Dependency-Based Automatic Evaluation for Machine Translation

Karolina Owczarzak **Josef van Genabith** **Andy Way**
National Centre for Language Technology
School of Computing, Dublin City University
Dublin 9, Ireland
{owczarzak,josef,away}@computing.dcu.ie

Abstract

We present a novel method for evaluating the output of Machine Translation (MT), based on comparing the dependency structures of the translation and reference rather than their surface string forms. Our method uses a treebank-based, wide-coverage, probabilistic Lexical-Functional Grammar (LFG) parser to produce a set of structural dependencies for each translation-reference sentence pair, and then calculates the precision and recall for these dependencies. Our dependency-based evaluation, in contrast to most popular string-based evaluation metrics, will not unfairly penalize perfectly valid syntactic variations in the translation. In addition to allowing for legitimate syntactic differences, we use paraphrases in the evaluation process to account for lexical variation. In comparison with other metrics on 16,800 sentences of Chinese-English newswire text, our method reaches high correlation with human scores. An experiment with two translations of 4,000 sentences from Spanish-English Europarl shows that, in contrast to most other metrics, our method does not display a high bias towards statistical models of translation.

1 Introduction

Since their appearance, string-based evaluation metrics such as BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) have been the standard tools used for evaluating MT quality. Both score a candidate translation on the basis of the number of *n*-grams shared with one or more reference translations. Automatic measures are indispensable in the development of MT systems, because they allow MT developers to conduct frequent, cost-effective, and fast evaluations of their evolving models.

These advantages come at a price, though: an automatic comparison of *n*-grams measures only the string similarity of the candidate translation to one or more reference strings, and will penalize any divergence from them. In effect, a candidate translation expressing the source meaning accurately and fluently will be given a low score if the lexical and syntactic choices it contains, even though perfectly legitimate, are not present in at least one of the references. Necessarily, this score would differ from a much more favourable human judgement that such a translation would receive.

The limitations of string comparison are the reason why it is advisable to provide multiple references for a candidate translation in BLEU- or NIST-based evaluations. While Zhang and Vogel (2004) argue that increasing the size of the test set gives even more reliable system scores than multiple references, this still does not solve the inadequacy of BLEU and NIST for sentence-level or small set evaluation. In addition, in practice even a number of references do not capture the whole potential variability of the translation. Moreover, when designing a statistical MT system, the need for large amounts of training data limits the researcher to collections of parallel corpora such as Europarl (Koehn, 2005), which provides only one reference, namely the target text; and the cost of creating additional reference translations of the test set, usually a few thousand sentences long, is often prohibitive. Therefore, it would be desirable to find an evaluation method that accepts legitimate syntactic and lexical differences

between the translation and the reference, thus better mirroring human assessment.

In this paper, we present a novel method that automatically evaluates the quality of translation based on the dependency structure of the sentence, rather than its surface form. Dependencies abstract away from the particulars of the surface string (and CFG tree) realization and provide a "normalized" representation of (some) syntactic variants of a given sentence. The translation and reference files are analyzed by a treebank-based, probabilistic Lexical-Functional Grammar (LFG) parser (Cahill et al., 2004), which produces a set of dependency triples for each input. The translation set is compared to the reference set, and the number of matches is calculated, giving the precision, recall, and f-score for that particular translation.

In addition, to allow for the possibility of valid lexical differences between the translation and the references, we follow Kauchak and Barzilay (2006) and Owczarzak et al. (2006) in adding a number of paraphrases in the process of evaluation to raise the number of matches between the translation and the reference, leading to a higher score.

Comparing the LFG-based evaluation method with other popular metrics: BLEU, NIST, General Text Matcher (GTM) (Turian et al., 2003), Translation Error Rate (TER) (Snover et al., 2006)[1], and METEOR (Banerjee and Lavie, 2005), we show that combining dependency representations with paraphrases leads to a more accurate evaluation that correlates better with human judgment.

