Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation

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

Abstract

In this paper we present a novel method for deriving paraphrases during automatic MT evaluation using only the source and reference texts, which are necessary for the evaluation, and word and phrase alignment software. Using target language paraphrases produced through word and phrase alignment a number of alternative reference sentences are constructed automatically for each candidate translation. The method produces lexical and low-level syntactic paraphrases that are relevant to the domain in hand, does not use external knowledge resources, and can be combined with a variety of automatic MT evaluation system.

1 Introduction

Since their appearance, BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) have been the standard tools used for evaluating the quality of machine translation. They both score candidate translations on the basis of the number of n-grams it shares with one or more reference translations provided. Such automatic measures are indispensable in the development of machine translation systems, because they allow the 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 choices and syntactic structure it contains, even though perfectly legitimate, are not present in at least one of the references. Necessarily, this score would not reflect a much more favourable human judgment 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 the BLEU- or NIST-based evaluation in the first place. 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. On the other hand, in practice even a number of references do not capture the whole potential variability of the translation. Moreover, often it is the case that multiple references are not available or are too difficult and expensive to produce: when designing a statistical machine translation system, the need for large amounts of training data limits the researcher to collections of parallel corpora like 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, often exceeds the resources available. Therefore, it would be desirable to find a way to automatically generate legitimate translation alternatives not present in the reference(s) already available.

Proceedings of the Workshop on Statistical Machine Translation, pages 86–93,
New York City, June 2006. ©2006 Association for Computational Linguistics

In this paper, we present a novel method that automatically derives paraphrases using only the source and reference texts involved in for the evaluation of French-to-English Europarl translations produced by two MT systems: statistical phrase-based Pharaoh (Koehn, 2004) and rule-based Logomedia.[1] In using what is in fact a miniature bilingual corpus our approach differs from the mainstream paraphrase generation based on monolingual resources. We show that paraphrases produced in this way are more relevant to the task of evaluating machine translation than the use of external lexical knowledge resources like thesauri or WordNet[2], in that our paraphrases contain both lexical equivalents and low-level syntactic variants, and in that, as a side-effect, evaluation bitext-derived paraphrasing naturally yields domain-specific paraphrases. The paraphrases generated from the evaluation bitext are added to the existing reference sentences, in effect creating multiple references and resulting in a higher score for the candidate translation. Our hypothesis, confirmed by the experiments in this paper, is that the scores raised by additional references produced in this way will correlate better with human judgment than the original scores.

The remainder of this paper is organized as follows: Section 2 describes related work; Section 3 describes our method and presents examples of derived paraphrases; Section 4 presents the results of the comparison between the BLUE and NIST scores for a single-reference translation and the same translation using the paraphrases automatically generated from the bitext, as well as the correlations between the scores and human judgment; Section 5 discusses ongoing work; Section 6 concludes.

2 Related work

2.1 Word and phrase alignment

Several researchers noted that the word and phrase alignment used in training translation models in Statistical MT can be used for other purposes as well. (Diab and Resnik, 2002) use second language alignments to tag word senses. Working on an assumption that separate senses of a L1 word can be distinguished by its different translations in L2, they also note that a set of possible L2 translations for a L1 word may contain many synonyms. (Bannard and Callison-Burch, 2005), on the other hand, conduct an experiment to show that paraphrases derived from such alignments can be semantically correct in more than 70% of the cases.

2.2 Automatic MT evaluation

The insensitivity of BLEU and NIST to perfectly legitimate 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 problem also noted by (Och et al., 2003) and (Russo-Lassner et al., 2005). 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 data 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 machine translation metrics. Some of them concentrate mainly on the word reordering aspect, like Maximum Matching String (Turian et al., 2003) or Translation Error Rate (Snover et al., 2005). 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; METEOR (Banerjee and Lavie, 2005), which uses stemming and WordNet synonymy; and a 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.

A closer examination of these metrics suggests that the accommodation of lexical equivalence is as difficult as the appropriate treatment of syntactic variation, in that it requires considerable external knowledge resources like WordNet, verb class databases, and extensive text preparation: stemming, tagging, etc. The advantage of our method is that it produces relevant paraphrases with nothing more than the evaluation bitext and a widely available word and phrase alignment software, and therefore can be used with any existing evaluation metric.

