Predicting Human Translation Quality

Lucia Specia l.specia@sheffield.ac.uk
Kashif Shah kashif.shah@sheffield.ac.uk
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello Street, S4 1DP, UK

Abstract

We present a first attempt at predicting the quality of translations produced by human, professional translators. We examine datasets annotated for quality at sentence- and word-level for four language pairs and provide experiments with prediction models for these datasets. We compare the performance of such models against that of models built from machine translations, highlighting a number of challenges in estimating quality and detecting errors in human translations.

1 Introduction

Metrics for translation quality estimation (QE) (Blatz et al., 2004; Specia et al., 2009) aim at providing an estimate on the quality of a translated text. Such metrics have no access to reference translations, as they are intended for translation systems in use. QE has shown promising results in several applications in the context of Machine Translation (MT), such as improving post-editing efficiency by filtering out low quality segments which would require more effort to correct than translating from scratch (Specia et al., 2009; Specia, 2011), selecting high quality segments to be published as they are, without post-editing (Soricut and Echihabi, 2010), ranking or selecting the best translation from multiple MT systems (Specia et al., 2010; Hildebrand and Vogel, 2013; Avramidis, 2013; Avramidis and Popović, 2013), or between translations from either an MT system or a translation memory (He et al., 2010), and highlighting sub-segments that need revision (Bach et al., 2011).

Generally speaking, QE models are built using supervised machine learning algorithms from examples of translations at a given granularity level (e.g. sentences). For training, these examples are annotated with quality labels and described by a number of features that can approximate quality (or errors). "Quality" is therefore defined according to the problem at hand and the labelled data, for example, post-editing time for a sentence or word-level errors. For an overview of various algorithms and features we refer the reader to the WMT12-14 shared tasks on QE (Callison-Burch et al., 2012; Bojar et al., 2013, 2014).

So far, QE has only been applied to machine translated texts. However, the above mentioned applications are also valid in the context of human translation. In particular, in scenarios where translations produced by humans may be of variable or questionable levels of reliability (e.g. crowdsourcing), it becomes important to estimate translation quality to, for example, select among multiple options of human translations (or even a mix of human and machine translations). In addition, even with professionally created translations, quality assurance is a common process and an estimation method could be useful, for example, to sample the lowest quality cases for checking/revision.

Even though it is known that human translations are generally different from machine translations, we put forward the hypothesis that it is possible and useful to have automated

metrics to estimate translation quality of both human and machine translations. In this paper we analyse existing human translations annotated for quality and errors and contrast them to machine translations. We use this data to experiment with an existing framework for quality estimation to predict quality in human translations. More specifically, we aim at answering the following questions:

1 Can we automatically distinguish machine from human translations?

2 Do professional human translators make mistakes?

3 Are human translation errors the same as machine translation errors?

4 Can quality estimation approaches capture issues in human translations?

We discuss each of these questions in Sections 3, 4, 5, and 6, respectively. Before that, we introduce the datasets and settings used in our experiments in Section 2.

2 Datasets and experimental settings

2.1 Datasets

Our datasets are those used for the WMT14 shared task on quality estimation[1] and were produced in the context of the QTLaunchPad project.[2] They contain *news* 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 language pair dataset 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 **MT-1**, **MT-2**, **MT-3**) which participated in the translation shared task in 2103, plus the professional translation provided by WMT as reference (**HT**). In addition, for the word-level analysis, for all language pairs except English→Spanish, which already had enough sentences, we included some customer data (mostly technical documentation) provided and annotated by language service providers as part of the QTLaunchPad project.

This data is very different from existing corpora of human translations annotated for quality. Existing resources contain translations from students, while ours only contain translations produced by professional translators, and annotated by (other) professional translators. In addition, our data contains translations from multiple state of the art MT systems, also annotated by professional translators. For comparison purposes, in the remaining of the paper we report statistics for the human versus all MT data together.