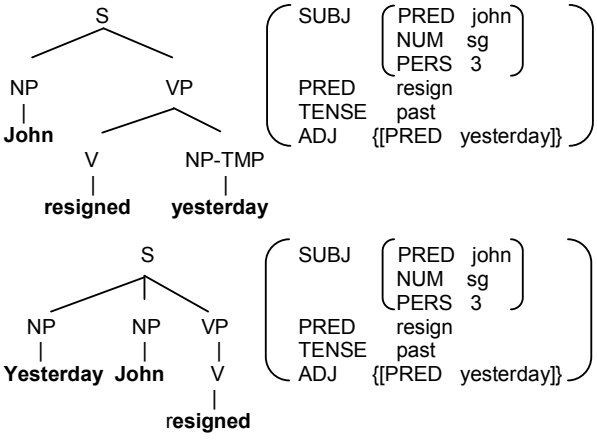
The remainder of this paper is organized as follows: Section 2 gives a basic introduction to LFG; Section 3 describes related work; Section 4 describes our method and gives results of two experiments on different sets of data: 4,000 sentences from Spanish-English Europarl and 16,800 sentences of Chinese-English newswire text from the Linguistic Data Consortium's (LDC) Multiple Translation project; Section 5 discusses ongoing work; Section 6 concludes.

2 Lexical-Functional Grammar

In Lexical-Functional Grammar (Bresnan, 2001) sentence structure is represented in terms of c(onstituent)-structure and f(unctional)-structure. C-structure represents the surface string word order and the hierarchical organisation of phrases in terms of CFG trees. F-structures are recursive feature (or attribute-value) structures, representing abstract grammatical relations, such as *subj*(ect), *obj*(ect), *obl*(ique), *adj*(unct), approximating to predicate-argument structure or simple logical forms. C-structure and f-structure are related in terms of functional annotations (attribute-value structure equations) in c-structure trees, describing f-structures.

While c-structure is sensitive to surface word order, f-structure is not. The sentences *John resigned yesterday* and *Yesterday, John resigned* will receive different tree representations, but identical f-structures, shown in (1).

(1) *C-structure:* *F-structure:*



Notice that if these two sentences were a translation-reference pair, they would receive a less-than-perfect score from string-based metrics. For example, BLEU with add-one smoothing[2] gives this pair a score of barely 0.3781.

The f-structure can also be described as a flat set of triples. In triples format, the f-structure in (1) could be represented as follows: {*subj*(resign, john), *pers*(john, 3), *num*(john, sg), *tense*(resign,

[1] As we focus on purely automatic metrics, we omit HTER (Human-Targeted Translation Error Rate) here.

[2] We use smoothing because the original BLEU gives zero points to sentences with fewer than one four-gram.

past), *adj*(resign, yesterday), *pers*(yesterday, 3), *num*(yesterday, sg)}.

Cahill et al. (2004) presents Penn-II Treebank-based LFG parsing resources. Her approach distinguishes 32 types of dependencies, including grammatical functions and morphological information. This set can be divided into two major groups: a group of predicate-only dependencies and non-predicate dependencies. Predicate-only dependencies are those whose path ends in a predicate-value pair, describing grammatical relations. For example, for the f-structure in (1), predicate-only dependencies would include: {*subj*(resign, john), *adj*(resign, yesterday)}.[3]

In parser evaluation, the quality of the f-structures produced automatically can be checked against a set of gold standard sentences annotated with f-structures by a linguist. The evaluation is conducted by calculating the precision and recall between the set of dependencies produced by the parser, and the set of dependencies derived from the human-created f-structure. Usually, two versions of f-score are calculated: one for all the dependencies for a given input, and a separate one for the subset of predicate-only dependencies.

In this paper, we use the parser developed by Cahill et al. (2004), which automatically annotates input text with c-structure trees and f-structure dependencies, reaching high precision and recall rates. [4]

3 Related work

The insensitivity of BLEU and NIST to perfectly legitimate syntactic and lexical variation has been raised, among others, in Callison-Burch et al. (2006), but the criticism is widespread. Even the creators of BLEU point out that it may not correlate particularly well with human judgment at the sentence level (Papineni et al., 2002). A side effect of this phenomenon is that BLEU is less reliable for smaller data sets, so the advantage it provides in the speed of evaluation is to some extent counterbalanced by the time spent by developers on producing a sufficiently large test set in order to obtain a reliable score for their system.

