Quality Estimation in MT Evaluation: A Survey

Paramveer Choudhary, Pushpak Bhattacharyya

Department of Computer Science and Engineering
CFILT, Indian Institute of Technology Bombay
Mumbai, India
{paramc, pb}@cse.iitb.ac.in

Abstract

Machine Translation evaluation by human annotators is a slow and rigorous task. The common automatic MT evaluation techniques depend on one or more reference translations to evaluate the output of MT systems. Quality estimation removes this dependence on human-generated reference translations by estimating the quality translation pair given just the source and translated text. In this paper, we present an extensive analysis of the quality estimation task. We establish the background necessary to understand quality estimation as a task and where it fits into the machine translation pipeline. We explore the different approaches to quality estimation, architectures used in QE systems and some of the challenges to quality estimation. This paper explores the various complexities in quality estimation and provides a valuable starting point for any researcher looking to advance the field of quality estimation.

1 Introduction

Technology in the last few decades has helped overcome geographical barriers creating a global community. One barrier that still remains is the different languages of communication. The innovations in the field of machine translation tackle this exact problem. Theoretical proposals from as far back as 1949 by Warren Weaver could finally be practically implemented once the computation and machine-readable data limitations were lifted with time. Advancements in machine translation have been substantial in the past decade. Automatic machine translation evaluation goes hand in hand with machine translation. An accurate prediction of the correctness of machine translation can help build user confidence in machine translation systems.

Most of the techniques of automatic MT evaluation require one or more reference translations which are human-generated. To overcome this dependence on manual translators there is a reference-less automatic MT evaluation task referred to as Quality Estimation (QE).

1.1 Motivation

The research area covered in this report is in the field of quality estimation. While reading different research papers on MT systems it was evident that for a deeper insight into the performance of systems, researchers had to manually judge the translation outputs from MT systems. The BLEU score helped establish a general performance metric for the MT systems. However, the insights into where the system is lacking and what linguistic features are leading to errors needed a better evaluation metric. Most researchers did this manually by finding out examples from the test dataset. This posed serious demand for the researchers to be well-versed in multiple languages or depend on other professional translators. This is where quality estimation could prove to be very useful. Quality estimation can help to identify patterns and linguistic features that are negatively impacting the performance of MT systems.

2 Background: Terminology and Definitions

This section explores some of the background knowledge necessary to understand the report. It covers machine translation, machine translation evaluation, quality estimation, its different granularities and the metrics that are used.

2.1 The Task of Machine Translation

Machine translation is the task of an automatic translation from one language to another. Depending on human translators is simply too ex-

pensive and time-consuming to be a feasible solution in the modern day. Machine Translation, often abbreviated as MT, brings the benefits of computational speed and cost-effectiveness to the task of translation. Once trained, an MT system can translate entire documents in a matter of seconds. Combined with the progress made in speech recognition and generation systems, end-to-end speech-to-speech MT systems are now a reality.

Example 2.1 *En: John plays the flute well.*
Hi: जॉन अच्छी बांसुरी बजाता है।

2.2 Four Paradigms of Machine Translation

Machine translation progress has gone through four different paradigms. Each paradigm has its own unique technique that it uses to tackle the task of machine translation. We explore each of the four paradigms individually now.

2.2.1 Rule-based MT

Rule-based machine translation takes the traditional deterministic and algorithmic approach to machine translation. This being the early years in MT evolution had rules that dictated the analysis, transfer, and generation steps of machine translation. An extensive set of rules govern the entire process of machine translation. The MT system in this paradigm is only as good as the rules that it contains and hence these rules are created by linguistic experts requiring in-depth knowledge about both the source and target language. The rule set is created with the goal to resolve the ambiguities across the Analysis-Transfer-Generation steps of machine translation. The deterministic approach to a problem as vast and contextual as MT has its disadvantages. The shortcoming of rule-based machine translation is the fact that language translation has an enormous amount of ambiguities that necessitate contextual information for resolution and covering all of these with a finite rule set is simply an infeasible endeavour. Also, the rules are the source and target language pair specific which makes the MT system language pair specific any scope of multilingualism is not possible.