Sentence-level data At sentence-level, the details about the datasets are given in Table 1. All translations for each source sentence were annotated by a single professional translator (and that one translator annotated all sentences for a given language pair) using the following three options representing the translator's perception on the effort that would be needed to post-edit such a sentence:

- **1** = Perfect translation, no post-editing needed at all.
- **2** = Near-miss translation: translation contains a maximum of 2-3 errors, and possibly additional errors that can be easily fixed (capitalisation, punctuation, etc.).
- **3** = Very low quality translation, cannot be easily fixed.

[1] http://www.statmt.org/wmt14/quality-estimation-task.html
[2] http://www.qt21.eu/launchpad/

# Source	# HT+MTs	# Target
1,104 English	4	4,416 Spanish
500 English	4	2,000 German
500 German	3	1,500 English
500 Spanish	3	1,500 English

Table 1: Number of source and target sentences labelled for post-editing effort at sentence-level.

Word-level data For word-level annotation, a subset of sentences of type "2" (near-miss) from MT systems and from human translators (Table 2) were annotated with core issue types (errors) of the Multidimensional Quality Metric (MQM),[3] as shown in Figure 1. In addition to the 16 fine-grained labels, two levels of labels were automatically generated by climbing up the MQM hierarchy: Accuracy versus Fluency, and OK (no issue) versus BAD (any issue). Each translation was annotated by 1-5 professional translators. For translations annotated by more than one translator, only one annotation was randomly selected and used in our analysis. For a discussion on annotator agreement within these datasets, see (Lommel et al., 2014).

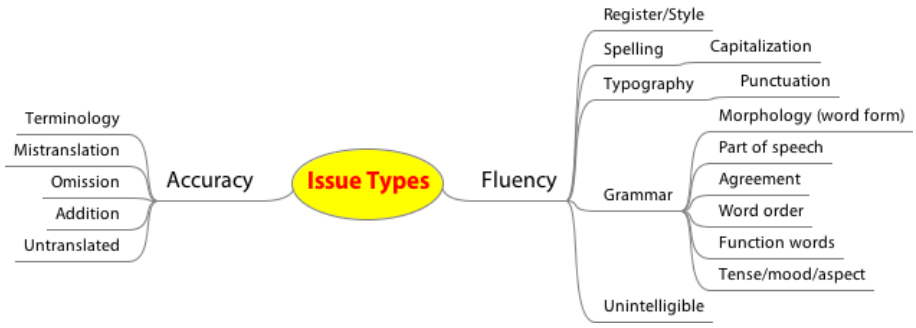


Figure 1: MQM core issue types used for the word-level annotation task.

Source→target	# WMT (news)	# Technical
English→Spanish	2,339	-
English→German	467	398
German→English	250	200
Spanish→English	440	610

Table 2: Number of sentences labelled at word-level, from news and technical domains.

2.2 Settings

Prediction models are only built for sentence-level, given the small number of human translations labelled at word-level (at most 294, for en-es). Our word-level analysis focuses on error distributions. For the building and evaluation of sentence-level prediction models (as described in Sections 3 and 6), we use the following settings.

Dataset splits We use the standard training and test splits as distributed by WMT14: each MT system or HT dataset is split into 70% for training and 30% for test.

[3] http://www.qt21.eu/launchpad/content/background-and-principles

Learning algorithms We use the Support Vector Machines (SVM) implementation within the `QuEst` toolkit for quality estimation[4] (Specia et al., 2013; Shah et al., 2013) to perform classification (SVC) (Section 3) and regression (SVR) (Section 6) with Radial Basis Function as kernel and parameters optimised using grid search.

Evaluation metrics To evaluate our models, we use standard metrics for regression (MAE: mean absolute error) and classification (precision, recall and F1). In all tables, bold-faced figures are significantly better (paired t-test with $p \leq 0.05$) wrt the baseline for the given language pair. As baseline for the regression models, we consider the **Mean** of the training data, i.e., simply outputting the average value of the training set to all test instances. Similarly, as baseline for the classification models, we consider assigning the most frequent class (**MC**) in the training set to all test instances.

