Packing it all up in search for a language independent MT quality measure tool

Kimmo Kettunen

Kymenlaakso University of Applied Sciences
Kouvola, Finland
kimmo.kettunen@kyamk.fi

Abstract

This study describes usage of a particular implementation of Normalized Compression Distance (NCD) as a machine translation quality evaluation tool. NCD has been introduced and tested for clustering and classification of different types of data and found a reliable and general tool. As far as we know NCD in its Complearn implementation has not been evaluated as a MT quality tool yet, and we wish to show that it can also be used for this purpose. We show that NCD scores given for MT outputs in different languages correlate highly with scores of a state-of-the-art MT evaluation metrics, METEOR 0.6. Our experiments are based on translations between one source and three target languages with a smallish sample that has available reference translations, UN's Universal Declaration of Human Rights. Results of the paper are preliminary, but very promising. We have also begun a large scale evaluation of NCD as an MT metric with WMT-08 Shared Task Evaluation Data.

Keywords: normalized compression distance, evaluation of machine translation

1. Introduction

Automatic evaluation of machine translation program output has been developed and used for about a decade. As a result of work done we have now available several MT evaluation systems or metrics, such as BLEU (Papineni et al, 2002), NIST (Doddington, 2002), METEOR (Lavie, Agarwal, 2007), IQ_{MT} (Giménez , 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. Many of the programs have a quite high degree of correlation with human judgements of translations and they have been a valuable tool in making especially the statistical MT systems better.

It is also well known that all the present MT evaluation programs have limitations. They may, e.g., be language dependent, i.e. they need to be tuned for specific language pairs to be able to perform or use language specific tools (stemmers, Wordnets). Also more severe concerns about MT metrics have been stated. Callison-Burch, Osborne and Koehn (2006) showed 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". MT metrics' correlation with human judgements of translations has also been disputed (Turian et al., 2003).

We show in this paper that an alternative language independent measure for MT evaluation can be obtained from a general classification and clustering tool called Normalized Compression Distance, NCD (Cilibrasi, Vitanyi, 2005, 2007; Li et al, 2004 ; the Complearn software package is available from http://www.complearn.org/download.html). As the Results section shows, the scores given by Complearn implementation of NCD to translations correlate very highly with METEOR 0.6 scores in three different target languages with 10-12 MT systems for each language pair.

2. Research setting

We evaluated En → {De, Es, Fr} translations of UN's Universal Declaration of Human Rights (UDHR) (http://www.un.org/Overview/rights.html) of 10-12 MT programs that were either freely available or could be used with an evaluation license. The used MT programs were Promt, Google Translate Beta, Babelfish, Translate It!, LEC Translate2Go, SDL Enterprise Translation Server, Systran, InterTran, Translated, Hypertrans, MZ-Win Translator, Dictionary.com, and Translendium. MZ-Win and Translate It! translated from English to German only, all others to all the three target languages. Translations were performed in late March 2009. If the web service of the MT system had limitations in the number of words to be translated, the text was split to smaller chunks, e.g. 5-10 articles. UN's Universal Declaration of Human Rights has 30 numbered articles and its length in English original is 1451 words and 60 sentences (7.3 Kb without spaces). Lengths of the articles vary quite a lot, some having only one sentence and others several sentences or even paragraphs. The used text is quite short and homogeneous textually, but can be considered to be long and representative enough for our preliminary experiments.

To get a baseline of the translation quality of the MT programs we evaluated the translation results of the MT systems with a state-of-the-art machine translation evaluation metrics, METEOR 0.6 (Lavie, Agarwal, 2007; Banerjee, Lavie, 2005). METEOR is based on a BLEU like evaluation idea: output of the MT program is compared to a given reference translation, which is usually a human translation. METEOR's most significant difference to BLEU like systems is, that it emphasizes more recall than precision of translations (Lavie, Sagae, Jayarman, 2004). The evaluation metric was run with exact match, where translations are compared to reference translation as such. Basically "METEOR evaluates a translation by computing a score based on explicit word-to-word matches between the translation

and a given reference translation". When "given a pair of strings to be compared, METEOR creates a word alignment between the two strings. An alignment is a mapping between words, such that every word in each string maps to most one word in the other string. This alignment is incrementally produced by a sequence of word-mapping modules. The 'exact' module maps two words if they are exactly the same." METEOR has been shown to outperform commonly used metrics BLEU and NIST in terms of correlations with human judgements of translation quality (Lavie, Agarwal, 2007).

