Translation Quality and Productivity: A Study on Rich Morphology Languages

Lucia Specia   l.specia@sheffield.ac.uk
Department of Computer Science, University of Sheffield, Sheffield, S1 4DP, UK

Kim Harris   kim_harris@textform.com
text & form GmbH, 10179 Berlin, Germany

Frédéric Blain   f.blain@sheffield.ac.uk
Department of Computer Science, University of Sheffield, Sheffield, S1 4DP, UK

Aljoscha Burchardt   aljoscha.burchardt@dfki.de
Viviven Macketanz   vivien.macketanz@dfki.de
Language Technology Lab, DFKI, Alt-Moabit 91c, 10559 Berlin, Germany

Inguna Skadiņa   inguna.skadina@tilde.lv
Tilde, Vienibas gatve 75a, Riga, LV 1004, Latvia

Matteo Negri   negri@fbk.eu
Marco Turchi   turchi@fbk.eu
Fondazione Bruno Kessler, Trento, 38100, Italy

Abstract

This paper introduces a unique large-scale machine translation dataset with various levels of human annotation combined with automatically recorded productivity features such as time and keystroke logging and manual scoring during the annotation process. The data was collected as part of the EU-funded QT21 project and comprises 20,000–45,000 sentences of industry-generated content with translation into English and three morphologically rich languages: English–German/Latvian/Czech and German–English, in either the information technology or life sciences domain. Altogether, the data consists of 176,476 tuples including a source sentence, the respective machine translation by a statistical system (additionally, by a neural system for two language pairs), a post-edited version of such translation by a native-speaking professional translator, an independently created reference translation, and information on post-editing: time, keystrokes, Likert scores, and annotator identifier. A subset of 2,000 sentences from this data per language pair and system type was also manually annotated with translation errors for deeper linguistic analysis. We describe the data collection process, provide a brief analysis of the resulting annotations and discuss the use of the data in quality estimation and automatic post-editing tasks.

1 Introduction

Data-driven approaches to machine translation (MT) rely largely on datasets of source sentences and their corresponding translations previously created by humans, so-called parallel corpora. MT systems, be they statistical or neural, are built in static fashion and (if at all) updated from time to time as more translations become available. With the popularisation of post-editing

(PE), a natural question is whether the corrected version of the MT output could be used in feedback loops to improve the current system via model retraining, model tuning or the addition of explicit model components. Additionally, by studying PE data, one can get insights on the errors made by the MT system to try and remedy them in different ways. PE data can also be used to build and benchmark metrics for the automatic evaluation of MT output, as well as quality estimation metrics and automatic PE systems.

To facilitate research in these and related areas, we have created a unique large-scale dataset with various levels of human annotation combined with automatically recorded productivity features. The data comprises 20,000–45,000 sentences of industry-generated content for English from or into three morphologically rich languages and was collected as part of the EU-funded QT21 project. The PE of all four language pairs was performed using a tool to record detailed process and product information at the sentence level during PE, including time, keystrokes, actual edits and Likert scores for the PE effort as given by the translator immediately after completion of the editing.

Most of the data was translated by a phrase-based statistical MT (PBMT) system. In addition, subsets of 15,000–20,000 sentences for EN–DE and EN–LV – respectively – were also translated using a neural MT (NMT) engine that was trained on exactly the same data used to train the original PBMT system. The PE of identical input data for both the PBMT and NMT systems facilitates large-scale direct comparisons between the actual output of these systems, as well as between process cues. For example, PE productivity can be calculated and compared using the time and keystroke information recorded during PE. The "preference" of translators can be compared through the scores given to the *perceived* quality of the output by such translators. A number of other comparative analyses and benchmarking in both research and industry scenarios become possible with this data.

Finally, a subset of 2,000 sentences was selected for each language pair and MT system type and manually annotated with word-level errors for deeper linguistic analysis. Both PE and error annotations were performed by professional translators.

While other datasets with PE data have been created in the past and also released for research purposes, these are limited in either their scale (e.g. see those used for the WMT13–14 shared tasks on quality estimation[1]), have been post-edited by non-professional translators (Wisniewski et al., 2013; Bojar et al., 2015), or make only the actual post-edits available, providing no additional information on the process and no explicit annotations. The most notable example of the latter is the Autodesk dataset (Zhechev, 2012). It contains sentences predominantly belonging to Autodesk software user manuals, covering 13 language pairs with English as the source language. The source sentence, its machine translation and its post-edit are provided. The translated sentences are produced by an MT system or are translation memory suggestions with a fuzzy match score larger than 75%.

In the remainder of this paper we first describe our data sources (Section 2) and the MT systems built (Section 3) to translate this data. We introduce the PE process and its results in Section 4, and the error annotation in Section 5. In Section 6 we present two uses of the dataset.

2 Data

The post-edited and annotated data described in this paper belongs to two specific domains: information technology (IT) and life sciences. These domains were chosen because of the high demand for this type of content in multiple languages due to its economic impact on businesses active on global markets where language is key. The use of this data in research can therefore play a significant role in building the necessary bridges between the constituencies most interested in achieving progress in the field of MT: research and industry.

---

[1] http://www.statmt.org/wmt14/

Language pair	# sentences	# source tokens	# target tokens	Domain	Data provider
EN–DE	80,874	1,322,775	1,312,975	IT	Adobe
EN–CS	81,352	1,332,654	1,175,463	IT	Adobe
EN–LV	231,028	3,713,803	3,168,740	Pharma	EMEA
DE–EN	193,637	3,120,482	3,228,761	Pharma	EMEA

Table 1: Domain-specific datasets: number of sentences and source and target tokens.