Features We use the `QuEst` toolkit to extract two feature sets for each dataset:

- Baseline features (**BL**): 17 features used as baseline in the WMT shared tasks on QE. Examples of baseline features for sentence-level include the following:
  - no. of tokens in the source & target texts
  - average source token length
  - average **no. of occurrences** of target words in target text
  - no. of punctuation marks in source & target texts
  - language model probability of source & target texts using LMs built from large source/target language corpora of human texts
  - avg. no. of translations per source word built using lexical tables from the IBM 1 model thresholded such that $P(t|s) > 0.2$)
  - % of 1-grams, 2-grams & 3-grams in frequency quartiles 1 & 4 (lower/higher frequency) in a large corpus of the source language
  - % of 1-grams in source text seen in a large corpus of the source language
- All features (**AF**): 80 common MT system-independent features (superset of **BL**).

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

3 Can we distinguish machine from human translations?

In this experiment we train an SVM classifier to distinguish human translations from machine translations at sentence-level. We put together all MT and human translations for each language pair, label all human translations as 1, and all system translations as 0. We then train a binary classifier to distinguish them. Results are given in Table 3, where MC stands for "majority class" (always picking MT). 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 (only 25% of the examples are HT), while for the remaining datasets, only two MT systems are available, and therefore the data distribution is less skewed (33% of the examples are HT). Nevertheless, figures for **en-es** are substantially better than those for **en-de**, possibly because of the larger size of the **en-es** dataset.

[4] `http://www.quest.dcs.shef.ac.uk/`

With similar classifiers (albeit different datasets), Gamon et al. (2005) reported as trivial the problem of distinguishing human translations from machine translations. However, our results seem to indicate that this is now a harder problem than some years ago, possibly pointing in the direction that MT systems produce more translations that are better in quality, and therefore closer to human translation nowadays. Moreover, human translations also contain errors, which gives us a further motivation for modelling the prediction of quality in human translations (see Figure 2).

	Model	#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: Performance of classifier to distinguish between human translations and machine translations (all MT systems together). "MC" corresponds to always picking machine translation (most frequent) as label.

4 Do professional human translators make mistakes?

In order to answer this question, we look at the distribution of the 1-3 scores at sentence-level (Figure 2) and the distribution of OK versus BAD word-level labels (Figure 3). Both sets of distributions show that, for all language pairs, human translations (HT), albeit professionally produced, contain errors. In the sentence-level figures, the first set of bars for all language pairs show that in the best case only about 80% of the human translations are labelled "1" (perfect). While – not surprisingly – very low quality translations (label "3") are virtually non-existent (maximum 1.2%), many cases of near-misses are found for all language pairs. For English→Spanish, 27% of the translations are considered near-misses, whereas for other languages pairs this rate is between 15 and 20%. The bars for MT systems essentially show the inverse behaviour: very few perfect translations (less than 10% for all language pairs except Spanish→English), predominantly near-miss translations for English↔Spanish, and a mostly even distribution between very low quality and near-miss translations for German↔English.

It is worth noticing that the translators annotating datasets for errors received explicit guidelines to consider only *true errors* for the annotation. They were instructed not to label any segment/word as incorrect or near-miss because of *preferential changes*, i.e., because they would simply have preferred a different translation. They were also instructed to consider a segment/word *correct* when they were not sure about such a segment/word because of lack of context, style guidelines, etc. Some examples of near-miss human translations (with issues highlighted and identified) are shown in Table 4.

Looking at the distribution of OK and BAD word-level annotations (Figure 3), we see that even though both HT and MT segments had already been pre-labelled as near-misses (i.e., as

Figure 2: Percentage of 1-3 scores given as labels at sentence-level data for human (HT) and each machine (MT-i) translation system.

