# Quality estimation for translation selection

**Kashif Shah** and **Lucia Specia**
Department of Computer Science
University of Sheffield
S1 4DP, UK
{kashif.shah,l.specia}@sheffield.ac.uk

## Abstract

We describe experiments on quality estimation to select the best translation among multiple options for a given source sentence. We consider a realistic and challenging setting where the translation systems used are unknown, and no relative quality assessments are available for the training of prediction models. Our findings indicate that prediction errors are higher in this blind setting. However, these errors do not have a negative impact in performance when the predictions are used to select the best translation, compared to non-blind settings. This holds even when test conditions (text domains, MT systems) are different from model building conditions. In addition, we experiment with quality prediction for translations produced by both translation systems and human translators. Although the latter are on average of much higher quality, we show that automatically distinguishing the two types of translation is not a trivial problem.

## 1 Introduction

Quality Estimation (QE) [Blatz et al., 2004, Specia et al., 2009] has several applications in the context of Machine Translation (MT), considering the use of translations for both inbound (e.g. gisting) and outbound (e.g. post-editing) purposes. To date, research on quality estimation has been focusing mostly on predicting absolute single-sentence quality scores. However, for certain applications an absolute score may not be necessary. Our goal is to model quality estimation by contrasting the output of several translation sources for the same input sentence against each other. The outcome of this process is a ranking of alternative translations based on their predicted quality. For our application, we are only interested in the top-ranked translation, which could for example be provided to a human post-editor for revision.

Previous research on this task has focused on ranking translations from multiple MT systems where system identifiers are known beforehand. Based on such identifiers, individual quality prediction models can be trained for each MT system [Specia et al., 2010], and the predicted (absolute) scores for translations of a given source sentence across multiple MT systems used to rank them. Alternatively, quality prediction models can be built to directly output a ranking of alternative translations based on training data annotated with relative quality scores, using for example pairwise ranking algorithms [Avramidis, 2013, Avramidis and Popović, 2013].

In this paper we model translation selection considering a scenario where translations are produced by multiple MT systems, but the identifiers of the MT systems are not given, i.e., we assume a blind setting where the sources of the translations are not known. While ranking approaches to system selection could also be used in this blind setting, they require training data labelled with comparative assessments of translations produced by multiple sources. In our experiments, we assume a more general scenario where the labelling of training data is produced for individual translation segments in absolute terms, independently and regardless of their origin. In addition, we also experiment with predicting the quality for human translations. Although human translations are on average of much higher quality than machine translations, we show that this is not always the case and that automatically distinguishing the two types of translation is not a trivial problem.

We present experiments with four language pairs and various prediction models in blind and non-blind settings, as well as with the use of the resulting predictions for translation selection. We show that while prediction errors are higher in blind settings, this does not have a negative impact in performance when using predictions in the task of translation selection. Our best result in terms of the quality scores of the selected translation sets are obtained with prediction models where all available translations are polled together in a system-agnostic way. Finally, we show that these system-agnostic models have good performance when predicting quality for out-of-domain translations, produced by other MT systems.

## 2 Related work

A handful of system ranking and selection techniques have been proposed in recent years. For an overview of various related approaches we refer the reader to the WMT13 shared task on QE [Bojar et al., 2013]. This shared task included a system ranking track aimed at predicting 5-way rankings for translations produced by five MT systems and ranked by humans for model bulding. All related work relies on either knowing the system identifiers or having access to relative rankings of translations at training time.

MT system selection was first proposed by Specia et al. [2010]. QE models are trained independently for each MT system, and the translation option with highest prediction score is used. 77% of the sentences with the highest QE score also have the highest score according to humans. In contrast, 54% of accuracy was found when selecting translations from the best MT system on average.

He et al. [2010] focus on the ranking between translations from either an MT system or a translation memory for post-editing. Classifiers showed promising results in selecting the option with the lowest estimated edit distance.

Hildebrand and Vogel [2013] use an classic n-best list re-ranking approach based on predicting BLEU scores. A feature set where all features that are solely based on the source sentence were removed showed the best results.

