# Labelled Dependencies in Machine Translation Evaluation

**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 method for evaluating the quality of Machine Translation (MT) output, using labelled dependencies produced by a Lexical-Functional Grammar (LFG) parser. Our dependency-based method, in contrast to most popular string-based evaluation metrics, does not unfairly penalize perfectly valid syntactic variations in the translation, and the addition of WordNet provides a way to accommodate lexical variation. In comparison with other metrics on 16,800 sentences of Chinese-English newswire text, our method reaches high correlation with human scores.

## 1 Introduction

Since the creation of BLEU (Papineni et al., 2002) and NIST (Doddington, 2002), the subject of automatic evaluation metrics for MT has been given quite a lot of attention. Although widely popular thanks to their speed and efficiency, both BLEU and NIST have been criticized for inadequate accuracy of evaluation at the segment level (Callison-Burch et al., 2006). As string based-metrics, they are limited to superficial comparison of word sequences between a translated sentence and one or more reference sentences, and are unable to accommodate any legitimate grammatical variation when it comes to lexical choices or syntactic structure of the translation, beyond what can be found in the multiple references. A natural next step in the field of evaluation was to introduce metrics that would better reflect our human judgement by accepting synonyms in the translated sentence or evaluating the translation on the basis of what syntactic features it shares with the reference.

Our method follows and substantially extends the earlier work of Liu and Gildea (2005), who use syntactic features and unlabelled dependencies to evaluate MT quality, outperforming BLEU on segment-level correlation with human judgement. Dependencies abstract away from the particulars of the surface string (and syntactic tree) realization and provide a "normalized" representation of (some) syntactic variants of a given sentence.

While Liu and Gildea (2005) calculate n-gram matches on non-labelled head-modifier sequences derived by head-extraction rules from syntactic trees, we automatically evaluate the quality of translation by calculating an f-score on labelled dependency structures produced by a Lexical-Functional Grammar (LFG) parser. These dependencies differ from those used by Liu and Gildea (2005), in that they are extracted according to the rules of the LFG grammar and they are labelled with a type of grammatical relation that connects the head and the modifier, such as *subject*, *determiner*, etc. The presence of grammatical relation labels adds another layer of important linguistic information into the comparison and allows us to account for partial matches, for example when a lexical item finds itself in a correct relation but with an incorrect partner. Moreover, we use a number of best parses for the translation and the reference, which serves to decrease the amount of noise that can be introduced by the process of parsing and extracting dependency information.

The translation and reference files are analyzed by a treebank-based, probabilistic 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



*Proceedings of the Second Workshop on Statistical Machine Translation*, pages 104–111,
Prague, June 2007. 

precision, recall, and f-score for each 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) in adding a number of synonyms in the process of evaluation to raise the number of matches between the translation and the reference, leading to a higher score.

In an experiment on 16,800 sentences of Chinese-English newswire text with segment-level human evaluation from the Linguistic Data Consortium's (LDC) Multiple Translation project, we compare the LFG-based evaluation method with other popular metrics like 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), and we show that combining dependency representations with synonyms leads to a more accurate evaluation that correlates better with human judgment. Although evaluated on a different test set, our method also outperforms the correlation with human scores reported in Liu and Gildea (2005).

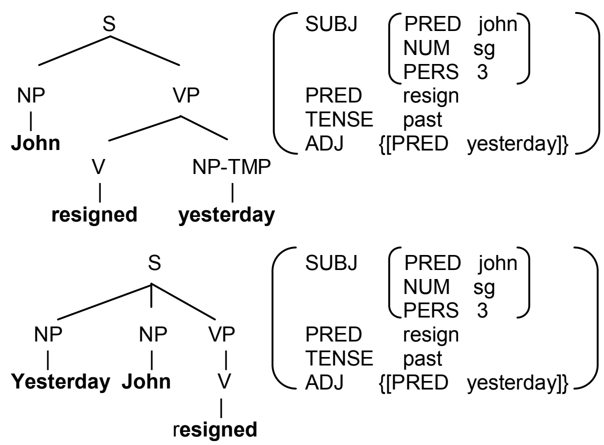
The remainder of this paper is organized as follows: Section 2 gives a basic introduction to LFG; Section 3 describes related work; Section 4 describes our method and gives results of the experiment on the Multiple Translation data; Section 5 discusses ongoing work; Section 6 concludes.