2.2.2 Statistical MT

Statistical machine translation is an approach to overcome the hurdle of developing a robust rule set by probabilistically tackling the ambiguities that make machine translation difficult. Parallel corpus, a collection of translated sentence pairs, is used to learn the patterns of translation between the source and target language. These patterns are phrase-level mapping from the source language to the target language. A phrase table is built to capture all these patterns for reference by the MT system. Learning here simply means to score the relative likelihoods of each pattern translation which is captured by the probability scores of each phrase mapping. After building the phrase table with probability scores, the task of decoding comes next. Decoding involves finding the set of possible mappings and picking out the one with the highest probability score. The capability of a statistical machine translation system relies on the quality of the parallel corpus and the effectiveness of the machine learning techniques.

2.2.3 Example-based MT

Example-based machine translation is a combination of a data-driven and rule-based approach to machine translation. In this paradigm, a database of stored examples of translation is maintained. If the source sentence has an exact match in the database of stored examples of translation then the translated sentence is directly picked up as the resultant translation. If an exact match is not found then the source sentence is broken up into phrases, these phrases are translated and then combined to generate the target translated sentence. Example-based machine translation thus draws its inspiration from both statistical and rule-based machine translation.

2.2.4 Neural MT

The modern-day state-of-the-art machine translation systems are end-to-end where we feed a source sentence as input to a neural network and the translated sentence is returned as the output. The rise to prominence of such neural machine translation systems goes hand-in-hand with the increase in the abundant parallel corpus and the ever-improving neural network architectures. The emergence of architectures such as recurrent neural networks which can process sequential data allowed the neural networks to learn sequence-to-sequence learning

tasks such as machine translation. From the outside neural networks may seem like a black box approach that no longer requires knowledge and insight into linguistic features of languages. However, the fact remains that the success of attention-based models that relate closely with fundamental concepts like an alignment for translation still shows that techniques inspired by linguistic knowledge improve the neural network architectures for better performance and explainability. Moreover, knowledge infusion into neural networks has shown promise in improving their performance.

2.3 Machine Translation Evaluation

Machine translation systems generate a translated sentence in the target language given an input sentence in the source language. The evaluation of machine translation requires judging the quality of the translated sentence. There are two factors that are a must for any machine translation output to be considered a high-quality translation at the sentence level.

- **Adequacy:** Adequacy is concerned with how much of the meaning conveyed by the source sentence is retained in the translated sentence.
- **Fluency:** Fluency is the notion of acceptability of translation by a native speaker. It is concerned with the register, word choice, and word choice.

Satisfying one of these factors does not guarantee the other being satisfied. MT systems often output fluent yet inadequate translations and adequate yet not fluent translations.

2.4 Automatic MT Evaluation

These two factors establish the goals for any automatic MT evaluation system. Having human translators evaluate output after every new tweak during the development of a machine translation system is clearly infeasible. Automatic MT evaluation systems are categorised under two different settings. We explore these next.

2.4.1 Reference-based Evaluation

For this setting of automatic evaluation, we explore the possibility of a matching or comparing-based system. The system evaluates machine translation output using reference translation to give scores that correlate closely with human evaluation. The same sentence can be formed in multiple fluent ways while retaining adequacy. Thus we often have multiple reference translations for a single source sentence. Matching at different n-gram levels and combining these matching scores is one method that is followed by metrics such as BLEU (Bilingual Evaluation Understudy) score. Having multiple reference translation help the task of automatic evaluation by facilitating comparison with multiple correct reference translations in such simple yet effective and robust approaches.

If we were given an ideal correct translation could we utilise it to compare our MT output with it and judge the similarity between the two sentences? This question forms the basis of different metrics that rely on reference translations for MT evaluation.

Now the next question that follows is how would we estimate this similarity between two sentences in the same language? This is where different metrics differ. Here we can again look at two categories to which different metrics belong. The two categories along with one example metric for each are discussed ahead.

2.4.1.1 N-gram Matching-based Scores

The approach here is straightforward, match different lengths ($N \rightarrow (1, 2, .., |S|)$ N-grams and use this to arrive at an evaluation score. The effectiveness of the method lies in the fact that varying length (N) N-grams are used and this combined with multiple reference translations does end up finding a measure of similarity between sentences. One popular metric that follows this scheme with some modifications is that of BLEU score (Papineni et al., 2002) discussed next.

