavie. 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, Ann Arbor, Michigan.

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17.

Jill Burstein. 2003. The E-rater scoring engine: Automated essay scoring with natural language processing. In Mark D. Shermis and Jill Burstein, editors, *Automated Essay Scoring: A Cross-Disciplinary Perspective*, chapter 7, pages 113–122. Lawrence Erlbaum Associates.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 19–27, Boulder, Colorado.

Bruno Cartoni, Andrea Gesmundo, James Henderson, Cristina Grisot, Paola Merlo, Thomas Meyer, Jacques Moeschler, Sandrine Zufferey, and Andrei Popescu-Belis. 2011. Improving MT coherence through text-level processing of input texts: The COMTIS project. In *Tralogy*, Paris.

Samuel W. K. Chan. 2004. Extraction of sailent textual patterns: Synergy between lexical cohesion and contextual coherence. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 34(2):205–218.

Elisabet Comelles, Jesus Giménez, Lluìs Márquez, Irene Castellòn, and Victoria Arranz. 2010. Document-level automatic MT evaluation based on discourse representations. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics-MATR*, pages 333–338, Uppsala.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based document-level statistical machine translation. In *EMNLP 2011*, pages 909–919, Edinburgh, Scotland.

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

Masaki Itagaki, Takako Aikawa, and Xiaodong He. 2007. Automatic validation of terminology translation consistency with statistical method. In *MT Summit XI*, pages 269–274.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: Kluwer.

Margaret King, Andrei Popescu-Belis, and Eduard Hovy. 2003. FEMTI: Creating and using a framework for MT evaluation. In *MT Summit IX*, pages 224–231, New Orleans.

Jing Li, Le Sun, Chunyu Kit, and Jonathan Webster. 2007. A query-focused multi-document summarizer based on lexical chains. In *DUC 2007*, Rochester, New York.

Kazem Lotfipour-Saedi. 1997. Lexical cohesion and translation equivalence. *Meta*, 42(1):185–192.

Xiaoyi Ma. 2006. Multiple-Translation Chinese (MTC) part 4. Linguistic Data Consortium.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Keith J. Miller and Michelle Vanni. 2001. Scaling the ISLE taxonomy: Development of metrics for the multi-dimensional characterisation of machine translation quality. In *MT Summit VIII*, pages 229–238.

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

Masaki Murata and Makoto Nagao. 1993. Determination of referential property and number of nouns in Japanese sentences for machine translation into English. In *TMI 1993*, pages 218–225, Kyoto.

Masaki Murata, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara. 2001. A machine-learning approach to estimating the referential properties of Japanese noun phrases. In *CICLING 2001*, pages 142–153, Mexico-City.

Hiromi Nakaiwa and Satoru Ikehara. 1992. Zero pronoun resolution in a machine translation system by using Japanese to English verbal semantic attributes. In *ANLP 1992*, pages 201–208.

Hiromi Nakaiwa and Satoshi Shirai. 1996. Anaphora resolution of Japanese zero pronouns with deictic reference. In *COLING 1996*, pages 812–817, Copenhagen.

Hiromi Nakaiwa, Satoshi Shirai, Satoru Ikehara, and Tsukasa Kawaok. 1995. Extrasentential resolution of Japanese zero pronouns using semantic and pragmatic constraints. In *Proceedings of the AAAI 1995 Spring Symposium Series: Empirical Methods in Discourse Interpretation and Generation*, pages 99–105, Stanford.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *ACL 2002*, pages 311–318.

Jesús Peral, Manuel Palomar, and Antonio Ferràndez. 1999. Coreference-oriented interlingual slot structure and machine translation. In *Proceedings of the ACL Workshop on Coreference and its Applications*, pages 69–76, College Park, MD.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Mark Przybocki, Kay Peterson, and Sébastien Bronsart. 2009. 2008 NIST metrics for machine translation (MetricsMATR08) development data. Linguistic Data Consortium.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *AMTA 2006*, pages 223–231.

Matthew Snover, Nitin Madnani, Bonnie J. Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the 4th Workshop on Statistical Machine Translation*, pages 259–268, Athens.

Georges van Slype. 1979. Critical Study of Methods for Evaluating the Quality of Machine Translation. Technical report, Bureau Marcel van Dijk / European Commission, Brussels.

Muriel Vasconcellos. 1989. Cohesion and coherence in the presentation of machine translation products. In James E. Alatis, editor, *Georgetown University Round Table on Languages and Linguistics 1989: Language Teaching, Testing, and Technology: Lessons from the Past with a View Toward the Future*, pages 89–105. Georgetown University Press.

Eric M. Visser and Masaru Fuji. 1996. Using sentence connectors for evaluating MT output. In *COLING 1996*, pages 1066–1069.

Yorick Wilks. 1978. The Value of the Monolingual Component in MT Evaluation and its Role in the Battelle. Report on Systran, Luxembourg CEC Memorandum.

Billy T. M. Wong, Cecilia F. K. Pun, Chunyu Kit, and Jonathan J. Webster. 2011. Lexical cohesion for evaluation of machine translation at document level. In *NLP-KE 2011*, pages 238–242, Tokushima.

Billy Tak-Ming Wong. 2010. Semantic evaluation of machine translation. In *LREC 2010*, pages 2884–2888, Valletta.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *ACL 1994*, pages 133–138, Las Cruces.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level consistency verification in machine translation. In *MT summit XIII*, pages 131–138, Xiamen.