Evaluating Machine Translations using mNCD

Marcus Dobrinkat and **Tero Tapiovaara** and **Jaakko Väyrynen**
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

This paper introduces mNCD, a method for automatic evaluation of machine translations. The measure is based on normalized compression distance (NCD), a general information theoretic measure of string similarity, and flexible word matching provided by stemming and synonyms. The mNCD measure outperforms NCD in system-level correlation to human judgments in English.

1 Introduction

Automatic evaluation of machine translation (MT) systems requires automated procedures to ensure consistency and efficient handling of large amounts of data. In statistical MT systems, automatic evaluation of translations is essential for parameter optimization and system development. Human evaluation is too labor intensive, time consuming and expensive for daily evaluations. However, manual evaluation is important in the comparison of different MT systems and for the validation and development of automatic MT evaluation measures, which try to model human assessments of translations as closely as possible. Furthermore, the ideal evaluation method would be language independent, fast to compute and simple.

Recently, normalized compression distance (NCD) has been applied to the evaluation of machine translations. NCD is a general information theoretic measure of string similarity, whereas most MT evaluation measures, e.g., BLEU and METEOR, are specifically constructed for the task. Parker (2008) introduced BADGER, an MT evaluation measure that uses NCD and a language independent word normalization method. BADGER scores were directly compared against the scores of METEOR and word error rate (WER). The correlation between BADGER and METEOR were low and correlations between BADGER and WER high. Kettunen (2009) uses the NCD directly as an MT evaluation measure. He showed with a small corpus of three language pairs that NCD and METEOR 0.6 correlated for translations of 10–12 MT systems. NCD was not compared to human assessments of translations, but correlations of NCD and METEOR scores were very high for all the three language pairs.

Väyrynen et al. (2010) have extended the work by including NCD in the ACL WMT08 evaluation framework and showing that NCD is correlated to human judgments. The NCD measure did not match the performance of the state-of-the-art MT evaluation measures in English, but it presented a viable alternative to de facto standard BLEU (Papineni et al., 2001), which is simple and effective but has been shown to have a number of drawbacks (Callison-Burch et al., 2006).

Some recent advances in automatic MT evaluation have included non-binary matching between compared items (Banerjee and Lavie, 2005; Agarwal and Lavie, 2008; Chan and Ng, 2009), which is implicitly present in the string-based NCD measure. Our motivation is to investigate whether including additional language dependent resources would improve the NCD measure. We experiment with relaxed word matching using stemming and a lexical database to allow lexical changes. These additional modules attempt to make the reference sentences more similar to the evaluated translations on the string level. We report an experiment showing that document-level NCD and aggregated NCD scores for individual sentences produce very similar correlations to human judgments.

Proceedings of the ACL 2010 Conference Short Papers, pages 80–85,
Uppsala, Sweden, 11-16 July 2010.

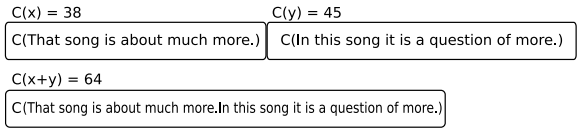

Figure 1: An example showing the compressed sizes of two strings separately and concatenated.

2 Normalized Compression Distance

Normalized compression distance (NCD) is a similarity measure based on the idea that a string x is similar to another string y when both share substrings. The description of y can reference shared substrings in the known x without repetition, indicating shared information. Figure 1 shows an example in which the compression of the concatenation of x and y results in a shorter output than individual compressions of x and y.

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

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

NCD computes the distance as a score closer to one for very different strings and closer to zero for more similar strings.

NCD is an approximation of the uncomputable normalized information distance (NID), a general measure for the similarity of two objects. NID is based on the notion of Kolmogorov complexity $K(x)$, a theoretical measure for the information content of a string x, 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 is an upper bound of the Kolmogorov complexity $K(x)$.

3 mNCD

Normalized compression distance was not conceived with MT evaluation in mind, but rather it is a general measure of string similarity. Implicit non-binary matching with NCD is indicated by preliminary experiments which show that NCD is less sensitive to random changes on the character level than, for instance, BLEU, which only counts the exact matches between word n-grams. Thus comparison of sentences at the character level could account better for morphological changes.