BLEU Score

BLEU is an abbreviation for BiLingual Evaluation Understudy. It is a metric based on the N-gram matching method that has provided excellent results in terms of correlation to human evaluation. At the heart of its effectiveness are two scores and each is discussed below.

- **Modified N-gram Precision:** N-gram precision that relies on a simple ratio of

matched N-grams to total N-grams fails for some ill-formed sentences that just have a few common occurring words repeated. This is because MT systems tend to over-generate certain reasonable words. For instance, consider the following example:

Example 2.2 *Reference Sentence: The sun is about to set.*
MT output: The the the the.

For this example, a simple N-gram matching-based precision would yield a score of $4/4 = 1$ score which is clearly incorrect. BLEU score overcomes this shortcoming by using a clipped count where a word once matched in reference translation exhausts in reference sentence so no more repeated words will match with it. As such the modified precision used in the BLEU score is given in equation 1:

So the count of matches is limited to max occurrence in reference and not solely based on the count in the candidate sentence.

- **Brevity Penalty:** The reasoning for this factor is that certain short sentences such as 'of the' will usually get high scores even with modified n-gram precision although they fail on both adequacy and fluency measures. For this purpose, a brevity penalty is introduced in the BLEU score. In the equation below c is the candidate sentence length and r is the reference sentence length.

$$BP = \begin{cases} 1 & \text{if } c > r, \\ e^{1-r/c} & \text{if } c \leq r \end{cases} \tag{3}$$

Combining these we get the final BLEU score metric as below:

$$BLEU_{score} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{4}$$

2.4.2 Edit Distance-based Methods

This method is based on how far the candidate translation is from the reference translation. This measure of distance is captured by using how many deletions, insertions, and substitutions are needed in the candidate sentence to make it the same as the reference sentence.

Word Error Rate

Word Error Rate (WER) is a simple normalised measure of edits needed to get the reference translation from the candidate translation. The WER is calculated as the number of edits divided by the number of reference words (N). The edits include insertions (I), deletions (D), and substitutions (S).

$$WER = \frac{C + D + I}{N} \tag{5}$$

2.4.3 Referenceless Evaluation

This task, commonly referred to as Quality Estimation (QE), evaluates the machine translation system without any reference translations (Specia et al., 2018). The evaluation is just given source text and translated text and the output is a translation quality score.

3 Quality Estimation

Quality Estimation (QE) is an automatic machine translation evaluation technique. What separates it from some of the other common metrics such as the BLEU (Papineni et al., 2002) score is that it does not require any reference translations (Specia et al., 2018). The only input required is source text and translated text.

3.1 Different Granularities for QE

QE metrics vary depending on the granularity at which it is applied. Most metrics work at the sentence level whereas QE has the flexibility to work at the word, sentence, and document levels (Ive et al., 2018). Depending on the granularity different metrics are used. We explore each of these in some detail now.

3.1.1 Word-Level QE

Quality Estimation can be used at the word level to predict the correctness of each word in the translation. The input is a sentence in the source language and a sentence in the target language. The output is a sequence of 'OK' and 'BAD' tokens. The output also covers gaps

$$P_n = \frac{\sum_{C \in Candidates} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C \in Candidates} \sum_{n-gram' \in C} Count_{clip}(n-gram')} \quad (1)$$

$$Count_{clip}(n-gram) = min(count, max_ref_count) \quad (2)$$

between words. Let us have a look at what the output tag represents under different cases:

- Source sentence words: The 'BAD' tag represents words that lead to incorrect translations in the target sentence and the 'OK' tag represents words that lead to correct translations in the target sentence.
- Target sentence words: The 'BAD' tag represents words that are incorrectly translated and the 'OK' tag represents words correctly translated.
- Target sentence gaps: The 'BAD' tag represents gaps that have missing translated words and the 'OK' tag represents gaps that do not have any missing translated words.

