Normalized Compression Distance as automatic MT evaluation metric

Jaakko Väyrynen[1], Tero Tapiovaara[1], Kimmo Kettunen[2], Marcus Dobrinkat[1]

jjvayryn@ics.tkk.fi, ttapiova@ics.tkk.fi, kimmo.kettunen@kyamk.fi, mdobrink@ics.tkk.fi

[1]Helsinki University of Technology, Adaptive Informatics Research Centre,
P.O. Box 5400, FI-02015 TKK, Finland

[2]Kymenlaakso University of Applied Sciences,
P.O. Box 9, FI-48401 Kotka, Finland

Abstract

This paper evaluates a new automatic MT evaluation metric, Normalized Compression Distance (NCD), which is a general tool for measuring similarities between binary strings. We provide system-level correlations and sentence-level consistencies to human judgements and comparison to other automatic measures with the WMT'08 dataset. The results show that the general NCD metric is at the same level as some of the currently widely used metrics defined for the particular task of MT evaluation. We discuss the possible reasons for this.

1 Introduction

Automatic evaluation of machine translation program output has been developed and used for about a decade. There are several MT evaluation systems or metrics, such as BLEU (Papineni et al., 2001), NIST (Doddington, 2002), METEOR (Lavie and Agarwal, 2007), IQMT (Giménez and Amigó, 2006) and several others not mentioned here. Most of the evaluation metrics are based on similar features, e.g., use of string level comparison of texts, recall and precision of translations, different penalty scores etc. The metrics have been a valuable tool in the development of automatic MT systems.

It is well known that all the present automatic MT evaluation methods have limitations. Many recent, better performing MT metrics are language dependent, as they use language specific resources such as syntactic parsers, synonym databases or stemming. Often, these resources are available only for a few languages, or include training and optimizing models. Therefore, a general and robust framework to be used without concerns of language pair dependencies would be useful. Also other concerns about MT metrics have been stated. Callison-Burch et al. (2006) show in a detailed analysis that BLEU's coarse model of allowable variation in word order of translations ”can mean that an improved BLEU score is not sufficient to reflect a genuine improvement in translation quality”. Turian et al. (2003) claim that the most popular MT evaluation metrics, BLEU and NIST, fail to correlate well with human judgements of translation quality.

We show in this paper that a language independent measure for MT quality evaluation can be obtained from a general classification and clustering tool called Normalized Compression Distance, NCD (Cilibrasi and Vitanyi, 2005; Li et al., 2004; Vitanyi et al., 2009). NCD has not been evaluated much in the context of MT evaluation, the only preliminary trials as far as we know are Parker (2008) and Kettunen (2009). Parker has introduced an MT metric named BADGER that utilizes NCD as one part of the metric with additional enhancements, such as a language independent word normalization method. Parker benchmarks BADGER against METEOR and word error rate (WER) metrics with Arabic to English translations. The correlation of BADGER results to those of METEOR are low and correlation to WER high. On the other hand, Kettunen (2009) has made preliminary testing of NCD with translations from English to German, Spanish and French, where the results showed that both NCD and METEOR were able to pick the best and worst MT systems for each language pair. Furthermore, the scores of NCD correlated very highly with the scores of METEOR.

In this study we broaden the scope of testing of NCD as an MT metric. As our test material we use the freely available WMT'08 Shared Task Evaluation Data (see Callison-Burch et al., 2008), which we extend to include the NCD metric. This enables us to compare the performance of NCD, and several other MT metrics, to human evaluations of automatic translations. We provide initial results with different compressors in NCD with respect to MT evaluation. We introduce a simple solution for avoiding the compressor window size problem with some compressors in NCD. (Cebrian et al., 2005)

2 Materials and Methods

2.1 Evaluation data

The NCD metric is evaluated using the shared task data and results of the 2008 ACL Workshop on Statistical Machine Translation (see Callison-Burch et al., 2008), which includes translations from a total of 30 MT systems between English and {Spanish, German, French, Hungarian and Czech} as well as evaluations of the translations with both manual human judgements and several automatic evaluation metrics. The translations are divided into tasks, which define the source language, target language and the text domain.

