A Dataset for Assessing Machine Translation Evaluation Metrics

Lucia Specia§, Nicola Cancedda† and Marc Dymetman†

§ Research Group in Computational Linguistics, University of Wolverhampton
Stafford Street, Wolverhampton, WV1 1SB, UK
L.Specia@wlv.ac.uk

†Xerox Research Centre Europe
6 Chemin de Maupertuis, Meylan, 38240, France
Nicola.Cancedda@xerox.com, Marc.Dymetman@xerox.com

Abstract

We describe a dataset containing 16,000 translations produced by four machine translation systems and manually annotated for quality by professional translators. This dataset can be used in a range of tasks assessing machine translation evaluation metrics, from basic correlation analysis to training and test of machine learning-based metrics. By providing a standard dataset for such tasks, we hope to encourage the development of better MT evaluation metrics.

1. Introduction

Automatic evaluation of Machine Translation (MT) is a long standing issue. Most metrics are based on reference translations to compute some form of overlapping between n-grams in the MT system output and in one or more human translations. This is the basic framework of the most commonly used metrics, including BLEU (Papineni et al., 2002) and NIST (Doddington, 2002). More complex reference-based metrics replace or complement n-gram matching with alternative lexical features, such as lemma- or synonym-based matching (Lavie and Agarwal, 2007), and sometimes using syntactic and semantic features, such as the matching of dependency relations (Gimenez and Marquez, 2008). In general, the performance of metrics is measured by computing the correlation between their scores and scores given by humans. The ultimate goal is to have a metric that produces scores as close as possible to human ones.

In order to maximise the correlation levels, some metrics use of machine learning techniques to learn quality estimates directly from data annotated with human scores (Quirk, 2004; Specia et al., 2009). Such metrics, sometimes called "Confidence Estimation" metrics (Blatz et al., 2004), have shown to correlate significantly better with human evaluation than standard metrics like BLEU and NIST (see Section 3.). In our particular experimental setup, we prefer to use the term "Quality Estimation", since we aim to estimate a quality indicator within a given range, as opposed to binary "bad" / "good" judgments estimated in previous work on "Confidence Estimation" (Quirk, 2004; Blatz et al., 2004).

Besides yielding better correlation with human scores, an advantage of quality estimation metrics is that they do not necessarily rely on reference translations. Quality indication features can be extracted given only the source and translation text, and optionally monolingual and bilingual corpora or information about the MT system used to produce the translations. Once a model has been learnt based on human annotated data for a certain language pair, it can be used to estimate the quality of any number of translations for that language pair and direction. While learning models for specific text domains and MT systems may allow more accurate estimates, the correlation obtained by models designed for other domains and systems still outperforms standard metrics (see Section 3.).

Another benefit of quality estimation metrics is that they can be more reliably applied at the sentence level, a well known limitation of n-gram matching based-metrics like BLEU or NIST, which usually correlate well with human judgments at the corpus or system level only.

The only requirement of such metrics is human annotation at training time. While manual annotations of quality have been made available through shared evaluation tasks like those promoted by the Workshops on Statistical MT (WMT) (Callison-Burch et al., 2007; Callison-Burch et al., 2008; Callison-Burch et al., 2009), these are usually very small: maximum of a few hundred sentences per language pair and MT system, which is not sufficient for training a machine learning algorithm. Moreover, such datasets are created as result of tasks comparing translations produced by different MT systems for the same source sentence. Therefore, the human scores are useful for ranking the alternative translations, but are not absolute indicators of quality, that is, a given score may be assigned to translations of various levels of quality for different source sentences. Additionally, agreement analysis performed in several editions of the shared task showed that acceptable levels of agreement were only found in ranking tasks (Callison-Burch et al., 2007). Finally, the annotations were created by volunteers, not necessarily experts in the language pair, and not trained for the task. Therefore, different annotators may interpret scores in a different way.