For the N words in a translated sentence, we have $2N + 1$ tags for the gaps and words. For M words in the source sentence, we have M tags for the words. The data for training of supervised models in word-level QE includes a dataset with these tags. Hence the task in word-level QE can be modelled as a sequence-to-sequence learning task.

3.1.2 Sentence-Level QE

At the sentence-level QE generates a score for a pair of source and translated sentences. The score can be a DA (Direct Assessment) score or HTER (Human-targeted Translation Error Rate) score. Let us look at each of these metrics:

- DA (Direct Assessment): A subjective score by a human annotator on a scale of 1-100. Usually, multiple human annotators' scores are normalised to obtain a z-score. This score is used in the training data to train the QE systems.
- HTER (Human Translation Error Rate): An edit distance measure that includes the number of edits required to convert the translated sentence into a correct one. The actual formula for the HTER score is given in equation 6/

3.1.3 Document-Level QE

A QE score referred to as a Multidimensional Quality Metric (MQM) score can be obtained at the document level as well. The score here is based on errors in the document and the severity of these errors. The document is annotated with errors as per the following characteristics:

- Word Span: Number of words in a given error. The words need not be contiguous.
- Severity: A error is categorised into three categories according to its severity. The categories are as follows:
  - Minor: Errors that do not distort the meaning during translation.
  - Major: Errors that change the meaning during translations.
  - Critical: Errors that change the meaning and contain a type of implication or lead to offensive translations.
- Type: A indication of error type such as missing words, wrong word order, agreement, etc.

Based on these errors we use the equation 7 to get the QE score at the document level:

4 Different Approaches to QE Systems

Quality Estimation systems have evolved similarly to MT systems over the years. Most of the early state-of-the-art systems were statistical machine learning based systems. Eventually with the advancements in neural networks and the introduction of sequence-to-sequence learning models (Papineni et al., 2002) such as Recurrent Neural Networks the research community shifted to neural QE systems.

4.1 Statistical Quality Estimation

In the early years of QE research neural networks were not yet explored in depth for many

$$\text{HTER} = \frac{\text{Substitutions} + \text{Insertions} + \text{Deletions} + \text{Shifts}}{\text{Reference Words}} \tag{6}$$

$$\text{MQM Score} = 100 - \frac{IssuesMinor + 5 \times IssuesMajor + 10 \times IssuesCritical}{Sentencelength} * 100 \tag{7}$$

natural language processing tasks. Most of the QE systems from this time were based on statistical approaches that relied heavily on linguistically inspired feature engineering to process the input text. Two of the popular statistical QE systems are QuEst by (Specia et al., 2013) and QuEst++ by (Specia et al., 2015). The extracted features were then fed into statistically supervised machine learning algorithms such as randomised decision trees and support vector machines.

4.2 Neural Quality Estimation

With the introduction of sequence-to-sequence learning models (Sutskever et al., 2014), the adoption of neural approaches to various natural language processing tasks started gathering attention. Tasks such as word-level quality estimation are inherently sequence-to-sequence learning tasks and these needed a different neural network architecture. Once neural network based started achieving impressive performance across different tasks in natural language processing, it was only a matter of time before such architectures would come to dominate QE systems as well. Today the state-of-the-art systems are all neural network-based architectures. We explore these in detail in the next section.

4.3 State-of-the-Art QE Systems

Neural network-based systems dominate the state-of-the-art architectures in most natural language processing tasks. Quality estimation is no exception to this trend and most of the state-of-the-art QE systems are based on neural networks. The recent advancements in the large pre-trained language representation models and their adoption in various NLP tasks eventually resulted in their adoption for QE systems as well. We explore their role in QE systems next.

4.3.1 Role of Pre-Trained Language Representation Models

Ever since the introduction of large pre-trained transformer-based language representation models such as BERT (Devlin et al., 2019), many of the NLP tasks have benefitted from them. Neural networks need numerical vectors as input. Converting the words into numerical vectors that capture their semantics is important for the performance of a neural network-based system. These vectors are referred to as word embeddings. These pre-trained language representation models are trained on huge monolingual data to generate word embeddings. Adding a separate task-specific output layer helps fine-tune these language representation models.