Lang.	Source	Target	Issues
de-en	Deutsche Welle: Anfang der Woche hatte Deutschland zunächst signalisiert, dass es gegen den Antrag auf einen Beobachterstatus der Palästinenser bei den Vereinten Nationen stimmen würde.	Deutsche Welle: At the beginning of the week, Germany had initially signalled that **it** would vote against the Palestinians' application for observer status within the United Nations.	**agreement**
en-de	So I had plenty of time to think about the subject of boredom.	So hatte ich viel Zeit, um an das Thema **der** Langeweile zu denken.	**grammar**
en-es	People assume we are like the Bullingdon Club without meeting us.	La gente supone que **parezcamos** *al* Club Bullingdon sin vernos	**mistranslation**, *function words*, mistranslation
es-en	La princesa D'Arenberg guarda sus vestidos de fiesta del modisto con "los máximos cuidados... porque un vestido no es solamente un vestido, también es el conjunto de recuerdos que conlleva".	Princess D'Arenberg looks after her couturier **gowns** with "the utmost care... because a dress *not* just a dress, it's also the many memories that go with it.	**terminology**, *omission*

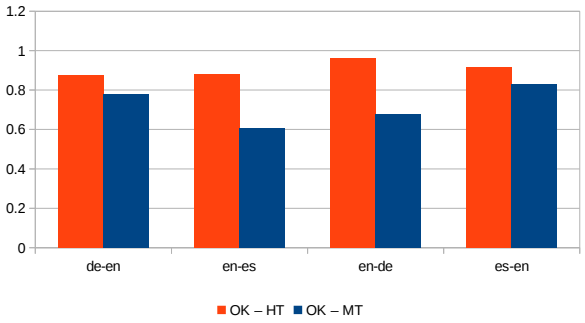
Table 4: Examples of near-miss human translations. Issues are highlighted and listed in order.

containing 1-3 errors that are easy to fix), as expected, MT segments contain more errors for all language pairs. Only up to 10% of the words in HT segments contain errors. For MT, this percentage reaches 40% for English→Spanish.

5 Are human translation errors the same as machine translation errors?

To answer this question we look at the distribution of specific issue types coming from the word-level annotation. In Figure 4 we show the distribution of errors in HT and MT grouped by fluency and accuracy types (here we ignore the "OK" category for clarity purposes). Once again, these statistics only consider segments that had already been pre-labelled as near-misses. For all language pairs except English→German, MT segments tend to contain considerably more words labelled as having fluency issues than as containing accuracy issues. In human translations, however, fluency issues are more frequent in language pairs involving Spanish, whereas accuracy issues are more frequent in language pairs involving German, although English→German shows a close distribution between accuracy and fluency issues. This seems to indicate that the types of errors in translations may be more dependent on the language pair than on the type of translation (MT or HT).

A more detailed view on the types of errors by HT and MT is given in Figure 5. Here we look at percentages of specific issues (again ignoring the "OK" category) in human and machine (a mixture of all MT systems) near-miss translations. Given the limited size of the datasets, some issues are not observed for certain language pairs. Overall, the distributions of specific issue types are very distinct in HT and MT segments, as well as across language pairs. Mistranslation is by far the most frequent error type in human translations for German↔English. For Spanish↔English, fluency errors are the most frequent. We note that the latter are not a combination of all errors under "fluency" in Figure 1. Instead, they are a more general category that annotators were asked to use when they could not flag the specific fluency issue with the word.



Lang	Words tagged	
	HT	MT
de-en	808	7420
en-es	8933	48089
en-de	1241	12406
es-en	1206	21818

Figure 3: Percentage of words labelled as OK versus BAD in human (HT) and machine (MT) near-miss translations (MT contains a mixture of all MT systems). The table shows the number of words tagged for issues, including the "OK" tag, which in fact means that no issue was found for the word.