In this paper we describe four datasets created in a controlled environment to guarantee the quality of the annotations, annotated by professional translators trained on the task and based on clearly defined guidelines about the interpretation of the quality scores (Section 2.). We also show some results obtained with a quality estimation metric trained on such datasets (Section 3.). These datasets can be used for assessment and comparison of MT evaluation metrics and also for the investigation of new metrics based on human annotation.

2. Construction of the datasets

2.1. Source and target sentences

The dataset consists of 4,000 English (source) sentences, their corresponding reference translations, and four alternative translations for each of them into Spanish (target), produced by four statistical MT systems, amounting to 16,000 source-target-reference triples.

The source and reference sentences were extracted from Europarl, the European Parliament corpus (Koehn, 2005), more specifically, they were randomly selected from the test and development sets of WMT-2008 (Callison-Burch et al., 2008).

The four MT systems used to produce the translations are statistical: Matrax (Simard et al., 2005), Portage (Johnson et al., 2006), Sinuhe (Kääriäinen, 2009) and MMR (Maximum Margin Regression) (Saunders, 2008). Portage and Matrax are standard phrase-based SMT systems, with the exception that Matrax allows for gaps in phrases. Sinuhe is also a phrase-based system, but differs from standard systems by allowing phrases to overlap during decoding, and by training individual phrase weights applying a regularized conditional random fields on the full parallel aligned corpus. MMR is a rather distinct approach to MT based on using predictions with structured output. It is an end-to-end translation system that does not rely on traditional word alignment, phrase extraction or language modeling techniques. Features based on global similarities of words are first calculated using a minimum-distance approach on sentence pairs. Then, a structured-learning approach is used to compute the alignment of words to phrases. Decoding is performed by dynamic programming and is guided by a heuristic based on overlapping bigrams. Sinuhe and MMR were still at initial development stages when used to produce the translations. In following sections we anonymise these systems by arbitrarily naming them System1-System4.

Each of these SMT systems was trained with approximately 1.2M sentences pairs extracted from the European Parliament corpus, more specifically, the “training” data provided by WMT-2008.

2.2. Annotation process

Judging the quality of translations is an inherently complex and subjective task, even for professional translators. To minimise these issues, we consider a criterion commonly used by translators to provide quality judgments for human or machine translations: amount of post-editing necessary to make the translations ready for publishing.

The translations were annotated by professional translators trained to produce assessments of quality according to the requirements of a language service provider. Translators were given the source sentence in English and its translation into Spanish, as produced by each of the four MT systems, and asked to assign to such translation one of the following four scores:

- 1: requires complete retranslation.
- 2: a lot of post editing needed (but quicker than retranslation).
- 3: a little post editing needed.
- 4: fit for purpose.

Before submitting the translation data to be annotated for quality, we performed a pilot annotation task in order to verify the agreement among annotators. The same 50 translations produced by one of the systems were given to three translators. We then measured agreement using the $Kappa$ coefficient (Cohen, 1960), which is defined as:

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the proportion of times that the annotators agree, and $P(E)$ the proportion of times they would agree by chance. There is no standard interpretation for Kappa values, but according to (Landis and Koch, 1977), 0 - 0.2 is slight, 0.2 - 0.4 is fair, 0.4 - 0.6 is moderate, 0.6-0.8 is substantial and the rest close to perfect agreement.

The $Kappa$ score obtained was 0.65, considerably higher than the agreement achieved in the WMT human evaluation tasks (around 0.32, even for ranking tasks). In order to further assure the consistency within a given dataset, the complete set of sentences produced by each MT system was annotated by a single translator, and therefore, four translators participated in the annotation task.

Table 1 shows some figures resulting from the annotation process: the average and median of the scores assigned by the translators to all the sentences in each MT system's dataset.