The human judgements are divided into three classes. The 'Rank' class contains rankings comparing the output of five different, randomly selected MT systems. In contrast to the 'Rank' class, in which annotators rank sentences, the 'Const' class contains rankings for short phrases (or constituents), and the class 'Yes/No' contains binary answers whether a given short phrase is an acceptable translation or not. For a complete explanation of the data, see Callison-Burch et al. (2008).

2.2 Normalized compression distance

The Normalized Compression Distance (NCD) was developed by Cilibrasi and Vitanyi (2005). It is an approximation of the incomputable Normalized Information Distance (NID), which is based on the theoretical foundations of Algorithmic Information Theory and the notion of Kolmogorov complexity. We provide a brief introduction to the topic.

2.2.1 Algorithmic information distance

Kolmogorov complexity $K(x)$, or algorithmic entropy, is a theoretical measure for information content of a string x, and is defined as the length of the shortest Universal Turing Machine that prints x (and stops). (Solomonoff, 1964)

We are interested in measures that compare two distinct objects. One such measure is algorithmic information distance, defined as the length of the shortest program that computes x from y, and y from x. Bennett et al. (1998) have shown that the algorithmic information distance equals

$$E(x,y) = \max\{K(x \mid y), K(y \mid x)\} \tag{1}$$

up to an additive $O(\log(\max\{K(x \mid y), K(y \mid x)\}))$ term, where the conditional Kolmogorov complexity $K(x \mid y)$ is defined as the length of the shortest program that can output x if the input string y is given on an auxiliary tape. The chain rule $K(x \mid y) = K(x,y) - K(y)$ is true up to an additive logarithmic term.

2.2.2 Normalized information distance

We want to measure the similarity between any two strings, therefore we are probably more interested in a relative measure than an absolute measure. The Normalized Information Distance (NID) is defined by Cilibrasi and Vitanyi (2005) as

$$NID(x,y) = \frac{E(x,y)}{\max\{K(x), K(y)\}} = \frac{\max\{K(x \mid y), K(y \mid x)\}}{\max\{K(x), K(y)\}} \tag{2}$$

in which the conditional Kolmogorov complexities can be rewritten using the chain rule to get

$$NID(x,y) = \frac{\max\{K(x,y) - K(y), K(x,y) - K(x)\}}{\max\{K(x), K(y)\}} = \frac{K(x,y) - \min\{K(x), K(y)\}}{\max\{K(x), K(y)\}} \tag{3}$$

up to an additive logarithmic term.

2.2.3 Normalized Compression Distance

There is a need for a practically applicable form of the distance, as NID is incomputable. Universal compression algorithms set an upper bound to the Kolmogorov complexity; therefore by approximating the Kolmogorov complexities with a compression algorithm in Equation 3, we arrive at the definition of Normalized Compression Distance (Cilibrasi and Vitanyi, 2005)

$$NCD(x,y) = \frac{C(x,y) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}} \tag{4}$$

where $C(x)$ 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.

2.3 Evaluation of machine translation metrics

We evaluated the performance of NCD as an automatic evaluation metric for machine translation with different translation tasks by including NCD to the MT metric evaluation in Callison-Burch et al. (2008). The MT metrics are compared to human judgements of translations on both system-level and sentence-level.

2.3.1 Measuring system level correlation

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

$$\rho = 1 - \frac{6\sum d_i}{n(n^2 - 1)} \tag{5}$$

where for each MT 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 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 judgement.

2.3.2 Measuring sentence-level consistency

The sentence-level consistency was measured as the fraction of cases in which the metric gave scores consistent with the annotator ranking. This was calculated as the number of cases where the annotator's ranking of any two sentences matched the ranking derived from the scores the metric assigned to each sentence, divided by the total number of comparisons. All cases where the annotator had judged ties were excluded.

3 Experiments and Results

The NCD experiments were performed with several different universal compressors (zlib, bzip2, gzip and ppmz) to evaluate the impact of the compressor choice.

To overcome the problems introduced by the sliding window size in some of the compressors (Cebrian et al., 2005), we interleave the translated sentences $x = (x_1, x_2, \ldots)$ and the respective reference translations $y = (y_1, y_2, \ldots)$ in the calculation of $C(x, y)$ in Equation 4, such that the corresponding sentences x_j and y_j are adjacent. This operation allows the use of texts exceeding the sliding window size without significant changes of the NCD score.