---

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

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

3 Contextual bitext-derived paraphrases

The method presented in this paper rests on a combination of two simple ideas. First, the components necessary for automatic MT evaluation like BLEU or NIST, a source text and a reference text, constitute a miniature parallel corpus, from which word and phrase alignments can be extracted automatically, much like during the training for a statistical machine translation system. Second, target language words e_{i1}, …, e_{in} aligned as the likely translations to a source language word f_i are often synonyms or near-synonyms of each other. This also holds for phrases: target language phrases ep_{i1}, …, ep_{in} aligned with a source language phrase fp_i are often paraphrases of each other. For example, in our experiment, for the French word *question* the most probable automatically aligned English translations are *question*, *matter*, and *issue*, which in English are practically synonyms. Section 3.2 presents more examples of such equivalent expressions.

3.1 Experimental design

For our experiment, we used two test sets, each consisting of 2000 sentences, drawn randomly from the test section of the Europarl parallel corpus. The source language was French and the target language was English. One of the test sets was translated by Pharaoh trained on 156,000 French-English sentence pairs. The other test set was translated by Logomedia, a commercially available rule-based MT system. Each test set consisted therefore of three files: the French source file, the English translation file, and the English reference file.

Each translation was evaluated by the BLEU and NIST metrics first with the single reference, then with the multiple references for each sentence using the paraphrases automatically generated from the source-reference mini corpus. A subset of a 100 sentences was randomly extracted from each test set and evaluated by two independent human judges with respect to accuracy and fluency; the human scores were then compared to the BLEU and NIST scores for the single-reference and the automatically generated multiple-reference files.

3.2 Word alignment and phrase extraction

We used the GIZA++ word alignment software[3] to produce initial word alignments for our miniature bilingual corpus consisting of the source French file and the English reference file, and the refined word alignment strategy of (Och and Ney, 2003; Koehn et al., 2003; Tiedemann, 2004) to obtain improved word and phrase alignments.

For each source word or phrase f_i that is aligned with more than one target words or phrases, its possible translations e_{i1}, ..., e_{in} were placed in a list as equivalent expressions (i.e. synonyms, near-synonyms, or paraphrases of each other). A few examples are given in (1).

```
(1) agreement - accordance
    adopted - implemented
    matter - lot - case
    funds - money
    arms - weapons
    area - aspect
    question - issue - matter
    we would expect - we cer-
    tainly expect
    bear on - are centred
    around
```

Alignment divides target words and phrases into equivalence sets; each set corresponds to one source word/phrase that was originally aligned with the target elements. For example, for the French word *citoyens* three English words were deemed to be the most appropriate translations: *people*, *public*, and *citizens*; therefore these three words constitute an equivalence set. Another French word *population* was aligned with two English translations: *population* and *people*; so the word *people* appears in two equivalence set (this gives rise to the question of equivalence transitivity, which will be discussed in Section 3.3). From the 2000-sentence evaluation bitext we derived 769 equivalence sets, containing in total 1658 words or phrases. Each set contained on average two or three elements. In effect, we produced at least one equivalent expression for 1658 English words or phrases.

An advantage of our method is that the target paraphrases and words come ordered with respect to their likelihood of being the translation of the source word or phrase – each of them is assigned a probability expressing this likelihood, so we are able to choose only the most likely translations, according to some experimentally established threshold. The experiment reported here was conducted without such a threshold, since the word and phrase alignment was of a very high quality.

---

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

3.3 Domain-specific lexical and syntactic paraphrases

It is important to notice here how the paraphrases produced are more appropriate to the task at hand than synonyms extracted from a general-purpose thesaurus or WordNet. First, our paraphrases are contextual - they are restricted to only those *relevant to the domain* of the text, since they are derived from the text itself. Given the context provided by our evaluation bitext, the word *area* in (1) turns out to be only synonymous with *aspect*, and not with *land*, *territory*, *neighbourhood*, *division*, or other synonyms a general-purpose thesaurus or WordNet would give for this entry. This allows us to limit our multiple references only to those that are likely to be useful in the context provided by the source text. Second, the phrase alignment captures something neither a thesaurus nor WordNet will be able to provide: a certain amount of syntactic variation of paraphrases. Therefore, we know that a string such as *we would expect* in (1), with the sequence *noun-aux-verb*, might be paraphrased by *we certainly expect*, a sequence of *noun-adv-verb*.

