METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments

Satanjeev Banerjee
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
banerjee+@cs.cmu.edu

Alon Lavie
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
alavie@cs.cmu.edu

Abstract

We describe METEOR, an automatic metric for machine translation evaluation that is based on a generalized concept of unigram matching between the machine-produced translation and human-produced reference translations. Unigrams can be matched based on their surface forms, stemmed forms, and meanings; furthermore, METEOR can be easily extended to include more advanced matching strategies. Once all generalized unigram matches between the two strings have been found, METEOR computes a score for this matching using a combination of unigram-precision, unigram-recall, and a measure of fragmentation that is designed to directly capture how well-ordered the matched words in the machine translation are in relation to the reference. We evaluate METEOR by measuring the correlation between the metric scores and human judgments of translation quality. We compute the Pearson R correlation value between its scores and human quality assessments of the LDC TIDES 2003 Arabic-to-English and Chinese-to-English datasets. We perform segment-by-segment correlation, and show that METEOR gets an R correlation value of 0.347 on the Arabic data and 0.331 on the Chinese data. This is shown to be an improvement on using simply unigram-precision, unigram-recall and their harmonic F1 combination. We also perform experiments to show the relative contributions of the various mapping modules.

proposed a combination of f1 and recall instead of the traditional single approaches

1 Introduction

Automatic Metrics for machine translation (MT) evaluation have been receiving significant attention in the past two years, since IBM's BLEU metric was proposed and made available (Papineni et al 2002). BLEU and the closely related NIST metric (Doddington, 2002) have been extensively used for comparative evaluation of the various MT systems developed under the DARPA TIDES research program, as well as by other MT researchers. The utility and attractiveness of automatic metrics for MT evaluation has consequently been widely recognized by the MT community. Evaluating an MT system using such automatic metrics is much faster, easier and cheaper compared to human evaluations, which require trained bilingual evaluators. In addition to their utility for comparing the performance of different systems on a common translation task, automatic metrics can be applied on a frequent and ongoing basis during system development, in order to guide the development of the system based on concrete performance improvements.

Evaluation of Machine Translation has traditionally been performed by humans. While the main criteria that should be taken into account in assessing the quality of MT output are fairly intuitive and well established, the overall task of MT evaluation is both complex and task dependent. MT evaluation has consequently been an area of significant research in itself over the years. A wide range of assessment measures have been proposed, not all of which are easily quantifiable. Recently developed frameworks, such as FEMTI (King et al, 2003), are attempting to devise effective platforms for combining multi-faceted measures for MT evaluation in effective and user-adjustable ways. While a single one-dimensional numeric metric cannot hope to fully capture all aspects of MT

evaluation, such metrics are still of great value and utility.

In order to be both effective and useful, an automatic metric for MT evaluation has to satisfy several basic criteria. The primary and most intuitive requirement is that the metric have very high correlation with quantified human notions of MT quality. Furthermore, a good metric should be as sensitive as possible to differences in MT quality between different systems, and between different versions of the same system. The metric should be consistent (same MT system on similar texts should produce similar scores), reliable (MT systems that score similarly can be trusted to perform similarly) and general (applicable to different MT tasks in a wide range of domains and scenarios). Needless to say, satisfying all of the above criteria is extremely difficult, and all of the metrics that have been proposed so far fall short of adequately addressing most if not all of these requirements. Nevertheless, when appropriately quantified and converted into concrete test measures, such requirements can set an overall standard by which different MT evaluation metrics can be compared and evaluated.

characteristics of a good MT evaluator

In this paper, we describe METEOR[1], an automatic metric for MT evaluation which we have been developing. METEOR was designed to explicitly address several observed weaknesses in IBM's BLEU metric. It is based on an explicit word-to-word matching between the MT output being evaluated and one or more reference translations. Our current matching supports not only matching between words that are identical in the two strings being compared, but can also match words that are simple morphological variants of each other (i.e. they have an identical stem), and words that are synonyms of each other. We envision ways in which this strict matching can be further expanded in the future, and describe these at the end of the paper. Each possible matching is scored based on a combination of several features. These currently include unigram-precision, unigram-recall, and a direct measure of how out-of-order the words of the MT output are with respect to the reference. The score assigned to each individual sentence of MT output is derived from the best scoring match among all matches over all reference translations. The maximal-scoring matching is then also used in order to calculate an aggregate score for the MT system over the entire test set. Section 2 describes the metric in detail, and provides a full example of the matching and scoring.