The average correlations over different translation tasks between automatic MT evaluation metrics and the human judgements are shown in Tables 1 and 2 for system-level analysis for each of the three human judgement classes. Sentence-level analysis measured by consistencies against the Rank class are shown in Table 3, averaged over tasks which include translations either from or into English. These correspond to Tables 8–11 in Callison-Burch et al. (2008) with the same language pairs and test sets. We have re-calculated all the values using the extended WMT'08 evaluation data. The sentence-level consistencies in Callison-Burch et al. (2008) were incorrect due to a programming error.

	Rank	Const	Yes/No	Overall
DP	**.81**	.66	.74	.73
ULCh	.80	.68	.77	.75
DR	.79	.54	.64	.66
meteor-ranking	.78	.55	.62	.65
ULC	.77	.72	**.80**	**.76**
SR	.75	.66	.76	.72
posbleu	.75	.69	.78	.74
meteor-baseline	.74	.60	.63	.66
posF4gram-gm	.74	.61	.70	.68
posF4gram-am	.74	.59	.69	.67
NCD-ppmz	.60	.67	.72	.66
NCD-bz2	.59	.65	.70	.65
NCD-zlib	.57	.71	.76	.68
NCD-gzip	.57	.71	.76	.68
mbleu	.50	**.75**	.71	.65
bleu	.50	.72	.75	.65
mter	.38	.73	.68	.60
svm-rank	.37	.10	.22	.23

Table 1: Average system-level correlations for the automatic evaluation metrics on translations into English.

	Rank	Const	Yes/No	Overall
posbleu	**.75**	.78	.80	**.78**
posF4gram-gm	.74	**.80**	.79	**.78**
posF4gram-am	.74	**.80**	.79	**.78**
bleu	.68	.79	.79	.75
NCD-bz2	.66	.75	.77	.73
svm-rank	.66	.73	.80	.73
NCD-ppmz	.64	.75	.80	.73
NCD-zlib	.64	.74	.80	.73
NCD-gzip	.63	.74	**.81**	.72
mbleu	.63	**.80**	**.81**	.75
meteor-baseline	.58	.78	.76	.71
meteor-ranking	.55	.74	.74	.68
mter	.52	.69	.73	.65

Table 2: Average system-level correlations for the automatic evaluation metrics on translations from English into French, German and Spanish.

	Into English	From English
ULC	**.65**	-
ULCh	.64	-
NCD-zlib	.63	**.61**
NCD-ppmz	.63	.60
NCD-gzip	.63	.61
svm-human-ref	.62	-
alignment-prob	.62	-
NCD-bz2	.61	.58
DP	.60	-
posF4gram-am	.60	.59
DR	.59	-
svm-pseudo-ref	.59	-
svm-rank	.58	.57
meteor-ranking	.56	.54
meteor-baseline	.56	.54
mbleu	.55	.53
SR	.55	-
posbleu	.50	.51
posF4gram-gm	.49	.50
mter	.48	.45

Table 3: Sentence-level analysis as the fraction of time that each automatic evaluation metric was consistent with human Rank judgements.

Results in Table 1 give NCD the highest correlation against other MT metrics that do not use language specific resources in the Rank and Yes/No classes, as well as in the overall average. In Table 2, NCD has slightly smaller correlations than BLEU. Sentence-level analysis results in Table 3 show NCD in the top together with the most sophisticated metrics. NCD performs best for translations from English, where several other top performing metrics can not be used due to missing special language resources. In the sentence-level analysis, NCD shows its strength as language independent method despite the demanding task to compress a very short text segment.

We present two main conclusions from these initial results: 1) The choice of the compressor does not seem to influence the performance of NCD as an MT evaluation metric. This holds only if the sentence lists are interleaved rather than concatenated in the joint compression. This requires more comprehensive testing with more compressors, but is an interesting result nonetheless. 2) The NCD metric performs roughly at the same level as, e.g., BLEU and METEOR in system-level analysis and slightly better in sentence-level analysis. A break-down of the results into different tasks (not shown here because of limited space) supports these conclusions.