## 2 Lexical-Functional Grammar

In Lexical-Functional Grammar (Kaplan and Bresnan, 1982; Bresnan, 2001) sentence structure is represented in terms of c(onstituent)-structure and f(unctional)-structure. C-structure represents the word order of the surface string 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), etc., 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 rearrangement of constituents, f-structure abstracts away from the particulars of the surface realization. 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:*




Note that if these 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. This is because, although all three unigrams from the "translation" (*John*; *resigned*; *yesterday*) are present in the reference, which contains four items including the comma (*Yesterday*; *,*; *John*; *resigned*), the "translation" contains only one bigram (*John resigned*) that matches the "reference" (*Yesterday ,*; *, John*; *John resigned*), and no matching trigrams.

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

[1] We omit HTER (Human-Targeted Translation Error Rate), as it is not fully automatic and requires human input.

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

Cahill et al. (2004) presents a set of Penn-II Treebank-based LFG parsing resources. Their 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)}.

Other predicate-only dependencies include: *apposition*, *complement*, *open complement*, *coordination*, *determiner*, *object*, *second object*, *oblique*, *second oblique*, *oblique agent*, *possessive*, *quantifier*, *relative clause*, *topic*, and *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*, and *infinitival clause*.

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, obtaining high precision and recall rates. [3]

## 3 Related work

### 3.1 String-based metrics

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).

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., 2006), which computes the number of substitutions, insertions, 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[4] 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 Shieber (2004), on the other hand, train a Support Vector Machine using features such as proportion of *n*-gram matches and word error rate to judge a given translation's distance from human-level quality.

### 3.2 Dependency-based metric

The metrics described above use only string-based comparisons, even while taking into consideration reordering. By contrast, Liu and Gildea (2005) present three metrics that use syntactic and unlabelled dependency information. Two of these metrics are based on matching syntactic subtrees between the translation and the reference, and one

[3] A demo of the parser can be found at http://lfg-demo.computing.dcu.ie/lfgparser.html

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

is based on matching headword chains, i.e. sequences of words that correspond to a path in the unlabelled dependency tree of the sentence. Dependency trees are created by extracting a headword for each node of the syntactic tree, according to the rules used by the parser of Collins (1999), where every subtree represents the modifier information for its root headword. The dependency trees for the translation and the reference are converted into flat headword chains, and the number of overlapping *n*-grams between the translation and the reference chains is calculated. Our method, extending this line of research with the use of labelled LFG dependencies, partial matching, and *n*-best parses, allows us to considerably outperform Liu and Gildea's (2005) highest correlations with human judgement (they report 0.144 for the correlation with human fluency judgement, 0.202 for the correlation with human overall judgement), although it has to be kept in mind that such comparison is only tentative, as their correlation is calculated on a different test set.

## 4 LFG f-structure in MT evaluation

LFG-based automatic MT evaluation reflects the same process that underlies the evaluation of parser-produced f-structure quality against a gold standard: we parse the translation and the reference, and then, for each sentence, we check the set of labelled translation dependencies against the set of labelled 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.

### 4.1 Determining parser noise

Because we are comparing two outputs that were produced automatically, there is a possibility that the result will not be noise-free, even if the parser fails to provide a parse only in 0.1% of cases.

To assess the amount of noise that the parser introduces, Owczarzak et al. (2006) conducted an experiment where 100 English sentences were hand-modified so that the position of adjuncts was changed, but the sentence remained grammatical and the meaning was not influenced. This way, an ideal parser should give both the source and the modified sentence the same f-structure, similarly to the example 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, and the dependency triples obtained from the "translation" were compared against the dependency triples for the "reference", calculating the f-score. Additionally, the same "translation-reference" set was scored with other metrics (TER, METEOR, BLEU, NIST, and GTM). The results, including 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.[5] The other column lists 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.

We propose that the problem of parser noise can be alleviated by introducing a number of best parses into the comparison between the translation and the reference. Table 2 shows how increasing the number of parses available for comparison brings our method closer to an ideal noise-free parser.

[5] Two things have to be noted here: (1) in the case of NIST the perfect score differs from text to text, which is why the percentage points are provided along the numerical score, and (2) in the 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.

| | dependency f-score |
|---|---|
| **1 best** | 96.56 |
| **2 best** | 97.31 |
| **5 best** | 97.90 |
| **10 best** | 98.31 |
| **20 best** | 98.59 |
| **30 best** | 98.74 |
| **50 best** | 98.79 |
| **baseline** | 100 |

**Table 2. Dependency f-scores for sentences with reordered adjuncts with n-best parses available**

It has to be noted, however, that increasing the number of parses beyond a certain threshold does little to further improve results, and at the same time it considerably decreases the efficiency of the method, so it is important to find the right balance between these two factors. In our opinion, the optimal value would be 10-best parses.