4.3.2 Common Architectures

Now that we the know importance of large pre-trained language representation models in the QE systems' architecture, we can lay out the general architecture followed by some of the state-of-the-art QE systems. The input is usually fed by a pre-processing layer followed by a transformer such as BERT. An output layer is then added which helps fine-tune the transformer for QE output.

4.3.2.1 TransQuest Quality Estimation System

One such state-of-the-art system is TransQuest by (Ranasinghe et al., 2020a), which was the winner of all 8 tasks of direct assessment sentence level QE shared task organised at WMT (Workshop on Machine Translation) 2020. The framework included two architectures Mono-TransQuest and Siamese-TransQuest as can be observed from Figure 1 [1]. The basic idea

[1]Figure from: Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation Quality Estimation with Cross-lingual Transformers. In Proceedings of the 28th International Conference on Computational Linguistics, pages 5070–5081, Barcelona, Spain (Online). International Committee

is to feed the two sentences into an XLM-R transformer and use the embedding from the transformer output to predict the QE score. The embedding is generated for the [CLS] token, [SEP] token and for each input word by the XLM-R transformer. There are three pooling strategies based on which the final embedding is taken from these output embeddings, these are listed below.

1. CLS Strategy: The output of embedding of [CLS] token is taken as actual output.
2. MEAN Strategy: The mean of all the embeddings from input words is considered as actual output.
3. MAX Strategy: A max-over-time of the output vectors of input words is taken.

The output from the pooling layer is then fed into a softmax layer which then predicts a QE metric, either HTER or DA. In Mono-TransQuest (MTransQuest) both sentences are fed into a single transformer separated by [SEP] toke. While in the Siamese TransQuest (STransQuest) each sentence is fed into a separate transformer and the cosine similarity between the outputs from the pooling layers of both transformers is used to predict the QE score.

4.4 Challenges to Quality Estimation

The current state-of-the-art systems perform well on the sentence level but struggle at the word level. Even at the sentence level, the performance suffers for low-resource languages. We delve deeper into some of the difficulties in quality estimation.

4.4.1 Difficulties in Predicting Adequacy

As we have seen how many QE systems are based on pre-trained language representation models, the QE systems tend to be more concerned with fluency as opposed to adequacy. Many pre-trained transformer-based QE systems focus on source sentence complexity and target sentence fluency as major factors contributing to QE scores. This was explored in depth by (Sun et al., 2020). The adequacy of the translated sentence is not adequately captured in the QE score predictions. This issue gets worse for low-resource languages.

4.4.2 Dataset Limitations

The dataset for quality estimation is released by the Workshop on Machine Translation (WMT). The WMT shared task on sentence-level direct assessment task contains training data for only eight language pairs. This hinders the development of multilingual QE systems. The performance suffers for low-resource languages. The amount of sentences in training data is also limited to 7000 sentences for seven out of eight language pairs. For the development of robust multilingual QE systems, more datasets in multiple language pairs are required.

4.4.3 Transfer Learning for Low-Resource Languages

The number of language pairs for sentence-level direct assessment is limited to eight language pairs only. This requires more transfer learning approaches for language pairs that do not have training data available. Language divergence makes the zero-shot learning approaches difficult. Usage of cross-lingual transformer-based language-representation models such as XLM-RoBERTa (XML-R) by (**?**) in QE systems such as TransQuest by (Ranasinghe et al., 2020b) have shown impressive transfer learning performance. However, better cross-lingual embeddings and more training data are still required for better transfer learning performance.

4.5 Multitask Learning

Multitask learning is a technique that has been used in two of the research studies during my tenure as a research student at IIT Bombay. In machine learning, we usually have a goal of minimising one loss function that captures the objective of our task. Multitask learning (Crawshaw, 2020) allows us to focus on multiple tasks at once. So we can have two separate tasks that may have their own objective functions but utilise the same data. We combine their losses into a combined loss function and use this to train our model. Such an approach benefits from numerous advantages such as better data efficiency, less overfitting, smaller model sizes, and faster overall learning times. In multitask learning we learn shared representations of data for multiple tasks. This provides the potential of having complementary tasks that might help each other during the learning phase. As an example, the tasks named entity

