Comparative Evaluation of Term Informativeness Measures in Machine Translation Evaluation Metrics

Billy Wong and **Chunyu Kit**
Department of Chinese, Translation and Linguistics
City University of Hong Kong
83 Tat Chee Avenue, Kowloon, Hong Kong SAR, PR China
{tmwong, ctckit}@cityu.edu.hk

Abstract

Most metrics in use for automatic evaluation of machine translation in use equally weigh different matched words in candidate and reference translations, ignoring the fact that each word contributes a different amount of information to the meaning of a sentence. We experiment with ten measures of term informativeness for the purpose of examining their performance in rating the information loads of words in machine translated texts. The information theoretic measure *information gain* is found to bring about a nearly 12% improvement in correlation with human judgments of translation quality under an optimal setting. We also assess how various parameters may affect the performance of these measures, among which data size turns out to be the most influential factor. A dataset of around 80 documents, 700 segments, with 4 versions of reference translation, is found to offer the most desirable performance.

1 Introduction

The practicality and prevalence of automatic evaluation metrics in machine translation (MT) evaluation has led to a *de facto* standard way of assessing MT performance in the last decade. The quality of an MT output is characterized in terms of its similarity to corresponding professional human translations (HT), which is qualified according to the literal/linguistic features adopted by the evaluation metric in question.

The most widely used feature in all existing metrics is the matched words between MT and HT, which form a cornerstone of both the robustness and reliability of a metric in general. This also provides a common ground for other evaluation features to operate, like counting n-grams of various lengths and parsing syntactic structures. With other factors similar, the number of matched words is certainly a critical indicator for the quality of an MT output, in sense that the more words matched with HT, the better.

Bearing in mind its importance, we move on to looking into the significance of the informativeness of each match. We know that each word carries a different amount of information contributing to the meaning of a sentence. It is reasonable in MT evaluation that a higher weight is assigned to a more informative word. As given in Example 1 below, while both candidates contain seven words that match with *Ref*, the matches in *C1* carrying more information than those in *C2* should give *C1* a higher quality weighting than *C2*. However, most existing MT evaluation metrics assign an equal weight to every matched word and, consequently, exclude this possibility. What is needed here to remedy this shortcoming is a suitable measure to properly capture the weight for the information load of each word in MT outputs.

Example 1
C1: it was <u>not</u> <u>first</u> <u>time</u> that <u>prime</u> <u>minister</u> confronts <u>northern</u> <u>league</u> …
C2: <u>this</u> <u>is</u> <u>not</u> <u>the</u> primary <u>the</u> operation <u>has</u> <u>the</u> north …
Ref: this is not the first time the prime minister has faced the northern league …

Aimed at this purpose, this paper will compare a number of popular measures for term informativeness, so as to examine their positive effect on automatic MT evaluation. The measurement of term informativeness has been thoroughly studied and successfully applied to various NLP tasks such as information retrieval and automatic summarization. We will illustrate how these measures affect the performance of MT evaluation metrics in different settings of evaluation, aiming to identify the most suitable one among them.

2 Previous Works

Rating MT outputs in terms of their informativeness can be traced back as early as in the ALPAC report (ALPAC, 1966), before it was popularized through the DARPA series of MT evaluation (White et al., 1994). In general, it measures the amount of semantic information conveyed in an MT output that users can identify, so as to indirectly assess its quality and understandability as a translation. The informativeness is found to be highly correlated with the adequacy of MT outputs (White, 2003). Its measurement gives a valid indication of translation quality and is hence adopted as an important criterion for translation assessment.

In the studies of automatic evaluation metrics, however, there are only a few attempts that adopt term informativeness as an evaluation feature. As a variant of BLEU (Papineni et al., 2001), the NIST metric (Doddington, 2002) is proposed to include several variations of n-gram scoring, one of which introduces information weights to different n-gram counts in the way that n-grams of fewer occurrences in the reference translation set are weighted more heavily. Babych and Hartley (2004) extend BLEU with frequency weightings that use the standard *tf-idf* measure and their own S-score initially designed for information extraction. These works show that information-weighted metrics bring in observable improvement on correlation with human judgments, particularly in terms of adequacy ratings.

