MT Evaluation: Human-like vs. Human Acceptable

Enrique Amigó † , Jesús Giménez ‡ , Julio Gonzalo † , and Lluís Màrquez ‡

† Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
Juan del Rosal, 16, E-28040, Madrid
{enrique,julio}@lsi.uned.es

‡ TALP Research Center, LSI Department
Universitat Politècnica de Catalunya
Jordi Girona Salgado, 1–3, E-08034, Barcelona
{jgimenez,lluism}@lsi.upc.edu

Abstract

We present a comparative study on Machine Translation Evaluation according to two different criteria: Human Likeness and Human Acceptability. We provide empirical evidence that there is a relationship between these two kinds of evaluation: Human Likeness implies Human Acceptability but the reverse is not true. From the point of view of automatic evaluation this implies that metrics based on Human Likeness are more reliable for system tuning.

Our results also show that current evaluation metrics are not always able to distinguish between automatic and human translations. In order to improve the descriptive power of current metrics we propose the use of additional syntax-based metrics, and metric combinations inside the QARLA Framework.

1 Introduction

Current approaches to Automatic Machine Translation (MT) Evaluation are mostly based on metrics which determine the quality of a given translation according to its similarity to a given set of reference translations. The commonly accepted criterion that defines the quality of an evaluation metric is its level of correlation with human evaluators. High levels of correlation (Pearson over 0.9) have been attained at the system level (Eck and Hori, 2005). But this is an average effect: the degree of correlation achieved at the sentence level, crucial for an accurate error analysis, is much lower.

We argue that there is two main reasons that explain this fact:

Firstly, current MT evaluation metrics are based on shallow features. Most metrics work only at the lexical level. However, natural languages are rich and ambiguous, allowing for many possible different ways of expressing the same idea. In order to capture this flexibility, these metrics would require a combinatorial number of reference translations, when indeed in most cases only a single reference is available. Therefore, metrics with higher descriptive power are required.

Secondly, there exists, indeed, two different evaluation criteria: (i) Human Acceptability, i.e., to what extent an automatic translation could be considered acceptable by humans; and (ii) Human Likeness, i.e., to what extent an automatic translation could have been generated by a human translator. Most approaches to automatic MT evaluation implicitly assume that both criteria should lead to the same results; but this assumption has not been proved empirically or even discussed.

In this work, we analyze this issue through empirical evidence. First, in Section 2, we investigate to what extent current evaluation metrics are able to distinguish between human and automatic translations (Human Likeness). As individual metrics do not capture such distinction well, in Section 3 we study how to improve the descriptive power of current metrics by means of metric combinations inside the QARLA Framework (Amigó et al., 2005), including a new family of metrics based on syntactic criteria. Second, we claim that the two evaluation criteria (Human Acceptability and Human Likeness) are indeed of a different nature, and may lead to different results (Section 4). However, translations exhibiting a high level of Human Likeness obtain good results in human judges. Therefore, automatic evaluation metrics based on similarity to references should be

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 17–24,
Sydney, July 2006.

optimized over their capacity to represent Human Likeness. See conclusions in Section 5.

2 Descriptive Power of Standard Metrics

In this section we perform a simple experiment in order to measure the descriptive power of current state-of-the-art metrics, i.e., their ability to capture the features which characterize human translations with respect to automatic ones.

2.1 Experimental Setting

We use the data from the *Openlab 2006* Initiative[1] promoted by the TC-STAR Consortium[2]. This test suite is entirely based on European Parliament Proceedings[3], covering April 1996 to May 2005. We focus on the Spanish-to-English translation task. For the purpose of evaluation we use the development set which consists of 1008 sentences. However, due to lack of available MT outputs for the whole set we used only a subset of 504 sentences corresponding to the first half of the development set. Three human references per sentence are available.

We employ ten system outputs; nine are based on Statistical Machine Translation (SMT) systems (Giménez and Màrquez, 2005; Crego et al., 2005), and one is obtained from the free Systran[4] on-line rule-based MT engine. Evaluation results have been computed by means of the IQ$_{\text{MT}}$[5] Framework for Automatic MT Evaluation (Giménez and Amigó, 2006).

We have selected a representative set of 22 metric variants corresponding to six different families: BLEU (Papineni et al., 2001), NIST (Doddington, 2002), GTM (Melamed et al., 2003), mPER (Leusch et al., 2003), mWER (Nießen et al., 2000) and ROUGE (Lin and Och, 2004a).