	Training data			
	EN–DE	EN–CS	EN–LV	DE–EN
# sentences	21,873	32,352	204,528	135,884
# source words	0.53	0.59	3.19	2.41

Table 2: Statistics on the in-domain training data. The number of words is reported in millions.

Four sets of parallel data in four language combinations (English– German/Latvian/Czech and German–English) were selected from the web. English Adobe software manuals translated into German and Czech were chosen for the IT domain, and a subset of the European Medicines Agency (EMEA) corpus was selected for the life sciences domain (which we also refer to as "pharma") to cover the English–Latvian and German–English language pairs.[2]

To create datasets that can satisfy different research needs and thus increase their usability, a set of criteria was applied to data selection and pre-processing. For English–German/Czech and German–English, sentences that did not end with a punctuation mark or contained less than three or more than 35 words were discarded, and duplicate sentences were removed. These strategies reduced the number of sentence pairs by approx. 45%. For English–Latvian, a part of parallel sentences were obtained by extracting textual sentences from PDF files in the EMEA repository. First, we used Adobe Acrobat v10 Professional to convert PDF files to HTML format, as this preserved most of the original document structure. Then we ran customised scripts to convert the HTML files to plain text and clean the data. The Microsoft Bilingual Sentence Aligner (Moore, 2002) was used for sentence alignment of the parallel plain text files. Duplicate sentence pairs and sentences with less than three or more than 35 words were removed. This sentence size filtering only marginally affected the size of the final corpus. The statistics of the final sets are reported in Table 1.

For each language pair, we selected a subset of data for annotation (see Table 5), and used the remaining sentence pairs as in-domain training data to build the MT systems (Section 3). This remaining data was split into training (see Table 2), development (2,000) and test (2,000) sets.

3 MT Engine Building

3.1 Training Data

A crucial aspect for creating a set of reliable post-edited sentences and error annotations is the availability of domain-adapted translations. This is necessary because a generic translation system is not able to correctly translate domain-specific terms or expressions, which would, in turn, cause translators to rewrite translations from scratch, rendering accurate error annotation

---

[2]The German–English dataset was created by taking the available English–German data and then inverting the language direction. This is not ideal; however, very little domain-specific data exists for under-resourced language pairs, including those whose source language is German.

	EN–DE		EN–CS		EN–LV		DE–EN	
	Parallel	Mono	Parallel	Mono	Parallel	Mono	Parallel	Mono
In-domain	7.2	-	-	-	0.181	-	2.09	2.35
Out-domain	12.7	-	50.34	51.46	-	-	-	-

Table 3: External resources collected to train the MT systems. The reported numbers represent millions of sentences.

impossible.

When building a domain-adapted MT system we rely on different external resources depending on the size of the in-domain data. For the language pairs for which there are less than 100,000 in-domain sentence pairs (*i.e.* EN–DE and EN–CS), a large collection of in- and out-of-domain monolingual and parallel corpora was gathered from the web, while for the remaining languages (EN–LV and DE–EN) only in-domain corpora were used. This process resulted in:

- EN–DE: Over 20 million generic and in-domain sentence pairs obtained by merging the datasets available in the OPUS (Tiedemann, 2012), TAUS, WMT and JRC [3] repositories (e.g. Europarl, CDEP, CommonCrawl, etc.);

- EN–CS: Over 51 million generic and in-domain sentence pairs available in the CzEng 1.6 dataset (Bojar et al., 2016b).[4] In addition, translating into a language with free word order suggests the use of a large collection (more than 50M sentences) of monolingual generic data obtained from the Translation task at WMT16;

- EN–LV: Over 385,000 parallel medical sentences from the EMEA corpus available in OPUS and the most recent documents from the EMEA website (years 2009-2014);

- DE–EN: Over 2 million in-domain sentence pairs collected from OPUS and the data released for the medical translation task at WMT14 (Bojar et al., 2014). These resources include MuchMore, PatTr, and the Wikipedia parallel titles. In addition to these parallel sentences, monolingual data (approx. 2 million) obtained from the medical translation task at WMT14.

A summary of the external resources used to train the MT system is shown in Table 3.

3.1.1 Data Selection

In MT literature, it has been shown that when large generic datasets and a small in-domain corpus exist, the use of data selection techniques can help improve translation quality (Eetemadi et al., 2015). To optimally leverage a domain-specific corpus, we used cross-entropy-based selection for monolingual data (Moore and Lewis, 2010), its extended version for bilingual texts proposed by Axelrod et al. (2011) and the latent-domain translation method (Cuong and Simaan, 2014).

Entropy-based method: Originally proposed by Gao and Zhang (2002), entropy-based approaches consist in computing the perplexity score of each sentence of a generic corpus against both an in-domain language model (LM) and an LM trained on the generic corpus. The sentences are then ranked according to the difference between their two perplexity scores. Once all of the generic sentences have been ranked, the size of the subset to extract is determined by minimising the perplexity of a development set against an LM trained on an increasing amount of the sorted corpus (e.g. 5%, 10%, ...). According to (Moore and Lewis, 2010), perplexity

---

[3] https://ec.europa.eu/jrc/en/language-technologies
[4] http://ufal.mff.cuni.cz/czeng/czeng16pre

decreases when less but more relevant data is used. We used the freely available open-source tool XenC (Rousseau, 2013).