Wong and Kit (2010) integrate the *tf-idf* measure into the ATEC metric, following two ideas to fit this informativeness measure into the special context of automatic MT evaluation. Firstly, *tf-idf* scores are computed at the segment level, the level of the basic text unit in MT evaluation. This gets around the problem that an evaluation dataset consists of only one or a few long documents or of collections of isolated segments from different sources, and thus allows a sensitive measure for words in different segments. Secondly, the *tf-idf* measure is applied to both matched and unmatched words, to also account for the missing information in MT outputs. This is certainly more in line with human evaluation that lower scores are assigned to MT outputs with more missing informative words.

So far, however, only a limited number of informativeness measures have been ever used in the practice of MT evaluation, leaving many others untested. We will examine a selective set of them, which are presented in the next section.

3 Term Informativeness Measures

The term informativeness measures selected for our experiment are all frequency-based ones to estimate the relative informative load of a term in a text collection. The most fundamental one in this aspect is the occurrence frequency of a word in a text collection, i.e., *term frequency* (*tf*) or *collection frequency*. It was used as early as in Luhn’s work (1958) to locate the main topics in a text, based on the observation that writers tend to repeat certain words when referring to the same or related ideas. It is a nice indicator of word significance when high frequency words, which are mostly function words, are filtered out by a stoplist.

Another measure, *inverse document frequency* (*idf*), first defined in Spärck Jones (1972), concerns the specificity of a term according to its distribution over documents. Its underlying assumption is that the specificity of a term is inversely related to its probability of occurring in a particular document, meaning that the fewer documents containing a term, the more informative it is, and vice versa. It is formulated as

$$idf(i) = \log \frac{df(i)}{D}$$

where $df(i)$ is the number of documents containing word i and D the total number of documents in the text collection in question.

As noted in Church and Gale (1995a), *tf* and *idf* of a word are highly correlated to each other in general but also fundamentally different in that the former puts aside the density of distribution over documents. An observation is that we have $tf(i) \gg df(i)$ for many words whose multiple occurrences “burst” within a small number of documents. This is typically true for content-bearing words, e.g., “boycott”. On the other hand, we have $tf(i) \approx df(i)$ for many words that are evenly distributed, e.g., “somewhat”. They tend to occur in almost every document and hence are less informative. Such a correlation between the term frequency and document frequency of a word in a text collection can be captured by the measure *burstiness* (*bur*) (Church and Gale, 1995b) or *term clustering*, defined as

$$bur(i) = \frac{tf(i)}{df(i)}$$

In our experiment, we use its reverse version $1/bur(i)$ as a goodness measure.

Church and Gale (1995a) elaborate the notion of uneven distribution of words in documents with another measure *variance* (*var*). It compares the actual number of occurrences of a word with its expected frequency in a document j that is estimated by Poisson distribution, assuming that the occurrences of a word in different documents follow a statistical distribution pattern. As used in Kireyev (2009), the mean expected word frequency rate is straightforwardly estimated as

$$e(i) = \frac{tf(i)}{D}$$

and correspondingly,

$$var(i) = \frac{1}{D-1}\sum_{j=1}^{D}(tf(i,j) - e(i))^2$$

It reflects that a larger variance indicates a greater deviation from the expected frequency of occurrence in a document, meaning that the word in question is more salient in terms of its information load.

Church and Gale (1995a) also introduce another measure, namely *residual idf* (*ridf*), to quantify the notion of deviation by comparing the actual *idf* of a word with its predicted *idf*. It is defined as

$$ridf(i) = idf(i) - \log(1 - p(0; e(i)))$$

where p is the Poisson distribution with parameter $e(i)$. Then $1 - p(0; e(i))$ is the Poisson probability of word i appearing at least once in a document. This measure is again based on the observation that the Poisson model can only fairly predict the distribution of non-content words. Therefore the deviation from Poisson can be used to predict term informativeness.