Recently a number of attempts to remedy these shortcomings have led to the development of other automatic MT evaluation metrics. Some of them concentrate mainly on word order, like General Text Matcher (Turian et al., 2003), which calculates precision and recall for translation-reference pairs, weighting contiguous matches more than non-sequential matches, or Translation Error Rate (Snover et al., 2005), which computes the number of substitutions, inserts, deletions, and shifts necessary to transform the translation text to match the reference. Others try to accommodate both syntactic and lexical differences between the candidate translation and the reference, like CDER (Leusch et al., 2006), which employs a version of edit distance for word substitution and reordering; or METEOR (Banerjee and Lavie, 2005), which uses stemming and WordNet synonymy. Kauchak and Barzilay (2006) and Owczarzak et al. (2006) use paraphrases during BLEU and NIST evaluation to increase the number of matches between the translation and the reference; the paraphrases are either taken from WordNet[5] in Kauchak and Barzilay (2006) or derived from the test set itself through automatic word and phrase alignment in Owczarzak et al. (2006). Another metric making use of synonyms is the linear regression model developed by Russo-Lassner et al. (2005), which makes use of stemming, WordNet synonymy, verb class synonymy, matching noun phrase heads, and proper name matching. Kulesza and Schieber (2004), on the other hand, train a Support Vector Machine using features like proportion of n-gram matches and word error rate to judge a given translation's distance from human-level quality.

Nevertheless, these metrics use only string-based comparisons, even while taking into consideration reordering. By contrast, our dependency-based method concentrates on utilizing linguistic structure to establish a comparison between translated sentences and their reference.

[3] Other predicate-only dependencies include: *apposition*, *complement*, *open complement*, *coordination*, *determiner*, *object*, *second object*, *oblique*, *second oblique*, *oblique agent*, *possessive*, *quantifier*, *relative clause*, *topic*, *relative clause pronoun*. The remaining non-predicate dependencies are: *adjectival degree*, *coordination surface form*, *focus*, complementizer forms: *if*, *whether*, and *that*, *modal*, *number*, *verbal particle*, *participle*, *passive*, *person*, *pronoun surface form*, *tense*, *infinitival clause*.

[4] http://lfg-demo.computing.dcu.ie/lfgparser.html

[5] http://wordnet.princeton.edu/

4 LFG f-structure in MT evaluation

The process underlying the evaluation of f-structure quality against a gold standard can be used in automatic MT evaluation as well: we parse the translation and the reference, and then, for each sentence, we check the set of translation dependencies against the set of reference dependencies, counting the number of matches. As a result, we obtain the precision and recall scores for the translation, and we calculate the f-score for the given pair. Because we are comparing two outputs that were produced automatically, there is a possibility that the result will not be noise-free.

To assess the amount of noise that the parser may introduce we conducted an experiment where 100 English Europarl sentences were modified by hand in such a way that the position of adjuncts was changed, but the sentence remained grammatical and the meaning was not changed. This way, an ideal parser should give both the source and the modified sentence the same f-structure, similarly to the case presented in (1). The modified sentences were treated like a translation file, and the original sentences played the part of the reference. Each set was run through the parser. We evaluated the dependency triples obtained from the "translation" against the dependency triples for the "reference", calculating the f-score, and applied other metrics (TER, METEOR, BLEU, NIST, and GTM) to the set in order to compare scores. The results, inluding the distinction between f-scores for all dependencies and predicate-only dependencies, appear in Table 1.

	baseline	modified
TER	0.0	6.417
METEOR	1.0	0.9970
BLEU	1.0000	0.8725
NIST	11.5232	11.1704 (96.94%)
GTM	100	99.18
dep f-score	100	96.56
dep_preds f-score	100	94.13

Table 1. Scores for sentences with reordered adjuncts

The baseline column shows the upper bound for a given metric: the score which a perfect translation, word-for-word identical to the reference, would obtain.[6] In the other column we list the scores that the metrics gave to the "translation" containing reordered adjunct. As can be seen, the dependency and predicate-only dependency scores are lower than the perfect 100, reflecting the noise introduced by the parser.

