# Ranking Translations using Error Analysis and Quality Estimation

**Mark Fishel**
Institute of Computational Linguistics
University of Zurich, Switzerland
fishel@cl.uzh.ch

## Abstract

We describe TerrorCat, a submission to this year's metrics shared task. It is a machine learning-based metric that is trained on manual ranking data from WMT shared tasks 2008–2012. Input features are generated by applying automatic translation error analysis to the translation hypotheses and calculating the error category frequency differences. We additionally experiment with adding quality estimation features in addition to the error analysis-based ones. When evaluated against WMT'2012 rankings, the system-level agreement is rather high for several language pairs.

## 1 Introduction

Recently a couple of methods of automatic analysis of translation errors have been described (Zeman et al., 2011; Popović and Ney, 2011). Both of these compare a hypothesis translation to a reference and draw detailed conclusions from the differences between the two.

TerrorCat, a metric submitted to the metrics shared task of WMT'2012 (Callison-Burch et al., 2012) used the output of those two error analysis methods as input features, which yielded mildly promising results (Fishel et al., 2012). However the submitted version of TerrorCat was language pair-specific, which means that the classifier model used by the metric has to be retrained on new manual pairwise ranking data for every new language pair. This in turn complicates its usage.

Our main aim in this work is to make TerrorCat usable out-of-the-box. We compare models specific to the language pair (baseline), target language and a universal model for all languages. The updated metric is applied to the WMT'13 metrics shared task.

An additional modification to the metric uses input features from quality estimation. Using the resources of the quality estimation shared task of WMT'13 the modified model is applied to the English–Spanish language pair.

We start by briefly re-introducing the TerrorCat metric.

## 2 Baseline

The baseline TerrorCat metric is a machine learning-based metric: it uses manually ranked translation hypothesis pairs to train a classifier model. The trained model is then used to predict a ranking for new sentence pairs that have not been ranked yet.

To convert the binary comparisons between translation hypothesis pairs into a numeric score per translation hypothesis the wins per hypothesis are summed together. Previous year's work has shown (Fishel et al., 2012) that weighting the wins with the classifier's confidence for the summed score improves agreement with human judgements.

The input features for learning and classification are obtained by

1. applying translation error analysis software to the compared hypotheses,
2. getting the frequencies of every error type, i.e. the ratio of words marked with that error type to the hypothesis sentence length,
3. and using each error type's frequency differences between the two hypotheses as input features.

Relative frequencies are used on both system and segment level: i.e. the ratios of words marked with a particular error type to the hypothesis translation length. This guarantees that feature values lie in the $[-1, 1]$ range.



*Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 405–407,
Sofia, Bulgaria, August 8-9, 2013 

Translation error analysis is done with two tools: Addicter (Zeman et al., 2011) and Hjerson (Popović and Ney, 2011). Both perform error analysis by comparing the hypothesis and reference translations on word level and treating each difference as an error of one or the other kind. Translation error taxonomies as well as the way word differences and their contexts are interpreted differ between the two tools. In order to enable independent input from both tools the feature vectors obtained from them both tools are concatenated.

To increase the level of detail the frequencies of each error category are counted separately for each part-of-speech separately. As a result, e.g. instead of having the information of order errors having a particular frequency, the classifier will separately see the frequencies of misplaced nouns, adjectives, particles, etc.

## 3 Experiments

The usage of part-of-speech tags improves agreement with human judgements (Fishel et al., 2012); however, it also introduces language dependency for the metric. In the first set of experiments we try to remove this imposed dependency without losing the achieved benefit.

### 3.1 Common Settings

We focused on six language pairs: between English and German, French and Spanish. Manual ranking data for training was taken from WMT shared task evaluations 2008–2011; data from WMT'2012 was used as a development set to assess the performance of metric variations.

Final models for the WMT'2013 shared task were re-trained on the whole set of manual rankings, from WMT 2008–2012.

The classifier used by TerrorCat is an SVM with a linear kernel; more powerful kernels, such as radial basis function-based ones scaled poorly to the high number of features and thus were not tested.

PoS-tagging was done using TreeTagger (Schmid, 1995) with the pre-trained models for English, German, French and Spanish.