Dataset	Average score	Median score
System1	2.835	3
System2	2.558	3
System3	2.508	3
System4	1.338	1

Table 1: Average and median scores resulting from the annotation of the 4,000 translations produced by each of the four MT systems

Clearly, the figures differ considerably from system to system. For the sake of curiosity, we computed standard 4-grams BLEU (Papineni et al., 2002), 5-grams NIST (Doddington, 2002), METEOR (Lavie and Agarwal, 2007) (using exact match and default values for its parameters) and TER (Snover et al., 2006) for each of these systems. At the system-level, particularly BLEU and METEOR correlate well with the human annotation, as shown in Table 2. As we discuss in Section 3., the task of predicting the quality for translations produced by a given system is likely to be easier for systems performing on average very well or very poorly.

3. Using the dataset for quality estimation

We present in what follows some results from experiments using the datasets described in Section 2. for estimating the quality of English-Spanish translations produced by a given system.

Dataset	BLEU	NIST	METEOR	TER
System1	0.3880	8.8586	0.3998	47.090
System2	0.3521	8.3985	0.3704	49.624
System3	0.3241	8.4029	0.3531	49.556
System4	0.1954	6.5605	0.2707	62.808

Table 2: System level scores according to MT evaluation metrics for the 4,000 translations produced by each of the four MT systems

We extracted 84 features from source and translation sentences, as well as from parallel and monolingual corpora. The same features are used for all four MT system datasets and can be summarized in the following groups:

- source & target sentence lengths and their ratios
- source & target sentence trigram language model probabilities and perplexities
- source & target sentence type/token ratio
- average source word length
- percentage of unigrams, bigrams and trigrams in the source sentence belonging to each frequency quartile of a monolingual corpus
- number of mismatching opening/closing brackets and quotation marks in the target sentence
- average number of occurrences of all target words within the target sentence
- alignment score (IBM-4) for source and target sentences and percentage of different types of word alignments, as given by GIZA++ using the actual SMT training data (∼1 million sentences) plus the QE sentences
- average number of translations per source word in the sentence (as given by probabilistic dictionaries produced by GIZA++), unweighted or weighted by the (inverse) frequency of the words
- percentages of numbers, content- & non-content words in the source & target sentences
- percentages and number of mismatches of each of the following superficial constructions between the source and target sentences: brackets, punctuation symbols, numbers
- trigram target language model probability trained on a corpus of POS-tags of words.

For each dataset, we applied a machine learning technique called Partial Least Squares (PLS) (Wold et al., 1984) for feature selection and model learning, as described in (Specia et al., 2009). Table 3 shows the performance obtained by the systems in terms of Root Mean Squared Prediction Error (RMSPE):

$$\sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_j - \widehat{y}_j)^2}$$

where N is the number of test examples, $\widehat{y}$ is the prediction obtained by PLS and y is the true value for the test case. Therefore, RMSPE quantifies the amount by which the estimator differs from the expected score.

Dataset	RMSPE
System4	0.603 ± 0.262
System1	0.653 ± 0.114
System3	0.706 ± 0.059
System2	0.718 ± 0.144

Table 3: RMSPE for each dataset

The models produced for different MT systems deviate from 0.6 to 0.72 points when predicting the sentence-level 1-4 score, which we believe is an acceptable deviation. For example, one sentence that should be classified as “fit for purpose” (score 4) would very rarely (if ever) be classified as “requires complete retranslation” (score 1) and discarded as a consequence. Although the errors for the different systems are not directly comparable, the better error scores obtained by System1 and System4 may be due to the fact that the quality of their translations is relatively easier to predict, since they are usually scored very high and very low, respectively.

We also computed the sentence level absolute Pearson's correlation between human annotation and the most common MT evaluation metrics, smoothed BLEU (with bigrams only to avoid 0 scores at the sentence level) (Lin and Och, 2004), standard NIST, METEOR and TER. The last column (**QE**) shows the correlation obtained using the score predicted by our quality estimation metric.