Biçici [2013] uses language and MT system independent features to predict F1 scores with regression algorithms. A threshold for judging if two translations are equal over the predicted F1 scores was learned from data.

Avramidis [2013] and Avramidis and Popović [2013] decompose rankings into pairwise decisions, with the best translation for each candidate pair predicted using logistic regression. A number of features of the source and target languages, including pseudo-references, are used. A similar pairwise ranking approach was used by Formiga et al. [2013], but with random forest classifiers.

## 3 Experimental settings

**Datasets** Our datasets contain news domain texts in four language pairs (Table 1): English-Spanish (**en-es**), Spanish-English (**es-en**), English-German (**en-de**), and German-English (**de-en**). Each contains a different number of source sentences and their human translations, as well as 2-3 versions of machine translations: by a statistical (SMT) system, a rule-based (RBMT) system and, for en-es/de only, a hybrid system. Source sentences were extracted from tests sets of WMT13 and WMT12, and the translations were produced by top MT systems of each type (SMT, RBMT and hybrid - hereafter **system2**, **system3**, **system4**) which participated in the translation shared task, plus the additional professional translation given as reference (**system1**). These are the official datasets used for the WMT14 Task 1.1 on QE.[1]

| Languages | # Training Src/Tgt | # Test Src/Tgt |
|---|---|---|
| **en-es** | 954/3,816 | 150/600 |
| **en-de** | 350/1,400 | 150/600 |
| **de-en** | 350/1,050 | 150/450 |
| **es-en** | 350/1,050 | 150/450 |

Table 1: Number of training and test source (Src) and target (Tgt) sentences.

Each translation in this dataset has been labelled by a professional translator with [1-3] scores for "perceived" post-editing effort, where:

- **1** = perfect translation, no editing needed.
- **2** = near miss translation: maximum of 2-3 errors, and possibly additional errors that can be easily fixed (capitalisation, punctuation).
- **3** = very low quality translation, cannot be easily fixed.

[1] `http://www.statmt.org/wmt14/quality-estimation-task.html`

The distribution of true scores in both training and test sets is given in Figures 1 and 2, for each language pair, and for each language pair and translation source, respectively.

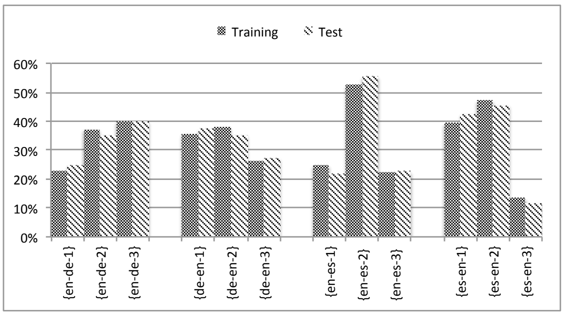

Figure 1: Distribution of true scores by lang. pair.

**Out-of-domain test sets** For three language pairs, we also experiment with out-of-domain test sets (Table 2) provided by translation companies (also made available by WMT14) and annotated in the same way as above by a translation company (i.e., one professional translator). These were generated using the companies' own source data (different domains than news), and own MT system (different from the three used in our main datasets).

| ID | Languages | # Test |
|---|---|---|
| **LSP**$_1$ | **en-es** | 233 |
| **LSP**$_2$ | **en-es** | 738 |
| **LSP**$_3$ | **en-de** | 297 |
| **LSP**$_4$ | **es-en** | 388 |
| **LSP**$_5$ | **es-en** | 677 |

Table 2: Number of out-of-domain test sentences.

**Features** We use the `QuEst` toolkit [Specia et al., 2013, Shah et al., 2013] to extract two feature sets for each dataset:

- **BL**: 17 features used as baseline in the WMT shared tasks on QE.
- **AF**: 80 common MT system-independent features (superset of **BL**).

The resources used to extract these features (language models, etc.) are also available as part of the WMT14 shared task on QE.

**Learning algorithms** We use the Support Vector Machines implementation within `QuEst` to perform either regression (SVR) or classification (SVC) with Radial Basis Function as kernel and parameters optimised using grid search. For SVC, we consider the "one-against-all" approach for multi-class classification with all classes are weighted equally.