To show the difference between the scoring based on LFG dependencies and other metrics in an ideal situation, we created another set of a hundred sentences with reordered adjuncts, but this time selecting only those reordered sentences that were given the same set of dependencies by the parser (in other words, we simulated having the ideal parser). As can be seen in Table 2, other metrics are still unable to tolerate legitimate variation in the position of adjuncts, because the sentence surface form differs from the reference; however, it is not treated as an error by the parser.

	baseline	modified
TER	0.0	7.841
METEOR	1.0	0.9956
BLEU	1.0000	0.8485
NIST	11.1690	10.7422 (96.18%)
GTM	100	99.35
dep f-score	100	100
dep_preds f-score	100	100

Table 2. Scores for sentences with reordered adjuncts in an ideal situation

4.1 Initial experiment – Europarl

In the first experiment, we attempted to determine whether the dependency-based measure is biased towards statistical MT output, a problem that has been observed for *n*-gram-based metrics like BLEU and NIST. Callison-Burch et al. (2006) report that BLEU and NIST favour *n*-gram-based MT models such as Pharaoh (Koehn, 2004), so the translations produced by rule-based systems score lower on the automatic evaluation, even though human judges consistently rate their output higher than Pharaoh's translation. Others repeatedly

[6] Two things have to be noted here: (1) in case of NIST the perfect score differs from text to text, which is why we provide the percentage points as well, and (2) in case of TER the lower the score, the better the translation, so the perfect translation will receive 0, and there is no upper bound on the score, which makes this particular metric extremely difficult to directly compare with others.

observed this tendency in previous research as well; in one experiment, reported in Owczarzak et al. (2006), where the rule-based system Logomedia[7] was compared with Pharaoh, BLEU scored Pharaoh 0.0349 points higher, NIST scored Pharaoh 0.6219 points higher, but human judges scored Logomedia output 0.19 points higher (on a 5-point scale).

4.1.1 Experimental design

In order to check for the existence of a bias in the dependency-based metric, we created a set of 4,000 sentences drawn randomly from the Spanish-English subset of Europarl (Koehn, 2005), and we produced two translations: one by a rule-based system Logomedia, and the other by the standard phrase-based statistical decoder Pharaoh, using alignments produced by GIZA++[8] and the refined word alignment strategy of Och and Ney (2003). The translations were scored with a range of metrics: BLEU, NIST, GTM, TER, METEOR, and the dependency-based method.

4.1.2 Adding synonyms

Besides the ability to allow syntactic variants as valid translations, a good metric should also be able to accept legitimate lexical variation. We introduced synonyms and paraphrases into the process of evaluation, creating new best-matching references for the translations using either paraphrases derived from the test set itself (following Owczarzak et al. (2006)) or WordNet synonyms (as in Kauchak and Barzilay (2006)).

Bitext-derived paraphrases

Owczarzak et al. (2006) describe a simple way to produce a list of paraphrases, which can be useful in MT evaluation, by running word alignment software on the test set that is being evaluated. Paraphrases derived in this way are specific to the domain at hand and contain low-level syntactic variants in addition to word-level synonymy.

Using the standard GIZA++ software and the refined word alignment strategy of Och and Ney (2003) on our test set of 4,000 Spanish-English sentences, the method generated paraphrases for just over 1100 items. These paraphrases served to create new individual best-matching references for the Logomedia and Pharaoh translations. Due to the small size of the paraphrase set, only about 20% of reference sentences were actually modified to better reflect the translation. This, in turn, led to little difference in scores.