2.2 Measuring Descriptive Power of Evaluation Metrics

Our main assumption is that if an evaluation metric is able to characterize human translations, then, human references should be closer to each other than automatic translations to other human references. Based on this assumption we introduce two measures (ORANGE and KING) which analyze the descriptive power of evaluation metrics from diferent points of view.

ORANGE Measure

ORANGE compares automatic and manual translations one-on-one. Let A and R be the sets of automatic and reference translations, respectively, and $x(a, R)$ an evaluation metric which outputs the quality of an automatic translation $a \in A$ by comparison to R. ORANGE measures the descriptive power as the probability that a human reference r is more similar than an automatic translation a to the rest of human references:

$$ORANGE_{A,R}(x) =$$
$$P(r \in R, a \in A : x(r, R-\{r\}) \geq x(a, R-\{r\}))$$

ORANGE was introduced by Lin and Och (2004b)[6] for the meta-evaluation of MT evaluation metrics. The $ORANGE$ measure provides information about the average behavior of automatic and manual translations regarding an evaluation metric.

KING Measure

However, ORANGE does not provide information about how many manual translations are discernible from automatic translations. The $KING$ measure complements the ORANGE, tackling these two issues by universally quantifying on variable a:

$$KING_{A,R}(x) =$$
$$P(r \in R, \forall a \in A : x(r, R-\{r\}) \geq x(a, R-\{r\}))$$

KING represents the probability that, for a given evaluation metric, a human reference is more similar to the rest of human references than *any* automatic translation [7].

KING does not depend on the distribution of automatic translations, and identifies the cases for

[1] http://tc-star.itc.it/openlab2006/

[2] http://www.tc-star.org/

[3] http://www.europarl.eu.int/

[4] http://www.systransoft.com.

[5] The IQ$_{\text{MT}}$ Framework may be freely downloaded at http://www.lsi.upc.edu/~nlp/IQMT.

[6] They defined this measure as the average rank of the reference translations within the combined machine and reference translations list.

[7] Originally KING is defined over the evaluation metric QUEEN, satisfying some restrictions which are not relevant in our context (Amigó et al., 2005).

which the given metric has been able to discern human translations from automatic ones. That is, it measures how many manual translations can be used as gold-standard for system evaluation/improvement purposes.

2.3 Results

Figure 1 shows the descriptive power, in terms of the ORANGE and KING measures, over the test set described in Subsection 2.1.

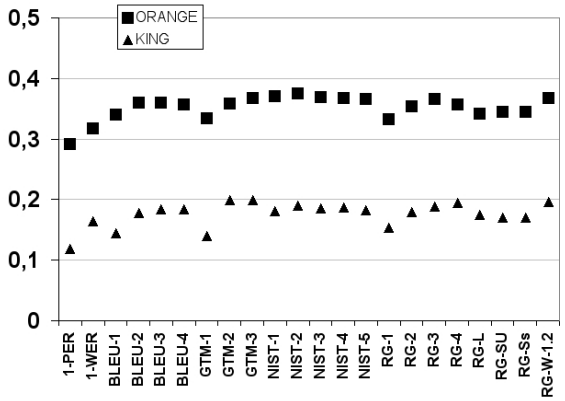

Figure 1: ORANGE and KING values for standard metrics.

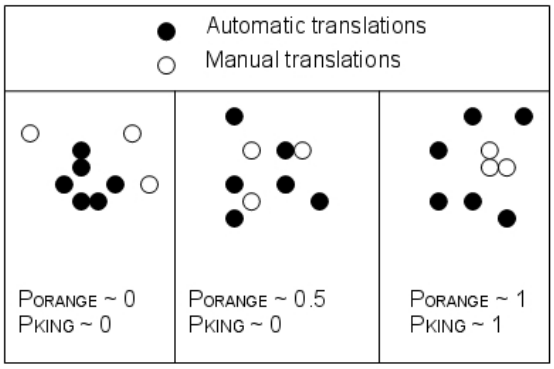

Figure 2: ORANGE and KING behavior.

ORANGE Results

All values of the ORANGE measure are lower than 0.5, which is the ORANGE value that a random metric would obtain (see central representation in Figure 2). This is a rather counterintuitive result. A reasonable explanation, however, is that automatic translations behave as centroids with respect to human translations, because they somewhat average the vocabulary distribution in the manual references; as a result, automatic translations are closer to each manual summary than manual summaries to each other (see leftmost representation in Figure 2).