## 4.2 Correlation with human judgement – MultiTrans

### 4.2.1 Experimental design

To evaluate the correlation with human assessment, we used the data from the Linguistic Data Consortium Multiple Translation Chinese (MTC) Parts 2 and 4, which 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 our labelled dependency-based method.

### 4.2.2 Labelled dependency-based method

We examined a number of modifications of the dependency-based method in order to find out which one gives the highest correlation with human scores. The correlation differences between immediate neighbours in the ranking were often too small to be statistically significant; however, there is a clear overall trend towards improvement.

Besides the plain version of the dependency f-score, we also looked at the f-score calculated on predicate dependencies only (ignoring "atomic" features such as *person*, *number*, *tense*, etc.), which turned out not to correlate well with human judgements.

Another addition was the use of 2-, 10-, or 50-best parses of the translation and reference sentences, which partially neutralized parser noise and resulted in increased correlations.

We also created a version where predicate dependencies of the type *subj*(resign,John) are split into two parts, each time replacing one of the elements participating in the relation with a variable, giving in effect *subj*(resign,x) and *subj*(y,John). This lets us score partial matches, where one correct lexical object happens to find itself in the correct relation, but with an incorrect "partner".

Lastly, we added WordNet synonyms into the matching process to accommodate lexical variation, and to compare our WordNet-enhanced method with the WordNet-enhanced version of METEOR.

### 4.2.3 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 3. 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. This finding has been previously reported, among others, in Liu and Gildea (2005). 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.

As to the correlation with human evaluation of translation accuracy, our method currently falls short of METEOR. This is caused by the fact that METEOR assign relatively little importance to the position of a specific word in a sentence, therefore rewarding the translation for content rather than linguistic form. Interestingly, while METEOR, with or without WordNet, considerably outperforms all other metrics when it comes to the correlation with human judgements of translation accuracy, it falls well behind most versions of our dependency-based method in correlation with human scores of translation fluency.

Surprisingly, adding partial matching to the dependency-based method resulted in the greatest increase in correlation levels, to the extent that the partial-match versions consistently outperformed versions with a larger number of parses available but without the partial match. The most interesting effect was that the partial-match versions (even those with just a single parse) offered results comparable to or higher than the addition of WordNet to the matching process when it comes to accuracy and overall judgement.

| fluency | | accuracy | | average | |
|---|---|---|---|---|---|
| **d_50+WN** | 0.177 | **M+WN** | 0.294 | **M+WN** | 0.255 |
| **d+WN** | 0.175 | **M** | 0.278 | **d_50_var** | 0.252 |
| **d_50_var** | 0.174 | **d_50_var** | 0.273 | **d_50+WN** | 0.250 |
| **GTM** | 0.172 | **NIST** | 0.273 | **d_10_var** | 0.250 |
| **d_10_var** | 0.172 | **d_10_var** | 0.273 | **d_2_var** | 0.247 |
| **d_50** | 0.171 | **d_2_var** | 0.270 | **d+WN** | 0.244 |
| **d_2_var** | 0.168 | **d_50+WN** | 0.269 | **d_50** | 0.243 |
| **d_10** | 0.168 | **d_var** | 0.266 | **d_var** | 0.243 |
| **d_var** | 0.165 | **d_50** | 0.262 | **M** | 0.242 |
| **d_2** | 0.164 | **d_10** | 0.262 | **d_10** | 0.242 |
| **d** | 0.161 | **d+WN** | 0.260 | **NIST** | 0.238 |
| **BLEU** | 0.155 | **d_2** | 0.257 | **d_2** | 0.237 |
| **M+WN** | 0.153 | **d** | 0.256 | **d** | 0.235 |
| **M** | 0.149 | **d_pr** | 0.240 | **d_pr** | 0.216 |
| **NIST** | 0.146 | **GTM** | 0.203 | **GTM** | 0.208 |
| **d_pr** | 0.143 | **BLEU** | 0.199 | **BLEU** | 0.197 |
| **TER** | -0.133 | **TER** | -0.192 | **TER** | -0.182 |

**Table 3. Pearson's correlation between human scores and evaluation metrics. Legend: d = dependency f-score, _pr = predicate-only f-score, 2, 10, 50 = n-best parses; var = partial-match version; M = METEOR, WN = WordNet[6]**

## 5 Current and future work

Fluency and accuracy are two very different aspects of translation quality, each with its own set of conditions along which the input is evaluated. Therefore, it seems unfair to expect a single automatic metric to correlate highly with human judgements of both at the same time. This pattern is very noticeable in Table 3: if a metric is (relatively) good at correlating with fluency, its accuracy correlation suffers (GTM might serve as an example here), and the opposite holds as well (see METEOR's scores). It does not mean that any improvement that increases the method's correlation with one aspect will result in a decrease in the correlation with the other aspect; but it does suggest that a possible way of development would be to target these correlations separately, if we want our automated metrics to reflect human scores better. At the same time, string-based metrics might have already exhausted their potential when it comes to increasing their correlation with human evaluation; as has been pointed out before, these metrics can only tell us that two strings differ, but they cannot distinguish legitimate grammatical variance from ungrammatical variance. As the quality of MT improves, the community will need metrics that are more sensitive in this respect. After all, the true quality of MT depends on producing grammatical output which describes the same concept as the source utterance, and the string identity with a reference is only a very selective approximation of this goal.