Our suggested new MT quality measurement tool, Normalized Compression Distance, is based on the work of Rudi Cilibrasi, Paul Vitányi and others (Cilibrasi, Vitányi, 2005, 2007; Li et al, 2004). The method is the outcome of mathematical developments, that are based on the notion of *Kolmogorov complexity*. Informally, for any sequence of symbols, the Kolmogorov complexity of the sequence is the length of the shortest algorithm that will exactly generate the sequence (and then stop). In other words, the more predictable the sequence, the shorter the algorithm needed is and thus the Kolmogorov complexity of the sequence is also lower (Li et al, 2004; Li, Vitányi, 1997).

Kolmogorov complexity itself is uncomputable, but file compression programs can be used to approximate the Kolmogorov complexity of a given file. A more complex string (in the sense of Kolmogorov complexity) will be less compressible. From this approach grew first normalized information distance, NID, (Li et al, 2004) and as its approximation NCD using a real compressor.

NCD's basic formula for counting the distance (and thus similarity) between two files is as follows (What is NCD, 2009):

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

Here C is the compressor, C(x) denotes the length of the compressed version of a string x, and C(xy) is a compressed concatenation of the pair (x,y). As a result, NCD gives a score between [0,1] for the strings (files) that are compared. Smaller numbers represent more similar files.

In plain words, NCD uses the lengths of the compressed hypothesis and reference strings, computing a ratio of the compressed length of the concatenated reference and hypothesis (minus the length of the shorter compressed reference or hypothesis sequence) to the length of the longer compressed reference or hypothesis sequence. The basic idea is that if the two sequences A and B are more similar, then B will compress with a smaller 'description' when combined with A than it would when compressed separately. Since compression 'descriptions' for text are typically based on frequencies of character sequences, the compression lengths can serve as a similarity measure for the words in the hypothesis and reference(s).

Parker (2008) has earlier introduced an MT metric named BADGER that utilizes also NCD as one part of the metric. BADGER does not use Complearn's NCD package, but implements NCD using the Burrows Wheeler Transformation as compressor, which enables the system to take into account more sentence context. BADGER uses also some language independent word normalization methods, such as Holographic Reduced representation, which utilizes binary vectors and relative distance counting with cosine similarity. Thus BADGER is more advanced than a bare NCD metric. Parker benchmarked BADGER against METEOR and word error rate metrics (WER). The correlation of BADGER results to those of METEOR were low and correlations to WER high. The used test set was Arabic to English translations. Author considers the results preliminary and wishes to do more testing with the software.

3. Results

Translations of MT systems were compared to one human reference translation with both METEOR 0.6 and Complearn NCD. In our case the reference translations were the French, German and Spanish translations of the Universal Declaration of Human rights from UN's web page (http://www.ohchr.org/EN/UDHR/Pages/SearchByLang.aspx). Table 1 shows a short example of results of METEOR's evaluations for three of the English → German MT outputs in their raw form. The compared sequence with METEOR was one article, and the overall system score in the table gives a unified score for all the 30 sequences.

	Google	**Babelfish**	**Promt**
Overall system score	0.66	0.21	0.25
Precision	0.82	0.54	0.56
Recall	0.82	0.57	0.60
Fmean	0.82	0.56	0.59
Penalty	0.20	0.64	0.58

Table 1. Example results of METEOR translation evaluation for En → De translations

The meanings of the METEOR scores in Table 1 are as follows:

1 *Overall system score* gives a combined figure for the result. It is computed as follows (Lavie, Sagae, Jayarman, 2004): Score = Fmean * (1- Penalty).

2 (Unigram) *Precision* = unigram precision 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.

3 (Unigram) *Recall* = unigram recall 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.