In practice, *idf* and *ridf* are often used in conjunction with term frequency in the following way, forming the *tf-idf* and *tf-ridf* measures:

$$tf\text{-}idf(i) = tf(i) \cdot idf(i) \text{ and}$$

$$tf\text{-}ridf(i) = tf(i) \cdot ridf(i)$$

In this way the virtue of term frequency to locate keywords in a document is combined with the discriminative power of *idf* and *ridf* in order to filter out high frequency non-content words. In contrast to the *tf-idf* that has become a popular informativeness measure in multiple domains, the *tf-ridf* has rather limited use, although it is been shown to be a better choice in specific applications such as automatic summarization (Orăsan, 2009).

Papineni (2001) presents an information theoretic measure, namely *gain*, which is defined as

$$gain(i) = \frac{df(i)}{D}\left(\frac{df(i)}{D} - 1 - \log\frac{df(i)}{D}\right)$$

This measure is formulated as a response to one of the main criticisms that *idf* overwhelmingly favors words of extremely low frequencies. Accordingly, it tends to assign low values to both very high- and low- frequency words, and treats the mid-frequency words as having the strongest “resolving power”.

The information theoretic measures illustrated in Mladenić and Grobelnik (1998) for document categorization are also applicable to term-weighting, according to Orăsan (2009). Among them, the mutual information (*mi*) to measure the amount of information of a word about a set of documents is formulated as

$$mi(i) = \sum_{j=1}^{D} P(d_j) \cdot log\frac{P(i|d_j)}{P(i)}$$

where $P(d_j)$, $P(i|d_j)$ and $P(i)$ are probabilities of document d_j, word i in d_j, and word i, estimated respectively as $1/D$, $df(i)/tf(i)$ and $tf(i)/T$ with T to be the total number of words in the text collection in question.

Another useful term-weighting measure proposed in Mladenić and Grobelnik (1998) is *information gain* (*ig*). It is formulated as follows to quantify the difference between the entropies of a document set with and without the word in question.

$$ig(i) = P(i) \cdot \sum_{j=1}^{D} P(d_j|i) \cdot log \frac{P(d_j|i)}{P(d_j)}$$
$$+ P(i') \cdot \sum_{j=1}^{D} P(d_j|i') \cdot log \frac{P(d_j|i')}{P(d_j)}$$

where i' indicates the absence of word i, $P(i)$ and $P(i')$ are the probabilities of i occurring and not occurring in the text collection, and $P(d_j|i)$ and $P(d_j|i')$ the probabilities of document d_j given and not given i, estimated as $df(i)/D$ and $1 - df(i)/D$, respectively.

4 Experiment

4.1 Data

We use the MetricsMATR08 development data (Przybocki et al., 2009) in our primary experiment. It consists of 25 documents with a total of 249 segments. For each segment there are eight different versions of MT output and four versions of human reference translation. The MT outputs are assessed by humans according to adequacy of translation and their preference. The former is adopted as a criterion for this work.

Another dataset, the Multiple-translation Chinese part 2 (MTC2) (Huang et al., 2003) from LDC, is used in another experiment to examine the influence of data size upon the performance of a metric. This dataset contains 100 news documents of 878 segments in total. There are three versions of MT output and four versions of reference. Its adequacy assessment data is used in our another experiment. The text genre of both datasets is newswire.

4.2 Metric

A term informativeness measure is to be integrated into a fundamental MT evaluation metric based on harmonic F-measure f of unigram matches between an MT output c and its reference translation t. The precision p and recall r are formulated as follows in terms of (1) the information value $info(c,t)$ given by an informativeness measure in question for the matches and (2) the total information values $info(c)$ of the MT output and $info(t)$ of the reference translation.

$$p(c,t) = \frac{info(c,t)}{info(c)} \qquad r(c,t) = \frac{info(c,t)}{info(t)}$$
$$f(c,t) = \frac{2pr}{p+r}$$

This ensures that the metric is sensitive to word choice only, disregarding all other features such as word order or syntax that should not interfere into our examination of word informativeness. This metric also accounts for the importance of missing information. Note that an unmatched informative word may outweigh a number of matched words if they are less significant. All words in a dataset are reduced to their stems with the Porter stemmer (Porter, 1980), in order to group the morphological variants of a word all together into one for a reliable calculation of their information values.