**Evaluation metrics** To evaluate our models, we use standard metrics for regression (MAE: mean absolute error; RMSE: root mean squared error) and classification (precision, recall and F1). For each Table and dataset, bold-faced figures represent results that are significantly better (paired t-test with $p \leq 0.05$) with respect to the baseline.

## 4 Classification experiments

Our main motivation to use classifiers is the need to distinguish human from machine translations to isolate the former for the system selection task, since in most settings they are not available. We are also interested in measuring the performance of classification-based QE in system selection.

In the experiments to distinguish human translations from machine translations, we pool all MT and human translations together for each language pair, and build binary classifiers where we label all human translations as 1, and all system translations as 0. Results are given in Table 3, where MC stands for "majority class". They show a large variation across language pairs, although MC is outperformed in all cases in terms of F1. The lower performance for **en-es** and **en-de** may be because here translations from three MT systems are put together, while for the remaining datasets, only two MT systems are available. Nevertheless, figures for **en-es** are substantially better than those for **en-de**. This could also be due to the fact that more high quality **human translations** are available for **es-en** and **de-en** (see Figure 2). On the the other hand, for language combination datasets where more low quality human translations or more high quality machine translations are found, distinguishing between these sets becomes a more difficult task. With similar classifiers (albeit different datasets), Gamon et al. [2005] reported as trivial the problem of distinguishing human translations from machine translations. Overall, our results could indicate that this is a harder problem nowadays than some years ago, possibly pointing in the direction that MT systems produce translations that are closer to human translation nowadays.

Results with SVC in the task of classifying instances with 1-3 labels (including human translations) are shown in Table 4. The performance of

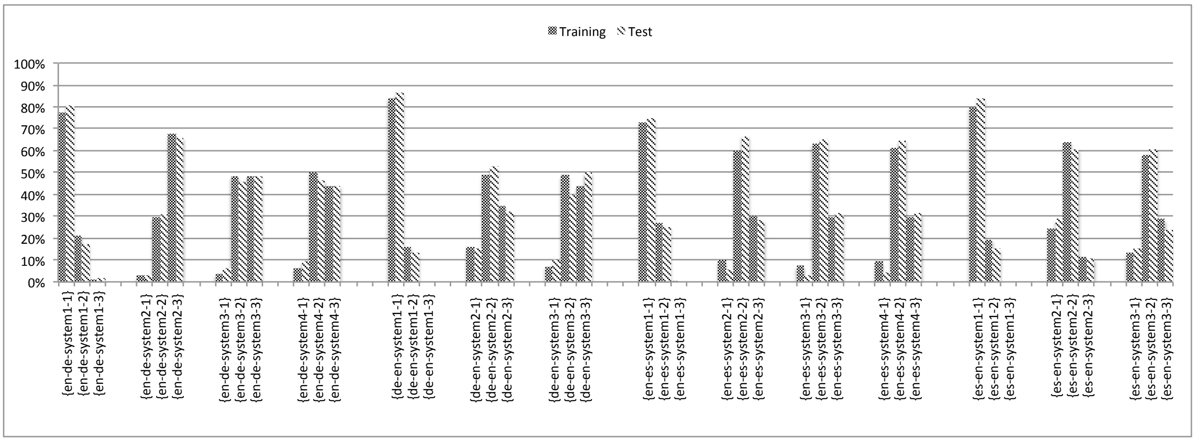

Figure 2: Distribution of true scores for each MT system and language pair.

| | System | #feats | Precision | Recall | F1 |
|---|---|---|---|---|---|
| en-de | MC | - | 0.3041 | 0.1316 | 0.1566 |
| | BL | 17 | **0.3272** | 0.1200 | **0.1756** |
| | AF | 80 | **0.3281** | 0.1193 | **0.1801** |
| de-en | MC | - | 0.5041 | 0.2416 | 0.2961 |
| | BL | 17 | **0.5420** | **0.2321** | **0.3262** |
| | AF | 80 | **0.5468** | **0.2333** | **0.3271** |
| en-es | MC | - | 0.6541 | 0.1521 | 0.2312 |
| | BL | 17 | **0.7012** | 0.1524 | **0.2561** |
| | AF | 80 | **0.7188** | 0.1533 | **0.2527** |
| es-en | MC | - | 0.7311 | 0.3513 | 0.4625 |
| | BL | 17 | **0.7665** | **0.3651** | **0.4942** |
| | AF | 80 | **0.7639** | **0.3667** | **0.4954** |