Latent-domain translation method: This technique is able to give priors to different domains that comprise the generic data set. The goal is to estimate the probability of whether a sentence pair belongs to the in- or out-of-domain data, using in-domain corpus statistics as prior. The Expectation-Maximisation training algorithm is derived and used to estimate the out-of-domain models (given only in and mixed-domain data). This technique provides the selected data directly without the need to choose a cut-off point in the ranked list of sentence pairs.

Both methods were first tested on the EN–DE language pair, and the best performing method was applied to EN–CS. In our experiments, we used the data shown in Table 1 as in-domain and the concatenation of the data in Table 3 as out-of-domain data. Although an in-domain corpus exists for EN–DE in the additional resources, it represents a mix of datasets resulting from a different distribution compared to the training data in Table 1. For this reason, all corpora in the additional resources are considered out-of-domain data.

The perplexity computed on the target side of the development set using all available data is 207. When applying both data selection methods, it significantly decreased to 150, indicating that selecting data in this fashion can be advantageous. The entropy-based method achieved a perplexity of 150, and selecting only the top 15% of the ranked sentences resulted in 3.3 million sentence pairs. The latent domain method obtained a similar perplexity (157) but selected a larger number of sentences. For this reason, the entropy-based technique is also used for EN–CS. In this case, the perplexity is higher than for EN–DE (1900), but using the top 5% of the ranked data (2.5 million sentences) allowed us to significantly reduce it to 1300. These high perplexity values stem from the fact that the external resources for EN–CS do not contain any IT data.

3.2 MT Systems

Different systems were built for each language pair using the selected and the in-domain data for EN–DE and EN–CS and the in-domain data for the other language pairs.

- EN–DE: Two different MT systems were created: a PBMT and a NMT system. The PBMT system was trained on all of the selected parallel training data. The phrase table was adapted to the in-domain data using the approach proposed in (Niehues and Waibel, 2012). To deal with complex reordering in the German language, this system uses a pre-reordering technique (Herrmann et al., 2013) in combination with lexical reordering. In addition, it takes advantage of two word-based n-gram language models and three additional non-word language models, namely, two automatic word class-based (Och, 1999) language models using 100 and 1,000 word classes, and a POS-based language model using fine-grained POS tags (Schmid and Laws, 2008). For the NMT system, we trained the Nematus toolkit (Sennrich et al., 2017) which is an implementation of the attentional encoder-decoder architecture (Bahdanau et al., 2014). To handle large vocabulary, the training data was previously segmented using the byte-pair encoding compression algorithm (Sennrich et al., 2016), resulting in a vocabulary of 40,000 sub-word units for both languages. We used mini-batches of 100, word embeddings of 500 dimensions, and gated recurrent unit layers of 1,024 units. The maximum sentence length was set to 50. The models were trained using Adam and by reshuffling the training set at each epoch. The NMT system was trained on the selected data and then fine-tuned on the in-domain data.

- EN–CS: The PBMT system was trained using Moses (Koehn et al., 2007) combined with TectoMT (Žabokrtskỳ et al., 2008). This was done by adding the source development and test sentences and their translations obtained by TectoMT as additional (synthetic) parallel

EN–DE		EN–CS	EN–LV		DE–EN
PBMT	NMT	PBMT	PBMT	NMT	PBMT
35.9	45.8	38.7	46.5	38.4	53.4

Table 4: BLEU score of the PBMT and NMT systems on different language pairs.

data to the Moses system previously trained on the selected data. This new corpus and the in-domain data were used to train separated phrase tables. At test time, we ran Moses using all of the phrase tables and we corrected its output using Depfix (Rosa et al., 2012). In addition, we trained a 7-gram LM on surface forms from all monolingual resources. Similar to the EN–DE system, two additional LMs over morphological tags were built to help maintain morphological coherence in the translation output. The system is described in (Tamchyna et al., 2016).

- EN–LV: The PBMT system was trained on Tilde's MT platform (Vasiļjevs et al., 2012). The system is based on the Moses toolkit using the standard components. Nematus with sub-word units was used to train the NMT system with a vocabulary size of 40,000 sub-words. The models were trained with a projection (embedding) layer of 500 dimensions, recurrent units of 1024 dimensions, a batch size of 20 and dropout enabled. All other parameters were set to their default values.

- DE–EN: The PBMT system was trained using the same components and adaptation techniques as those used for the EN–DE model.

The results of the different systems for each of the language pairs are reported in Table 4 according to BLEU (Papineni et al., 2002). The parameters of the models were optimised on the development set and the final results computed on the test set. When comparing the PBMT and NMT performance, we noticed that when using a large collection of training data (*i.e.* EN–DE) the NMT system can significantly outperform the PBMT as shown in several evaluation campaigns. However, when the training data is limited (*i.e.* EN–LV) , the PBMT performs better than the NMT. The language pairs with the lowest out-of-vocabulary rate (EN–LV: 0.2 and DE–EN: 0.5) achieve the best BLEU score values. The DE–EN system obtains better performance compared to EN–LV because it can leverage more in-domain training data.