2 good considerations over strict word matching

In previous work (Lavie et al., 2004), we compared METEOR with IBM's BLEU metric and it's derived NIST metric, using several empirical evaluation methods that have been proposed in the recent literature as concrete means to assess the level of correlation of automatic metrics and human judgments. We demonstrated that METEOR has significantly improved correlation with human judgments. Furthermore, our results demonstrated that recall plays a more important role than precision in obtaining high-levels of correlation with human judgments. The previous analysis focused on correlation with human judgments at the *system level*. In this paper, we focus our attention on improving correlation between METEOR score and human judgments at the *segment level*. High-levels of correlation at the segment level are important because they are likely to yield a metric that is sensitive to minor differences between systems and to minor differences between different versions of the same system. Furthermore, current levels of correlation at the sentence level are still rather low, offering a very significant space for improvement. The results reported in this paper demonstrate that all of the individual components included within METEOR contribute to improved correlation with human judgments. In particular, METEOR is shown to have statistically significant better correlation compared to unigram-precision, unigram-recall and the harmonic F1 combination of the two.

We are currently in the process of exploring several further enhancements to the current METEOR metric, which we believe have the potential to significantly further improve the sensitivity of the metric and its level of correlation with human judgments. Our work on these directions is described in further detail in Section 4.

2 The METEOR Metric

2.1 Weaknesses in BLEU Addressed in METEOR

The main principle behind IBM's BLEU metric (Papineni et al, 2002) is the measurement of the

[1] METEOR: **M**etric for **E**valuation of **T**ranslation with **E**xplicit **OR**dering

overlap in unigrams (single words) and higher order n-grams of words, between a translation being evaluated and a set of one or more reference translations. The main component of BLEU is *n-gram precision*: the proportion of the matched n-grams out of the total number of n-grams in *the evaluated translation*. Precision is calculated separately for each n-gram order, and the precisions are combined via a geometric averaging. BLEU does not take *recall* into account directly. Recall – the proportion of the matched n-grams out of the total number of n-grams in *the reference translation*, is extremely important for assessing the quality of MT output, as it reflects to what degree the translation covers the entire content of the translated sentence. BLEU does not use recall because the notion of recall is unclear when matching simultaneously against a set of reference translations (rather than a single reference). To compensate for recall, BLEU uses a Brevity Penalty, which penalizes translations for being "too short". The NIST metric is conceptually similar to BLEU in most aspects, including the weaknesses discussed below.

why recall is important

BLEU and NIST suffer from several weaknesses, which we attempt to address explicitly in our proposed METEOR metric:

The Lack of Recall: We believe that the fixed brevity penalty in BLEU does not adequately compensate for the lack of recall. Our experimental results strongly support this claim.

Use of Higher Order N-grams: Higher order N-grams are used in BLEU as an indirect measure of a translation's level of grammatical well-formedness. We believe an explicit measure for the level of grammaticality (or word order) can better account for the importance of grammaticality as a factor in the MT metric, and result in better correlation with human judgments of translation quality.

word freq is not a good way to determine grammatical correctness. translated sent can map in word freq burt still not mean anything.

Lack of Explicit Word-matching Between Translation and Reference: N-gram counts don't require an explicit word-to-word matching, but this can result in counting incorrect "matches", particularly for common function words.

Use of Geometric Averaging of N-grams: Geometric averaging results in a score of "zero" whenever one of the component n-gram scores is zero. Consequently, BLEU scores at the sentence (or segment) level can be meaningless. Although BLEU was intended to be used only for aggregate counts over an entire test-set (and not at the sentence level), scores at the sentence level can be useful indicators of the quality of the metric. In experiments we conducted, a modified version of BLEU that uses equal-weight arithmetic averaging of n-gram scores was found to have better correlation with human judgments.

2.2 The METEOR Metric

METEOR was designed to explicitly address the weaknesses in BLEU identified above. It evaluates a translation by computing a score based on explicit word-to-word matches between the translation and a reference translation. If more than one reference translation is available, the given translation is scored against each reference independently, and the best score is reported. This is discussed in more detail later in this section.