Table 3: SVC to distinguish between human translations and machine translations (all MT systems). MC corresponds to always picking machine translation (most frequent) as label.

| | System | #feats | Precision | Recall | F1 |
|---|---|---|---|---|---|
| en-de | MC | - | 0.1521 | 0.4231 | 0.2072 |
| | BL | 17 | 0.1600 | 0.4000 | 0.2285 |
| | AF | 80 | **0.3401** | **0.4316** | **0.3078** |
| de-en | MC | - | 0.1121 | 0.3521 | 0.1672 |
| | BL | 17 | **0.1248** | 0.3533 | **0.1844** |
| | AF | 80 | **0.1267** | 0.3512 | **0.1851** |
| en-es | MC | - | 0.2911 | 0.5561 | 0.4014 |
| | BL | 17 | 0.3080 | 0.5550 | 0.3961 |
| | AF | 80 | 0.3092 | 0.5542 | 0.3972 |
| es-en | MC | - | 0.1941 | 0.4516 | 0.2677 |
| | BL | 17 | **0.2075** | 0.4555 | **0.2851** |
| | AF | 80 | **0.2071** | 0.4541 | **0.2855** |

Table 4: SVC to predict 1-3 labels for each language pair, with all translations pooled together. MC corresponds to applying the most frequent class of the training set to all test instances.

the classifiers is compared to the standard baseline of the majority class in the training set (MC). The classifiers perform better than MC for all language pairs except **en-es**, particularly in terms of recall and F1. Since this dataset is significantly larger, the majority class is likely to be more representative of the general data distribution. Overall, the classification results are not very positive, and this corroborates the findings of previous work contrasting classification and regression [Specia et al., 2010].

Overall, the use of all features (**AF**) instead of baseline features (**BL**) only leads to slight improvements in some cases.

## 5 Regression experiments

Here we train models to estimate a continuous score within [1,3], as opposed to discrete 1-3 scores. We compare prediction error for models trained (and tested) on pooled translations from all MT systems (and humans) together (Table 5) – which would be comparable to the settings used to generate Table 4 – against models trained on data from each MT system (or human translation) individually (i.e., system identifier known). For the latter, we consider two settings at test time:

- The system (or human) used to produce the translation is unknown (Table 6 blind setting), and therefore all models are applied, one by one, to predict the quality of this translation and the average prediction is used.
- The system (or human) is known and thus the model for the same translation system/human can be used for prediction (Table 6 non blind setting).

These two variants may be relevant depending on the application scenario. We consider very realistic a scenario where system identifiers are known by developers at model building time, but unknown at test time, e.g. if QE is provided as a web ser-

|  | System | #feats | MAE | RMSE |
|---|---|---|---|---|
| en-de | Mean | - | 0.6831 | 0.7911 |
|  | BL | 17 | **0.6416** | **0.7620** |
|  | AF | 80 | **0.6303** | **0.7616** |
| de-en | Mean | - | 0.6705 | 0.7979 |
|  | BL | 17 | **0.6524** | **0.7791** |
|  | AF | 80 | **0.6518** | **0.7682** |
| en-es | Mean | - | 0.4585 | 0.6678 |
|  | BL | 17 | 0.5240 | 0.6590 |
|  | AF | 80 | 0.5092 | 0.6442 |
| es-en | Mean | - | 0.5825 | 0.6718 |
|  | BL | 17 | 0.5736 | 0.6788 |
|  | AF | 80 | **0.5662** | **0.6663** |

Table 5: SVR to build predictions models for each language pair combination, with all translation sources (including human) pooled together.

vice with pre-trained models (Table 6). Users may request predictions using translations produced by any sources, and for out-of-domain data (Table 7). In all tables, **Mean** represents a strong baseline: assigning the mean of the training set labels to all test set instances.