4   Post-Editing Process

Post-editing was performed using the PET tool (Aziz et al., 2012). This is a simple and freely available, open-source tool that tracks PE using a number of indicators. Figure 1 shows a screenshot of the tool with an English–German PE task. The tool tracks the process of PE, records PE time per sentence, and logs all keystrokes pressed by the annotator. This allows us to reproduce the PE activity, which can be useful for research on topics such as PE process, productivity gains, and automatic PE. The following information was recorded during PE:

- Editing time: time spent translating or editing a unit.
- Keystrokes: number of keys pressed during the PE according to type of keys (deletion, alpha-numeric, etc.).
- HTER: edit distance between the draft translation and its post-edited version.
- Evaluation: quality assessment based on a pre-defined set. We ask a question about the usefulness of the draft translation for PE (top left corner in Figure 1).
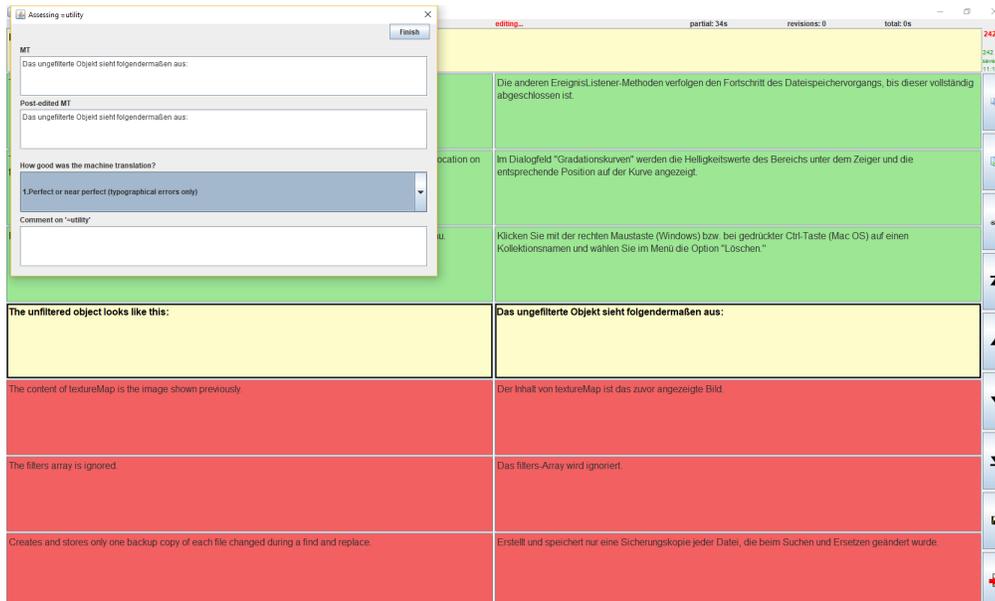
Figure 1: Example of project in the PET tool.

Time, one of the most important indicators collected by the tool, is computed from the moment the target box of the unit is clicked to the moment the task is completed (either the job is closed or the navigation button "next" is pressed). The tool allows for multiple revisions, where the annotator can go back to the same sentence and edit it again. For the statistics reported here, we take the aggregation of PE time and keystrokes, and compute the edit distance between the last version and draft MT output. The outcome of a job is also stored in an XML file.

A set of PE and annotation guidelines created by the QTLaunchpad project were adapted for the PE of our data. To ensure that the quality of the post-edits was consistent and reflected the requirements of the research to be performed on the resulting data, agreement was reached on the level of editing to be done on the data. Based on the previous experience of the language partners involved, the following general rules were defined:[5]

- Use as much of the raw MT output as possible.
- Aim for grammatically and syntactically correct translations.
- Ensure that no information has been accidentally added or omitted.
- Edit any offensive, inappropriate or culturally unacceptable content.
- Ensure proper and appropriate spelling.
- Do not restructure or change word order solely to improve the flow of the text unless dictated by grammar or domain standards.

Additionally, the following domain-specific rules for software localisation were used:

- Ensure that domain-specific terminology is correctly translated.
- Ensure that standard domain and language-specific style issues are followed.
- If formatting is used, ensure that it is correct.

---

[5]For details: qt21-wiki.dfki.de/index.php?title=Post-editing_guidelines

	# sentences		# words SRC	# words MT		# words PE	
Lang.	PBMT	NMT	-	PBMT	NMT	PBMT	NMT
DE–EN	45,000	–	14.66	15.54	–	15.58	–
EN–DE	30,000	15,000[6]	14.47	14.61	14.77	14.61	14.56
EN–LV	20,738	20,738	15.91	13.50	13.42	13.52	13.42
EN–CS	45,000	–	15.04	13.16	–	13.23	–

Table 5: General statistics of the post-edited data: Total number of sentences, average number of words in source, translation and post-edited sentences.

Avg. utility score		
Lang.	PBMT	NMT
DE–EN	1.62	–
EN–DE	1.98	1.40
EN–LV	1.64	1.84
EN–CS	2.17	–

Table 6: PE utility scores: the lower the score, the more useful the MT output.

	Avg. TER MT-PE		Avg. TER REF-PE	
Lang.	PBMT	NMT	PBMT	NMT
DE–EN	0.17	–	0.36	
EN–DE	0.25	0.08	0.40	0.37
EN–LV	0.15	0.23	0.29	0.34
EN–CS	0.32	–	0.34	–

Table 7: Average edit distance between PE and original MT (HTER), and between PE and independent reference. The higher the distance, the more edits performed.

The guidelines were made available to all language teams and pre-editing meetings were held to avoid communication issues. Consistency in the application of these rules was critical, which is why professional translators were employed and thorough consultations were performed prior to PE.

Professional translators performed PE on every language pair. Six translators were involved in the PE for EN–DE, 4 for DE–EN, 8 for EN–LV, and 5 for EN–CS. For the evaluation score, the following options were given to the translator after the post-editing of each sentence:

- **1.** Perfect or near perfect (typographical errors only).
- **2.** Very good, could be post-edited quickly.
- **3.** Poor, required significant post-editing.
- **4.** Very poor, required retranslation.

