Normalized Compression Distance Based Measures for MetricsMATR 2010

Marcus Dobrinkat and **Jaakko Väyrynen** and **Tero Tapiovaara**
Adaptive Informatics Research Centre
Aalto University School of Science and Technology
P.O. Box 15400, FI-00076 Aalto, Finland
{marcus.dobrinkat,jaakko.j.vayrynen,tero.tapiovaara}@tkk.fi

Kimmo Kettunen
Kymenlaakso University of Applied Sciences
P.O. Box 9, FI-48401 Kotka, Finland
Kimmo.kettunen@kyamk.fi

Abstract

We present the MT-NCD and MT-mNCD machine translation evaluation metrics as submission to the machine translation evaluation shared task (MetricsMATR 2010). The metrics are based on normalized compression distance (NCD), a general information theoretic measure of string similarity, and evaluated against human judgments from the WMT08 shared task. The experiments show that 1) our metric improves correlation to human judgments by using flexible matching, 2) segment replication is effective, and 3) our NCD-inspired method for multiple references indicates improved results. Generally, the proposed MT-NCD and MT-mNCD methods correlate competitively with human judgments compared to commonly used machine translations evaluation metrics, for instance, BLEU.

1 Introduction

The quality of automatic machine translation (MT) evaluation metrics plays an important role in the development of MT systems. Human evaluation would no longer be necessary if automatic MT metrics correlated perfectly with manual judgments. Besides high correlation with human judgments of translation quality, a good metric should be language independent, fast to compute and sensitive enough to reliably detect small improvements in MT systems.

Recently there have been some experiments with normalized compression distance (NCD) as a method for automatic evaluation of machine translation. NCD is a general string similarity measure that has been useful for clustering in various tasks (Cilibrasi and Vitanyi, 2005).

Parker (2008) introduced BADGER, a machine translation evaluation metric that uses NCD together with a language independent word normalization method. Kettunen (2009) independently applied NCD to the direct evaluation of translations. He showed with a small corpus of three language pairs that the scores of NCD and METEOR (v0.6) from translations of 10–12 MT systems were highly correlated.

Väyrynen et al. (2010) have extended the work by showing that NCD can be used to rank translations of different MT systems so that the ranking order correlates with human rankings at the same level as BLEU (Papineni et al., 2001). For translations into English, NCD had an overall system-level correlation of 0.66 whereas the best method, ULC had an overall correlation of 0.76, and BLEU had an overall correlation of 0.65. NCD presents a viable alternative to the de facto standard BLEU. Both metrics are language independent, simple and efficient to compute. However, NCD is a general measure of similarity that has been applied in many domains. More advanced methods achieve better correlation with human judgments, but typically use additional language specific linguistic resources. Dobrinkat et al. (2010) experimented with relaxed word matching, adding language specific resources to NCD. The metric called mNCD, which works similarly to mBLEU (Agarwal and Lavie, 2008), showed improved correlation to human judgments in English, the only language where a METEOR synonym module was used.

The motivation for this challenge submission is to evaluate the MT-NCD and MT-mNCD metric performance in an open competition with state-of-

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 343–348,
Uppsala, Sweden, 15-16 July 2010.

the-art MT evaluation metrics. Our experiments and submission build on NCD and mNCD. We expand NCD to handle multiple references and report experimental results for replicating segments as a preprocessing step that improves the NCD as an MT evaluation metric.

2 NCD-based MT evaluation metrics

NCD-based MT evaluation metrics build on the idea that a string x is similar to another string y, when both share common substrings. When describing y, common substrings do not have to be repeated, but can be referenced to x. This is done when compressing the concatenation of x and y, which results in smaller output when more information of y is already included in x.