Comparing the two variants of the blind setting (Tables 5 - blind training and test; and Table 6, blind test only), we see that pooling the data from multiple translation systems for blind model training leads to significantly better results than training models for individual translation sources but testing them in blind settings. This is likely to be due to the larger quantities of data available in the pooled models. In fact, the best results are observed with **en-es**, the largest dataset overall.

Comparing scores between blind versus non-blind test setting in Table 6, we observe a substantial difference in the scores for each of the individual translation system. This shows that the task is much more challenging when QE models are trained independently, but the identifiers of the systems producing the test instances are not known.

There is also a considerable difference in the performance of individual models for different translation systems, which can be explained by the different distribution of scores (and also indicated by the performance of the **Mean** baseline). However, in general the prediction performance of the individual models seems less stable, and worse than the baseline in several cases. Interestingly, the individual models trained on human translations only (system1) do even worse than individual models for MT systems. This can be an indication that the features used for quality prediction are not sufficient to model human translations.

In all cases, the use of all features (**AF**) instead of baseline features (**BL**) comparable or better results.

Table 7 shows the results for SVR models trained on pooled models for each language pair (i.e., models in Table 5) when applied to predict the quality of the out-of-domain datasets. This is an extremely challenging task, as the only constant between training and test data is the language pair. The text domain is different, and so are MT systems used to produce the translations. In addition, no human translation is available in the out-of-domain test sets. Surprisingly, the prediction errors are low, even lower than those observed for the in-domain test sets. This is true for all except two out-of-domain test sets: $\text{LSP}_5$, which contains unusual texts (such as URLs and markup tags), and $\text{LSP}_2$. Manual inspection of these source and translation segments show many extremely short segments (1-2 words), which may render most features useless.

| WMT14 | System | #features | MAE | RMSE |
|---|---|---|---|---|
| $\text{LSP}_1$ (en-es) | Mean | - | 0.2715 | 0.4311 |
|  | BL | 17 | **0.2524** | **0.4116** |
|  | AF | 80 | **0.2419** | **0.4076** |
| $\text{LSP}_2$ (en-es) | Mean | - | 0.8119 | 0.9703 |
|  | BL | 17 | 0.8094 | **0.9470** |
|  | AF | 80 | **0.8062** | **0.9453** |
| $\text{LSP}_3$ (en-de) | Mean | - | 0.4315 | 0.5914 |
|  | BL | 17 | 0.4270 | **0.5500** |
|  | AF | 80 | 0.4262 | **0.5463** |
| $\text{LSP}_4$ (es-en) | Mean | - | 0.5012 | 0.6711 |
|  | BL | 17 | **0.4847** | **0.6412** |
|  | AF | 80 | **0.4812** | **0.6392** |
| $\text{LSP}_5$ (es-en) | Mean | - | 0.7112 | 0.8881 |
|  | BL | 17 | **0.6862** | **0.8447** |
|  | AF | 80 | **0.6828** | **0.8472** |

Table 7: Results with SVR pooled models tested on out-of-domain datasets.

## 6 System selection results

In what follows we turn to using the predictions from SVR and SVC models showed before for system selection. The task consists in selecting, for each source segment, the best *machine* translation among all available: two or three depending on the language pair. For this experiments, we eliminated the human translations – as they do not tend to be represented in settings for system selection. Given the low performance of our classifiers in Table 3, we ruled out human translations based on the metadata available, without using these classifiers. Another reason to rule out human translations from the selection is that they are used as references to