In other words, automatic translations tend to share (lexical) features with most of the references, but not to match exactly any of them. This is a combined effect of:

- The nature of MT systems, mostly statistical, which compute their estimates based on the number of occurrences of words, tending to rely more on events that occur more often. Consequently, automatic translations typically consist of frequent words, which are likely to appear in most of the references.
- The shallowness of current metrics, which are not able to identify the common properties of manual translations with regard to automatic translations.

KING Results

KING values, on the other hand, are slightly higher than the value that a random metric would obtain ($\frac{1}{|A|} = 0.1$). This means that every standard metric is able to discriminate a certain number of manual translations from the set of automatic translations; for instance, GTM-3 identifies 19% of the manual references. For the remaining 81% of the test cases, however, GTM-3 cannot make the distinction, and therefore cannot be used to detect and improve weaknesses of the automatic MT systems.

These results provide an explanation for the low correlation between automatic evaluation metrics and human judgements at the sentence level. The necessary conclusion is that new metrics with higher descriptive power are required.

3 Improving Descriptive Power

The design of a metric that is able to capture all the linguistic aspects that distinguish human translations from automatic ones is a difficult path to trace. We approach this challenge by following a 'divide and conquer' strategy. We suggest to build a set of specialized similarity metrics devoted to the evaluation of partial aspects of MT quality. The challenge is then how to combine a set of similarity metrics into a single evaluation measure of

MT quality. The QARLA framework provides a solution for this challenge.

3.1 Similarity Metric Combinations inside QARLA

The QARLA Framework permits to combine several similarity metrics into a single quality measure (QUEEN). Besides considering the similarity of automatic translations to human references, the QUEEN measure additionally considers the distribution of similarities among human references.

The QUEEN measure operates under the assumption that a good translation must be similar to human references (R) according to all similarity metrics. QUEEN(a) is defined as the probability, over $R \times R \times R$, that for every metric x in a given metric set X the automatic translation a is more similar to a human reference than two other references to each other:

$$\text{QUEEN}_{X,R}(a) = P(\forall x \in X : x(a,r) \geq x(r',r''))$$

where a is the automatic translation being evaluated, $\langle r, r', r'' \rangle$ are three different human references in R, and $x(a,r)$ stands for the similarity of r to a.

In the case of Openlab data, we can count only on three human references per sentence. In order to increase the number of samples for QUEEN estimation we can use reference similarities $x(r',r'')$ between manual translation pairs from other sentences, assuming that the distances between manual references are relatively stable across examples.

3.2 Similarity Metrics

We begin by defining a set of 22 similarity metrics taken from the list of standard evaluation metrics in Subsection 2.1. Evaluation metrics can be tuned into similarity metrics simply by considering only one reference when computing its value.

Secondly, we explore the possibility of designing complementary similarity metrics that exploit linguistic information at levels further than lexical. Inspired in the work by Liu and Gildea (2005), who introduced a series of metrics based on constituent/dependency syntactic matching, we have designed three subgroups of syntactic similarity metrics. To compute them, we have used the dependency trees provided by the MINIPAR dependency parser (Lin, 1998). These metrics compute the level of word overlapping (unigram precision/recall) between dependency trees associated to automatic and reference translations, from three different points of view:

TREE-X overlapping between the words hanging from non-terminal nodes of type X of the tree. For instance, the metric TREE_PRED reflects the proportion of word overlapping between subtrees of type 'pred' (predicate of a clause).

GRAM-X overlapping between the words with the grammatical category X. For instance, the metric GRAM_A reflects the proportion of word overlapping between terminal nodes of type 'A' (Adjective/Adverbs).

LEVEL-X overlapping between the words hanging at a certain level X of the tree, or deeper. For instance, LEVEL-1 would consider overlapping between all the words in the sentences.

In addition, we also consider three coarser metrics, namely TREE, GRAM and LEVEL, which correspond to the average value of the finer metrics corresponding to each subfamily.

3.3 Metric Set Selection

We can compute KING over combinations of metrics by directly replacing the similarity metric $x(a,r)$ with the QUEEN measure. This corresponds exactly to the KING measure used in QARLA:

$$\text{KING}_{A,R}(X) = P(r \in R, \forall a \in A : \text{QUEEN}_{X,R-\{r\}}(r) \geq \text{QUEEN}_{X,R-\{r\}}(a))$$

KING represents the probability that, for a given set of human references R, and a set of metrics X, the QUEEN quality of a human reference is greater than the QUEEN quality of any automatic translation in A.