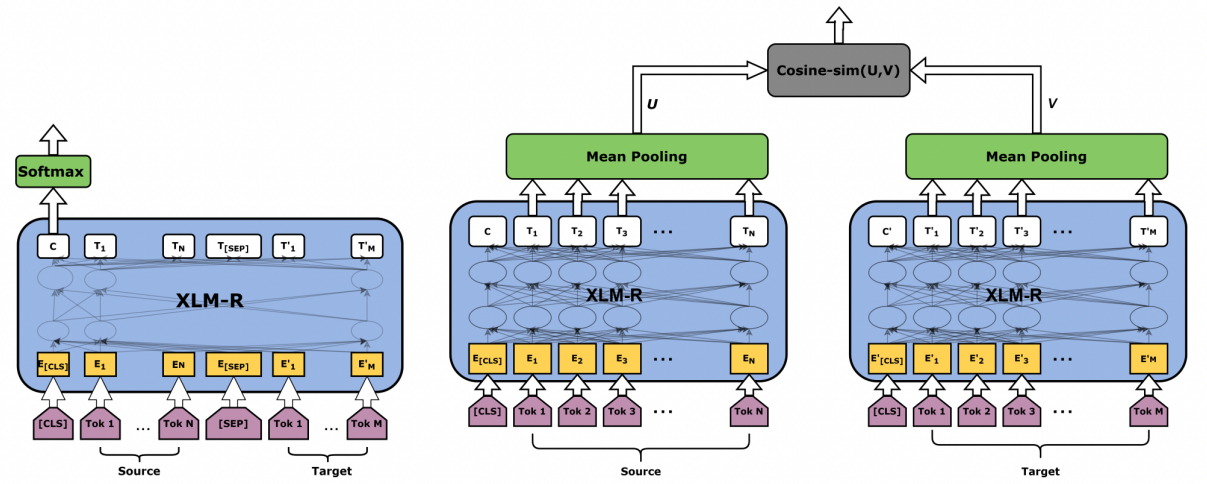

Figure 1: Mono-TransQuest (left) and Siamese-Transquest (right) architectures

recognition and part of speech tagging are two such tasks.

5 Perturbations

Perturbation in natural language processing terms can be defined as a small change in a sentence. Perturbation has been extensively used as a data augmenting technique in various researches in NLP. We use the perturbations described by (Kanojia et al., 2021) for our research into using perturbation and multi-task learning to sentence-level QE performance. The example 5.1 shows one of the perturbations in which negation words are removed if present.

Example 5.1 *Original sentence: Rainfalls are not necessarily preceded by storms.*
Perturbed sentence: Rainfalls are necessarily preceded by storms.

We use the work by Diptesh Kanojia (Kanojia et al., 2021) to generate perturbations. We generate meaning-preserving perturbations (MPPs) and meaning-altering perturbations (MAPs).

5.1 Meaning Preserving Perturbations

The meaning-preserving perturbations are the ones that change the sentence without altering the overall meaning of the sentence. We include six meaning-preserving perturbations to augment the dataset. The six MPPs as described below:

1. **Removal of Punctuations (MPP-1)**: In this perturbation, punctuations are removed from the sentence using Python's standard *string* library.

2. **Replacement of Punctuations (MPP-2)**: In this perturbation, each punctuation is replaced by a different random punctuation.

3. **Removal of Determiners (MPP-3)**: We used Spacy's[2] Part-of-speech (POS) tagger to identify and remove the determiners.

4. **Replacing Determiners (MPP-4)**: In this perturbation, we use *spaCy's* POS tagger to identify the determiners and then replace them with a randomly chosen determiner from a list.

5. **Change in Word-casing to UPPERCASE (MPP-5)**: For MPP-5 perturbations we randomly selected content words and converted them to UPPERCASE.

6. **Change in Word-casing to lowercase (MPP-6)**: For MPP-6 perturbations we randomly selected content words and converted them to lowercase.

An example showing all six MPPs from (Kanojia et al., 2021) is displayed in Table 1.