### 3.2 Language Independence

It is natural to expect error categories to have varying importance on the quality comparison between two translation candidates. For instance, one might expect order differences between translations into functional languages (e.g. English, Chinese) to have a greater importance than in case of languages with a more flexible word order (e.g. German, Russian); on the other hand inflection errors are likely to do more damage to the meaning in morphologically complex languages (e.g. Russian, Finnish) than in languages with simpler morphology (e.g. English, French). However, we want to see whether we can train a classifier that would generalize over all language pairs.

The main obstacle for training a general model on all language pairs are the different taxonomies of part-of-speech tags for different target languages: the arity of the input feature vectors is different for different target languages, which makes the data incompatible between them.

To overcome the difference we define a mapping from every used taxonomy to a common general set of PoS-tags, which is supposed to cover any language. It consists of general part-of-speech categories (such as noun, verb, particle, etc., a total of 15), without any morphological information (tense, case, person, etc.).

By using the same set of generalized PoS-tags for every language we ensure that the used TerrorCat classifier model is language-independent; the PoS-tagging step is naturally language-dependent still.

Tables 1 and 2 present system-level and segment-level correlations of TerrorCat based on this common PoS-tag set and three models, specific to the language pair, target language only and a general model for any language. Both sets of results show that using a language-independent model neither improves nor worsens the performance.

### 3.3 Quality Estimation for Ranking

To further improve the agreement between TerrorCat and human assessment we experimented with adding input features from quality estimation.

The input features were adopted from this year's shared task on quality estimation. We selected the smaller set of black-box features, which included the sentence lenghts, their language model probabilities, average numbers of translations per word, percentages of uni-, bi- and tri-grams in the different frequency quartiles, etc. All resources were taken from the shared task, which also meant that this modified model was applied only to English–Spanish.

| | de-en | en-de | es-en | en-es | fr-en | en-fr |
|---|---|---|---|---|---|---|
| Language pair-specific | 0.94 | 0.56 | 0.94 | 0.59 | 0.85 | 0.82 |
| Target language-specific | 0.92 | 0.56 | 0.97 | 0.59 | 0.84 | 0.82 |
| Language-independent | 0.93 | 0.71 | 0.94 | 0.66 | 0.84 | 0.88 |
| BLEU | 0.67 | 0.22 | 0.87 | 0.40 | 0.81 | 0.71 |
| METEOR | 0.89 | 0.18 | 0.95 | 0.45 | 0.84 | 0.82 |
| TER | 0.62 | 0.41 | 0.92 | 0.45 | 0.82 | 0.66 |

Table 1: System-level correlation between TerrorCat and human ranking. Correlations of BLEU, METEOR and TER scores are given for comparison.

| | de-en | en-de | es-en | en-es | fr-en | en-fr |
|---|---|---|---|---|---|---|
| Language pair-specific | 0.31 | 0.18 | 0.24 | 0.21 | 0.23 | 0.20 |
| Target language-specific | 0.31 | 0.18 | 0.28 | 0.21 | 0.23 | 0.20 |
| Language-independent | 0.28 | 0.20 | 0.27 | 0.22 | 0.24 | 0.21 |

Table 2: Segment-level correlation between TerrorCat and human ranking.

Training the model on quality estimation features alone yields a system-level score of 0.56. Although this is lower than the TerrorCat baseline, it beats the correlations of BLEU, TER and METEOR (see Table 1). The segment-level correlation is -0.01.

Next we combined features from error analysis and quality estimation by concatenating them into a single input feature vector. As a result system-level correlation improved to 0.72, which is higher than all TerrorCat variants so far (best correlation: 0.66). Segment-level correlation remained practically the same (0.22).

## 4 Conclusion

We have applied TerrorCat to the shared metrics task of WMT'2013. Just like last year, the results are mildly promising.

We were successful at achieving language independence, provided that PoS-tagging is done before applying the metric.

The trained model as well as the metric implementation with all the necessary scripts is available online[1].

It remains to be tested, whether quality estimation features fit well with the language-independent models. As the extracted feature values are based on completely different, language-specific resources, this does not seem to be a likely outcome.

[1] https://github.com/fishel/TerrorCat

## References


Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada.

Mark Fishel, Rico Sennrich, Maja Popović, and Ondřej Bojar. 2012. Terrorcat: a translation error categorization-based mt quality metric. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 64–70, Montréal, Canada.

Maja Popović and Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland.

Daniel Zeman, Mark Fishel, Jan Berka, and Ondřej Bojar. 2011. Addicter: What is wrong with my translations? *The Prague Bulletin of Mathematical Linguistics*, 96:79–88.