The similarity metrics based on standard evaluation measures together with the two new families of similarity metrics form a set of 104 metrics. Our goal is to obtain the subset of metrics with highest descriptive power; for this, we rely on the KING probability. A brute force exploration of all possible metric combinations is not viable. In order to

perform an approximate search for a local maximum in KING over all the possible metric combinations defined by X, we have used the following greedy heuristic:

1. Individual metrics are ranked by their KING value.

2. In decreasing rank order, metrics are individually added to the set of optimal metrics if, and only if, the global KING is increased.

After applying the algorithm we have obtained the optimal metric set:

{GTM-1, NIST-2, GRAM_A, GRAM_N, GRAM_AUX, GRAM_BE, TREE, TREE_AUX, TREE_PNMOD, TREE_PRED, TREE_REL, TREE_S and TREE_WHN}

which has a KING value of 0.29. This is significantly higher than the maximum KING obtained by any individual standard metric (which was 0.19 for GTM-3).

As to the probability ORANGE that a reference translation attains a higher score than an automatic translation, this metric set obtains a value of 0.49 vs. 0.42. This means that still the metrics are, on average, unable to discriminate between human references and automatic translations. However, the proportion of sentences for which the metrics are able to discriminate (KING value) is significantly higher.

The metric set with highest descriptive power contains metrics at different linguistic levels. For instance, GTM-1 and NIST-2 reward n-gram matches at the lexical level. GRAM_A, GRAM_N, GRAM_AUX and GRAM_BE capture word overlapping for nouns, auxiliary verbs, adjectives and adverbs, and auxiliary uses of the verb 'to be', respectively. TREE, TREE_AUX, TREE_PNMOD, TREE_PRED, TREE_REL, TREE_S and TREE_WHN reward lexical overlapping over different types of dependency subtrees: surface subjects, relative clauses, predicates, auxiliary verbs, postnominal modifiers, and whn-elements at C-spec positions, respectively.

These results are a clear indication that features from several linguistic levels are useful for the characterization of human translations.

4 Human-like vs. Human Acceptable

In this section we analyze the relationship between the two different kinds of MT evaluation presented: (i) the ability of MT systems to generate human-like translations, and (ii) the ability of MT systems to generate translations that look acceptable to human judges.

4.1 Experimental Setting

The ideal test set to study this dichotomy inside the QARLA Framework would consist of a large number of human references per sentence, and automatic outputs generated by heterogeneous MT systems.

4.2 Descriptive Power vs. Correlation with Human Judgements

We use the data and results from the IWSLT04 Evaluation Campaign[8]. We focus on the evaluation of the Chinese-to-English (CE) translation task, in which a set of 500 short sentences from the Basic Travel Expressions Corpus (BTEC) were translated (Akiba et al., 2004). For purposes of automatic evaluation, 16 reference translations and outputs by 20 different MT systems are available for each sentence. Moreover, each of these outputs was evaluated by three judges on the basis of adequacy and fluency (LDC, 2002). In our experiments we consider the sum of adequacy and fluency assessments.

However, the BTEC corpus has a serious drawback: sentences are very short (8 word length in average). In order to consider a sentence adequate we are practically forcing it to match exactly some of the human references. To alleviate this effect we selected sentences consisting of at least ten words. A total of 94 sentences (of 13 words length in average) satisfied this constraint.

Figure 3 shows, for all metrics, the relationship between the power of characterization of human references (KING, horizontal axis) and the correlation with human judgements (Pearson correlation, vertical axis). Data are plotted in three different groups: original standard metrics, single metrics inside QARLA (QUEEN measure), and the optimal metric combination according to KING. The optimal set is:

{GRAM_N, LEVEL_2, LEVEL_4, NIST-1, NIST-3, NIST-4, and 1-WER}

This set suggests that all kinds of n-grams play an important role in the characterization of human

[8] http://www.slt.atr.co.jp/IWSLT2004/

translations. The metric GRAM_N reflects the importance of noun translations. Unlike the Openlab corpus, levels of the dependency tree (LEVEL_2 and LEVEL_4) are descriptive features, but dependency relations are not (TREE metrics). This is probably due to the small average sentence length in IWSLT.

Metrics exhibiting a high level of correlation outside QARLA, such as NIST-3, also exhibit a high descriptive power (KING). There is also a tendency for metrics with a KING value around 0.6 to concentrate at a level of Pearson correlation around 0.5.