[2] https://spacy.io/

Source	**În alegerile europarlamentare din 2014, UKIP, partid de extremă dreaptă, a obținut peste 20 de locuri in parlamentul european.**	
Reference	**In the 2014 European Parliamentary elections, UKIP, a right-wing party, obtained more than 20 seats in the European Parliament.**	**S1**
Translation	**In the 2014 European Parliamentary elections, UKIP, party of extreā dreaptă, obtained more than 20 seats in the European Parliament.**	0.81
MPP1	In the 2014 European Parliamentary elections UKIP party of extreā dreaptă obtained more than 20 seats in the European Parliament	0.79
MPP2	In the 2014 European Parliamentary elections**!** UKIP(party of extreā dreaptă. obtained more than 20 seats in the European Parliament**?**	0.69
MPP3	In 2014 European Parliamentary elections, UKIP, party of extreā dreaptă, obtained more than 20 seats in European Parliament.	0.80
MPP4	In **such** 2014 European Parliamentary elections , UKIP , party of extreā dreaptă , obtained more than 20 seats in **those** European Parliament.	0.69
MPP5	**IN** the 2014 **EUROPEAN** Parliamentary **ELECTIONS**, UKIP, party of extreā **DREAPTĀ**, **OBTAINED** more **THAN 20 SEATS** in **THE EUROPEAN PARLIAMENT**.	0.76
MPP6	in the 2014 European parliamentary elections, **ukip**, party of extreā dreaptă, obtained more than 20 seats in the European Parliament.	0.75

Table 1: Meaning Preserving Perturbations example. S1 column includes the predicted Z-standardised DA score.

5.2 Meaning Altering Perturbations

The meaning-preserving perturbations are the ones that change the sentence along with altering the overall meaning of the sentence. For Meaning altering sentences we generated the following eight perturbations. The eight MAPs as described below:

1. **Removal of Negation Markers (MAP-1)**: In this perturbation, negation markers such as "no", "not", "n't" etc. are removed from the sentences.

2. **Removal of Random Content Words (MAP-2)**: In this perturbation, each punctuation is replaced by a different random punctuation.

3. **Duplication of Random Content Words (MAP-3)**: In this MAP a random content word from the translation is chosen and added at the immediate next position.

4. **Insertion of Random Words (MAP-4)**: From a vocabulary of words from the complete set of translations in our data set, a random word is inserted at a random position in the sentence making sure the previous and next words if present are not similar.

5. **Replacing Random Content Words (MAP-5)**: For MAP-5 we replace a random content word from the translation with another word from the vocabulary created of all words in the data set.

6. **BERT-based Sentence Replacement (MAP-6)**: We used sentence replacements, based on the BERT-base model (Devlin et al., 2019), with the help of a data augmentation library[3]. This library generates a sentence synonymous with the input provided by using a word replacement approach proposed by (Kobayashi, 2018). However, we observed that these synonymous sentences replace content words thus altering the inherent meaning of the input sentence. Hence, we treat this perturbation as MAP.

7. **Replacing Words with Antonyms (MAP7)**: For this perturbation, we generate perturbed translations where we replace random words in the sentence with their antonyms from the English Wordnet (Miller et al., 1990) using the data augmentation library[4].

8. **Source Sentence as Target (MAP8)**: For the final meaning altering perturbation we simply replace the translation with the source side sentence.

An example showing all eight MAPs from (Kanojia et al., 2021) is displayed in Table 2.

6 Dataset

The WMT dataset contains both sentence-level and word-level data for quality estimation. The WMT21 (Specia et al., 2021), and WMT22 (Zerva et al., 2022) Quality Estimation Shared tasks provide the training, development, and test sets for both word-level and sentence-level quality estimation tasks. The data consists of three low-resource language pairs: Nepali-English (Ne-En), English-Marathi (En-Mr), Sinhalese-English (Si-En); three medium-resource language pairs: Romanian-English (Ro-En), Estonian-English (Et-En), Russian-English (Ru-En); and one high-resource lan-

[3] https://github.com/makcedward/nlpaug
[4] https://github.com/makcedward/nlpaug