|  | System | #feats | blind |  | non-blind |  |
|---|---|---|---|---|---|---|
|  |  |  | MAE | RMSE | MAE | RMSE |
| en-de-system1 | Mean | - | 1.0351 | 1.2133 | 0.3552 | 0.4562 |
|  | BL | 17 | 1.0487 | 1.2348 | **0.3350** | 0.4540 |
|  | AF | 80 | 1.0510 | 1.2375 | **0.3325** | 0.4545 |
| en-de-system2 | Mean | - | 0.7780 | 0.9339 | 0.4857 | 0.5487 |
|  | BL | 17 | **0.7006** | 0.9499 | **0.3615** | **0.4634** |
|  | AF | 80 | **0.6924** | **0.9124** | **0.3570** | **0.4644** |
| en-de-system3 | Mean | - | 0.7369 | 0.8426 | 0.5577 | 0.6034 |
|  | BL | 17 | **0.6354** | **0.7950** | **0.4535** | **0.5363** |
|  | AF | 80 | **0.6572** | 0.8127 | **0.4482** | **0.5245** |
| en-de-system4 | Mean | - | 0.7231 | 0.8215 | 0.5782 | 0.6433 |
|  | BL | 17 | **0.6438** | **0.7842** | **0.4912** | **0.5834** |
|  | AF | 80 | **0.6386** | **0.7905** | **0.4818** | **0.5741** |
| de-en-system1 | Mean | - | 0.8594 | 1.0882 | 0.2506 | 0.3409 |
|  | BL | 17 | 0.8747 | 1.1299 | **0.2123** | 0.3421 |
|  | AF | 80 | 0.8747 | 1.1299 | **0.2065** | 0.3415 |
| de-en-system2 | Mean | - | 0.7321 | 0.8484 | 0.5412 | 0.6678 |
|  | BL | 17 | **0.6897** | 0.8330 | **0.4745** | **0.5931** |
|  | AF | 80 | 0.7122 | 0.8509 | **0.4604** | **0.5850** |
| de-en-system3 | Mean | - | 0.8137 | 0.9253 | 0.6000 | 0.6640 |
|  | BL | 17 | **0.7472** | **0.8903** | **0.4965** | **0.6011** |
|  | AF | 80 | **0.7629** | 0.9300 | **0.4828** | **0.5901** |
| en-es-system1 | Mean | - | 0.8542 | 0.9923 | 0.3883 | 0.4353 |
|  | BL | 17 | 0.8956 | 1.0480 | **0.3633** | 0.4390 |
|  | AF | 80 | 0.8957 | 1.0480 | **0.3519** | 0.4381 |
| en-es-system2 | Mean | - | 0.5567 | 0.6952 | 0.4232 | 0.5314 |
|  | BL | 17 | **0.5275** | 0.6827 | **0.3812** | **0.4951** |
|  | AF | 80 | **0.5302** | 0.6884 | **0.3730** | **0.4893** |
| en-es-system3 | Mean | - | 0.5653 | 0.6998 | 0.4288 | 0.5213 |
|  | BL | 17 | **0.5155** | **0.6711** | **0.3821** | **0.4844** |
|  | AF | 80 | **0.5184** | **0.6704** | **0.3714** | **0.4761** |
| en-es-system4 | Mean | - | 0.5573 | 0.6955 | 0.4300 | 0.5321 |
|  | BL | 17 | **0.5103** | **0.6680** | **0.4022** | **0.5162** |
|  | AF | 80 | **0.5206** | **0.6727** | **0.3902** | **0.5016** |
| es-en-system1 | Mean | - | 0.6617 | 0.8307 | 0.3026 | 0.3916 |
|  | BL | 17 | 0.6617 | 0.8307 | 0.3022 | 0.3917 |
|  | AF | 80 | 0.6617 | 0.8308 | 0.3023 | 0.3915 |
| es-en-system2 | Mean | - | 0.5637 | 0.6931 | 0.4494 | 0.6027 |
|  | BL | 17 | 0.5588 | 0.7023 | **0.4384** | 0.6061 |
|  | AF | 80 | 0.5567 | 0.7026 | **0.4309** | 0.6053 |
| es-en-system3 | Mean | - | 0.6602 | 0.8129 | 0.4720 | 0.6245 |
|  | BL | 17 | 0.7233 | 0.8621 | 0.4993 | 0.6220 |
|  | AF | 80 | 0.6973 | 0.8435 | 0.4974 | 0.6198 |

Table 6: SVR to build individual predictions models for each language pair and translation source.

compute BLEU scores of the selected sets of sentences, as explained below.