Variation in language leads to several acceptable translations for each source sentence, which is why multiple reference translations are preferred in evaluation. Unfortunately, it is typical to have only one reference translation. Paraphrasing techniques can produce additional translation variants (Russo-Lassner et al., 2005; Kauchak and Barzilay, 2006). These can be seen as new reference translations, similar to pseudo references (Ma et al., 2007).

The proposed method, mNCD, works analogously to M-BLEU and M-TER, which use the flexible word matching modules from METEOR to find relaxed word-to-word alignments (Agarwal and Lavie, 2008). The modules are able to align words even if they do not share the same surface form, but instead have a common stem or are synonyms of each other. A similarized translation reference is generated by replacing words in the reference with their aligned counterparts from the translation hypothesis. The NCD score is computed between the translations and the similarized references to get the mNCD score.

Table 1 shows some hand-picked German–English candidate translations along with a) the reference translations including the 1-NCD score to easily compare with METEOR and b) the similarized references including the mNCD score. For comparison, the corresponding METEOR scores without implicit relaxed matching are shown.

4 Experiments

The proposed mNCD and the basic NCD measure were evaluated by computing correlation to human judgments of translations. A high correlation value between an MT evaluation measure and human judgments indicates that the measure is able to evaluate translations in a more similar way to humans.

Relaxed alignments with the METEOR modules `exact`, `stem` and `synonym` were created for English for the computation of the mNCD score. The `synonym` module was not available with other target languages.

4.1 Evaluation Data

The 2008 ACL Workshop on Statistical Machine Translation (Callison-Burch et al., 2008) shared task data includes translations from a total of 30 MT systems between English and five European languages, as well as automatic and human trans-

	Candidate C/ Reference R/ Similarized Reference S	1-NCD	METEOR
C	There is no effective means to stop a Tratsch, which was already included in the world.		
R	There is no `good way` to `halt` gossip that has already begun to spread.	.41	.31
S	There is no `effective means` to `stop` gossip that has already begun to spread.	.56	.55
C	Crisis, not only in America		
R	A Crisis `Not Only` in the `U.S.`	.51	.44
S	A Crisis `not only` in the `America`	.72	.56
C	Influence on the whole economy should not have this crisis.		
R	Nevertheless, the crisis should not have `influenced` the entire economy.	.60	.37
S	Nevertheless, the crisis should not have `Influence` the entire economy.	.62	.44
C	Or the lost tight meeting will be discovered at the hands of a gentlemen?		
R	Perhaps you `see` the pen you thought you lost lying on your colleague's desk.	.42	.09
S	Perhaps you `meeting` the pen you thought you lost lying on your colleague's desk.	.40	.13

Table 1: Example German–English translations showing the effect of relaxed matching in the 1-mNCD score (for rows S) compared with METEOR using the `exact` module only, since the modules `stem` and `synonym` are already used in the similarized reference. Replaced words are emphasized.

lation evaluations for the translations. There are several tasks, defined by the language pair and the domain of translated text.

The human judgments include three different categories. The RANK category has human quality rankings of five translations for one sentence from different MT systems. The CONST category contains rankings for short phrases (constituents), and the YES/NO category contains binary answers if a short phrase is an acceptable translation or not.

For the translation tasks into English, the relaxed alignment using a `stem` module and the `synonym` module affected $7.5\,\%$ of all words, whereas only $5.1\,\%$ of the words were changed in the tasks from English into the other languages.

The data was preprocessed in two different ways. For NCD we kept the data as is, which we called real casing (rc). Since the used METEOR align module lowercases all text, we restored the case information in mNCD by copying the correct case from the reference translation to the similarized reference, based on METEOR's alignment. The other way was to lowercase all data (lc).

4.2 System-level correlation

We follow the same evaluation methodology as in Callison-Burch et al. (2008), which allows us to measure how well MT evaluation measures correlate with human judgments on the system level.

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

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

where for each system i, d_i is the difference between the rank derived from annotators' input and the rank obtained from the measure. 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 judgment.

We computed system-level correlations for tasks with English, French, Spanish and German as the target language[1].

5 Results

We compare mNCD against NCD and relate their performance to other MT evaluation measures.

5.1 Block size effect on NCD scores

Väyrynen et al. (2010) computed NCD between a set of candidate translations and references at the same time regardless of the sentence alignments, analogously to document comparison. We experimented with segmentation of the candidate translations into smaller blocks, which were individually evaluated with NCD and aggregated into a single value with arithmetic mean. The resulting system-level correlations between NCD and human judgments are shown in Figure 2 as a function of the block size. The correlations are very similar with all block sizes, except for Spanish, where smaller block size produces higher correlation. An experiment with geometric mean produced similar results. The reported results with mNCD use maximum block size, similar to Väyrynen et al. (2010).