Source	На слушании в декабре Блэквуд сказал, что не имел намерения оскорбить буддизм, когда размещал изображение, а после того, как осознал, что оно вызвало массовое возмущение, удалил его и опубликовал извинение.	
Reference	**At a hearing in December, Blackwood said he had not intended to offend Buddhism when he posted the image, and after realizing it had caused widespread outrage, deleted it and issued an apology.**	**S1**
Translation	**At a hearing in December, Blackwood said he had not intended to offend Buddhism when he posted the image, and after realizing it had caused widespread outrage, deleted it and issued an apology.**	0.83
MAP1	At a hearing in December, Blackwood said he **had intended** to offend Buddhism when he posted the image, and after realizing it had caused widespread outrage, deleted it and issued an apology.	0.82
MAP2	At a hearing **in,** Blackwood said he had not intended to offend Buddhism when he posted the image, and after realizing it had caused widespread outrage, deleted it and issued an apology.	0.82
MAP3	At a hearing in December, Blackwood said he had not intended to offend Buddhism when he posted the image, and after realizing **realizing** it had caused widespread outrage, deleted it and issued an apology.	0.81
MAP4	At a hearing in December, Blackwood said he had not intended to offend Buddhism **party** when he posted the image, and after realizing it had caused widespread outrage, deleted it and issued an apology.	0.82
MAP5	At a hearing in December, Blackwood said he had not intended to offend Buddhism when he posted the image, and after realizing it had caused widespread **Ferris**, deleted it and issued an apology.	0.80
MAP6	**at** a hearing in **japan**, **bailey admitted graham did** not intended to offend **buddhism** when **buddhist** posted the **video**, and after realizing he has caused widespread outrage, deleted it and issued **her** apology.	0.77
MAP7	At a hearing in December, Blackwood said he **lack** not intended to **keep** Buddhism when he posted the image, and after realizing it **refuse** caused widespread outrage, **record** it and **recall** an apology.	0.76
MAP8 (Russian)	На слушании в декабре Блэквуд сказал, что не имел намерения оскорбить буддизм, когда размещал изображение, а после того, как осознал, что оно вызвало массовое возмущение, удалил его и опубликовал извинение.	0.83

Table 2: Meaning Altering Perturbations example. S1 column includes the predicted Z-standardised DA score.

Language Pair	Train Set Sentences	Dev Set Sentences	Test Set Sentences
En-De	7,000	1,000	1,000
Et-En	7,000	1,000	1,000
Ro-En	7,000	1,000	1,000
Ru-En	7,000	1,000	1,000
En-Mr	26,000	1,000	1,000
Ne-En	7,000	1,000	1,000
Si-En	7,000	1,000	1,000

Table 3: Dataset used for Quality Estimation from WMT.

guage pair: English-German (En-De). We list the language pair[5] and the number of sentences for each language pair in the table 3.

7 Summary

This survey paper describes the task of Quality Estimation. We began with a brief introduction followed by the motivation behind the research into QE. Then we looked at the background covering machine translation, its four paradigms, and machine translation evaluation. Then we studied the task of QE in Machine translation and its different granularities. We then looked at the different approaches to QE, some architectures, and challenges to QE followed by an overview of multitask learning. We then explored perturbation in detail before moving on to the dataset description.

[5]The language name abbreviations are as follows: En: English, De: German, Zh: Chinese, Et: Estonian, Ro: Romanian, Ru: Russian, Ne: Nepalese, Si: Sinhala, and Mr: Marathi

References

Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey. *CoRR*, abs/2009.09796.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Julia Ive, Frédéric Blain, and Lucia Specia. 2018. deepQuest: A framework for neural-based quality estimation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3146–3157, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Diptesh Kanojia, Marina Fomicheva, Tharindu Ranasinghe, Frédéric Blain, Constantin Orăsan, and Lucia Specia. 2021. Pushing the right buttons: Adversarial evaluation of quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 625–638, Online. Association for Computational Linguistics.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. 1990. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020a. TransQuest at WMT2020: Sentence-level direct assessment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1049–1055, Online. Association for Computational Linguistics.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020b. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with QuEst++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria. Association for Computational Linguistics.

Shuo Sun, Francisco Guzmán, and Lucia Specia. 2020. Are we estimating or guesstimating translation quality? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6262–6267, Online. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.