But the main point is the fact that the QUEEN measure obtained by the metric combination with highest KING does not yield the highest level of correlation with human assessments, which is obtained by standard metrics outside QARLA (0.5 vs. 0.7).

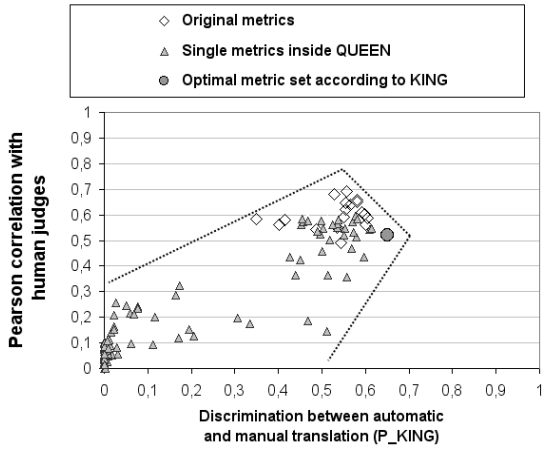

Figure 3: Human characterization vs. correlation with human judgements for IWSLT'04 CE translation task.

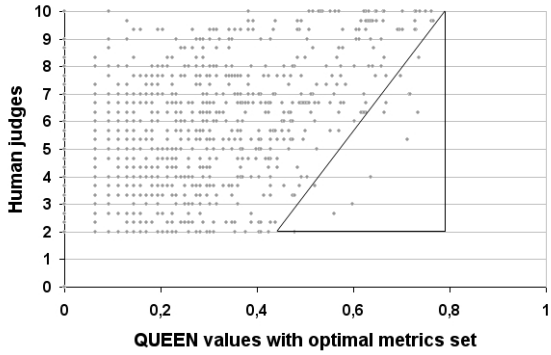

Figure 4: QUEEN values vs. human judgements for IWSLT'04 CE translation task.

4.3 Human Judgements vs. Similarity to References

In order to explain the above results, we have analyzed the relationship between human assessments and the QUEEN values obtained by the best combination of metrics for every individual translation.

Figure 4 shows that high values of QUEEN (i.e., similarity to references) imply high values of human judgements. But the reverse is not true. There are translations acceptable to a human judge but not similar to human translations according to QUEEN. This fact can be understood by inspecting a few particular cases. Table 1 shows two cases of translations exhibiting a very low QUEEN value and very high human judgment score. The two cases present the same kind of problem: there exists some word or phrase absent from all human references. In the first example, the automatic translation uses the expression "seats" to make a reservation, where humans invariably choose "table". In the second example, the automatic translation users "rack" as the place to put a bag, while humans choose "overhead bin", "overhead compartment", but never "rack".

Therefore, the QUEEN measure discriminates these automatic translations regarding to all human references, thus assigning them a low value. However, human judges find the translation still acceptable and informative, although not strictly human-like.

These results suggest that inside the set of human acceptable translations, which includes human-like translations, there is also a subset of translations unlikely to have been produced by a human translator. This is a drawback of MT evaluation based on human references when the evaluation criteria is Human Acceptability. The good news are that when Human Likeness increases, Human Acceptability increases as well.

5 Conclusions

We have analyzed the ability of current MT evaluation metrics to characterize human translations (as opposed to automatic translations), and the relationship between MT evaluation based on Human Acceptability and Human Likeness.

The first conclusion is that, over a limited number of references, standard metrics are unable to identify the features that characterize human translations. Instead, systems behave as centroids with

respect to human references. This is due, among other reasons, to the combined effect of the shallowness of current MT evaluation metrics (mostly lexical), and the fact that the choice of lexical items is mostly based on statistical methods. We suggest two complementary ways of solving this problem. First, we introduce a new family of syntax-based metrics covering partial aspects of MT quality. Second, we use the QARLA Framework to combine multiple metrics into a single measure of quality. In the future we will study the design of new metrics working at different linguistic levels. For instance, we are currently developing a new family of metrics based on shallow parsing (i.e., part-of-speech, lemma, and chunk information).

Second, our results suggest that there exists a clear relation between the two kinds of MT evaluation described. While Human Likeness is a sufficient condition to get Human Acceptability, Human Acceptability does not guarantee Human Likeness. Human judges may consider acceptable automatic translations that would never be generated by a human translator.

Considering these results, we claim that improving metrics according to their descriptive power (Human Likeness) is more reliable than improving metrics based on correlation with human judges. First, because this correlation is not granted, since automatic metrics are based on similarity to models. Second, because high Human Likeness ensures high scores from human judges.