[1]The English-Spanish news task was left out as most measures had negative correlation with human judgments.

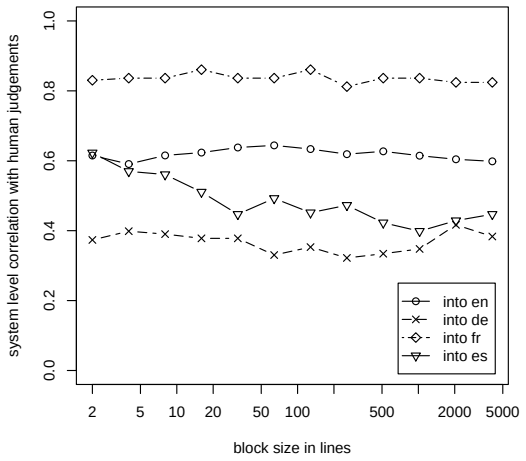

Figure 2: The block size has very little effect on the correlation between NCD and human judgments. The right side corresponds to document comparison and the left side to aggregated NCD scores for sentences.

5.2 mNCD against NCD

Table 2 shows the average system level correlation of different NCD and mNCD variants for translations into English. The two compressors that worked best in our experiments were PPMZ and bz2. PPMZ is slower to compute but performs slightly better compared to bz2, except for the lowercased CONST category.

Method	Parameters		RANK	CONST	YES/NO	Mean
mNCD	PPMZ	rc	**.69**	**.74**	**.80**	**.74**
NCD	PPMZ	rc	.60	.66	.71	.66
mNCD	bz2	rc	**.64**	**.73**	**.73**	**.70**
NCD	bz2	rc	.57	.64	.69	.64
mNCD	PPMZ	lc	**.66**	**.80**	**.79**	**.75**
NCD	PPMZ	lc	.56	.79	.75	.70
mNCD	bz2	lc	**.59**	**.85**	**.74**	**.73**
NCD	bz2	lc	.54	.82	.71	.69

Table 2: Mean system level correlations over all translation tasks into English for variants of mNCD and NCD. Higher values are emphasized. Parameters are the compressor PPMZ or bz2 and the preprocessing choice lowercasing (lc) or real casing (rc).

			Target Lang Corr			
Method	Parameters		EN	DE	FR	ES
mNCD	PPMZ	rc	**.69**	.37	.82	.38
NCD	PPMZ	rc	.60	.37	**.84**	**.39**
mNCD	bz2	rc	**.64**	.32	.75	.25
NCD	bz2	rc	.57	**.34**	**.85**	**.42**
mNCD	PPMZ	lc	**.66**	.33	**.79**	**.23**
NCD	PPMZ	lc	.56	**.37**	.77	.21
mNCD	bz2	lc	**.59**	.25	**.78**	**.16**
NCD	bz2	lc	.54	**.26**	.77	.15

Table 3: mNCD versus NCD system correlation RANK results with different parameters (the same as in Table 2) for each target language. Higher values are emphasized. Target languages DE, FR and ES use only the `stem` module.

Table 2 shows that real casing improves RANK correlation slightly throughout NCD and mNCD variants, whereas it reduces correlation in the categories CONST, YES/NO as well as the mean.

The best mNCD (PPMZ rc) improves the best NCD (PPMZ rc) method by 15% in the RANK category. In the CONST category the best mNCD (bz2 lc) improves the best NCD (bz2 lc) by 3.7%. For the total average, the best mNCD (PPMZ rc) improves the the best NCD (bz2 lc) by 7.2%.

Table 3 shows the correlation results for the RANK category by target language. As shown already in Table 2, mNCD clearly outperforms NCD for English. Correlations for other languages show mixed results and on average, mNCD gives lower correlations than NCD.