Dataset	BLEU	NIST	METEOR	TER	**QE**
System2	0.296	0.254	0.337	0.268	**0.542**
System1	0.237	0.203	0.277	0.194	**0.556**
System3	0.209	0.195	0.240	0.168	**0.562**
System4	0.165	0.129	0.231	0.145	**0.524**

Table 4: Sentence-level Pearson's correlation of automatic metrics with human annotation

Table 4 shows that the correlation of the score predicted using our method is superior to that of any MT evaluation metric. The differences are statistically significant with 99.8% confidence, according to bootstrapping re-sampling (Koehn, 2004). Figures for Spearman's correlation, which considers only the ranking of the scores, are higher for all datasets, but the proportion of the differences between the **QE** score and other metrics remains.

The quality estimation task was designed here for predicting the scores for a given MT system. However, we also found very high correlation between human scores and the score estimated for a given system's dataset from models produced for other system's dataset. For example, a QE system trained on System3's dataset and used to predict scores for the other three datasets results in the following correlation scores: System2 = 0.517, System1 = 0.478 and System4 = 0.423. These correlation scores are still significantly higher than the scores of any MT evaluation metric.

4. Conclusions

We have presented a dataset of four sets of 4,000 source-translation-reference-score quadruples produced by English-Spanish SMT systems, which can be used for as-

sessing existing MT evaluation metrics and also investigating new metrics based on human evaluation.
The dataset can be downloaded from `http://pers-www.wlv.ac.uk/~in1316/resources/ce_dataset.rar` or `http://www.smart-project.eu/node/565`.

5. References

J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2004. Confidence Estimation for Machine Translation. In *20th Coling*, pages 315–321, Geneva.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *2nd Workshop on Statistical Machine Translation*, pages 136–158, Prague.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In *3rd Workshop on Statistical Machine Translation*, pages 70–106, Columbus.

C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *4th Workshop on Statistical Machine Translation*, pages 1–28.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, April.

G. Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *2nd Conference on Human Language Technology Research*, pages 138–145, San Diego.

J. Gimenez and L. Marquez. 2008. A Smorgasbord of Features for Automatic MT Evaluation. In *3rd Workshop on Statistical Machine Translation*, pages 195–198, Columbus, Ohio.

H. Johnson, F. Sadat, G. Foster, R. Kuhn, M. Simard, E. Joanis, and S. Larkin. 2006. Portage with Smoothed Phrase Tables and Segment Choice Models. In *Workshop on Statistical Machine Translation*, pages 134–137, New York.

M. Kääriäinen. 2009. Sinuhe – Statistical Machine Translation using a Globally Trained Conditional Exponential Family Translation Model. In *Conference on Empirical Methods in Natural Language Processing*, pages 1027–1036, Singapore.

P. Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Conference on Empirical Methods in Natural Language Processing*, Barcelona.

P. Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit*.

R. J. Landis and G. G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

A. Lavie and A. Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *2nd Workshop on Statistical Machine Translation*, pages 228–231, Prague.

C. Y. Lin and F. J. Och. 2004. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Coling-2004*, pages 501–507, Geneva.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Morristown.

C. B. Quirk. 2004. Training a Sentence-Level Machine Translation Confidence Measure. In *4th Conference on Language Resources and Evaluation*, pages 825–828, Lisbon.

C. Saunders. 2008. Application of Markov Approaches to Statistical Machine Translation. Technical report, SMART Project Deliverable 2.2.

M. Simard, N. Cancedda, B. Cavestro, M. Dymetman, E. Gaussier, C. Goutte, and K. Yamada. 2005. Translating with Non-contiguous Phrases. In *Conference on Empirical Methods in Natural Language*, pages 755–762, Vancouver.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *7th Conference of the Association for Machine Translation in the America*, pages 223–231, Cambridge, MA.

L. Specia, M. Turchi, N. Cancedda, M. Dymetman, and N. Cristianini. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. In *13th Conference of the European Association for Machine Translation*, pages 28–37, Barcelona.

S. Wold, A. Ruhe, H. Wold, and W. J. Dunn. 1984. The Covariance Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM Journal on Scientific Computing*, 5:735–743.