3.4 Open and closed class items

One important conclusion we draw from analysing the synonyms obtained through word alignment is that equivalence is limited mainly to words that belong to open word classes, i.e. nouns, verbs, adjectives, adverbs, but is unlikely to extend to closed word classes like prepositions or pronouns. For instance, while the French preposition *à* can be translated in English as *to*, *in*, or *at*, depending on the context, it is not the case that these three prepositions are synonymous in English. The division is not that clear-cut, however: within the class of pronouns, *he*, *she*, and *you* are definitely not synonymous, but the demonstrative pronouns *this* and *that* might be considered equivalent for some purposes. Therefore, in our experiment we exclude prepositions and in future work we plan to examine the word alignments more closely to decide whether to exclude any other words.

3.5 Creating multiple references

After the list of synonyms and paraphrases is extracted from the evaluation bitext, for each reference sentence a string search replaces every eligible word or phrase with its equivalent(s) from the paraphrase list, one at a time, and the resulting string is added to the array of references. The original string is added to the array as well. This process results in a different number of reference sentences for every test sentence, depending on whether there was anything to replace in the reference and how many paraphrases we have available for the original substring. One example of this process is shown in (2).

(2) *Original reference:*
i admire the **answer** mrs parly gave this morning **but** we have turned a blind eye to **that**
Paraphrase 1:
i admire the **reply** mrs parly gave this morning but we have turned a blind eye to that
Paraphrase 2:
i admire the answer mrs parly gave this morning **however** we have turned a blind eye to that
Paraphrase 3:
i admire the answer mrs parly gave this morning but we have turned a blind eye to **it**

Transitivity

As mentioned before, an interesting question that arises here is the potential transitivity of our automatically derived synonyms/paraphrases. It could be argued that if the word *people* is equivalent to *public* according to one set from our list, and to the word *population* according to another set, then *public* can be thought of as equivalent to *population*. In this case, the equivalence is not controversial. However, consider the following relation: if *sure* in one of the equivalence sets is synonymous to *certain*, and *certain* in a different set is listed as equivalent to *some*, then treating *sure* and *some* as synonyms is a mistake. In our experiment we do not allow synonym transitivity; we only use the paraphrases from equivalence sets containing the word/phrase we want to replace.

Multiple simultaneous substitution

Note that at the moment the references we are producing do not contain multiple simultaneous substitutions of equivalent expressions; for example, in (2) we currently do not produce the following versions:

(3) *Paraphrase 4:*
i admire the **reply** mrs parly
gave this morning **however** we
have turned a blind eye to
that
Paraphrase 5:
i admire the **answer** mrs parly
gave this morning **however** we
have turned a blind eye to **it**
Paraphrase 6:
i admire the **reply** mrs parly
gave this morning **but** we have
turned a blind eye to **it**

This can potentially prevent higher n-grams being successfully matched if two or more equivalent expressions find themselves within the range of n-grams being tested by BLEU and NIST. To avoid combinatorial problems, implementing multiple simultaneous substitutions could be done using a lattice, much like in (Pang et al., 2003).

4 Results

As expected, the use of multiple references produced by our method raises both the BLEU and NIST scores for translations produced by Pharaoh (test set PH) and Logomedia (test set LM). The results are presented in Table 1.

	BLEU	NIST
PH single ref	0.2131	6.1625
PH multi ref	0.2407	7.0068
LM single ref	0.1782	5.5406
LM multi ref	0.2043	6.3834

Table 1. Comparison of single-reference and multi-reference scores for test set PH and test set LM

The hypothesis that the multiple-reference scores reflect better human judgment is also confirmed. For 100-sentence subsets (Subset PH and Subset LM) randomly extracted from our test sets PH and LM, we calculated Pearson's correlation between the average accuracy and fluency scores that the translations in this subset received from two human judges (for each subset) and the single-reference and multiple-reference sentence-level BLEU and NIST scores.