2.1 Normalized Compression Distance

The normalized compression distance, as defined by Cilibrasi and Vitanyi (2005) is given in Equation 1, in which $C(x)$ is the length of the compression of x and $C(x, y)$ is the length of the compression of the concatenation of x and y.

$$\text{NCD}(x, y) = \frac{C(x, y) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}} \quad (1)$$

NCD computes the distance as a score closer to one for very different strings and closer to zero for more similar strings. Most MT evaluation metrics are defined as similarity measures in contrast to NCD, which is a distance measure. For easier comparison with other MT evaluation metrics, we define the NCD based MT evaluation similarity metric MT-NCD as $1 - \text{NCD}$.

NCD is a practically usable form of the uncomputable normalized information distance (NID), a general metric for the similarity of two objects. NID is based on the notion of Kolmogorov complexity $K(x)$, a theoretical measure for the algorithmic information content of a string x. It is defined as the shortest universal Turing machine that prints x and stops (Solomonoff, 1964). NCD approximates NID by the use of a compressor $C(x)$ that presents a computable approximation of the Kolmogorov complexity $K(x)$.

2.2 NCD with multiple references

Most ideas can be described with in different ways, therefore using only one reference translation for the evaluation of a candidate sentence is not ideal and the exploitation of knowledge in several different reference translations is helpful for automatic MT evaluation.

One simple way for handling multiple references is to evaluate against each reference individually and select the maximum score. Although this works, it is clearly not optimal. We developed the NCD_m metric, which is inspired by NCD. It considers all references simultaneously and the quality of a translation t against multiple references $R = \{r_1, \ldots, r_m\}$ is assessed as

$$\text{NCD}_m(t, R) = \frac{\max\{C(t|R), \min_{r \in R} C(r|t)\}}{\max\{C(t), \min_{r \in R} C(r)\}} \quad (2)$$

where $C(x|y) = C(x, y) - C(y)$ approximates conditional algorithmic information with the compressor C. The NCD_m similarity metric with a single reference ($m = 1$) is equal to NCD in Equation 1. Again, we define MT-NCD_m as $1 - \text{NCD}_m$.

Figure 1 shows how both, the MT-NCD_m and the BLEU metric change with a different number of references when the translation is varied from correct to a random sequence of words. The scores are computed with 249 sentences from the LDC2010E28Dev data set using the first reference as the correct translation. A higher score with multiple references against the correct translation indicates that the measure is able to take into account information from multiple references at the same time.

The words in the candidate translation are replaced with probability p with a word randomly selected with uniform probability from a lexicon created from all reference translations. This simulates partially correct translations. The words are changed in a simple way without deletions, insertions or word order permutations. The MT-NCD_m score increases with more than one reference translation and random changes to the sentence reduce the score roughly proportional to the number of changed words. With BLEU, the score is affected more by a small number of changes.

2.3 mNCD

One enhancement to the basic NCD as automatic evaluation metric is mNCD (Dobrinkat et al., 2010), which provides relaxed word matching based on the flexible matching modules of METEOR (Agarwal and Lavie, 2008).

What mNCD does is that it changes the reference sentence to be more similar to the candi-

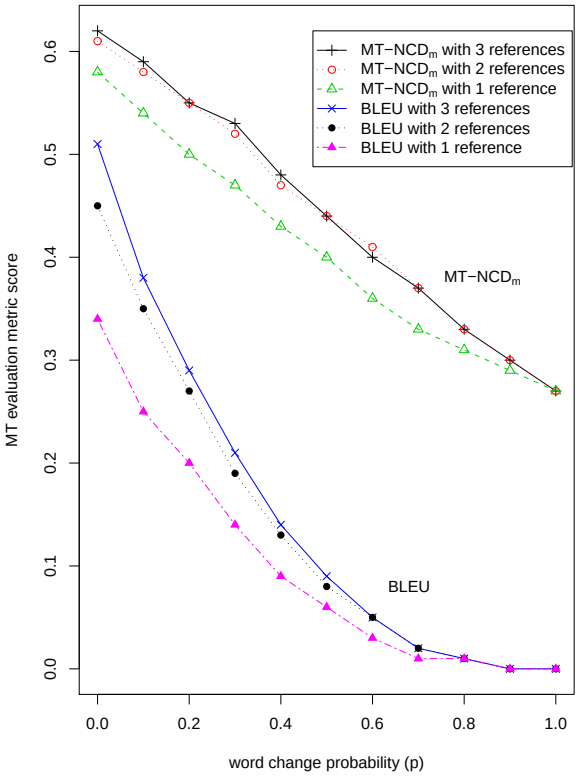

Figure 1: The MT-NCD$_m$ and BLEU scores with a different number of multiple references against correct translation with random word change probability (p).