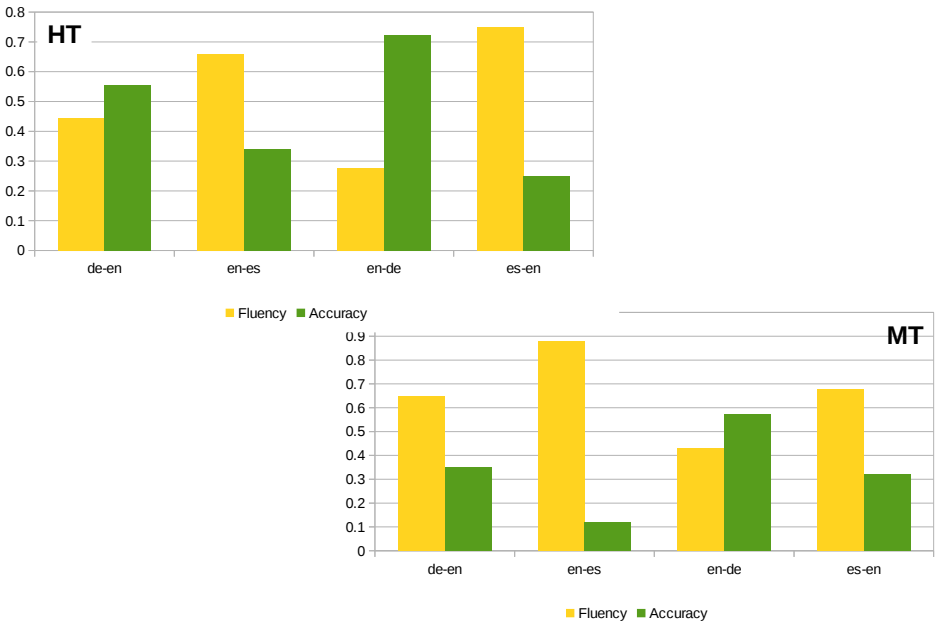


Figure 4: Percentage of words labelled as containing fluency versus accuracy issues in human (HT) and machine (MT) near-miss translations (MT contains a mixture of all MT systems).