Given a pair of translations to be compared (a system translation and a reference translation), METEOR creates an *alignment* between the two strings. We define an alignment as a mapping between unigrams, such that every unigram in each string maps to zero or one unigram in the other string, and to no unigrams in the same string. Thus in a given alignment, a single unigram in one string cannot map to more than one unigram in the other string. This alignment is incrementally produced through a series of *stages,* each stage consisting of two distinct phases.

In the first phase an external module lists all the possible unigram mappings between the two strings. Thus, for example, if the word "computer" occurs once in the system translation and twice in the reference translation, the external module lists two possible unigram mappings, one mapping the occurrence of "computer" in the system translation to the first occurrence of "computer" in the reference translation, and another mapping it to the second occurrence. Different modules map unigrams based on different criteria. The "exact" module maps two unigrams if they are exactly the same (e.g. "computers" maps to "computers" but not "computer"). The "porter stem" module maps two unigrams if they are the same *after* they are stemmed using the Porter stemmer (e.g.: "computers" maps to both "computers" and to "computer"). The "WN synonymy" module maps two unigrams if they are synonyms of each other.

mapping is still not exclusive when it comes to multi-lingual

In the second phase of each stage, the largest subset of these unigram mappings is selected such

that the resulting set constitutes an *alignment* as defined above (that is, each unigram must map to at most one unigram in the other string). If more than one subset constitutes an alignment, and also has the same cardinality as the largest set, METEOR selects that set that has the least number of unigram mapping *crosses*. Intuitively, if the two strings are typed out on two rows one above the other, and lines are drawn connecting unigrams that are mapped to each other, each line crossing is counted as a "unigram mapping cross". Formally, two unigram mappings (t_i, r_j) and (t_k, r_l) (where t_i and t_k are unigrams in the system translation mapped to unigrams r_j and r_l in the reference translation respectively) are said to *cross* if and only if the following formula evaluates to a negative number:

$$(pos(t_i) - pos(t_k)) * (pos(r_j) - pos(r_l))$$

where $pos(t_x)$ is the numeric position of the unigram t_x in the system translation string, and $pos(r_y)$ is the numeric position of the unigram r_y in the reference string. For a given alignment, every pair of unigram mappings is evaluated as a cross or not, and the alignment with the least total crosses is selected in this second phase. Note that these two phases together constitute a variation of the algorithm presented in (Turian et al, 2003).

Each stage only maps unigrams that have not been mapped to any unigram in any of the preceding stages. Thus the order in which the stages are run imposes different priorities on the mapping modules employed by the different stages. That is, if the first stage employs the "exact" mapping module and the second stage employs the "porter stem" module, METEOR is effectively preferring to first map two unigrams based on their surface forms, and performing the stemming only if the surface forms do not match (or if the mapping based on surface forms was too "costly" in terms of the total number of crosses). Note that METEOR is flexible in terms of the number of stages, the actual external mapping module used for each stage, and the order in which the stages are run. By default the first stage uses the "exact" mapping module, the second the "porter stem" module and the third the "WN synonymy" module. In section 4 we evaluate each of these configurations of METEOR.

Once all the stages have been run and a final alignment has been produced between the system translation and the reference translation, the METEOR score for this pair of translations is computed as follows. First unigram precision (P) is computed as the ratio of the number of unigrams in the system translation that are mapped (to unigrams in the reference translation) to the total number of unigrams in the system translation. Similarly, unigram recall (R) is computed as the ratio of the number of unigrams in the system translation that are mapped (to unigrams in the reference translation) to the total number of unigrams in the reference translation. Next we compute *Fmean* by combining the precision and recall via a harmonic-mean (van Rijsbergen, 1979) that places most of the weight on recall. We use a harmonic mean of P and 9R. The resulting formula used is:

$$Fmean = \frac{10PR}{R + 9P}$$

Precision, recall and Fmean are based on unigram matches. To take into account longer matches, METEOR computes a *penalty* for a given alignment as follows. First, all the unigrams in the system translation that are mapped to unigrams in the reference translation are grouped into the fewest possible number of *chunks* such that the unigrams in each chunk are in adjacent positions in the system translation, and are also mapped to unigrams that are in adjacent positions in the reference translation. Thus, the longer the n-grams, the fewer the chunks, and in the extreme case where the entire system translation string matches the reference translation there is only one chunk. In the other extreme, if there are no bigram or longer matches, there are as many chunks as there are unigram matches. The penalty is then computed through the following formula:

$$Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams\_matched} \right)^3$$

For example, if the system translation was "the president spoke to the audience" and the reference translation was "the president then spoke to the audience", there are two chunks: "the president" and "spoke to the audience". Observe that the penalty increases as the number of chunks increases to a maximum of *0.5*. As the number of chunks goes to 1, penalty decreases, and its lower bound is decided by the number of unigrams matched. The parameters if this penalty function were determined based on some experimentation with de-

good technique but what abt languages like

ENG: SVO
Hindi: SOV

??
Wont there be more cross-overs??

veopment data, but have not yet been trained to be optimal.

Finally, the METEOR *Score* for the given alignment is computed as follows:

$$Score = Fmean * (1 - Penalty)$$

This has the effect of *reducing* the Fmean by the maximum of 50% if there are *no* bigram or longer matches.

For a single system translation, METEOR computes the above score for each reference translation, and then reports the best score as the score for the translation. The overall METEOR score for a system is calculated based on aggregate statistics accumulated over the entire test set, similarly to the way this is done in BLEU. We calculate aggregate precision, aggregate recall, an aggregate penalty, and then combine them using the same formula used for scoring individual segments.

3 Evaluation of the METEOR Metric

3.1. Data

We evaluated the METEOR metric and compared its performance with BLEU and NIST on the DARPA/TIDES 2003 Arabic-to-English and Chinese-to-English MT evaluation data released through the LDC as a part of the workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, at the Annual Meeting of the Association of Computational Linguistics (2005). The Chinese data set consists of 920 sentences, while the Arabic data set consists of 664 sentences. Each sentence has four reference translations. Furthermore, for 7 systems on the Chinese data and 6 on the Arabic data, every sentence translation has been assessed by two separate human judges and assigned an *Adequacy* and a *Fluency Score*. Each such score ranges from one to five (with one being the poorest grade and five the highest). For this paper, we computed a *Combined Score* for each translation by averaging the adequacy and fluency scores of the two judges for that translation. We also computed an average *System Score* for each translation system by averaging the Combined Score for all the translations produced by that system. (Note that although we refer to these data sets as the "Chinese" and the "Arabic" data sets, the MT evaluation systems analyzed in this paper only evaluate English sentences produced by translation systems by comparing them to English reference sentences).

3.2 Comparison with BLEU and NIST MT Evaluation Algorithms

In this paper, we are interested in evaluating METEOR as a metric that can evaluate translations on a sentence-by-sentence basis, rather than on a coarse grained system-by-system basis. The standard metrics – BLEU and NIST – were however designed for system level scoring, hence computing sentence level scores using BLEU or the NIST evaluation mechanism is unfair to those algorithms. To provide a point of comparison however, table 1 shows the *system level* correlation between human judgments and various MT evaluation algorithms and sub components of METEOR over the Chinese portion of the Tides 2003 dataset. Specifically, these correlation figures were obtained as follows: Using each algorithm we computed one score per Chinese system by calculating the aggregate scores produced by that algorithm for that system. We also obtained the overall human judgment for each system by averaging all the human scores for that system's translations. We then computed the Pearson correlation between these system level human judgments and the system level scores for each algorithm; these numbers are presented in table 1.

System ID	Correlation
BLEU	0.817
NIST	0.892
Precision	0.752
Recall	0.941
F1	0.948
Fmean	0.952
METEOR	0.964

Table 1: Comparison of human/METEOR correlation with BLEU and NIST/human correlations

Observe that simply using Recall as the MT evaluation metric results in a significant improvement in correlation with human judgment over both the BLEU and the NIST algorithms. These correlations further improve slightly when precision is taken into account (in the F1 measure),

when the recall is weighed more heavily than precision (in the Fmean measure) and when a penalty is levied for fragmented matches (in the main METEOR measure).

3.3 Evaluation Methodology

As mentioned in the previous section, our main goal in this paper is to evaluate METEOR and its components on their translation-by-translation level correlation with human judgment. Towards this end, in the rest of this paper, our evaluation methodology is as follows: For each system, we compute the METEOR Score for every translation produced by the system, and then compute the correlation between these *individual* scores and the human assessments (average of the adequacy and fluency scores) for the same translations. Thus we get a single *Pearson R value* for each system for which we have human assessments. Finally we average the *R* values of all the systems for each of the two language data sets to arrive at the overall average correlation for the Chinese dataset and the Arabic dataset. This number ranges between -1.0 (completely negatively correlated) to +1.0 (completely positively correlated).