Tables 5–8 summarise the outcome of the PE process. Much more detailed information is available in the XML output files. Table 5 provides general statistics on numbers of sentences and words per language pair and MT system type. The average perceived PE effort scores are given in Table 6. Table 7 measures the edit distance between MT and PE, and between PE and the original reference (REF). Finally, 8 shows average PE time and keystrokes. As expected, PE time varies considerably for different sentences, even if outliers are removed. Therefore, Table 8 also shows standard deviations.

5   Error Annotation Process using MQM

Our error annotation process follows a 2-step workflow. After PE, the quality of each sentence is evaluated on a scale from 1–4 as explained in the previous section. A subset of sentences scored as 2 (very good) are then selected for the error annotation phase, during which all issues resolved during the PE phase are classified. The errors are annotated using the Multidimensional Quality Metrics (MQM) error annotation framework (Lommel et al., 2014), which is popular in industry and research, and actively supported by XTM, Trados Studio, and other commercial tools. We

	Avg. PE time		Avg. PE time w/o outliers		Avg. keystrokes	
Lang.	PBMT	NMT	PBMT	NMT	PBMT	NMT
DE–EN	42±80	–	36±45	–	24.71	–
EN–DE	51±78	46±602	46±39	32±36	15.55	13.89
EN–LV	27±77	43±406	23±28	36±39	18.91	26.08
EN–CS	44±43	–	42±35	–	45.78	–

Table 8: Post-editing time and keystrokes: average number of seconds per word, with and without outliers (plus standard deviation) and average number of keys pressed during post-editing of a sentence. Outliers are sentences that took more than four minutes to be edited.

used the open-source tool translate5[7] (see Figure 2), a database-driven tool with a GUI. Source texts, translations, post-edits, and error annotations are organised in a relational database. The tool, originally implemented as a proofreading and PE environment for the translation industry, has been recently extended to support MQM annotation.
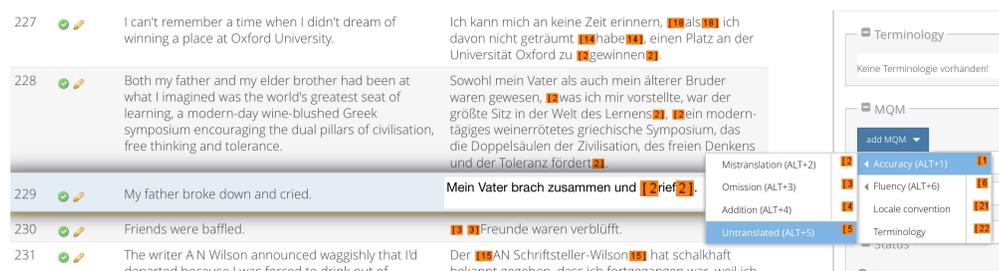

Figure 2: MQM error annotation in translate5 (excerpt of screenshot).

An error represents any issue that has been corrected during the PE step in the translated sentence. In the annotation step, a relevant error classification must be provided for all corrections made during PE according to a given list of errors. Error annotation is performed by experienced professional translators supported by detailed annotation guidelines.

The list of errors is divided into the main issue categories *accuracy*, *fluency* and *terminology*, which fold into a selection of more detailed categories from the MQM hierarchy. Figure 3 shows part of a decision tree that annotators used to select the most appropriate issue. The actual error categories used in the annotation are shown in Table 9.

Annotators are instructed to use the subcategories whenever possible and to resort to the more general category level only in case of doubt, for example, if the German term *Zoomfaktor* is incorrectly translated as *zoom shot factor*, and the annotator is unsure whether this represents a *mistranslation* or an *addition*. In this case, the error can be classified as an Accuracy error since it is unclear whether content has been added or a term mistranslated.

The annotation process has been completed for all languages and MT system types, resulting in 1,800 unique sentences per language pair and MT system type, with an additional 200 sentences doubly annotated for agreement analysis. The breakdown of error annotations for all 2,000 sentences per language pair and MT system type is shown in Table 9.

Table 10 shows an initial analysis on the agreement between pairs of annotators. Agreement was computed using Cohen's kappa (Cohen, 1960) at the word level in two ways: firstly, for each word we count an agreement whenever both annotators agree that it is incorrect (or correct), with agreement by chance = $1/2$; second, for each word we count an agreement whenever
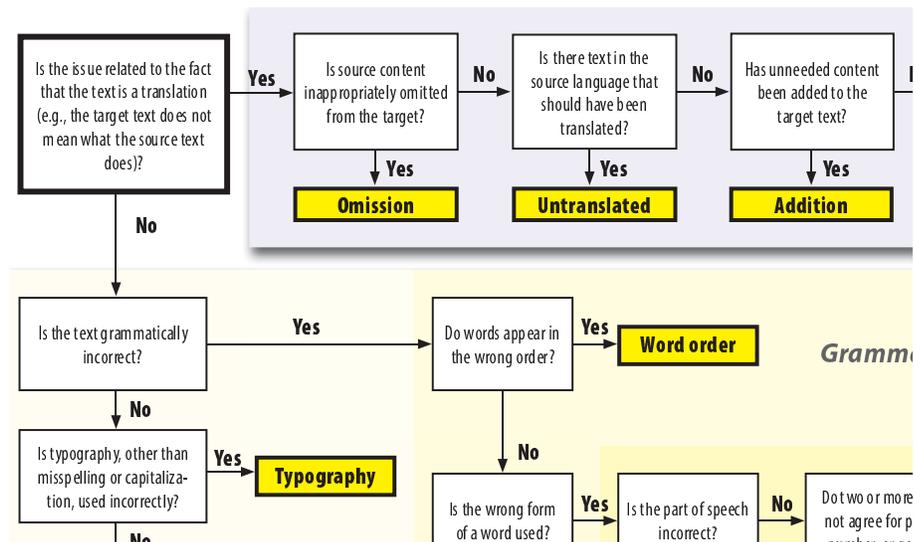
---
[7]http://translate5.net

Figure 3: Decision tree guiding error annotation (excerpt).

both annotators agree on the exact error type assigned to the word (or agree on the word being correct), considering all the 20 categories shown in Table 9 as equally likely (i.e. no distinction was made among different levels in the hierarchy), with agreement by chance = $1/21$.