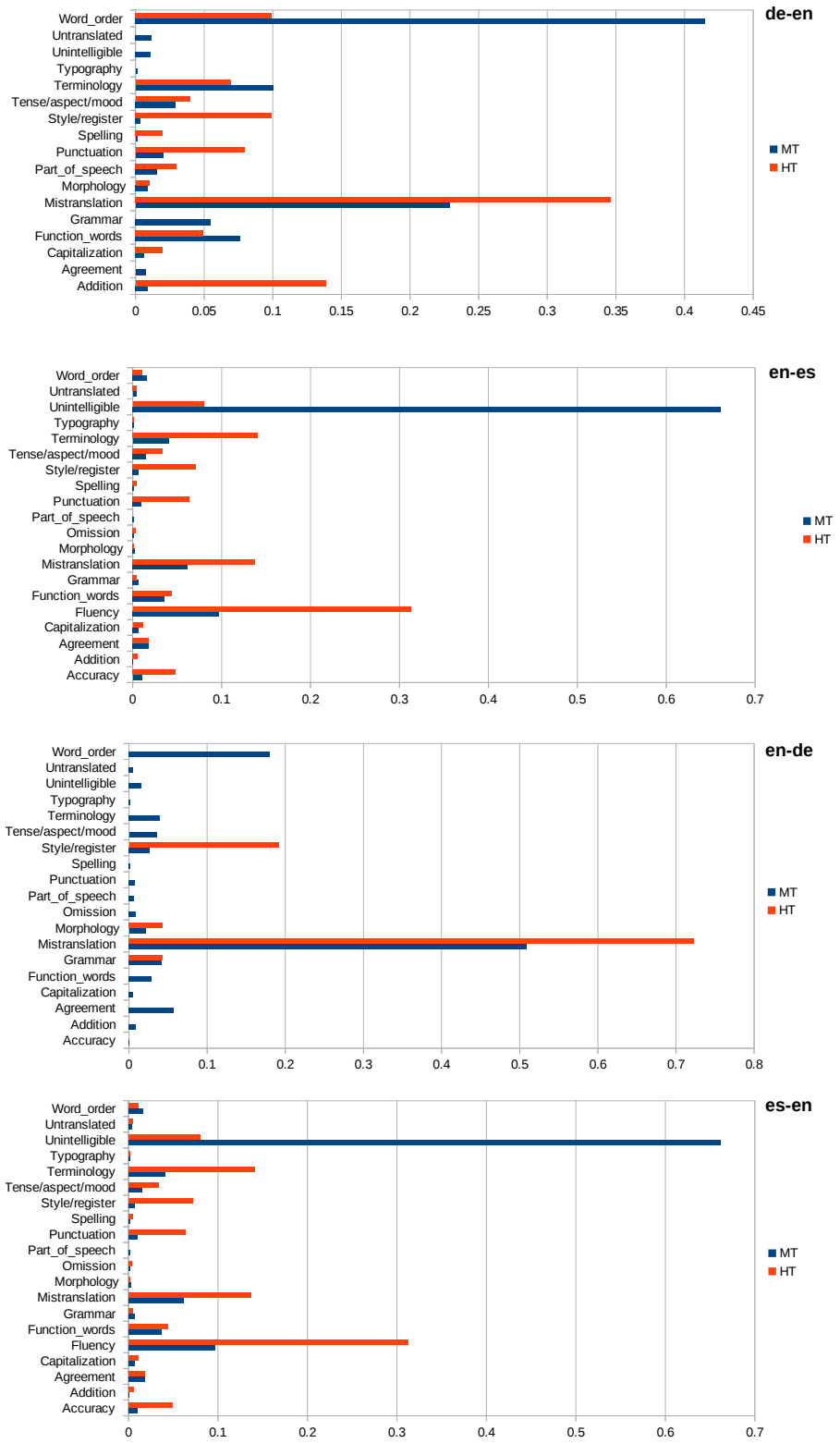


Figure 5: Percentage of words labelled with each type of MQM issue in human (HT) and machine (MT) near-miss translations.

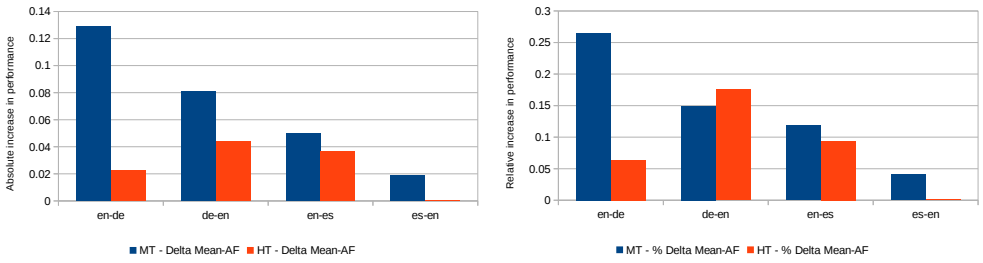


Figure 6: Absolute and relative improvement of prediction models over Mean baseline for machine (MT) and human (HT) translation data. Only the most predictable (lowest MAE score) MT system is shown for each language pair.