4 *Fmean:* precision and recall are combined via harmonic mean that places most of the weight on recall. The present formulation of Fmean is stated in Lavie, Agarwal and Jayarman (2004) as follows: Fmean = P * R/ α * P + (1- α) * R.

5 *Penalty:* This figure takes into account the extent to which the matched unigrams in the two strings are in the same word order.

Table 2 lists the scores given by METEOR 0.6 and NCD for all translations in German. It should be noted, that the scale of METEOR and NCD are reverse: bigger score in METEOR means better translation quality whereas smaller score in NCD means greater similarity and thus better translation.

En → De	**METEOR**	**NCD**
Babelfish	0,21	0,75
Dictionary.com	0,21	0,74
SDL Enterprise Translation Service	0,26	0,74
Google Translate Beta	**0,66**	**0,40**
Hypertrans	0,20	0,77
InterTran	**0,11**	**0,86**
LEC Translate2Go	0,27	0,74
MZ-Win	0,22	0,76
Promt	0,25	0,73
Systran	0,23	0,74
Translate It!	0,26	0,78
Translendum	0,24	0,75
Average	0,26	0,73
Standard deviation	0,11	0,13
Correlation co-efficient NCD vs. METEOR	-0,98	

Table 2**.** Scores for En →De MT translations compared with human reference translations

Google's German translation was given the best score (0.66) by METEOR and translation of Intertran the worst (0.11). All the others were given a score between 0.20 and 0.27. NCD's scores for translations follow the same pattern.

The last row of Table 2 shows that scores for both analyses for all translations correlate highly. The correlation seems negative, but if the scores are transformed to equal scale (this can be done by subtracting NCD score from 1, 1-NCD), the correlation is positive. Furthermore we see, that both measures indicate clearly the worst (InterTran) and best (Google) MT programs. The middle area is quite even, and there seems to be no big differences in the translation quality of other systems.

Tables 3 and 4 show results of Spanish and French translation evaluations.

En → Es	**METEOR**	**NCD**
Babelfish	0,26	0,72
Dictionary.com	0,26	0,71
SDL Enterprise Translation Service	0,27	0,72
Google Translate Beta	**0,54**	**0,39**
Hypertrans	0,22	0,77
InterTran	**0,19**	**0,82**
LEC Translate2Go	0,26	0,73
MZ-Win	N/A	N/A
Promt	0,28	0,69
Systran	0,27	0,70
Translate It!	N/A	N/A
Translendum	0,26	0,70
Average	0,28	0,70
Standard deviation	0,09	0,11
Correlation co-efficient NCD vs. METEOR	-0,995	

Table 3. Scores for En →Es MT translations compared with human reference translations

En → Fr	**METEOR**	**NCD**
Babelfish	0,17	0,71
Dictionary.com	0,17	0,69
SDL Enterprise Translation Service	0,11	0,74
Google Translate Beta	**0,45**	**0,38**
Hypertrans	0,12	0,76
InterTran	**0,06**	**0,85**
LEC Translate2Go	0,14	0,72
MZ-Win	N/A	N/A
Promt	0,13	0,72
Systran	0,18	0,69
Translate It!	N/A	N/A
Translendum	0,15	0,71
Average	0,17	0,70
Standard deviation	0,10	0,12
Correlation co-efficient NCD vs. METEOR	-0,99	

Table 4. Scores for En →Fr MT translations compared with human reference translations

For Spanish and French translations both METEOR and NCD were again able to distinguish the best and worst translations. Also scores for En → Es and En → Fr translations correlated highly: for Spanish translations the correlation was 0.995 and for French 0.99.

We also have other independent data that strengthens our case. Kettunen (2009a, 2009b) shows that METEOR, NIST and BLEU scores of MT output all correlate well with mean average precisions of Cross-language information retrieval runs, thus confirming the bond between translation quality and CLIR result achieved by others, e.g. Kishida (2008) and Zhu and Wang (2006).

The MAP figures for the CLIR runs of Kettunen (2009a) were now tested with NCD scores, and a high correlation of 0.91 was found between NCD scores for translations and achieved MAP results of 56 German title and topic translations from English to German. Thus, even though our data samples are small, two different evaluation settings show a high correlation between METEOR and NCD score sets both for three target language translations and one target language translation score set and MAP scores gained in CLIR evaluation of the translated queries.