We compare the correlation between human assessments and METEOR Scores produced above with that between human assessments and precision, recall and Fmean scores to show the advantage of the various components in the METEOR scoring function. Finally we run METEOR using different mapping modules, and compute the correlation as described above for each configuration to show the effect of each unigram mapping mechanism.

3.4 Correlation between METEOR Scores and Human Assessments

System ID	Correlation
ame	0.331
ara	0.278
arb	0.399
ari	0.363
arm	0.341
arp	0.371
Average	**0.347**

Table 2: Correlation between METEOR Scores and Human Assessments for the Arabic Dataset

We computed sentence by sentence correlation between METEOR Scores and human assessments (average of adequacy and fluency scores) for each translation for every system. Tables 2 and 3 show the Pearson R correlation values for each system, as well as the average correlation value per language dataset.

System ID	Correlation
E09	0.385
E11	0.299
E12	0.278
E14	0.307
E15	0.306
E17	0.385
E22	0.355
Average	**0.331**

Table 3: Correlation between METEOR Scores and Human Assessments for the Chinese Dataset

3.5 Comparison with Other Metrics

We computed translation by translation correlations between human assessments and other metrics besides the METEOR score, namely precision, recall and Fmean. Tables 4 and 5 show the correlations for the various scores.

Metric	Correlation
Precision	0.287
Recall	0.334
Fmean	0.340
METEOR	**0.347**

Table 4: Correlations between human assessments and precision, recall, Fmean and METEOR Scores, averaged over systems in the Arabic dataset

Metric	Correlation
Precision	0.286
Recall	0.320
Fmean	0.327
METEOR	**0.331**

Table 5: Correlations between human assessments and precision, recall, Fmean and METEOR Scores, averaged over systems in the Chinese dataset

We observe that recall by itself correlates with human assessment much better than precision, and that combining the two using the Fmean formula

described above results in further improvement. By penalizing the Fmean score using the chunk count we get some further marginal improvement in correlation.

3.6 Comparison between Different Mapping Modules

To observe the effect of various unigram mapping modules on the correlation between the METEOR score and human assessments, we ran METEOR with different sequences of stages with different mapping modules in them. In the first experiment we ran METEOR with only one stage that used the "exact" mapping module. This module matches unigrams only if their surface forms match. (This module does not match unigrams that belong to a list of "stop words" that consist mainly of function words). In the second experiment we ran METEOR with two stages, the first using the "exact" mapping module, and the second the "Porter" mapping module. The Porter mapping module matches two unigrams to each other if they are identical after being passed through the Porter stemmer. In the third experiment we replaced the Porter mapping module with the WN-Stem mapping module. This module maps two unigrams to each other if they share the same *base form* in WordNet. This can be thought of as a different kind of stemmer – the difference from the Porter stemmer is that the word stems are actual words when stemmed through WordNet in this manner. In the last experiment we ran METEOR with three stages, the first two using the exact and the Porter modules, and the third the WN-Synonymy mapping module. This module maps two unigrams together if at least one sense of each word belongs to the same synset in WordNet. Intuitively, this implies that at least one sense of each of the two words represent the same concept. This can be thought of as a poor-man's synonymy detection algorithm that does not disambiguate the words being tested for synonymy. Note that the METEOR scores used to compute correlations in the other tables (1 through 4) used exactly this sequence of stages.

Tables 6 and 7 show the correlations between METEOR scores produced in each of these experiments and human assessments for both the Arabic and the Chinese datasets. On both data sets, adding either stemming modules to simply using the exact matching improves correlations. Some further improvement in correlation is produced by adding the synonymy module.

Mapping module sequence used (Arabic)	Correlation
Exact	0.312
Exact, Porter	0.329
Exact, WN-Stem	0.330
Exact, Porter, WN-Synonym	0.347

Table 6: Comparing correlations produced by different module stages on the Arabic dataset.