4.3 Parameters

We also attempt to find out the optimal parameter combination that can maximize the performances of the informativeness measures in use.

Data sources for informative measures

A basic issue of using frequency-based informativeness measures is concerned with the source of training data. Although a desirable practice is to resort to a large external corpus, available MT evaluation data is mostly from special subject domains such as news or technical texts that form sublanguages suitable for MT to handle. The word frequency data from a general text corpus may not be as sensitive as in these subject domains to allow necessary differentiation between important terms and general words. Furthermore, a set of MT evaluation data may contain new terms coined by human translators preparing reference translations, such as transliterations of proper names, that are unlikely to occur elsewhere. Therefore, we have to trust the MT evaluation dataset to provide reliable frequency data for the informativeness measures.

Frequency from MT and human references

A problem with using the aforementioned evaluation datasets for our purpose is to differentiate between the roles of word frequency from different sources, namely, MT output vs. reference translation. It is conceptually reasonable to regard reference translations as correct and use the frequency data therefrom. However, using the word frequency from an MT output to evaluate itself seems questionable. It is also not sure whether it is suitable to compute precision using only frequency data from references. Doing so would unexpectedly bias the overall rating towards recall, instead of achieving a harmonic F-mean as expected. Another practical problem is that many words in an MT output do not occur in any reference, resulting in that they have no occurrence frequency for the computation of their informativeness. Therefore we opt to use the frequency in MT output for calculating precision and that in references for recall. Our result shows that the frequency in MT output gives a nice ranking for the relative informativeness of a word.

Translation variants in multiple references

The use of multiple references is highly beneficial in that it allows more than one legitimate translation variant to compare with an MT output. It also gives rise to a problem of how to deal with the informativeness of synonyms in different references. For example, in our evaluation dataset, there are two variants of a proper name, "Terje Roed-Larsen" and "Terry Rod Larsen", the former of which occurs in three versions of reference translation and the latter in the remaining one. If we simply use a bag-of-words approach to group different word variants together, then the informativeness of the former would outweigh the latter to a large extent if use a measure favoring high frequency or vice versa if use a measure devoting heavy weights to rarely used words. Neither of these is desirable for our purpose. Instead, the two variants are expected to be weighted in the same way, as they are identical in all aspects but spelling.

Two approaches are attempted to exploit multiple references in our experiment. One compares an MT output with each reference for its matching information value and then selects the reference with the highest value. The other combines variant words from different references for matching, in a way that the information value of a word is based on the reference containing the word. For the words occurring in more than one reference, their average occurrence frequencies are used.

Informativeness at document and segment levels

Usually we have an informativeness measure to work on documents, but the granularity of document may vary in terms of size. Wong and Kit (2010) practice to take advantage of the *tf-idf* measure at the segment level for MT evaluation, resulting in a version of ATEC metric of a higher sensibility to the significances of words in segments. Note that segment is the basic unit of MT evaluation that demands no document structure in the evaluation dataset. In this experiment, we compare the performance of different informativeness measures at the document and segment levels.

Number of documents/segments in dataset

Most measures for frequency-based informativeness calculate the relative information load of each word in terms of its distribution in different documents/segments. The number of documents/segments in a dataset thus has a significant influence upon the capability of a measure of weighting words. We assess the performance of the selected measures on datasets of various sizes, so as to find out a minimal number of documents/segments that best fits each measure.