References

Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Jun'ichi Tsujii. 2004. Overview of the IWSLT04 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–12, Kyoto, Japan.

Enrique Amigó, Julio Gonzalo, Anselmo Peñas, and Felisa Verdejo. 2005. QARLA: a Framework for the Evaluation of Automatic Sumarization. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics*, Michigan, June. Association for Computational Linguistics.

J.M. Crego, Costa jussà M.R., J.B. Mariño, and Fonollosa J.A.R. 2005. Ngram-based versus Phrase-based Statistical Machine Translation. In *Proceedings of the International Workshop on Spoken Language Technology (IWSLT'05)*.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd Internation Conference on Human Language Technology*, pages 138–145.

Matthias Eck and Chiori Hori. 2005. Overview of the IWSLT 2005 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, Carnegie Mellon University, Pittsburgh, PA.

Jesús Giménez and Enrique Amigó. 2006. IQMT: A Framework for Automatic Machine Translation Evaluation. In *Proceedings of the 5th LREC*.

Jesús Giménez and Lluís Màrquez. 2005. Combining Linguistic Data Views for Phrase-based SMT. In *Proceedings of the Workshop on Building and Using Parallel Texts, ACL*.

LDC. 2002. Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Chinese-English Translations Revision 1.0. Technical report, Linguistic Data Consortium. http://-www.ldc.upenn.edu/Projects/TIDES/Translation/-TransAssess02.pdf.

G. Leusch, N. Ueffing, and H. Ney. 2003. A Novel String-to-String Distance Measure with Applications to Machine Translation Evaluation. In *Proceedings of MT Summit IX*.

Chin-Yew Lin and Franz Josef Och. 2004a. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. In *Proceedings of ACL*.

Chin-Yew Lin and Franz Josef Och. 2004b. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of COLING*.

Dekang Lin. 1998. Dependency-based Evaluation of MINIPAR. In *Proceedings of the Workshop on the Evaluation of Parsing Systems*.

Ding Liu and Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and Recall of Machine Translation. In *Proceedings of HLT/NAACL*.

S. Nießen, F.J. Och, G. Leusch, and H. Ney. 2000. Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation, IBM Research Report, RC22176. Technical report, IBM T.J. Watson Research Center.

Automatic Translation:	my name is endo i ’ve reserved **seats** for nine o’clock
Human Reference 1:	this is endo i booked a **table** at nine o’clock
2:	i reserved a **table** for nine o’clock and my name is endo
3:	my name is endo and i made a reservation for a **table** at nine o’clock
4:	i am endo and i have a reservation for a **table** at nine pm
5:	my name is endo and i booked a **table** at nine o’clock
6:	this is endo i reserved a **table** for nine o’clock
7:	my name is endo and i reserved a **table** with you for nine o’clock
8:	i ’ve booked a **table** under endo for nine o’clock
9:	my name is endo and i have a **table** reserved for nine o’clock
10:	i ’m endo and i have a reservation for a **table** at nine o’clock
11:	my name is endo and i reserved a **table** for nine o’clock
12:	the name is endo and i have a reservation for nine
13:	i have a **table** reserved for nine under the name of endo
14:	hello my name is endo i reserved a **table** for nine o’clock
15:	my name is endo and i have a **table** reserved for nine o’clock
16:	my name is endo and i made a reservation for nine o’clock

Automatic Translation:	could you help me put my bag on the **rack** please
Human Reference 1:	could you help me put my bag in the **overhead bin**
2:	can you help me to get my bag into the **overhead bin**
3:	would you give me a hand with getting my bag into the **overhead bin**
4:	would you mind assisting me to put my bag into the **overhead bin**
5:	could you give me a hand putting my bag in the **overhead compartment**
6:	please help me put my bag in the **overhead bin**
7:	would you mind helping me put my bag in the **overhead compartment**
8:	do you mind helping me put my bag in the **overhead compartment**
9:	could i get a hand with putting my bag in the **overhead compartment**
10:	could i ask you to help me put my bag in the **overhead compartment**
11:	please help me put my bag in the **overhead bin**
12:	would you mind helping me put my bag in the **overhead compartment**
13:	i ’d like you to help me put my bag in the **overhead compartment**
14:	would you mind helping get my bag up into the **overhead storage compartment**
15:	may i get some assistance getting my bag into the **overhead storage compartment**
16:	please help me put my into the **overhead storage compartment**

Table 1: Automatic translations with high score in human judgements and low QUEEN value.