Machine Translation Quality Estimation Adapted to the Translation Workflow

Sabine Hunsicker
euroscript Deutschland GmbH

Alexandru Ceausu
euroscript Luxembourg S.à r.l.

ABSTRACT

The varying quality of machine translation (MT) poses a problem for language service providers (LSPs) which want to use MT to make the translation production process more efficient. In this user study we describe the MT confidence score we developed. It predicts the quality of a segment translated by MT and it is fully integrated into the translation workflow.

1. Introduction

The varying quality of MT poses a problem in the translation workflow as it requires different levels of post-editing effort. As one sentence may be translated well, the next one may turn out to be completely unusable. There might be cases where discarding the MT suggestions and translating from scratch is going to be faster. This decision time also increases the post-editing time.

In order to be able to exploit the full potential of MT suggestions, they should be annotated with a score that is indicative for the translation quality. The translators are already familiar with such scores from the usage of translation memory (TM).

MT does not assign a score to its output. The decoder calculates internal scores to find the best hypothesis from the translation options, but these decoder scores cannot be used to estimate a level of quality.

For an LSP, a predictive score for MT quality would be very useful, as this would be in line with the way TMs are used in the workflow. Another important advantage is that it provides an upfront estimation of the cost for a given translation.

2. Confidence Score

The translation workflow at euroscript involves several MT-related stages. One of these stages contains the quality estimation component which we call confidence score, e.g. a component that would answer the question on ‘how confident is the MT that a particular sentence was well-translated? ‘

In order to reduce the annotation effort, we developed this score starting from the automatic scores. The approach can be easily automated so that it can be run immediately after training a new MT system. Another advantage is that there is no time lost in finding human annotators and for data creation before the new MT system is deployed in production.

The prediction model makes use of a combination of system-independent and system-dependent features. For example, the sentence lengths of the source and the MT candidate are taken into account. The system-dependent features vary on the MT system that should be evaluated. SMT systems usually provide different scores calculated during decoding.

Each training instance includes the source sentence, the target sentence, the MT candidate, the feature vector and the automatic score.

The training algorithm automatically chooses a well-distributed sample of training instances to train the prediction model. As the confidence score is integrated into the MT workflow, each MT request is automatically annotated with the confidence score.

The confidence score is optimized to predict which of the following levels of quality the current translation belongs to:

- good (no or little editing needed)
- usable (some editing is required)
- useless (discard)

3. Experiments

For exemplification, we present our confidence score experiments on the language direction English→Danish.

The texts used in our experiments come from the public domain. The training data was created by translating these texts with MT and then post-editing the results. As such, the translations are usually very close to the MT candidates, except where MT was of such a bad quality that it was discarded. The test set contains 1074 sentences.

During translation with MT, the automatic scores for the classifier were collected and a confidence model was trained on the resulting data. After integration into the translation workflow, the model was evaluated on unseen texts from the same domain.

We trained prediction models for the following three automatic scores:

- BLEU (Papineni et al., 2002)
- normalised Editing Distance
- Fuzzy Match

The editing distance (ED), based on the Levenshtein distance, allows us to draw conclusions concerning the post-editing effort—the lower the editing distance, the lower the effort for post-editing. In the original version a low score means that the MT candidate was close to the reference translation, a high score respectively that the MT candidate varied a lot from the reference. In contrast to BLEU, this distance is not contained in a closed interval, therefor we use

a normalised version that transposes the scores to the interval [0,1]. Additionally we reverse the score, so that 1 is the best score, analogous to the other scores used.

Fuzzy matching (FM) is another indicator of how close the reference and MT translation are. FM is usually used when evaluating a new text against a translation memory and works on the source sentence. In our experiments, we used the fuzzy matching algorithm implemented in the Okapi Framework[1]. The reference translation is set as the original translation (that would be saved in the TM) and the MT candidate is set as the new text.

Neither FM nor ED take into account the source sentence.

Each confidence score model is evaluated compared to the three scores: BLEU, editing distance (ED) and fuzzy matching (FM).

To compare the different metrics, we calculate three types of measures: the mean absolute error (MAE), the root mean squared error (RMSE) and Pearson's correlation coefficient (r).

Confidence Score	Score	RMSE	MAE	r
$conf_{BLEU}$	BLEU	0.4577	0.2894	-0.1615
	ED	0.6765	0.6469	-0.2414
	FM	0.5305	0.4618	0.2358*
$conf_{ED}$	BLEU	0.5743	0.5422	0.0063
	ED	**0.1878**	**0.1537**	**0.1980**
	FM	0.3611	0.2658	-0.1414
$conf_{FM}$	BLEU	0.3861	0.3257	0.4063*
	ED	0.4707	0.4348	0.2743
	FM	0.3858	0.3483	-0.1044

Table 1: Evaluation statistics for EN → DA confidence score correlation

Table 1 shows the evaluation results for all three prediction models. We see that the error rates differ considerably between the evaluation metrics and the error rates. The prediction model based on the editing distances performs quite well: it achieves the lowest error rates and correlates moderately with the score it tries to predict.

As predicting the full range of scores is a very complex task, we decided to scale down and only predict the three quality levels described in Section 2.

To determine the thresholds of these levels, we ran two experiments, one with a very high level (95% for good, 75% for usable) and the other with a moderate level (75% for good, 50% for usable).

From these experiments, we can tell that the moderate quality levels are easier to predict, as we achieve higher correlation values with them. The editing distance model performs well here as well: choosing the minimum thresholds, we achieve a correlation of 0.2588 of the confidence score to the actual editing distance score.

[1] http://okapi.opentag.com/

In a human evaluation, we used a random sample of the evaluation data to be judged by professional translators. Of 221 sentences, 102 scores were judged to be appropriate and 119 to be inappropriate, indicating that the scoring mechanism needs more fine-tuning, but that it is still usable.

4. Conclusion

We presented a practical example on how to incorporate the confidence score into the traditional translation workflow. The first prototype of our confidence score works well in our production set-up. To fine-tuning the predictions require more evaluation and data, both of which are created automatically during the translation post-editing in production.

Fine-tuning the predictions requires more evaluation and data, both of which are created automatically during the translation production. Our on-going work is to introduce more features for the prediction model, such as linguistic analysis. Another constraint for assessing MT, however, is the time required to do so, as the MT needs to be provided quickly. Our current model provides judgments in less than half a second, and further improvements need to scale accordingly.

References

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.