Use of stoplist

Some informativeness measures, like term frequency, are usually used in conjunction with a stoplist so as to skip counting any stopword in a dataset. In automatic MT evaluation, however, this would most likely reduce the number of matched words and inevitably the reliability of evaluation result. We investigate whether an enhancement of performance can be obtained for an evaluation metric at the cost of sacrificing some matching rate in exchange for a better scoring of term informativeness.

4.4 Results and Discussion

Table 1 shows the correlation results in terms of Pearson's R coefficients (Pearson, 1900) for the evaluation metric incorporating one of the 10 term informativeness measures under various evaluation settings using the human assessment data from MetricsMATR08 dataset. The "Highest" and "Average" columns correspond to our two approaches of using multiple references, "Document" and "Segment" to the two levels of informativeness, and "Stop" and "All" to the use of stoplist or not, respectively. We also provide the correlations without the use of informativeness measure as a baseline, in the "word" row at the bottom of the table. The correlations stronger than that of the

Informativeness measure	Single reference				Multiple references							
					Highest				Average			
	Document		Segment		Document		Segment		Document		Segment	
	Stop	All	Stop	All	Stop	All	Stop	All	Stop	All	Stop	All
tf	.508	.446	.536	.560	.617	.502	.641	.627	.640	.389	.674	.591
idf	.525	.555	.529	.582	.638	.659	.643	.670	.669	***.698***	.669	***.707***
ridf	.526	.560	.529	.584	.639	.659	.643	.673	.670	***.701***	.669	***.708***
tfidf	.497	.529	.527	.577	.602	.620	.636	.663	.625	***.654***	.667	***.706***
tfridf	.497	.538	.527	.582	.603	.627	.636	.667	.626	***.660***	.667	***.708***
bur	.511	.567	.538	***.599***	.613	.651	.648	***.685***	.640	***.685***	***.676***	***.708***
var	.461	.326	.450	.351	.540	.351	.528	.374	.530	.247	.535	.284
gain	.521	.580	.484	.540	.629	.657	.590	.619	.658	***.701***	.609	.622
mi	.531	.583	.536	.598	.642	.667	.647	***.685***	.673	***.712***	.675	***.714***
ig	.533	.588	.535	***.602***	.647	.677	.644	.682	***.678***	***.717***	.674	***.713***
word	.541	.598	.541	.598	.650	.681	.650	.682	.675	.648	.675	.648

Table 1. Correlations of the evaluation metric using different term informativeness measures in various settings

baseline in the same experiment setting are highlighted in bold, and the strongest correlations in the settings of single and multiple references are underlined.

In general, strong disparities are observed in the performances of different informativeness measures in various or even the same settings. Each measure seems to have its own optimal evaluation setting for its best performance. An unexpected finding is that the baseline, which assigns an equal weight to each word, outperforms most other informativeness measures. In the context of single reference, only *bur* and *ig* are better than the baseline with an insignificant gain in correlation, i.e., 0.001-0.004. Most measures perform their best with multiple references under the “Average” approach. The best one in this group is *ig* at the document level without a stoplist, achieving a correlation of 0.717, which is nearly 10% beyond the baseline 0.648.

It is worth noting that the use of a stoplist only favors a few measures, particularly, *tf* and *var*. For all others, stopwords bring in no advantage. A possible reason might be that these high-frequency stopwords can be utilized by an informativeness measure in calculating the relative distinctiveness of words. This also conforms to the above-mentioned analysis that skipping stopwords in automatic MT evaluation may result in a lower matching rate and hence a reduction in correlation.

Interestingly, all measures give a better performance at the segment level than at the document level, especially in the setting of single reference. The difference is smaller in the multiple references setting with the “Highest” approach and further smaller with the “Average” approach. For some measures like *mi* and *ig*, their difference is nearly neglectable for the “Average” approach.