There are two issues that need to be noted at this point. First, BLEU scored many of the sentences as zero, artificially leveling many of the weaker translations.[4] This explains the low, although still statistically significant (p value < 0.01[5]) correlation with BLEU for both single and multiple reference translations. Using a version of BLEU with add-one smoothing we obtain considerably higher correlations. Table 2 shows Pearson's correlation coefficient for BLEU, BLEU with add-one smoothing, NIST, and human judgments for Subsets PH. Multiple paraphrase references produced by our method consistently lead to a higher correlation with human judgment for every metric.[6]

Subset PH Metric	single ref	multi ref
H & BLEU	0.297	0.307
H & BLEU smoothed	0.396	0.404
H & NIST	0.323	0.355

Table 2. Pearson's correlation between human judgment and single-reference and multiple-reference BLEU, smoothed BLEU, and NIST for subset PH (of test set PH)

The second issue that requires explanation is the lower general scores Logomedia's translation received on the full set of 2000 sentences, and the extremely low correlation of its automatic evaluation with human judgment, irrespective of the number of references. It has been noticed (Callison-Burch et al., 2006) that BLEU and NIST favour n-gram based MT models such as Pharaoh, so the translation produced by Logomedia scored lower on the automatic evaluation, even though the human judges rated Logomedia output higher than Pharaoh's translation. Both human judges consistently gave very high scores to most sentences in subset LM (Logomedia), and as a consequence there was not enough variation in the scores assigned by them to create a good correlation with the BLEU and NIST scores. The average human scores for the subsets PH and LM and the coefficients of variation are presented in Table 3. It is easy to see that Logomedia's translation received a higher mean score (on a scale 0 to 5) from the human judges and with less variance than Pharaoh.

---

[4] BLEU uses a geometric average while calculating the sentence-level score and will score a sentence as 0 if it does not have at least one 4-gram.

[5] A critical value for Pearson's correlation coefficient for the sample size between 90 and 100 is 0.267, with $p < 0.01$.

[6] The significance of the rise in scores was confirmed in a resampling/bootstrapping test, with $p < 0.0001$.

	Mean score	**Variation**
Subset PH	3.815	19.1%
Subset LM	4.005	16.25%

Table 3. Human judgment mean scores and coefficients of variation for Subset PH and Subset LM

As a result of the consistently high human scores for Logomedia, none of the Pearson's correlations computed for Subset LM is high enough to be significant. The values are lower than the critical value 0.164 corresponding to $p < 0.10$.

Subset LM / **Metric**	**single ref**	**multi ref**
H & BLEU	0.046*	0.067*
H & BLEU smoothed	0.163*	0.151*
H & NIST	0.078*	0.116*

Table 4. Pearson's correlation between human judgment and single-reference and multiple-reference BLEU, smoothed BLEU, and NIST for subset LM (of test set LM). * denotes values with $p > 0.10$.

5 Current and future work

We would like to experiment with the way in which the list of equivalent expressions is produced. One possible development would be to derive the expressions from a very large training corpus used by a statistical machine translation system, following (Bannard and Callison-Burch, 2005), for instance, and use it as an external wider-purpose knowledge resource (rather than a current domain-tailored resource as in our experiment), which would be nevertheless improve on a thesaurus in that it would also include phrase equivalents with some syntactic variation. According to (Bannard and Callison-Burch, 2005), who derived their paraphrases automatically from a corpus of over a million German-English Europarl sentences, the baseline syntactic and semantic accuracy of the best paraphrases (those with the highest probability) reaches 48.9% and 64.5%, respectively. That is, by replacing a phrase with its one most likely paraphrase the sentence remained syntactically well-formed in 48.9% of the cases and retained its meaning in 65% of the cases.

In a similar experiment we generated paraphrases from a French-English Europarl corpus of 700,000 sentences. The data contained a considerably higher level of noise than our previous experiment on the 2000-sentence test set, even though we excluded any non-word entities from the results. Like (Bannard and Callison-Burch, 2005), we used the product of probabilities $p(f_i|e_{i1})$ and $p(e_{i2}|f_i)$ to determine the best paraphrase for a given English word e_{i1}. We then compared the accuracy across four samples of data. Each sample contained 50 randomly drawn words/phrases and their paraphrases. For the first two samples, the paraphrases were derived from the initial 2000-sentence corpus; for the second two, the paraphrases were derived from the 700,000-sentence corpus. For each corpus, one of the two samples contained only one best paraphrase for each entry, while the other listed all possible paraphrases. We then evaluated the quality of each paraphrase with respect to its syntactic and semantic accuracy. In terms of syntax, we considered the paraphrase accurate either if it had the same category as the original word/phrase; in terms of semantics, we relied on human judgment of similarity. Tables 5 and 6 summarize the syntactic and semantic accuracy levels in the samples.