4. Discussion and conclusions

The aim of this paper was to introduce usage of a general classifier and clustering program, Normalized Compression Distance, and its specific Complearn implementation, as an MT evaluation tool. For performance testing we used three target language translations of the Universal Declaration of Human Rights with 10-12 different available MT systems for each target language. We first evaluated the quality of the translations with a language specific state-of-the-art MT metrics, METEOR 0.6, by using available reference translations as comparison. After that we compared the translations to references with NCD. The scores given by METEOR and NCD correlated very highly in all three target languages (0.99-0.995). Based on the preliminary results of our study Complearn NCD seems to be as good as METEOR 0.6 in the sense that it can pick the outliers of MT systems, the best and the worst translators, very well. Google's outstanding translation performance with all the language pairs might have a very simple explanation: the parallel texts of UDHR are most obviously in the training corpus of the Google MT system, and thus it is able to outperform all the other systems with such a margin.

Those systems that produce midrange output are not very well separated by METEOR, and they are also given quite similar scores by NCD, as was seen in Tables 2-4. This, in turn, is a specific problem of MT metrics: we do not know what the real meaning of the differences in scores is, and thus we can not really say, if a small score difference really matters (Turian et al, 2003).

We believe that our results, although preliminary with respect to amount of data and language pairs, are very promising. It is clear, that further studies are needed with more data and more reference MT evaluation systems and human judgements of translation quality, but as the current MT quality metrics mostly work in quite a similar manner, we expect that correlation of NCD scores to other MT evaluation systems will also be clear. 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 feature- and parameter-free. As it works with character strings instead of word n-grams, it is also language independent and possibly more robust in regards to morphological variation in languages. It has already been shown to work in many real-world applications that range from bioinformatics to music clustering. (Vitányi et al, 2009). This gives it an advantage with respect to common MT evaluation systems that might need parameter setting and are usually n-gram based, and sometimes language dependent.

One of the problems of n-gram based metrics is the assumption "that a good translation will be similar to other good translations" (Culy, Riehemann, 2003). Everything an MT metric actually does is just to compare the sameness of the n-gram distribution of the translations to reference translation(s), and the result is thus basically a measure of document similarity. This, as such, may not tell much about the translation's goodness itself (Culy, Riehemann, 2003). Whether a low score given by NCD to translation really indicates the quality of translation, is also left somehow open. In this respect NCD is similar to common measures of MT quality.

Some of the possible problems of NCD are discussed in Cebrián et al. (2005). The authors show that the compressors used in NCD are strongly skewed with the size of the objects and window size, which causes deviation in the identity property of the distance if care is not taken that the objects to be compressed fit the windows. However, after testing NCD with different compressors with Calgary corpus, a well known benchmark for compression algorithms, the authors conclude that NCD is a very good distance measurement when used in a proper way (i.e. with a compressor the results of which do not depend on the size of the objects, such as PPMZ). We believe that larger MT translation data sets compared with more existent MT metrics and human judgement scores will also show the suitability of the approach. We have recently begun a large scale evaluation of the NCD as MT metrics with the WMT-08 Shared Task Evaluation Data available at http://www.statmt.org/wmt08/results.html (cf. Callison-Burch et al., 2008) as a joint project with Adaptive Informatics Research Centre of the Laboratory of Information and Computer Science at the Helsinki University of Technology. The first results of the evaluation are published in Väyrynen et al. (2009), and they show that NCD correlates also relatively well with human judgements of translations. We shall continue evaluation of NCD with the WMT-08 data and compare performance of NCD to several other MT metrics and human judgements of translations and also test different aspects of NCD, such as different compressors, more language pairs etc., with respect to MT evaluation. Some of the most interesting aspects of MT evaluation, influence of word order and morphological complexity of the language to NCD's performance will also be tackled. The possibility to use more than one reference sentence in NCD evaluation needs also to be considered.

References

Banerjee S., Lavie, A. (2005). METEOR: Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.* Ann Arbor, pp. 65-72.