The unexpected result that most measures underperform the baseline in most settings is found to be due to the insufficient data size for a valid estimation of informativeness. Table 2 presents the result of another experiment to examine the performances of the measures according to the number of documents, ranging from 20 to 100, using the MTC2 dataset. In contrast, the MetricsMATR08 dataset consists of 25 documents only. The figures in Table 2 are the correlation differences in percentage from the baseline. For example, *tf* is -10.74% below the baseline, under the setting of these parameters: single reference, document level informativeness and data size of 20 documents. The positive ones, marked in bold, are those beyond the baseline, and those within the top 10% in each of the four groups of single/multiple references and document/segment level are underlined. All figures are obtained with the “Average” approach, and no stoplist is used, except for *tf* and *var* which work better when stopwords are removed.

	Document level					Segment level				
Documents	20	40	60	80	100	20	40	60	80	100
Measure	Single reference									
*tf**	-10.74	-8.12	-6.63	-5.57	-5.94	-1.44	-0.78	**0.51**	**1.77**	**0.86**
idf	-7.92	-3.40	**0.14**	**2.45**	***4.63***	-4.40	-1.48	**1.50**	**2.36**	***3.29***
ridf	-6.43	-2.15	**1.43**	***3.28***	***5.13***	-4.22	-1.37	**1.59**	**2.44**	***3.36***
tfidf	-20.69	-13.66	-9.89	-7.33	-4.71	-3.75	-0.29	**2.33**	***3.11***	***3.93***
tfridf	-19.51	-12.42	-8.16	-6.24	-3.82	-3.42	-0.06	**2.49**	***3.27***	***4.15***
bur	-11.29	-7.27	-6.05	-6.40	-7.06	-0.71	**0.89**	**1.78**	**1.58**	**1.57**
*var**	-37.89	-44.76	-44.37	-43.58	-49.78	-29.13	-38.83	-39.29	-41.20	-51.35
gain	-4.41	-3.89	-4.40	-5.44	-7.04	-8.53	-12.66	-17.23	-21.20	-25.58
mi	-4.56	-1.18	**1.05**	**1.70**	**2.36**	-1.44	**0.78**	**2.73**	***3.13***	***3.62***
ig	-3.95	**0.21**	**2.51**	***3.74***	**4.87**	-0.90	**1.09**	**2.51**	**2.52**	**2.59**
Measure	Multiple references									
*tf**	-4.56	-0.34	**2.65**	**3.02**	**0.37**	**4.62**	**6.46**	**8.81**	**9.95**	**7.42**
idf	**2.64**	**5.84**	**9.00**	**10.19**	**10.76**	**6.29**	**8.57**	**10.73**	***11.47***	***11.07***
ridf	**3.93**	**6.89**	**9.95**	***10.81***	***11.24***	**6.42**	**8.60**	**10.71**	***11.45***	***11.10***
tfidf	-8.96	-2.33	**1.75**	**2.52**	**2.89**	**6.12**	**8.53**	**10.68**	***11.29***	**10.75**
tfridf	-7.71	-1.25	**2.73**	**3.24**	**3.49**	**6.25**	**8.40**	**10.50**	***11.13***	**10.77**
bur	**3.44**	**5.74**	**6.72**	**6.87**	**5.14**	**7.22**	**8.71**	**9.37**	**9.51**	**8.97**
*var**	-28.57	-37.19	-37.55	-37.69	-50.87	-22.99	-31.37	-32.95	-33.99	-48.45
gain	**5.64**	**7.00**	**6.88**	**6.13**	**3.02**	-2.03	-7.07	-11.41	-14.84	-19.40
mi	**7.09**	**9.45**	***11.08***	***11.47***	**10.64**	**7.72**	**9.56**	***11.03***	***11.51***	***11.10***
ig	**5.62**	**8.90**	***11.05***	***11.80***	***11.50***	**7.43**	**9.27**	**10.41**	**10.61**	***9.89***

Table 2. Correlation differences (in percentage %) from baseline in various data sizes (*used with a stoplist)

In general, a better performance is observed on a larger dataset. With single reference, most outperformances over the baseline start at 60 documents (~500 segments). With multiple references, however, the best performances are achieved around 80 documents (~700 segments). The best is achieved by *ig*, which is 11.8% beyond the baseline under the setting of document level, multiple references and 80 documents. Other comparable performances with this are achieved by *idf*, *ridf*, *tfidf*, *tfridf* and *mi*, all of which work better at the segment level.