Paraphrases / **Derived from**	**Best**	**All**
2000-sent. corpus	59%	60%
700,000-sent. corpus	70%	48%

Table 5. Syntactic accuracy of paraphrases

Paraphrases **Derived from**	**Best**	**All**
2000-sent. corpus	83%	74%
700,000-sent. corpus	76%	68%

Table 6. Semantic accuracy of paraphrases

Although it has to be kept in mind that these percentages were taken from relatively small samples, an interesting pattern emerges from comparing the results. It seems that the average syntactic accuracy of all paraphrases decreases with increased corpus size, but the syntactic accuracy of the one best paraphrase improves. This reflects the idea behind word alignment: the bigger the corpus, the more potential alignments there are for a given word, but at the same time the better their order in terms of probability and the likelihood to obtain the correct translation. Interestingly, the same pattern is not repeated for semantic accuracy, but again, these samples are quite small. In order to address this issue, we plan to repeat the experiment with more data.

Additionally, it should be noted that certain expressions, although not completely correct syntactically, could be retained in the paraphrase lists for the purposes of machine translation evaluation. Consider the case where our equivalence set looks like this:

`(4) abandon - abandoning - abandoned`

The words in (4) are all inflected forms of the verb *abandon*, and although they would produce rather ungrammatical paraphrases, those ungrammatical paraphrases still allow us to score our translation higher in terms of BLEU or NIST if it contains one of the forms of *abandon* than when it contains some unrelated word like *piano* instead. This is exactly what other scoring metrics mentioned in Section 2 attempt to obtain with the use of stemming or prefix matching.

6 Conclusions

In this paper we present a novel combination of existing ideas from statistical machine translation and paraphrase generation that leads to the creation of multiple references for automatic MT evaluation, using only the source and reference files that are required for the evaluation. The method uses simple word and phrase alignment software to find possible synonyms and paraphrases for words and phrases of the target text, and uses them to produce multiple reference sentences for each test sentence, raising the BLEU and NIST evaluation scores and reflecting human judgment better. The advantage of this method over other ways to generate paraphrases is that (1) unlike other methods, it does not require extensive parallel monolingual paraphrase corpora, but it extracts equivalent expressions from the miniature bilingual corpus of the source and reference evaluation files; (2) unlike other ways to accommodate synonymy in automatic evaluation, it does not require external lexical knowledge sources like thesauri or WordNet; (3) it extracts only synonyms that are relevant to the domain in hand; and (4) the equivalent expressions it produces include a certain amount of syntactic paraphrases.

The method is general and it can be used with any automatic evaluation metric that supports multiple references. In our future work, we plan to apply it to newly developed evaluation metrics like CDER and TER that aim to allow for syntactic variation between the candidate and the reference, therefore bringing together solutions for the two shortcomings of automatic evaluation systems: insensitivity to allowable lexical differences and syntactic variation.

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.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005): 597-604.*

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

Mona Diab and Philip Resnik. 2002. An unsupervised Method for Word Sense Tagging using Parallel Corpora. *Proceedings of the 40th Annual Meeting of the*

Association for Computational Linguistics, Philadelphia, PA.

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

Philipp Koehn, Franz Och and Daniel Marcu. 2003. Statistical Phrase-Based Translation. *Proceedings of the Human Language Technology Conference (HLT-NAACL 2003)*: 48-54.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. *Machine translation: From real users to research. 6th Conference of the Association for Machine Translation in the Americas (AMTA 2004)*: 115-124.

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

Gregor Leusch, Nicola Ueffing and Hermann Ney. 2006. CDER: Efficient MT Evaluation Using Block Movements. To appear in *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006).*

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

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2003. *Syntax for statistical machine translation.* Technical report, Center for Language and Speech Processing, John Hopkins University, Baltimore, MD.

Bo Pang, Kevin Knight and Daniel Marcu. 2003. Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. *Proceedings of Human Language Technology-North American Chapter of the Association for Computational Linguistics (HLT-NAACL) 2003: 181—188.*

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*: 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 and Ralph Weischedel. 2005. *A Study of Translation Error Rate with Targeted Human Annotation*. Technical Report LAMP-TR-126, CS-TR-4755, UMIACS-TR-2005-58, University of Maryland, College Park. MD.

Jörg Tiedemann. 2004. Word to word alignment strategies. *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*: 212-218.

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. *TMI-2004: Proceedings of the 10th Conference on Theoretical and Methodological Issues in Machine Translation*: 85-94.