Further analysis is required to verify which differences in the computed correlations and consistencies are significant. Preliminary investigation of confidence intervals for Spearman's rank correlation coefficient for the Rank class suggests that the correlations of the different MT metrics are overlapping considerably.

4 Discussion and Conclusions

We expanded the MT metric evaluation results in Callison-Burch et al. (2008) to include Normalized Compression Distance as one of the evaluation metrics. NCD has been shown to work in many real-world applications that range from bioinformatics to music clustering (Vitanyi et al., 2009). The results of its use as MT evaluation metric suggest that NCD correlates to human judgements approximately at the same level as the widely used BLEU metric. Further studies are required for measuring which differences are significant.

In the consideration of automatic MT evaluation, one should keep in mind how an MT metric computes its score and what the metric actually measures. An MT metric usually compares output of an MT system to one or several human reference translations by means of string level comparison, i.e., comparing aligned n-grams of MT output to aligned n-grams of the reference translation. The final score is a weighted combination over the whole sentence, possibly including heuristic regularization.

Basic ideas for these were given already by Thompson (1991) and later developed and varied in many systems now in use, such as BLEU, NIST and METEOR etc. What should specifically be kept in mind with respect to MT metrics is, that they are mainly tools for consistent MT system development, and may not have that much to do with real quality of translation. Culy and Riehemann (2003) state this nicely: ”A final important point is a reminder that the n-gram metrics are really document similarity measures rather than true translation quality measures.”

Therefore, this might explain why NCD, which is a general similarity measure, correlates so closely to n-gram based MT metrics such as BLEU, METEOR etc. used in the WMT’08 evaluations. We do not suggest that NCD overcomes all the difficulties related to automated MT metrics, but it offers clear benefits. The special advantage of NCD is that it is an information theoretic general measure of similarity. It is parameter free and works with character strings instead of word n-grams, and thus is also language independent and possibly more robust in regards to morphological variation in languages.

References

Bennett, C. H., Gács, P., Li, M., Vitanyi, P. M. B., and Zurek, W. H. (1998). Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2008). Further meta-evalutation of machine translation. *ACL Workshop on Statistical Machine Translation*.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. In EACL-2006: *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy.

Cebrian, M., Alfonseca, M., and Ortega, A. (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.

Cilibrasi, R. and Vitanyi, P. (2005). Clustering by compression. *IEEE Transactions on Information Theory*, 51:1523–1545.

Culy, C. and Riehemann, S. Z. (2003). The limits of n-gram translation evaluation metrics. *In Proceedings of MT Summit IX*, pages 71–78.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *Proceedings of Human Language Technology conference (HLT-2002)*, San Diego, California.

Giménez, J. and Amigó, E. (2006). IQMT: A framework for automatic machine translation evaluation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*.

Kettunen, K. (2009). Packing it all up in search for a language independent MT quality measure tool. In *LTC'09, 4th Language & Technology Conference*, Poznan.

Lavie, A. and Agarwal, A. (2007). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Li, M., Chen, X., Li, X., Ma, B., and Vitanyi, P. (2004). The similarity metric. *Information Theory, IEEE Transactions on*, 50(12):3250–3264.

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

Parker, S. (2008). Badger: A new machine translation metric. In *Metrics for Machine Translation Challenge 2008*, Waikiki, Hawai'i. AMTA.

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

Thompson, H. (1991). Automatic evaluation of translation quality: Outline of methodology and report on pilot experiment. In *Proceedings of the Evaluators' Forum*, pages 215–223.

Turian, J. P., Shen, L., and Melamed, I. D. (2003). Evaluation of machine translation and its evaluation. In *Proceedings of the Machine Translation Summit IX*, pages 386–393, New Orleans, USA.

Vitanyi, P. M. B., Balbach, F. J., Cilibrasi, R. L., and Li, M. (2009). *Information Theory and Statistical Learning*, chapter Normalized Information Distance, pages 45–82. Springer, first edition.

Acknowledgements

We thank Prof. Chris Callison-Burch for the helpful communications and for sharing the complete WMT'08 evaluation data with us.