To provide an indication of the average quality of each MT system, Table 8 presents the BLEU scores on the test and training sets for individual MT systems. The bold-face figures for each language test set indicate the (BLEU) quality that would be achieved for that test set if the "best" system were selected on the basis of the average (BLEU) quality of the training set (i.e., no system selection). There is a significant variance in terms of quality scores, as measured by BLEU, among the outputs of 2-3 MT systems for each language pair, with training set quality being a good predictor of test set quality for all but **en-es**, once again, the largest dataset.

We measure the performance of the selected sets in two ways: (i) by computing the BLEU scores of the entire sets containing the supposedly best translations, using the human translation available in the datasets as reference, and (ii) by computing the accuracy of the selection process against the human labels, i.e., by computing the proportion of times both system selection and human agree (based on the pre-defined 1-3 human labels) that the sentence selected is the best among the 2-3 options (2-3 MT systems). We compare the results obtained from building pooled (all MT systems)

| WMT14 | system2 | | system3 | | system4 | |
|---|---|---|---|---|---|---|
| | Test | Training | Test | Training | Test | Training |
| en-de | 15.39 | 12.79 | 13.75 | 13.83 | **17.04** | 16.19 |
| de-en | **27.96** | 24.03 | 22.66 | 20.19 | - | - |
| en-es | **25.89** | 34.13 | 32.68 | 28.42 | 29.25 | 31.97 |
| es-en | **37.83** | 40.01 | 23.55 | 25.07 | - | - |

Table 8: BLEU scores of individual MT systems, without system selection. Bold-faced figures indicate scores obtained when selecting best system on average (using BLEU scores for the training set).

against individual prediction models (one per MT system).

Table 9 and 10 show the selection results with various models trained on MT translations only:

- **Best-SVR(I):** Best translation selected with regression model trained on data from individual MT systems, where prediction models are trained per MT system, and the translation selected for each source segment is the one with the highest predicted score among these independent models. This requires knowing the source of the translations for training, but not for testing (blind test).

- **Best-SVR(P):** Best translation selected with regression model trained on pooled data from all MT systems. This assumes a *blind* setting where the source of the translations for both training and test sets is unknown, and thus pooling data is the only option for system selection.

- **Best-SVC(P):** Best translation selected with the classification model trained on pooled data from all MT systems as above. For SVC, only the pooled models were used as predicting exact 1-3 labels with independently trained models leads to an excessively number of ties (i.e., multiple translations with same score), making the decision between them virtually arbitrary.

Table 9 shows that the regression models trained on individual systems – **Best-SVR(I)** – with **AF** as feature set yield the best results, despite the fact that error scores (MAE and RMSE) for these individual systems are worse than for systems trained on pooled data. This is somewhat expected as knowing the system that produced the translation (i.e., training models for each MT system) adds a strong bias to the prediction problem towards the average quality of such a system, which is generally a decent quality predictor. We note however that the **Best-SVR(P)** models are not far behind in terms of performance, with the **Best-SVC(P)** following closely. In all cases, the gains with respect to the MC baseline are substantial. More important, we note the gains in BLEU scores as compared to the bold-face test set figures in Table 8, showing that our system selection approach leads to best translated test sets than simply picking the MT system with best average quality (BLEU).

Results in terms of accuracy in selecting the best translation (Table 10) are similar to those in terms of BLEU scores, with models trained independently performing best.

## 7 Remarks

We have presented a number of experiments showing the potential of a system selection techniques in scenarios where translations are given by multiple MT systems and system identifiers are unknown. System selection was performed based on predictions from classification and regression models. Results in terms of BLEU and accuracy of selected sets with an MT system-agnostic approach show improvements for system selection over strong baselines.

Overall – in bind test settings – although the prediction error of models trained on individual MT systems are worse than models trained on pooled data, when used for system selection, models trained on individual systems generally perform better.

## Acknowledgments

This work was supported by funding from the from European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 296347 (QTLaunchPad).