Callison-Burch, C., Osborne, M., Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In: *EACL* 2006, pp. 249-256.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. Schroeder, J. (2008). Further Meta-Evaluation of Machine Translation. In: *Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 70–106.

Cebrián, M., Alfonseca, M., Ortega, M. (2005). Common Pitfalls Using The Normalized Compression Distance: What To Watch Out For In A Compressor. *Communications In Information And Systems* 5, pp. 367-384.

Cilibrasi, R., Vitányi, P.M.B. (2005) Clustering by Compression. *IEEE Transactions on Information Theory 51*, pp. 1523–1545.

Cilibrasi, R., Vitányi P.M.B. (2007). The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering* 19, pp. 370–383.

Culy, C., Riehemann, S.Z. (2003). The Limits of N-gram Translation Metrics. In: *MT Summit IX*, New Orleans, USA, 23-27 September 2003; pp. 71-78.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics. In: *Proceedings of the Second International Conference on Human Language Technology Research,* pp. 138-145.

Giménez, J., Amigó, E. (2006). IQ_{MT}: A Framework for Automatic Machine Translation Evaluation. In: *Proceedings of the 5th LREC*.

Kettunen, K. (2009a). Choosing the best MT programs for CLIR purposes - can MT metrics be helpful? In: M. Boughanem et al. (Eds.): *ECIR 2009*, LNCS 5478, pp. 706–712.

Kettunen, K. (2009b**).** Facing The Machine Translation Babel in CLIR - Can MT Metrics Help in Choosing CLIR Resources? In: *Recent Advances in Intelligent Information Systems*, M. A. Klopotek, A. Przepiórkowski, S. T. Wierzchon, K. Trojanowski (Eds.), IIS 2009, pp. 103–116.

Kishida, K. (2008). Prediction of performance of cross-language information retrieval system using automatic evaluation of translation. *Library & Information Science Research* 30, pp. 138–144.

Lavie, A., Agarwal, A. (2007) METEOR: An automatic Metric for MT Evaluation with High Levels of Correlation with Human judgements. In*: Proceedings of the Second Workshop on Statistical Machine Translation,* Prague, June 2007, pp. 228–231.

Lavie, A., Agarwal, A. (2007) The METEOR Automatic Machine Translation Evaluation System. Retrieved from: http://www.cs.cmu.edu/~alavie/METEOR/. Access date: July 15, 2009.

Lavie, A., Sagae, K., Jayarman, S. (2004) The Significance of Recall in Automatic Metrics for MT Evaluation. In: *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004)*, Washington, DC, pp. 134–143.

Li, M., Chen, X., Li, X. et al (2004). The Similarity Metric. *IEEE Transactions on Information Theory* 50, pp. 3250–3264.

Li, M., Vitányi P.M.B. (1997). *An Introduction to Kolmogorov Complexity and its Applications. Second edition.* Springer Verlag, New York Berlin Heidelberg.

Papineni, K., Roukos, S, Ward, T., et al (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)*, pp. 311–318.

Parker, S. (2008). BADGER: A New Machine Translation Metric. Retrieved from: http://www.amtaweb.org/papers/badger_babblequest_matr08.pdf. Access date May 15, 2009.

Turian, J.P., Shen, L., Melamed, D. (2003). Evaluation of machine translation and its evaluation. In: *MT Summit IX,* New Orleans, USA, 23-27 September 2003, pp. 386-393.

What is NCD? (2009). Retrieved from: http://www.complearn.org/ncd.html. Access date July 15, 2009.

Vitányi, P.M.B., Balbach, F. J., Cilibrasi, R., Ming, L. (2009). Normalized Information Distance. In: F. Emmert-Streib, M. Dehmer (eds.), *Information Theory and Statistical Learning,* Springer, pp. 39–71.

Väyrynen, J., Tapiovaara, T., Kettunen, K., Dobrinkat, M. 2009. Normalized Compression Distance as an automatic MT evaluation metric. In: *Machine Translation Twenty-Five Years On,* Cranfield, U.K.

Zhu, J., Wang, H. (2006). The Effect of Translation Quality in MT-Based Cross-Language Information Retrieval. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th annual Meeting of the ACL*, pp. 593–600.