WordNet synonyms

To maximize the number of matches between a translation and a reference, Kauchak and Barzilay (2006) use WordNet synonyms during evaluation. In addition, METEOR also has an option of including WordNet in the evaluation process. As in the case of bitext-derived paraphrases, we used WordNet synonyms to create new best-matching references for each of the two translations. This time, given the extensive database containing synonyms for over 150,000 items, around 70% of reference sentences were modified: 67% for Pharaoh, and 75% for Logomedia. Note that the number of substitutions is higher for Logomedia; this confirms the intuition that the translation produced by Pharaoh, trained on the domain which is also the source of the reference text, will need fewer lexical replacements than Logomedia, which is based on a general non-domain-specific model.

4.1.3 Results

Table 3 shows the difference between the scores which Pharaoh's and Logomedia's translations obtained from each metric: a positive number shows by how much Pharaoh's score was higher than Logomedia's, and a negative number reflects Logomedia's higher score (the percentages are absolute values). As can be seen, all the metrics scored Pharaoh higher, inlcuding METEOR and the dependency-based method that were boosted with WordNet. The values in the table are sorted in descending order, from the largest to the lowest advantage of Pharaoh over Logomedia.

Interestingly, next to METEOR boosted with WordNet, it is the dependency-based method, and especially the predicates-only version, that shows the least bias towards the phrase-based translation. In the next step, we selected from this set smaller subsets of sentences that were more and more similar in terms of translation quality (as determined by a sentence's BLEU score). As the similarity of the translation quality increased, most metrics lowered their bias, as is shown in Table 4.

The first column shows the case where the sentences chosen differed at the most by 0.05

[7] http://www.lec.com/

[8] http://www.fjoch.com/GIZA++

points BLEU score; in the second column the difference was lowered to 0.01; and in the third column to 0.005. The numbers following the hash signs in the header row indicate the number of sentences in a given set.

metric	PH score – LM score
TER	1.997
BLEU	7.16%
NIST	6.58%
dep	4.93%
dep+paraphr	4.80%
GTM	3.89%
METEOR	3.80%
dep_preds	3.79%
dep+paraphr_preds	3.70%
dep+WordNet	3.55%
dep+WordNet_preds	2.60%
METEOR+WordNet	1.56%

Table 3. Difference between scores assigned to Pharaoh and Logomedia. Positive numbers show by how much Pharaoh's score was higher than Logomedia's. Legend: dep = dependency f-score, paraph = paraphrases, _preds = predicate-only f-score.

~ 0.05	#1692	~ 0.01	#567	~ 0.005	#335
NIST	2.29%	**NIST**	1.76%	**NIST**	1.48%
BLEU	0.95%	**BLEU**	0.42%	**BLEU**	0.59%
GTM	0.94%	**GTM**	0.29%	**GTM**	-0.09%
d+p	0.67%	**d**	0.04%	**d+p**	-0.15%
d	0.61%	**d+p**	0.02%	**d**	-0.24%
d+WN	-0.29%	**d+WN**	-0.78%	**d+WN**	-0.99%
d+p_pr	-0.70%	**M**	-0.99%	**d+p_pr**	-1.30%
d_pr	-0.75%	**d_pr**	-1.37%	**d_pr**	-1.43%
M	-1.03%	**d+p_pr**	-1.38%	**M**	-1.57%
d+WN_pr	-1.43%	**d+WN_pr**	-1.97%	**d+WN_pr**	-1.94%
M+WN	-2.51%	**M+WN**	-2.21%	**M+WN**	-2.74%
TER	-1.579	**TER**	-1.228	**TER**	-1.739

Table 4. Difference between scores assigned to Pharaoh and Logomedia for sets of increasing similarity. Positive numbers show Pharaoh's advantage, negative numbers show Logomedia's advantage. Legend: d = dependency f-score, p = paraphrases, _pr = predicate-only f-score, M = METEOR, WN = WordNet.

These results confirm earlier suggestions that the predicate-only version of the dependency-based evaluation is less biased in favour of the statistical MT system than the version that includes all dependency types. Adding a sufficient number of lexical choices reduces the bias even further; although again, paraphrases generated from the test set only are too few to make a significant difference. Similarly to METEOR, the dependency-based method shows on the whole lower bias than other metrics. However, we cannot be certain that the underlying scores vary linearly with each other and with human judgements, as we have no framework of reference such as human segment-level assessment of translation quality in this case. Therefore, the correlation with human judgement is analysed in our next experiment.