Mapping module sequence used (Chinese)	Correlation
Exact	0.293
Exact, Porter	0.318
Exact, WN-Stem	0.312
Exact, Porter, WN-Synonym	0.331

Table 7: Comparing correlations produced by different module stages, on the Chinese dataset

3.7 Correlation using Normalized Human Assessment Scores

One problem with conducting correlation experiments with human assessment scores at the sentence level is that the human scores are noisy – that is, the levels of agreement between human judges on the actual sentence level assessment scores is not extremely high. To partially address this issue, the human assessment scores were normalized by a group at the MITRE Corporation. To see the effect of this noise on the correlation, we computed the correlation between the METEOR Score (computed using the stages used in the 4th experiment in section 7 above) and both the raw human assessments as well as the normalized human assessments.

	Arabic Dataset	Chinese Dataset
Raw human assessments	0.347	0.331
Normalized human assessments	0.403	0.365

Table 8: Comparing correlations between METEOR Scores and both raw and normalized human assessments

Table 8 shows that indeed METEOR Scores correlate better with normalized human assessments. In other words, the noise in the human assessments hurts the correlations between automatic scores and human assessments.

4 Future Work

The METEOR metric we described and evaluated in this paper, while already demonstrating great promise, is still relatively simple and naïve. We are in the process of enhancing the metric and our experimentation in several directions:

Train the Penalty and Score Formulas on Data: The formulas for Penalty and METEOR score were manually crafted based on empirical tests on a separate set of development data. However, we plan to optimize the formulas by *training* them on a separate data set, and choosing that formula that best correlates with human assessments on the training data.

Use Semantic Relatedness to Map Unigrams: So far we have experimented with exact mapping, stemmed mapping and synonymy mapping between unigrams. Our next step is to experiment with different measures of semantic relatedness to match unigrams that have a related meaning, but are not quite synonyms of each other.

More Effective Use of Multiple Reference Translations: Our current metric uses multiple reference translations in a weak way: we compare the translation with each reference separately and select the reference with the best match. This was necessary in order to incorporate recall in our metric, which we have shown to be highly advantageous. As our matching approach improves, the need for multiple references for the metric may in fact diminish. Nevertheless, we are exploring ways in which to improve our matching against multiple references. Recent work by (Pang et al, 2003) provides the mechanism for producing semantically meaningful additional "synthetic" references from a small set of real references. We plan to explore whether using such synthetic references can improve the performance of our metric.

Weigh Matches Produced by Different Modules Differently: Our current multi-stage approach prefers metric imposes a priority on the different matching modules. However, once all the stages have been run, unigrams mapped through different mapping modules are treated the same. Another approach to treating different mappings differently is to apply different *weights* to the mappings produced by different mapping modules. Thus "computer" may match "computer" with a score of 1, "computers" with a score of 0.8 and "workstation" with a score of 0.3. As future work we plan to develop a version of METEOR that uses such weighting schemes.

Acknowledgements

We acknowledge Kenji Sagae and Shyamsundar Jayaraman for their work on the METEOR system. We also wish to thank John Henderson and William Morgan from MITRE for providing us with the normalized human judgment scores used for this work.

References

George Doddington. 2002. *Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics.* In Proceedings of 2nd Human Language Technologies Conference (HLT-02). San Diego, CA. pp. 128-132.

Margaret King, Andrei Popescu-Belis and Eduard Hovy. 2003. *FEMTI: Creating and Using a Framework for MT Evaluation.* In Proceedings of MT Summit IX, New Orleans, LA. Sept. 2003. pp. 224-231.

Alon Lavie, Kenji Sagae and Shyamsundar Jayaraman, 2004. *The Significance of Recall in Automatic Metrics for MT Evaluation*. In Proceedings of AMTA-2004, Washington DC. September 2004.

Bo Pang, Kevin Knight and Daniel Marcu. 2003. *Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences.* In Proceedings of HLT-NAACL 2003. Edmonton, Canada. May 2003.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation.* In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02). Philadelphia, PA. July 2002. pp. 311-318.

Joseph P. Turian, Luke Shen and I. Dan Melamed. 2003. *Evaluation of Machine Translation and its Evaluation.* In Proceedings of MT Summit IX, New Orleans, LA. Sept. 2003. pp. 386-393.

C. van Rijsbergen. 1979. *Information Retrieval.* Butterworths. London, England. 2nd Edition.