Translation Quality Assessment: A Brief Survey on Manual and Automatic Methods

Lifeng Han[1]**, Gareth J. F. Jones**[1]**,** and **Alan F. Smeaton**[2]

[1] ADAPT Research Centre
[2] Insight Centre for Data Analytics
School of Computing, Dublin City University, Dublin, Ireland
`lifeng.han@adaptcentre.ie`

Abstract

To facilitate effective translation modeling and translation studies, one of the crucial questions to address is how to assess translation quality. From the perspectives of accuracy, reliability, repeatability and cost, translation quality assessment (TQA) itself is a rich and challenging task. In this work, we present a high-level and concise survey of TQA methods, including both manual judgement criteria and automated evaluation metrics, which we classify into further detailed sub-categories. We hope that this work will be an asset for both translation model researchers and quality assessment researchers. In addition, we hope that it will enable practitioners to quickly develop a better understanding of the conventional TQA field, and to find corresponding closely relevant evaluation solutions for their own needs. This work may also serve inspire further development of quality assessment and evaluation methodologies for other natural language processing (NLP) tasks in addition to machine translation (MT), such as automatic text summarization (ATS), natural language understanding (NLU) and natural language generation (NLG). [1]

1 Introduction

Machine translation (MT) research, starting from the 1950s (Weaver, 1955), has been one of the main research topics in computational linguistics (CL) and natural language processing (NLP), and has influenced and been influenced by several other language processing tasks such as parsing and language modeling. Starting from rule-based methods to example-based, and then statis-

---

[1]authors GJ and AS in alphabetic order

tical methods (Brown et al., 1993; Och and Ney, 2003; Chiang, 2005; Koehn, 2010), to the current paradigm of neural network structures (Cho et al., 2014; Johnson et al., 2016; Vaswani et al., 2017; Lample and Conneau, 2019), MT quality continue to improve. However, as MT and translation quality assessment (TQA) researchers report, MT outputs are still far from reaching human parity (Läubli et al., 2018; Läubli et al., 2020; Han et al., 2020a). MT quality assessment is thus still an important task to facilitate MT research itself, and also for downstream applications. TQA remains a challenging and difficult task because of the richness, variety, and ambiguity phenomena of natural language itself, e.g. the same concept can be expressed in different word structures and patterns in different languages, even inside one language (Arnold, 2003).

In this work, we introduce human judgement and evaluation (HJE) criteria that have been used in standard international shared tasks and more broadly, such as NIST (LI, 2005), WMT (Koehn and Monz, 2006a; Callison-Burch et al., 2007a, 2008, 2009, 2010, 2011, 2012; Bojar et al., 2013, 2014, 2015, 2016, 2017, 2018; Barrault et al., 2019, 2020), and IWSLT (Eck and Hori, 2005; Paul, 2009; Paul et al., 2010; Federico et al., 2011). We then introduce automated TQA methods, including the automatic evaluation metrics that were proposed inside these shared tasks and beyond. Regarding Human Assessment (HA) methods, we categorise them into traditional and advanced sets, with the first set including intelligibility, fidelity, fluency, adequacy, and comprehension, and the second set including task-oriented, extended criteria, utilizing post-editing, segment ranking, crowd source intelligence (direct assessment), and revisiting traditional criteria.

Regarding automated TQA methods, we classify these into three categories including simple n-gram based word surface matching, deeper lin-

guistic feature integration such as syntax and semantics, and deep learning (DL) models, with the first two regarded as traditional and the last one regarded as advanced due to the recent appearance of DL models for NLP. We further divide each of these three categories into sub-branches, each with a different focus. Of course, this classification does not have clear boundaries. For instance some automated metrics are involved in both n-gram word surface similarity and linguistic features. This paper differs from the existing works (Dorr et al., 2009; EuroMatrix, 2007) by introducing recent developments in MT evaluation measures, the different classifications from manual to automatic evaluation methodologies, the introduction of more recently developed quality estimation (QE) tasks, and its concise presentation of these concepts.

We hope that our work will shed light and offer a useful guide for both MT researchers and researchers in other relevant NLP disciplines, from the similarity and evaluation point of view, to find useful quality assessment methods, either from the manual or automated perspective, inspired from this work. This might include, for instance, natural language generation (Gehrmann et al., 2021), natural language understanding (Ruder et al., 2021) and automatic summarization (Mani, 2001; Bhandari et al., 2020).

The rest of the paper is organized as follows: Sections 2 and 3 present human assessment and automated assessment methods respectively; Section 4 presents some discussions and perspectives; Section 5 summarizes our conclusions and future work. We also list some further relevant readings in the appendices, such as evaluating methods of TQA itself, MT QE, and mathematical formulas.[2]

2 Human Assessment Methods

In this section we introduce human judgement methods, as reflected in Fig. 1. This categorises these human methods as Traditional and Advanced.