4.2 Correlation with human judgement – MultiTrans

To calculate how well the dependency-based method correlates with human judgement, and how it compares to the correlation shown by other metrics, we conducted an experiment on Chinese-English newswire text.

4.2.1 Experimental design

We used the data from the Linguistic Data Consortium Multiple Translation Chinese (MTC) Parts 2 and 4. The data consists of multiple translations of Chinese newswire text, four human-produced references, and segment-level human scores for a subset of the translation-reference pairs. Although a single translated segment was always evaluated by more than one judge, the judges used a different reference every time, which is why we treated each translation-reference-human score triple as a separate segment. In effect, the test set created from this data contained 16,800 segments. As in the previous experiment, the translation was scored using BLEU, NIST, GTM, TER, METEOR, and the dependency-based method.

4.2.2 Results

We calculated Pearson's correlation coefficient for segment-level scores that were given by each metric and by human judges. The results of the correlation are shown in Table 5. Note that the correlation for TER is negative, because in TER zero is the perfect score, in contrast to other metrics where zero is the worst possible score; however, this time the absolute values can be easily compared to each other. Rows are ordered

by the highest value of the (absolute) correlation with the human score.

First, it seems like none of the metrics is very good at reflecting human fluency judgments; the correlation values in the first column are significantly lower than the correlation with accuracy. However, the dependency-based method in almost all its versions has decidedly the highest correlation in this area. This can be explained by the method's sensitivity to the grammatical structure of the sentence: a more grammatical translation is also a translation that is more fluent.

H_FL		H_AC		H_AVE	
d+WN	0.168	**M+WN**	0.294	**M+WN**	0.255
d	0.162	**M**	0.278	**d+WN**	0.244
d+WN_pr	0.162	**NIST**	0.273	**M**	0.242
BLEU	0.155	**d+WN**	0.266	**NIST**	0.238
d_pr	0.154	**GTM**	0.260	**d**	0.236
M+WN	0.153	**d**	0.257	**GTM**	0.230
M	0.149	**d+WN_pr**	0.232	**d+WN_pr**	0.220
NIST	0.146	**d_pr**	0.224	**d_pr**	0.212
GTM	0.146	**BLEU**	0.199	**BLEU**	0.197
TER	-0.133	**TER**	-0.192	**TER**	-0.182

Table 5. Pearson's correlation between human scores and evaluation metrics. Legend: d = dependency f-score, _pr = predicate-only f-score, M = METEOR, WN = WordNet, H_FL = human fluency score, H_AC = human accuracy score, H_AVE = human average score.[9]

Second, and somewhat surprisingly, in this detailed examination the relative order of the metrics changed. The predicate-only version of the dependency-based method appears to be less adequate for correlation with human scores than its non-restricted versions. As to the correlation with human evaluation of translation accuracy, our method currently falls short of METEOR and even NIST. This is caused by the fact that both METEOR and NIST assign relatively little importance to the position of a specific word in a sentence, therefore rewarding the translation for content rather than linguistic form. For our dependency-based method, the noise introduced by the parser might be the reason for low correlation: if even one side of the translation-reference pair contains parsing errors, this may lead to a less reliable score. An obvious solution to this problem, which we are examining at the moment, is to include a number of best parses for each side of the evaluation.

High correlation with human judgements of fluency and lower correlation with accuracy results in a high second place for our dependency-based method when it comes to the average correlation coefficient. The WordNet-boosted dependency-based method scores only slightly lower than METEOR with WordNet. These results are very encouraging, especially as we see a number of ways the dependency-based method could be further developed.

5 Current and future work

While the idea of a dependency-based method is a natural step in the direction of a deeper linguistic analysis for MT evaluation, it does require an LFG grammar and parser for the target language. There are several obvious areas for improvement with respect to the method itself. First, we would also like to adapt the process of translation-reference dependency comparison to include *n*-best parsers for the input sentences, as well as some basic transformations which would allow an even deeper logical analysis of input (e.g. passive to active voice transformation).