date, given that some of the words are synonyms or share the same stem. Subsequent analysis using any n-gram based automatic analysis should result in a larger similarity score in the hope that this reflects more than just the surface similarity between the candidate and the reference.

Given suitable Wordnet resources, mNCD should alleviate the problem of translation variability especially in absence of multiple reference translations. Our submission uses the default METEOR `exact stem synonym` modules, which provide synonyms only for English. We base our submission metric on the MT-NCD metric and therefore define MT-mNCD as $1 -$ mNCD.

3 MT Evaluation System Description

3.1 System Parameters

The system parameters for the submission metrics include how candidates and references are preprocessed, the choice of compressor for the NCD itself, as well as the granularity of how large segments are evaluated by NCD and how they are combined into a final score.

Partly due to time constraints we decided not to introduce language specific parameters, therefore we chose those parameter values that perform well in overall and are simple to compute.

3.1.1 Preprocessing

Character casing For MT-NCD, we did experiments without preprocessing and with lowercasing candidates and references. On average over all tasks for language pairs into English, lowercasing consistently decreased the RANK correlation scores but increased the CONST correlation scores. No consistent effect could be found for the language pairs from English. In our submission metrics we use no preprocessing.

For MT-mNCD the used METEOR matching module lower-cases the adapted words by default. After adapting a synonym in a reference, we tried to keep the casing as it was in the candidate, which we called real-casing. We use no real-casing for our submitted MT-mNCD metric as this did not improve results consistently over all task into English.

Segment Replication Compression algorithms may not work optimally with short strings, which would deteriorate the approximation of Kolmogorov complexity. Our hypothesis was that a replication of a string ("abc") multiple times ($3 \times$ "abc" $=$ "$abcabcabc$") could help the compression algorithm to produce a better estimate of the algorithmic information. This was tested in the MT evaluation framework, and correlation between MT-NCD and human judgments improved when the segments were replicated two times. Further replication did not produce improvements.

Results for the MT-NCD metric with replications one, two and three times are shown in Table 1. The results are averages over all used languages. With two compared to one replication, the details for each language show that RANK correlation is improved for the target languages English and French, but degrades for German and Spanish. CONST andYES/NO correlation improve for all languages except German. We did not use replication in our submissions.

3.1.2 Block size

The block size parameter governs the number of joined segments that are compared with NCD as a single string. On one extreme, with block size one,

		RANK	CONST	YES/NO	TOTAL
MT-NCD	rep 1	.61	.71	.73	.68
MT-NCD	rep 2	.62	.73	.75	.70
MT-NCD	rep 3	.61	.72	.74	.69

Table 1: Effect of the replication factor on MT-NCD correlation scores for the bz2 compressor with block size one as average over all languages.

each segment is evaluated separately and the segment scores are aggregated to a document score. This is similar to how other MT metrics, for example, BLEU, work. The other extreme is to join all segments together, with block size equal to the number of segments, and evaluate it as a single string, which is similar to document comparison. For block aggregation we experimented with arithmetic and geometric mean and obtained very similar results. We selected arithmetic mean for the submission metrics.

Figure 2 shows the block size effect on the correlation between MT-NCD and human judgments for different target languages. Except for Spanish, our experiments indicate that the block size value has little effect. Therefore, and given how other evaluation metrics work, we chose a block size of one for our submission metrics. We noticed inconsistencies with Spanish in other settings as well and will investigate these issues further.