2.1 Traditional Human Assessment

2.1.1 *Intelligibility and Fidelity*

The earliest human assessment methods for MT can be traced back to around 1966. They include the intelligibility and fidelity used by the au-

---

[2]This work is based on an earlier preprint edition (Han, 2016)
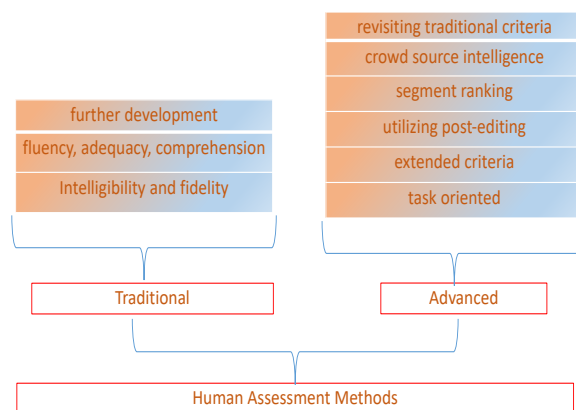

Figure 1: Human Assessment Methods

tomatic language processing advisory committee (ALPAC) (Carroll, 1966). The requirement that a translation is intelligible means that, as far as possible, the translation should read like normal, well-edited prose and be readily understandable in the same way that such a sentence would be understandable if originally composed in the translation language. The requirement that a translation is of high fidelity or accuracy includes the requirement that the translation should, as little as possible, twist, distort, or controvert the meaning intended by the original.

2.1.2 *Fluency, Adequacy and Comprehension*

In 1990s, the Advanced Research Projects Agency (ARPA) created a methodology to evaluate machine translation systems using the adequacy, fluency and comprehension of the MT output (Church and Hovy, 1991) which adapted in MT evaluation campaigns including (White et al., 1994).

To set upp this methodology, the human assessor is asked to look at each fragment, delimited by syntactic constituents and containing sufficient information, and judge its adequacy on a scale 1-to-5. Results are computed by averaging the judgments over all of the decisions in the translation set.

Fluency evaluation is compiled in the same manner as for the adequacy except that the assessor is to make intuitive judgments on a sentence-by-sentence basis for each translation. Human assessors are asked to determine whether the translation is good English without reference to the correct translation. Fluency evaluation determines whether a sentence is well-formed and fluent in context.

Comprehension relates to "Informativeness", whose objective is to measure a system's ability to produce a translation that conveys sufficient information, such that people can gain necessary information from it. The reference set of expert translations is used to create six questions with six possible answers respectively including, "none of above" and "cannot be determined".

2.1.3 Further Development

Bangalore et al. (2000) classified accuracy into several categories including simple string accuracy, generation string accuracy, and two corresponding tree-based accuracy. Reeder (2004) found the correlation between fluency and the number of words it takes to distinguish between human translation and MT output.

The "Linguistics Data Consortium (LDC)"[3] designed two five-point scales representing fluency and adequacy for the annual NIST MT evaluation workshop. The developed scales became a widely used methodology when manually evaluating MT by assigning values. The five point scale for adequacy indicates how much of the meaning expressed in the reference translation is also expressed in a translation hypothesis; the second five point scale indicates how fluent the translation is, involving both grammatical correctness and idiomatic word choices.

Specia et al. (2011) conducted a study of MT adequacy and broke it into four levels, from score 4 to 1: highly adequate, the translation faithfully conveys the content of the input sentence; fairly adequate, where the translation generally conveys the meaning of the input sentence, there are some problems with word order or tense/voice/number, or there are repeated, added or non-translated words; poorly adequate, the content of the input sentence is not adequately conveyed by the translation; and completely inadequate, the content of the input sentence is not conveyed at all by the translation.

2.2 Advanced Human Assessment

2.2.1 Task-oriented

White and Taylor (1998) developed a task-oriented evaluation methodology for Japanese-to-English translation to measure MT systems in light of the tasks for which their output might be used. They seek to associate the diagnostic scores as-

signed to the output used in the DARPA (Defense Advanced Research Projects Agency)[4] evaluation with a scale of language-dependent tasks, such as scanning, sorting, and topic identification. They develop an MT proficiency metric with a corpus of multiple variants which are usable as a set of controlled samples for user judgments. The principal steps include identifying the user-performed text-handling tasks, discovering the order of text-handling task tolerance, analyzing the linguistic and non-linguistic translation problems in the corpus used in determining task tolerance, and developing a set of source language patterns which correspond to diagnostic target phenomena. A brief introduction to task-based MT evaluation work was shown in their later work (Doyon et al., 1999).