The interpretation of the kappa coefficient is difficult, but it is generally believed that 0.4–0.6 is moderate, while 0.6–0.8 represents substantial agreement, with anything above 0.8 indicating perfect agreement (Landis and Koch, 1977). Considering the subjectivity of the task and the number of error categories and different levels in the hierarchy, we consider the moderate to high agreement found a very positive result towards validating the annotation of the data. In the near future, further quantitative and qualitative analysis will be performed to understand problematic categories and the reasons behind certain disagreements.

6 Examples of Uses of the Dataset

Subsets of the datasets collected have been used in the 2016 and 2017 editions of the WMT shared tasks on Quality Estimation and Automatic Post-editing (Bojar et al., 2016a, 2017).[8] In what follows we summarise some of the outcomes from these tasks.

6.1 Quality Estimation

Quality Estimation (QE) is the task of predicting the quality of the output of an MT system without the use of reference translations (Blatz et al., 2004; Specia et al., 2009). This is approached as a machine learning task, where training data with quality labels is needed. These labels can target different granularity levels: words, phrases, sentences or entire documents.

Early work in the area relied on proxies to quality labels generated using automatic evaluation metrics such as BLEU (Papineni et al., 2002) based on human translations. The task was thus framed as that of predicting an automatic evaluation metric score. This did not prove very successful because of the limitations of the automatic metrics themselves and the lack of a clear interpretation for the predictions (i.e. what does a BLEU score of 0.5 mean?).

Quality labels given by humans have been suggested in (Quirk, 2004) but only started to be

---

[8] http://www.statmt.org/wmt16/ and http://www.statmt.org/wmt17/.

Error type	DE–EN PBMT	EN-DE PBMT	EN-DE NMT	EN–LV PBMT	EN–LV NMT	EN–CS PBMT
Accuracy	3	0	0	39	50	0
Addition	539	332	167	277	268	385
Mistranslation	437	967	852	274	677	786
Omission	576	690	355	395	560	588
Untranslated	278	102	24	79	62	301
Fluency	3	0	0	233	210	234
Grammar	0	0	0	11	2	103
Function words	1	2	1	0	0	0
Extraneous	302	525	245	49	49	228
Incorrect	139	804	449	56	55	454
Missing	362	779	231	66	32	348
Word form	0	94	267	280	261	1401
Part of speech	20	128	132	38	35	147
Agreement	18	506	97	419	357	48
Tense/aspect/mood	63	184	51	60	46	397
Word order	218	868	309	336	152	1148
Spelling	118	126	132	324	387	638
Typography	282	553	249	823	387	1085
Unintelligible	0	33	0	10	14	30
Terminology	27	82	139	34	31	0
All categories	3386	6775	3700	3803	3635	8321

Table 9: MQM error categories and breakdown of annotations completed to data.

	DE–EN PBMT	EN-DE PBMT	EN-DE NMT	EN–LV PBMT	EN–LV NMT	EN–CS PBMT
# annotated words (A1/A2)	516/643	974/920	338/288	669/682	303/310	324/370
Kappa on annotated words	0.61	0.70	0.82	0.69	0.67	0.62
Kappa on error type	0.51	0.48	0.69	0.53	0.51	0.51

Table 10: Number of annotated words per language pair for each annotator (A1 and A2) and the Cohen's kappa measuring inter-annotator agreement for MQM error annotations.

used more recently (Specia et al., 2009). In particular, the use of objective labels derived from extrinsic uses of MT output, such as PE, have become popular (Specia, 2011). Labels of this type include normalised PE distance (HTER - Human Targeted Translation Error Rate (Snover et al., 2006)). These can be acquired as a by-product of PE in a translation workflow, are less subjective and less subject to biases such as the annotators' perception of MT.

The datasets described in this paper open many new avenues for research in QE. The main benefits with respect to previously collected labels include its scale, domain specificity and the availability of multiple types of (reliable) human annotation.

In the WMT16 QE shared task, a subset of the English-German IT domain post-edited data containing 15,000 sentences was used for the sentence, word and phrase-level tasks. The quality labels were automatically derived from the PE of the MT output, e.g. for sentence level, HTER scores were used. Bojar et al. (2016a) claim that, when compared to previous year – approx. 14,000 crowdsourced post-edited sentences – the results of the 2016 task were more conclusive. They attribute this to the higher quality of the new dataset and observe that:

Task	Baseline ↑	Best system ↑
2016 Training set		
Word-level QE	0.32	0.55
Phrase level QE	0.40	0.50
Sentence-level QE	0.35	0.53
2017 Training set		
Word-level QE	0.36	0.58
Phrase level QE	0.33	0.60
Sentence-level QE	0.39	0.71

Table 11: QE shared task results on the 2016 test set: baseline and winning systems in 2016 and 2017 (larger training set) for sentence (Pearson), word and phrase (F_1-mult = multiplication of F_1 for the GOOD and BAD classes) levels.