5.3 mNCD versus other methods

Table 4 presents the results for the selected mNCD (PPMZ rc) and NCD (bz2 rc) variants along with the correlations for other MT evaluation methods from the WMT'08 data, based on the results in Callison-Burch et al. (2008). The results are averages over language pairs into English, sorted by RANK, which we consider the most significant category. Although mNCD correlation with human evaluations improved over NCD, the ranking among other measures was not affected. Language and task specific results not shown here, reveal very low mNCD and NCD correlations in the Spanish-English news task, which significantly

Method	RANK	CONST	YES/NO	Mean
DP	.81	.66	.74	.73
ULCh	.80	.68	.78	.75
DR	.79	.53	.65	.66
meteor-ranking	.78	.55	.63	.65
ULC	.77	.72	.81	.76
posbleu	.75	.69	.78	.74
SR	.75	.66	.76	.72
posF4gram-gm	.74	.60	.71	.68
meteor-baseline	.74	.60	.63	.66
posF4gram-am	.74	.58	.69	.67
mNCD (PPMZ rc)	.69	.74	.80	.74
NCD (PPMZ rc)	.60	.66	.71	.66
mbleu	.50	.76	.70	.65
bleu	.50	.72	.74	.65
mter	.38	.74	.68	.60
svm-rank	.37	.10	.23	.23
Mean	.67	.62	.69	.66

Table 4: Average system-level correlations over translation tasks into English for NCD, mNCD and other MT evaluations measures

Method	Target Lang Corr			
	DE	FR	ES	Mean
posbleu	.75	.80	.75	.75
posF4gram-am	.74	.82	.79	.74
posF4gram-gm	.74	.82	.79	.74
bleu	.47	.83	.80	.68
NCD (bz2 rc)	.34	.85	.42	.66
svm-rank	.44	.80	.80	.66
mbleu	.39	.77	.83	.63
mNCD (PPMZ rc)	.37	.82	.38	.63
meteor-baseline	.43	.61	.84	.58
meteor-ranking	.26	.70	.83	.55
mter	.26	.69	.73	.52
Mean	.47	.77	.72	.65

Table 5: Average system-level correlations for the RANK category from English for NCD, mNCD and other MT evaluation measures.

degrades the averages. Considering the mean of the categories instead, mNCD's correlation of .74 is third best together with 'posbleu'.

Table 5 shows the results from English. The table is shorter since many of the better MT measures use language specific linguistic resources that are not easily available for languages other than English. mNCD performs competitively only for French, otherwise it falls behind NCD and other methods as already shown earlier.

6 Discussion

We have introduced a new MT evaluation measure, mNCD, which is based on normalized compression distance and METEOR's relaxed alignment modules. The mNCD measure outperforms NCD in English with all tested parameter combinations, whereas results with other target languages are unclear. The improved correlations with mNCD did not change the position in the RANK category of the MT evaluation measures in the 2008 ACL WMT shared task.

The improvement in English was expected on the grounds of the synonym module, and indicated also by the larger number of affected words in the similarized references. We believe there is potential for improvement in other languages as well if synonym lexicons are available.

We have also extended the basic NCD measure to scale between a document comparison measure and aggregated sentence-level measure. The rather surprising result is that NCD produces quite similar scores with all block sizes. The different result with Spanish may be caused by differences in the data or problems in the calculations.

After using the same evaluation methodology as in Callison-Burch et al. (2008), we have doubts whether it presents the most effective method exploiting all the given human evaluations in the best way. The system-level correlation measure only awards the winner of the ranking of five different systems. If a system always scored second, it would never be awarded and therefore be overly penalized. In addition, the human knowledge that gave the lower rankings is not exploited.

In future work with mNCD as an MT evaluation measure, we are planning to evaluate synonym dictionaries for other languages than English. The synonym module for English does not distinguish between different senses of words. Therefore, synonym lexicons found with statistical methods might provide a viable alternative for manually constructed lexicons (Kauchak and Barzilay, 2006).

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.

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, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL-2006*, pages 249–256.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christoph Monz, and Josh Schroeder. 2008. Further meta-evalutation of machine translation. *ACL Workshop on Statistical Machine Translation*.

Yee Seng Chan and Hwee Tou Ng. 2009. MaxSim: performance and effects of translation fluency. *Machine Translation*, 23(2-3):157–168.

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

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 455–462, Morristown, NJ, USA. Association for Computational Linguistics.

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

Yanjun Ma, Nicolas Stroppa, and Andy Way. 2007. Bootstrapping word alignment via word packing. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 304–311, Prague, Czech Republic, June. Association for Computational Linguistics.

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.

Grazia Russo-Lassner, Jimmy Lin, and Philip Resnik. 2005. A paraphrase-based approach to machine translation evaluation. Technical Report LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57, University of Maryland, College Park.

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.