Voss and Tate (2006) introduced tasked-based MT output evaluation by the extraction of *who*, *when*, and *where* three types of elements. They extended their work later into event understanding (Laoudi et al., 2006).

2.2.2 Extended Criteria

King et al. (2003) extend a large range of manual evaluation methods for MT systems which, in addition to the earlir mentioned accuracy, include *suitability*, whether even accurate results are suitable in the particular context in which the system is to be used; *interoperability*, whether with other software or hardware platforms; *reliability*, i.e., don't break down all the time or take a long time to get running again after breaking down; *usability*, easy to get the interfaces, easy to learn and operate, and looks pretty; *efficiency*, when needed, keep up with the flow of dealt documents; *maintainability*, being able to modify the system in order to adapt it to particular users; and *portability*, one version of a system can be replaced by a new version, because MT systems are rarely static and they tend to improve over time as resources grow and bugs are fixed.

2.2.3 Utilizing Post-editing

One alternative method to assess MT quality is to compare the post-edited correct translation to the original MT output. This type of evaluation is, however, time consuming and depends on the skills of the human assessor and post-editing performer. One example of a metric that is designed in such a manner is the human translation error rate (HTER) (Snover et al., 2006). This is based on

---

[3]https://www.ldc.upenn.edu

[4]https://www.darpa.mil

the number of editing steps, computing the editing steps between an automatic translation and a reference translation. Here, a human assessor has to find the minimum number of insertions, deletions, substitutions, and shifts to convert the system output into an acceptable translation. HTER is defined as the sum of the number of editing steps divided by the number of words in the acceptable translation.

2.2.4 *Segment Ranking*

In the WMT metrics task, human assessment based on segment ranking was often used. Human assessors were frequently asked to provide a complete ranking over all the candidate translations of the same source segment (Callison-Burch et al., 2011, 2012). In the WMT13 shared-tasks (Bojar et al., 2013), five systems were randomised for the assessor to give a rank. Each time, the source segment and the reference translation were presented together with the candidate translations from the five systems. The assessors ranked the systems from 1 to 5, allowing tied scores. For each ranking, there was the potential to provide as many as 10 pairwise results if there were no ties. The collected pairwise rankings were then used to assign a corresponding score to each participating system to reflect the quality of the automatic translations. The assigned scores could also be used to reflect how frequently a system was judged to be better or worse than other systems when they were compared on the same source segment, according to the following formula:

$$\frac{\#\text{better pairwise ranking}}{\#\text{total pairwise comparison} - \#\text{ties comparisons}} \quad (1)$$

2.2.5 *Crowd Source Intelligence*

With the reported very low human inter-agreement scores from the WMT segment ranking task, researchers started to address this issue by exploring new human assessment methods, as well as seeking reliable automatic metrics for segment level ranking (Graham et al., 2015).

Graham et al. (2013) noted that the lower agreements from WMT human assessment might be caused partially by the interval-level scales set up for the human assessor to choose regarding quality judgement of each segment. For instance, the human assessor possibly corresponds to the situation where neither of the two categories they were forced to choose is preferred. In light of this rationale, they proposed continuous measurement scales (CMS) for human TQA using fluency criteria. This was implemented by introducing the crowdsource platform Amazon MTurk, with some quality control methods such as the insertion of *bad-reference* and *ask-again*, and statistical significance testing. This methodology reported improved both intra-annotator and inter-annotator consistency. Detailed quality control methodologies, including statistical significance testing were documented in direct assessment (DA) (Graham et al., 2016, 2020).

2.2.6 *Revisiting Traditional Criteria*

Popović (2020a) criticized the traditional human TQA methods because they fail to reflect real problems in translation by assigning scores and ranking several candidates from the same source. Instead, Popović (2020a) designed a new methodology by asking human assessors to mark all problematic parts of candidate translations, either words, phrases, or sentences. Two questions that were typically asked of the assessors related to *comprehensibility* and *adequacy*. The first criteria considered whether the translation is understandable, or understandable but with errors; the second criteria measures if the candidate translation has different meaning to the original text, or maintains the meaning but with errors. Both criteria take into account whether parts of the original text are missing in translation. Under a similar experimental setup, Popović (2020b) also summarized the most frequent error types that the annotators recognized as misleading translations.

3 Automated Assessment Methods

Manual evaluation suffers some disadvantages such as that it is time-consuming, expensive, not tune-able, and not reproducible. Due to these aspects, automatic evaluation metrics have been widely used for MT. Typically, these compare the output of MT systems against human reference translations, but there are also some metrics that do not use reference translations. There are usually two ways to offer the human reference translation, either offering one single reference or offering multiple references for a single source sentence (Lin and Och, 2004; Han et al., 2012).