- for sentence level, the best Pearson correlation between the system prediction and true HTER in 2015 was 0.39 (against 0.14 of the baseline system). In 2016, the winning submission reached 0.52 Pearson correlation (against 0.35 of the same baseline system). One can speculate that the task was made somewhat "easier" by using high quality data, but the delta in the Pearson correlation between the baseline and winning submission is still substantial.
- for word level, 2016 systems performed much better: 0.56 against 0.43 F_1-BAD. The baseline systems are not comparable.

In order to further push progress in the QE field, the 2017 QE task was provided with an extended version of the 2016 dataset in addition to data from a different domain and a different language pair. For English-German, the 2016 dataset was extended to include a total of 28,000 sentence pairs. For German-English, 28,000 sentence pairs in the life sciences domain were made available for the task.

The two datasets are significantly larger than any dataset used before in QE shared tasks. The same data was used for the three subtasks: sentence, word and phrase levels. The results of this year's task (Bojar et al., 2017) show major improvements for all tasks over the 2016 results. In addition to general advances in the field, these can in part be attributed to the larger dataset provided. For the 2016 test set, also used in 2017 for comparison, Table 11 shows the results using the official metrics for the best system and the baseline system using the 2016 vs the 2017 training sets.

This data has proven useful for subsequent work in the field: for instance, (Forcada et al., 2017) focuses on the prediction of PE time at sentence level on the 2016 dataset, while (Martins et al., 2017) proposes an novel word-level QE approach using automatic PE techniques.

6.2 Automatic Post-Editing

Automatic Post-editing (APE) systems are usually trained on (*source, MT, human_post-edit*) triplets from which the appropriate corrections of systematic errors should be learned and possibly generalised. This supervised learning problem is addressed as a "monolingual translation" task in which rough MT output in a given target language has to be translated into a fluent and adequate translation of the original source text. BLEU and TER computed against reference human post-edits are the standard evaluation metrics for the task, and their respective improvements and reductions are usually compared against the baseline scores obtained by the original MT output that has been left untouched (*i.e.* rough, non post-edited translations).

Early APE systems (Allen and Hogan, 2000; Simard et al., 2007) were developed under the PBMT paradigm, that is, by learning from "parallel" data, either (*MT, human_post-edit*)

pairs or triplets including information from the source text (Béchara et al., 2011; Chatterjee et al., 2015). Recent solutions achieved larger and more significant improvements by exploiting neural methods (Junczys-Dowmunt and Grundkiewicz, 2016; Pal et al., 2016, 2017), which approach the task as a sequence to sequence learning problem.

Both paradigms suffer from drawbacks that have, to date, represented the main obstacles towards a wider adoption of APE technology. According to Bojar et al. (2015), one of the major problems lies in data sparsity, which limits the ability to exploit training data in order to learn correction patterns that can also be applied to test instances. Several factors contribute to raising this data sparsity issue, namely: *i)* the size of the data (although human post-edits are a by-product of industrial translation workflows, few corpora are available for research, *ii)* the domain of the data (general domains – like *news* – are definitely less repetitive than narrow ones – like *information technology*), and *iii)* the origin of the post-edits (professional post-editors are definitely more reliable and coherent than non-expert ones).

The datasets described in this paper aim to mitigate the problems related to data sparsity for reasons that are similar to those discussed in the previous section on QE. Indeed, their size, domain specificity and professional PE quality may explain the renewed interest and the impressive progress of APE research in the past few years. The following figures drawn from the WMT experience support our claims:

- Number of tasks and submitted runs. At WMT 2016, only one English-German translation task in the IT domain was organised, while 2017 saw two tasks: English-German (IT) and German-English (life sciences). The new corpora (more repetitive than *news* data edited by non-experts in 2015) motivated more teams to participate: from 7 submissions in 2016 to 20 in 2017.

- Improvements over the baseline. The switch to new data coincided with significant performance gains that prove the viability of APE in domain-specific settings. While in 2015 none of the participants was able to beat the baseline, the best English-German submissions in 2016 and 2017 improved over the baseline by up to 5.5 and 7.6 BLEU points.

- Improvements over the PBMT approach. While in 2015 all systems followed this paradigm, falling in the same range of performance, the combination of advancements in neural research and the provision of more suitable data resulted in impressive performance gains in the next two evaluation rounds. The same PBMT system used for comparison in all the evaluation rounds was significantly outperformed by most of the participants in 2016 (up to 3.2 BLEU points) and in 2017 (up to 7.1 BLEU points).

7   Conclusions

In this paper we introduced a large and unique set of data points derived from *industry data* that have been post-edited and annotated by *professional translators*. This allows for specific features and **novel combinations of features** to be used for a variety of research and user-oriented purposes, including establishing the actual PE effort by translators based on time and keystrokes and comparing these results to the *perceived* level of quality of the post-edited sentence, establishing correlations between certain characteristics such as sentence length and post-editing time, or post-editing time and human or automatic quality evaluation metrics. The datasets also measure post-editing productivity and can be used to detect error patterns in the MT output. In addition, the creation of MQM-annotated subsets of these post-edits for typical industry domains provide information about error patterns and support feature-oriented quality estimation and evaluation, among many other novel avenues for research. This dataset is freely available and can be downloaded from the project website: `http://www.qt21.eu/`.

Acknowledgements

References

Allen, J. and Hogan, C. (2000). Toward the Development of a Post Editing Module for Raw Machine Translation Output: A Controlled Language Perspective. In *Third International Controlled Language Applications Workshop (CLAW-00)*, pages 62–71.