6 Can quality estimation approaches capture issues in human translations?

In what follows we show the performance of regression models trained on HT and MT data independently (Table 5), for the sentence-level annotated data. The performance obtained for models trained on MT data is comparable to the state of the art, based on the results of the latest WMT14 shared task (Bojar et al., 2014). In absolute terms, the figures show that models trained on HT datasets are better (lower MAE) than models trained on any MT dataset, for all language pairs. That could be seen as indicative that tools used for MT quality estimation are also applicable for HT quality estimation. However, although all HT and MT models were trained on datasets of the same size, the distribution of scores in each of these datasets is very different (see Figure 2). Human translations are "perfect" in approximately 80% of the cases for all languages. Therefore, it becomes much harder to outperform the "Mean" baseline in HT models. This is reflected in the consistently lower MAE scores obtained by the Mean baseline on the HT data. Therefore, a better way of comparing the performance of models for HT against models for MT is to measure the improvement on the MAE scores between the Mean baseline and the best (**AF**) prediction model. The absolute and relative improvements for each language pair are shown in Figure 6. In terms of absolute improvement, the figures for MT are always more substantial than those for HT. This is also the case in relative terms, except for German→English, where the HT model achieves relatively better improvement over the Mean baseline than the MT models, although the difference is minor (18% improvement versus 15% improvement).

Our results seems to indicate that it is generally harder to predict human translation quality. In addition to the highly skewed data distribution, one reason for that could be that errors in human translations may be more subtle than in machine translations, requiring more sophisticated features than the ones used in current quality estimation approaches. In fact, another interesting finding from Table 5 is that there is zero or little gain for moving from the BL to the AF feature sets for HT, whereas the gain is evident for models built from MT data. This seems to indicate again that the features we resort to are not appropriate or sufficient to capture the quality of human translations.

To further inspect this problem, we take the MQM core issue types (see Figure 1) as guidance on the types of quality issues features should attempt to capture. We note that many issue types are not covered at all or only approximated by features in current quality estimation approaches. In what follows we provide a discussion for each issue type:[5]

[5]A detailed description of the issue types can be found on `http://www.qt21.eu/launchpad/content/list-mqm-issue-types`

en-de	Model	#feats	MAE
HT	Mean	-	0.3552
	BL	17	**0.3350**
	AF	80	**0.3325**
MT-1	Mean	-	0.4857
	BL	17	**0.3615**
	AF	80	**0.3570**
MT-2	Mean	-	0.5577
	BL	17	**0.4535**
	AF	80	**0.4482**
MT-3	Mean	-	0.5782
	BL	17	**0.4912**
	AF	80	**0.4818**

en-es	Model	#feats	MAE
HT	Mean	-	0.3883
	BL	17	**0.3633**
	AF	80	**0.3519**
MT-1	Mean	-	0.4232
	BL	17	**0.3812**
	AF	80	**0.3730**
MT-2	Mean	-	0.4288
	BL	17	**0.3821**
	AF	80	**0.3714**
MT-3	Mean	-	0.4300
	BL	17	**0.4022**
	AF	80	**0.3902**

de-en	Model	#feats	MAE
HT	Mean	-	0.2506
	BL	17	**0.2123**
	AF	80	**0.2065**
MT-1	Mean	-	0.5412
	BL	17	**0.4745**
	AF	80	**0.4604**
MT-2	Mean	-	0.6000
	BL	17	**0.4965**
	AF	80	**0.4828**

es-en	Model	#feats	MAE
HT	Mean	-	0.3026
	BL	17	0.3022
	AF	80	0.3023
MT-1	Mean	-	0.4494
	BL	17	**0.4384**
	AF	80	**0.4309**
MT-2	Mean	-	0.4720
	BL	17	0.4993
	AF	80	0.4974

Table 5: Error (MAE) scores for prediction models built for each language pair and translation system. **Mean** indicates a baseline that always outputs the average score of the training set. **BL** indicates the set of simple model using baseline features. **AF** indicates models built using all features.

Accuracy

- *Terminology*: Normative terminology infringed. This issue is not directly covered by current approaches to quality estimation. However, as a proxy to it, both monolingual (target) and bilingual terminology lists could be used for simple checks, such as whether all content words (or nouns) in the translation belong to the terminology list.
- *Mistranslation*: Incorrect word translation chosen (overly literal, false friend, should not have been translated, entity, date/time/number, unit conversion). This issue cannot be easily automated, apart from some mechanical checks on date/time/number format.
- *Omission*: Translation for source word is missing. Certain existing features approximate this issue type, e.g., source versus target segment word counts, counts of words with certain POS tags in both source and target segments, and language models of the target language, which can detect unusual constructions due to – among other things - omissions.
- *Addition*: Word that is not in the source segment is added to the translation. Existing features approximate this issue as in the case of "omission".
- *Untranslated*: A source word is left untranslated in the translation. This issue is currently covered by out-of-vocabulary features based on language model of the target language.