3.1.3 Compressor

There are several universal compressors that can be utilized with NCD, for instance, zlib/gzip, bz2 and PPMZ, which represent different approaches to compression. In terms of compression rate, PPMZ is the best of the mentioned methods, but it is considerably slower to compute compared to the other methods. In terms of correlation with human judgments, NCD using bz2 performs slightly worse than using PPMZ. Given much shorter compression times for bz2 with very little correlation performance degradation, our choice for the submission is the more standard bz2 compressor.

3.1.4 Segment Interleaving

Computation of NCD between longer texts (e.g. documents) may exceed the internal compressor window size that is present in some compression

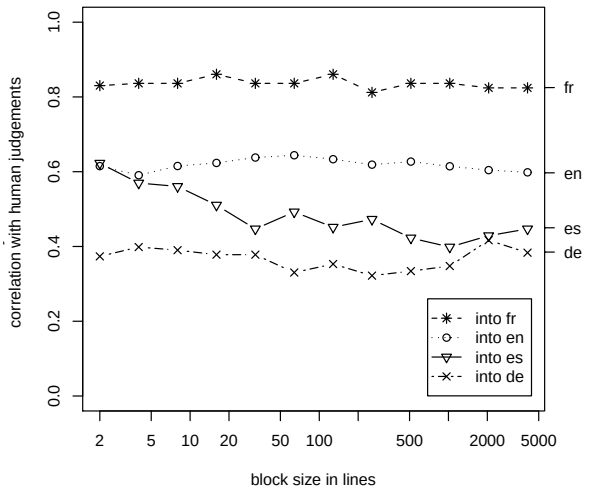

Figure 2: Effect of the block size on the correlation of MT-NCD to human judgments for the system level evaluation.

algorithms (Cebrian et al., 2005). In this case, only a part of the texts to be compared are visible at any time to the compressor and similarities to the text outside the window will be missed. One solution for the MT evaluation task is to use utilize the known parallel segments of candidate and reference translations. The two segment lists can be interleaved so that the corresponding segments are always adjacent and the compression window size is not exceeded for matching segments.

For our submission, we chose a block size of one, therefore every segment is evaluated individually. As a result, segment interleaving does not have any effect. Segment interleaving is affective in the block size evaluation and results shown in Figure 2.

3.2 Evaluation Experiments

We chose parameters and evaluated our metrics using the WMT08 part of the MetricsMATR 2010 development data, which contains human judgments of the 2008 ACL Workshop on Statistical Machine Translation (Callison-Burch et al., 2008) for translations from a total of 30 MT systems between English and five other European languages. There are human evaluations and several automatic evaluations for the translations, divided into several tasks defined by the language pair and the domain of the translated sentences. For each of these tasks, the WMT08 data contains about 2 000

reference sentences (segments) plus their aligned translations for 12 to 17 different translation systems, depending on the language pair.

The human judgments include three categories which contain evaluations for at most one segment at a time, not whole documents. In the RANK category, humans had to rank the output of five MT systems according to quality. The CONST category contains rankings for short phrases (constituents), and the YES/NO category contains binary answers to judge if a short phrase is an acceptable translation or not.

We report RANK, CONST and YES/NO system level correlations to human judgments as results of our metrics for French, Spanish and German both from and to English. The English–Spanish news task was left out as most metrics had negative correlation with human judgments.

The evaluation methodology used in Callison-Burch et al. (2008) allows us to measure how each MT evaluation metric correlates with human judgments on the system level, in which all translations from each MT system are aggregated into a single score. The system rankings based on the scores are compared to human judgments.

Spearman's rank correlation coefficient ρ was calculated between each MT metric and human judgment category using the simplified equation:

$$\rho = 1 - \frac{6 \sum_i d_i}{n(n^2 - 1)} \quad (3)$$

where for each system i, d_i is the difference between the rank derived from annotators' input and the rank obtained from the metric. From the annotators' input, the n MT systems were ranked based on the number of times each system's output was selected as the best translation divided by the number of times each system was part of a judgment.