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the conference on empirical methods in natural language processing*, pages 355–362. Association for Computational Linguistics.

Aziz, W., de Sousa, S. C. M., and Specia, L. (2012). Pet: a tool for post-editing and assessing machine translation. In *Eighth International Conference on Language Resources and Evaluation*, LREC, pages 3982–3987, Instanbul, Turkey.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Béchara, H., Ma, Y., and van Genabith, J. (2011). Statistical Post-Editing for a Statistical MT System. In *Proceedings of the 13th Machine Translation Summit*, pages 308–315, Xiamen, China.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence estimation for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, number 315 in COLING '04, Geneva, Switzerland.

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 *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016a). Findings of the 2016 conference on machine translation. In *First Conference on Machine Translation, Volume 2: Shared Task Papers*, WMT, pages 131–198, Berlin, Germany.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017). Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Copenhagen, Denmark.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal.

Bojar, O., Dušek, O., Kocmi, T., Libovický, J., Novák, M., Popel, M., Sudarikov, R., and Variš, D. (2016b). CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In *Text, Speech and Dialogue: 19th International Conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Proceedings*. Springer Verlag. In press.

Chatterjee, R., Weller, M., Negri, M., and Turchi, M. (2015). Exploring the Planet of the APEs: a Comparative Study of State-of-the-art Methods for MT Automatic Post-Editing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 156–161, Beijing, China.

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

Cuong, H. and Simaan, K. (2014). Latent domain translation models in mix-of-domains haystack. In *Proceedings of COLING*, pages 1928–1939.

Eetemadi, S., Lewis, W., Toutanova, K., and Radha, H. (2015). Survey of data-selection methods in statistical machine translation. *Machine Translation*, 29(3-4):189–223.

Forcada, M. L., Sánchez-Martínez, F., Esplà-Gomis, M., and Specia, L. (2017). Towards optimizing mt for post-editing effort: Can BLEU still be useful? *Prague Bull. Math. Linguistics*, 108:183–195.

Gao, J. and Zhang, M. (2002). Improving language model size reduction using better pruning criteria. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 176–182. Association for Computational Linguistics.

Herrmann, T., Niehues, J., and Waibel, A. (2013). Combining Word Reordering Methods on different Linguistic Abstraction Levels for Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, Altanta, Georgia, USA.

Junczys-Dowmunt, M. and Grundkiewicz, R. (2016). Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the First Conference on Machine Translation*, pages 751–758, Berlin, Germany.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Lommel, A. R., Burchardt, A., and Uszkoreit, H. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0(12):455–463.

Martins, A. F. T., Junczys-Dowmunt, M., Kepler, F., Astudillo, R., and Hokamp, C. (2017). Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics (to appear)*.

Moore, B. (2002). Fast and accurate sentence alignment of bilingual corpora. Springer-Verlag.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Proceedings of the ACL 2010 conference short papers*, pages 220–224. Association for Computational Linguistics.

Niehues, J. and Waibel, A. (2012). Detailed Analysis of Different Strategies for Phrase Table Adaptation in SMT. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas*, San Diego, CA, USA.

Och, F. J. (1999). An Efficient Method for Determining Bilingual Word Classes. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway.

Pal, S., Naskar, S. K., Vela, M., Liu, Q., and van Genabith, J. (2017). Neural automatic post-editing using prior alignment and reranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 349–355, Valencia, Spain.

Pal, S., Naskar, S. K., Vela, M., and van Genabith, J. (2016). A neural network based approach to automatic post-editing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 281–286, Berlin, Germany.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Quirk, C. (2004). Training a sentence-level machine translation confidence measure. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Rosa, R., Mareček, D., and Dušek, O. (2012). Depfix: A system for automatic correction of czech mt outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 362–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rousseau, A. (2013). XenC: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, 100:73–82.

Schmid, H. and Laws, F. (2008). Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. In *International Conference on Computational Linguistics (COLING 2008)*, Manchester, Great Britain.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Miceli Barone, A. V., Mokry, J., and Nadejde, M. (2017). Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics.

Simard, M., Goutte, C., and Isabelle, P. (2007). Statistical Phrase-Based Post-Editing. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT)*, pages 508–515, Rochester, New York.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA.

Specia, L. (2011). Exploiting objective annotations for measuring translation post-editing effort. In *15th Conference of the European Association for Machine Translation*, EAMT, pages 73–80, Leuven, Belgium.

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

Tamchyna, A., Sudarikov, R., Bojar, O., and Fraser, A. (2016). Cuni-lmu submissions in wmt2016: Chimera constrained and beaten. In *Proceedings of the First Conference on Machine Translation*, pages 385–390, Berlin, Germany. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *LREC*, volume 2012, pages 2214–2218.

Vasiļjevs, A., Skadiņš, R., and Tiedemann, J. (2012). Letsmt!: a cloud-based platform for do-it-yourself machine translation. In *Proceedings of the ACL 2012 System Demonstrations*, pages 43–48. Association for Computational Linguistics.

Wisniewski, G., Singh, A. K., Segal, N., and Yvon, F. (2013). Design and analysis of a large corpus of post-edited translations: Quality estimation, failure analysis and the variability of post-edition. In *Proceedings of the MT Summit*, pages 117–124.

Žabokrtskỳ, Z., Ptáček, J., and Pajas, P. (2008). Tectomt: Highly modular mt system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170. Association for Computational Linguistics.

Zhechev, V. (2012). Machine translation infrastructure and post-editing performance at autodesk. In *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, pages 87–96.