[6] In general terms, an increase of 0.022 or more between any two scores in the same column is significant with a 95% confidence interval. The statistical significance of correlation differences was calculated using Fisher's z' transformation and the general formula for confidence interval.

In order to maximize the correlation with human scores of fluency, we plan to look more closely at the parser output, and implement some basic transformations which would allow an even deeper logical analysis of input (e.g. passive to active voice transformation).

Additionally, we want to take advantage of the fact that the score produced by the dependency-based method is the proportional average of matches 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. In a preliminary experiment, for example, assigning a low weight to the *topic* dependency increases our correlations slightly (this particular case can also be seen as a transformation into a more basic logical form by removing non-elementary dependency types).

In a similar direction, we want to experiment more with the f-score calculations. Initial check shows that assigning a higher weight to recall than to precision improves results.

To improve the correlation with accuracy judgements, we would like to experiment using a paraphrase set derived from a large parallel corpus, 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.

## 6 Conclusions

In this paper we present a linguistically-motivated method for automatically evaluating the output of Machine Translation. Most currently used popular metrics rely on 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, and there exists a need in the community to develop more dependable metrics. Our method explores one such direction of development, comparing the sentences on the level of their grammatical structure, as exemplified by their f-structure labelled dependency triples produced by an LFG parser. 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 has not been extensively researched before (one exception here would be Liu and Gildea (2005)), 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 and Dan Melamed 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 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the Association for Computational Linguistics Conference 2005*: 65-73. Ann Arbor, Michigan.

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 Association for Computational Linguistics 2004*: 320-327. Barcelona, Spain.

Chris Callison-Burch, Miles Osborne and Philipp Koehn. 2006. Re-evaluating the role of BLEU in Machine Translation Research. *Proceedings of the European Chapter of the Association for Computational Linguistics 2006*: 249-256. Oslo, Norway.

Michael J. Collins. 1999. Head-driven Statistical Models for Natural Language Parsing. Ph.D. thesis, University of Pennsylvania, Philadelphia.

George Doddington. 2002. Automatic Evaluation of MT Quality using N-gram Co-occurrence Statistics. *Proceedings of Human Language Technology Conference 2002*: 138-145. San Diego, California.

Kaplan, R. M., and J. Bresnan. 1982. *Lexical-functional Grammar: A Formal System for Grammatical*

*Representation*. In J. Bresnan (ed.), The Mental Representation of Grammatical Relations. MIT Press, Cambridge.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. *Proceedings of Human Language Technology – North American Chapter of the Association for Computational Linguistics Conference 2006*: 45-462. New York, New York.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. *Proceedings of the Workshop on Machine Translation: From real users to research at the Association for Machine Translation in the Americas Conference 2004*: 115-124. Washington, DC.

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

Alex Kulesza and Stuart M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation. In *Proceedings of* the *Conference on Theoretical and Methodological Issues in Machine Translation 2004*: 75-84. Baltimore, Maryland.

Gregor Leusch, Nicola Ueffing and Hermann Ney. 2006. CDER: Efficient MT Evaluation Using Block Movements. *Proceedings of European Chapter of the Association for Computational Linguistics Conference 2006*: 241-248. Trento, Italy.

Ding Liu and Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization at the Association for Computational Linguistics Conference 2005.* Ann Arbor, Michigan.

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 Workshop on Statistical Machine Translation at the Human Language Technology – North American Chapter of the Association for Computational Linguistics Conference 2006*: 86-93. New York, New York.

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

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, Maryland.

Mathew Snover, Bonnie Dorr, Richard Schwartz, John Makhoul, Linnea Micciula. 2006. A Study of Translation Error Rate with Targeted Human Annotation. *Proceedings of the Association for Machine Translation in the Americas Conference 2006*: 223-231. Boston, Massachusetts.

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

Ying Zhang and Stephan Vogel. 2004. Measuring confidence intervals for the machine translation evaluation metrics. *Proceedings of Conference on Theoretical and Methodological Issues in Machine Translation 2004*: 85-94. Baltimore, Maryland.