5 Conclusion

We have investigated, through experiments, into a problem with most automatic MT evaluation metrics that the informativeness of words is simply discarded, inevitably leading to an underestimation of the quality of the MT outputs with many informative words and an overestimation of those with few. We have comparatively evaluated a number of term informativeness measures under various operational environments whether any of them can enhance MT evaluation performance in general, and found the *information gain* to be the best among them under our experimental setting.

Our experiments also show that not every informativeness measure can result in an improvement. Many of them in fact cause damage on the correlation with human judgments if used improperly. A successful application of them heavily relies on an appropriate setting of various parameters for an evaluation, among which a particularly critical one is data size.

These preliminary discoveries may serve as a basis for further exploration on applying informativeness measurement to enhance other well-established MT evaluation metrics, and also for further study of other related parameters in the quantification of translation quality such as the information structure of a text. What we have done is certainly a meaningful step towards a more accurate, content-based evaluation method for MT beyond those merely relying on textual similarity, which are not always reliable and even annoying, sometimes and somewhere, as shown in use in our current work.

Acknowledgment

The research described in this paper was supported by City University of Hong Kong through the SRG grants 7002267 and 7008003 and by the Research Grants Council (RGC) of HKSAR, China, through the GRF grant 9041597.

References

ALPAC (Automatic Language Processing Advisory Committee). 1966. *Report of the ALPAC; Language and Machines: Computers in Translation and Linguistics*. Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, D.C.

Bogdan Babych and Anthony Hartley. 2004. Extending the BLEU MT evaluation method with frequency weightings. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 621-628.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65-72.

Kenneth W. Church and William A. Gale. 1995a. Inverse document frequency (IDF): A measure of deviations from Poisson. In *Proceedings of Third the Workshop on Very Large Corpora*, pages 121-130.

Kenneth W. Church and William A. Gale. 1995b. Poisson mixtures. *Natural Language Engineering*, 1(2): 163-190.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Human Language Technology Conference: Proceedings of the Second International Conference on Human Language Technology Research*, pages 138-145.

Shudong Huang, David Graff, Kevin Walker, David Miller, Xiaoyi Ma, Chris Cieri, and George Doddington. 2003. *Multiple-translation Chinese (MTC) part 2*. Linguistic Data Consortium, Philadelphia.

Kirill Kireyev. 2009. Semantic-based estimation of term informativeness. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 530-538.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159-165.

Dunja Mladenić and Marko Grobelnik. 1998. Feature selection for classification based on text hierarchy. In *Conference on Automated Learning and Discovery (CONALD-98)*.

Constantin Orăsan. 2009. Comparative evaluation of term-weighting methods for automatic summarization. *Journal of Quantitative Linguistics*, 16(1):67-95.

Kishore Papineni. 2001. Why inverse document frequency? In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, pages 25-32.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2001. *BLEU: A Method for Automatic Evaluation of Machine Translation.* IBM Research Report RC22176 (W0109-022).

Karl Pearson. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(5):157-175.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130-137.

Mark Przybocki, Kay Peterson and Sébastien Bronsart. 2009. *2008 NIST Metrics for Machine Translation (MetricsMATR08) Development Data*. Linguistic Data Consortium, Philadelphia.

Karen Spärck-Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11-21.

John S. White. 2003. How to evaluate machine translation. In Somer (ed.) *Computers and Translation: A Translator's Guide*, pages 211-244, John Benjamins.

John S. White, Theresa A. O'Connell and Francis E. O'Mara. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the America*, pages 193-205.

Billy Wong and Chunyu Kit. 2010. ATEC: Automatic evaluation of machine translation via word choice and word order. *Machine Translation*, 23(2/3): 141-151.