## References

E. Avramidis. Sentence-level ranking with quality estimation. *Machine Translation*, 28:1–20,

|  | System | #feats | Best-SVR(I) | Best-SVR(P) | Best-SVC(P) |
|---|---|---|---|---|---|
| en-de | MC | - | 16.14 | 15.55 | 16.12 |
|  | BL | 17 | 17.20 | 17.05 | 17.33 |
|  | AF | 80 | **18.10** | 17.55 | 17.32 |
| de-en | MC | - | 25.81 | 25.17 | 25.07 |
|  | BL | 17 | 28.39 | 28.13 | 28.19 |
|  | AF | 80 | **28.75** | 28.43 | 28.21 |
| en-es | MC | - | 30.88 | 30.29 | 30.23 |
|  | BL | 17 | 32.92 | 32.81 | 32.74 |
|  | AF | 80 | **33.45** | 33.25 | 33.10 |
| es-en | MC | - | 30.13 | 29.70 | 29.53 |
|  | BL | 17 | 38.10 | 38.11 | 38.14 |
|  | AF | 80 | **38.73** | 38.41 | 38.15 |

Table 9: BLEU scores on best selected translations (I = Individual, P = Pooled).

|  | System | #feats | Best-SVR(I) | Best-SVR(P) | Best-SVC(P) |
|---|---|---|---|---|---|
| en-de | MC | - | 0.1823 | 0.1787 | 0.1793 |
|  | BL | 17 | 0.2017 | 0.2001 | 0.2033 |
|  | AF | 80 | **0.2155** | 0.2055 | 0.2065 |
| de-en | MC | - | 0.3511 | 0.3527 | 0.3559 |
|  | BL | 17 | 0.3733 | 0.3711 | 0.3713 |
|  | AF | 80 | 0.3915 | **0.3923** | 0.3821 |
| en-es | MC | - | 0.3698 | 0.3643 | 0.3659 |
|  | BL | 17 | 0.3912 | 0.3901 | 0.3892 |
|  | AF | 80 | **0.4102** | 0.4087 | 0.4051 |
| es-en | MC | - | 0.4073 | 0.4043 | 0.4059 |
|  | BL | 17 | 0.4321 | 0.4301 | 0.4290 |
|  | AF | 80 | **0.4513** | 0.4421 | 0.4423 |

Table 10: Accuracy in selecting the best translation for each dataset (I = Individual, P = Pooled).

2013.

E. Avramidis and M. Popović. Machine learning methods for comparative and time-oriented Quality Estimation of Machine Translation output. In *8th WMT*, pages 329–336, Sofia, 2013.

E. Biçici. Referential translation machines for quality estimation. In *8th WMT*, Sofia, 2013.

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

O. Bojar, C. Buck, C. Callison-Burch, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. Findings of the 2013 WMT. In *8th WMT*, pages 1–44, Sofia, 2013.

L. Formiga, M. González, A. Barrón-Cedeno, J. A. Fonollosa, and L. Màrquez. The TALP-UPC approach to system selection: Asiya features and pairwise classification using random forests. In *8th WMT*, pages 359–364, Sofia, 2013.

M. Gamon, A. Aue, and M. Smets. Sentence-level MT evaluation without reference translations: beyond language modeling. In *EAMT-2005*, Budapest, 2005.

Y. He, Y. Ma, J. van Genabith, and A. Way. Bridging smt and tm with translation recommendation. In *ACL-2010*, pages 622–630, Uppsala, Sweden, 2010.

S. Hildebrand and S. Vogel. MT quality estimation: The CMU system for WMT'13. In *8th WMT*, pages 373–379, Sofia, 2013.

K. Shah, E. Avramidis, E. Biçici, and L. Specia. Quest - design, implementation and extensions of a framework for machine translation quality estimation. *Prague Bull. Math. Linguistics*, 100: 19–30, 2013.

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

L. Specia, D. Raj, and M. Turchi. Machine translation evaluation versus quality estimation. *Machine Translation*, pages 39–50, 2010.

L. Specia, K. Shah, J. G. C. d. Souza, and T. Cohn. Quest - a translation quality estimation framework. In *ACL-2013 Demo Session*, pages 79–84, Sofia, 2013.