3.3 Results

The results for WMT08 data for our submitted metrics are shown in Table 2 and are sorted by the RANK category separately for language pairs from English and into English.

For tasks into English, the correlations show that MT-mNCD improves over the MT-NCD metric in all categories. Also the flexible matching seems to work better for NCD-based metrics than for BLEU, where mBLEU only improves the CONST correlation scores. For tasks from English, MT-mNCD shows slightly higher correlation compared to MT-NCD, except for the YES/NO category. The standard BLEU correlation score is best of the shown evaluation metrics. Relaxed matching using mBLEU does not improve BLEU's RANK correlation scores here either, but CONST and YES/NO correlation performs better relative to BLEU than MT-mNCD compared to MT-NCD.

		RANK	CONST	YES/NO	TOTAL
INTO EN	MT-mNCD	.61	.74	.75	.70
	MT-NCD	.57	.69	.71	.66
	mBLEU	.50	.76	.70	.65
	BLEU	.50	.72	.74	.65
FROM EN	BLEU	.68	.79	.79	.75
	MT-mNCD	.67	.76	.74	.72
	MT-NCD	.65	.73	.75	.71
	mBLEU	.63	.81	.81	.75

Table 2: Average system-level correlations for the WMT08 data sorted by RANK into English and from English for our submitted metrics MT-NCD and MT-mNCD and for BLEU and mBLEU

4 Conclusions

In our submissions, we applied MT-NCD and MT-mNCD metrics and extended the NCD MT evaluation metric to handle multiple references. The reported experiment indicate a possible improvement for the multiple references.

We showed that a replication of segments as a preprocessing step improves the correlation to human judgments. The string replication might alleviate problems in the compressor for short strings and thus could provide better estimates of the algorithmic information.

The results of our experiments show that relaxed matching in MT-mNCD works well with proper synonym dictionaries, but is less effective for tasks from English, which only use stemming.

MT-mNCD and MT-NCD are reasonably simple to compute and utilize standard and widely used resources, such as the bz2 compression algorithm and WordNet. The metrics perform comparable to the de facto standard BLEU. Improvements with language dependent resources, in particular relaxed matching using synonym dictionaries proved to be useful.

References

Abhaya Agarwal and Alon Lavie. 2008. METEOR, M-BLEU and M-TER: evaluation metrics for high-correlation with human rankings of machine translation output. In *StatMT '08: Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118, Morristown, NJ, USA. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *StatMT '08: Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Morristown, NJ, USA. Association for Computational Linguistics.

Manuel Cebrian, Manuel Alfonseca, and Alfonso Ortega. 2005. Common pitfalls using the normalized compression distance: What to watch out for in a compressor. *Communications in Information and Systems*, 5(4):367–384.

Rudi Cilibrasi and Paul Vitanyi. 2005. Clustering by compression. *IEEE Transactions on Information Theory*, 51:1523–1545.

Marcus Dobrinkat, Tero Tapiovaara, Jaakko J. Väyrynen, and Kimmo Kettunen. 2010. Evaluating machine translations using mNCD. In *Proceedings of the ACL-2010 (to appear)*, Uppsala, Sweden.

Kimmo Kettunen. 2009. Packing it all up in search for a language independent MT quality measure tool. In *Proceedings of LTC-09, 4th Language and Technology Conference*, pages 280–284, Poznan.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center.

Steven Parker. 2008. BADGER: A new machine translation metric. In *Metrics for Machine Translation Challenge 2008*, Waikiki, Hawai'i, October. AMTA.

Ray Solomonoff. 1964. Formal theory of inductive inference. Part I. *Information and Control*, 7(1):1–22.

Jaakko J. Väyrynen, Tero Tapiovaara, Kimmo Kettunen, and Marcus Dobrinkat. 2010. Normalized compression distance as an automatic MT evaluation metric. In *Proceedings of MT 25 years on*. To appear.