Second, we want to repeat both experiments using a paraphrase set derived from a large parallel corpus, rather than the test set, as described in Owczarzak et al. (2006). While retaining the advantage of having a similar size to a corresponding set of WordNet synonyms, this set will also capture low-level syntactic variations, which can increase the number of matches and the correlation with human scores.

Finally, we want to take advantage of the fact that the score produced by the dependency-based method is the proportional average of f-scores for a group of up to 32 (but usually far fewer) different dependency types. We plan to implement a set of weights, one for each dependency type, trained in such a way as to maximize the correlation of the final dependency f-score with human evaluation.

6 Conclusions

In this paper we present a novel way of evaluating MT output. So far, all metrics relied on

[9] In general terms, an increase of 0.015 between any two scores is significant with a 95% confidence interval.

comparing translation and reference on a string level. Even given reordering, stemming, and synonyms for individual words, current methods are still far from reaching human ability to assess the quality of translation. Our method compares the sentences on the level of their grammatical structure, as exemplified by their f-structure dependency triples produced by an LFG parser. The dependency-based method can be further augmented by using paraphrases or WordNet synonyms, and is available in full version and predicate-only version. In our experiments we showed that the dependency-based method correlates higher than any other metric with human evaluation of translation fluency, and shows high correlation with the average human score. The use of dependencies in MT evaluation is a rather new idea and requires more research to improve it, but the method shows potential to become an accurate evaluation metric.

Acknowledgements

This work was partly funded by Microsoft Ireland PhD studentship 2006-8 for the first author of the paper. We would also like to thank our reviewers for their insightful comments. All remaining errors are our own.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*: 65-73.

Joan Bresnan. 2001. *Lexical-Functional Syntax*, Blackwell, Oxford.

Aoife Cahill, Michael Burke, Ruth O'Donovan, Josef van Genabith, and Andy Way. 2004. Long-Distance Dependency Resolution in Automatically Acquired Wide-Coverage PCFG-Based LFG Approximations, In *Proceedings of ACL-04*: 320-327

Chris Callison-Burch, Miles Osborne and Philipp Koehn. 2006. Re-evaluating the role of BLEU in Machine Translation Research. *Proceedings of EACL 2006*: 249-256

George Doddington. 2002. Automatic Evaluation of MT Quality using N-gram Co-occurrence Statistics. *Proceedings of HLT 2002*: 138-145.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. *Proceedings of HLT-NAACL 2006*: 45-462.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. *Proceedings of the AMTA 2004 Workshop on Machine Translation: From real users to research*: 115-124.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. *Proceedings of MT Summit 2005*: 79-86.

Alex Kulesza and Stuart M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation. In *Proceedings of the TMI 2004*: 75-84.

Gregor Leusch, Nicola Ueffing and Hermann Ney. 2006. CDER: Efficient MT Evaluation Using Block Movements. *Proceedings of EACL 2006*: 241-248.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Modes. *Computational Linguistics*, 29:19-51.

Karolina Owczarzak, Declan Groves, Josef van Genabith, and Andy Way. 2006. Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation. *Proceedings of the HLT-NAACL 2006 Workshop on Statistical Machine Translation*: 86-93.

Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*: 311-318.

Grazia Russo-Lassner, Jimmy Lin, and Philip Resnik. 2005. A Paraphrase-based Approach to Machine Translation Evaluation. Technical Report LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57, University of Maryland, College Park, MD.

Mathew Snover, Bonnie Dorr, Richard Schwartz, John Makhoul, Linnea Micciula. 2006. A Study of Translation Error Rate with Targeted Human Annotation. *Proceedings of AMTA 2006*: 223-231.

Joseph P. Turian, Luke Shen, and I. Dan Melamed. 2003. Evaluation of Machine Translation and Its Evaluation. *Proceedings of MT Summit 2003*: 386-393.

Ying Zhang and Stephan Vogel. 2004. Measuring confidence intervals for the machine translation evaluation metrics. *Proceedings of TMI 2004*: 85-94.