Fluency

- *Register/style*: Incorrect use of words due to variants/slang, company style or style guide. This issue is not directly covered by existing approaches, but it is approximated by the target language model features, as long as this model is trained on documents with the correct register/style.
- *Spelling*: Incorrect word spelling due to capitalisation or diacritics. This issue is also approximated by language model features, which are trained on truecased models. Spell checkers could also be used.
- *Typography*: Incorrect use of punctuation, unpaired quote marks or brackets. These issues are captured by a number of features, such as those checking for missing closing brackets or quotation symbols in the target segment, and those contrasting the percentage of different punctuation symbols in the source and target languages.
- *Grammar*: The several grammar-related issues (morphology, part of speech, agreement, word order, function words, tense/mood/aspect) are captured partly by target language model features, and partly by advanced syntactic features based on probabilistic context free grammars, dependency structures and categorical combinatory grammar (Felice and Specia, 2012; Almaghout and Specia, 2013).
- *Unintelligible*: Parts of the translation are not understandable enough to be analysed. This issue is only approximated by language model features of the target language.

7 Conclusions

This paper has presented an analysis and experiments on quality prediction of professionally produced translations. The data analysis has shown that although intuitively we know that human translations differ significantly from machine translations, distinguishing them using automated methods is not a trivial task. In particular, it seems to be a harder problem nowadays then it was ten years ago. This is most likely due to overall improvements in the quality of machine translation systems over the time. In addition, the human translations analysed, albeit professionally created, contain errors in up to almost 30% of the cases. We have shown that the types of errors in human translations tend to be different from those in machine translations, but that larger differences are observed across language pairs.

Finally, we have shown that human translation quality seems harder to estimate than machine translation quality. We believe this is mostly due to two reasons: skewed label distribution (most human translations are labelled as perfect), and the limitations of existing features, which do not capture more subtle or complex issues present in human translations. Our on-going work is aimed at addressing these two challenges: we are collecting a larger dataset including more lower quality human translations (produced by less experienced translators) and designing more linguistically motivated features.

Acknowledgements

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

Almaghout, H. and Specia, L. (2013). A CCG-based quality estimation metric for statistical machine translation. In *MT Summit XIV*, pages 223–230, Nice, France.

Avramidis, E. (2013). Sentence-level ranking with quality estimation. *Machine Translation*, 28:1–20.

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

Bach, N., Huang, F., and Al-Onaizan, Y. (2011). Goodness: a method for measuring machine translation confidence. In *ACL-2011*, pages 211–219, Portland, Oregon.

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

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

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 workshop on statistical machine translation. In *9th WMT*, pages 12–58, Baltimore, Maryland.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 workshop on statistical machine translations. In *7th WMT*, pages 10–51, Montréal, Canada.

Felice, M. and Specia, L. (2012). Linguistic features for quality estimation. In *7th WMT*, pages 96–103, Montréal, Canada.

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

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

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

Lommel, A. R., Popovic, M., and Burchardt, A. (2014). Assessing inter-annotator agreement for translation error annotation. In *LREC Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*, Reykjavik, Iceland.

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

Soricut, R. and Echihabi, A. (2010). Trustrank: Inducing trust in automatic translations via ranking. In *ACL-2011*, pages 612–621, Uppsala, Sweden.

Specia, L. (2011). Exploiting objective annotations for measuring translation post-editing effort. In *EAMT-2011*, pages 73–80, Leuven.

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

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

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