Automated metrics often measure the overlap in words and word sequences, as well as word order and edit distance. We classify these kinds of
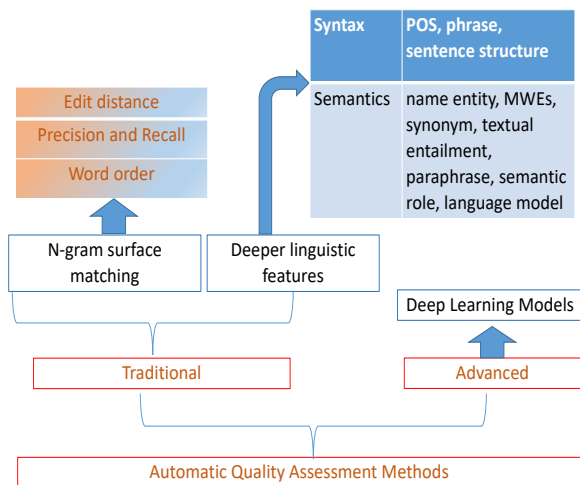
Syntax	POS, phrase, sentence structure
Semantics	name entity, MWEs, synonym, textual entailment, paraphrase, semantic role, language model

- Edit distance
- Precision and Recall
- Word order

N-gram surface matching | Deeper linguistic features

Deep Learning Models

Traditional | Advanced

Automatic Quality Assessment Methods

Figure 2: Automatic Quality Assessment Methods

metrics as "simple n-gram word surface matching". Further developed metrics also take linguistic features into account such as syntax and semantics, including POS, sentence structure, textual entailment, paraphrase, synonyms, named entities, multi-word expressions (MWEs), semantic roles and language models. We classify these metrics that utilize the linguistic features as "Deeper Linguistic Features (aware)". This classification is only for easier understanding and better organization of the content. It is not easy to separate these two categories clearly since sometimes they merge with each other. For instance, some metrics from the first category might also use certain linguistic features. Furthermore, we will introduce some recent models that apply deep learning into the TQA framework, as in Fig. 2. Due to space limitations, we present MT quality estimation (QE) task which does not rely on reference translations during the automated computing procedure in the appendices.

3.1 N-gram Word Surface Matching

3.1.1 *Levenshtein Distance*

By calculating the minimum number of editing steps to transform MT output to reference, Su et al. (1992) introduced the word error rate (WER) metric into MT evaluation. This metric, inspired by Levenshtein Distance (or edit distance), takes word order into account, and the operations include insertion (adding word), deletion (dropping word) and replacement (or substitution, replace one word with another), the minimum number of editing steps needed to match two sequences.

One of the weak points of the WER metric is

the fact that word ordering is not taken into account appropriately. The WER scores are very low when the word order of system output translation is "wrong" according to the reference. In the Levenshtein distance, the mismatches in word order require the deletion and re-insertion of the misplaced words. However, due to the diversity of language expressions, some so-called "wrong" order sentences by WER also prove to be good translations. To address this problem, the position-independent word error rate (PER) introduced by Tillmann et al. (1997) is designed to ignore word order when matching output and reference. Without taking into account of the word order, PER counts the number of times that identical words appear in both sentences. Depending on whether the translated sentence is longer or shorter than the reference translation, the rest of the words are either insertion or deletion ones.

Another way to overcome the unconscionable penalty on word order in the Levenshtein distance is adding a novel editing step that allows the movement of word sequences from one part of the output to another. This is something a human post-editor would do with the cut-and-paste function of a word processor. In this light, Snover et al. (2006) designed the translation edit rate (TER) metric that adds block movement (jumping action) as an editing step. The shift option is performed on a contiguous sequence of words within the output sentence. For the edits, the cost of the block movement, any number of continuous words and any distance, is equal to that of the single word operation, such as insertion, deletion and substitution.

3.1.2 *Precision and Recall*

The widely used evaluation BLEU metric (Papineni et al., 2002) is based on the degree of n-gram overlap between the strings of words produced by the MT output and the human translation references at the corpus level. BLEU calculates precision scores with n-grams sized from 1-to-4, together multiplied by the coefficient of brevity penalty (BP). If there are multi-references for each candidate sentence, then the nearest length as compared to the candidate sentence is selected as the effective one. In the BLEU metric, the n-gram precision weight λ_n is usually selected as a uniform weight. However, the 4-gram precision value can be very low or even zero when the test corpus is small. To weight more heavily those n-grams that are more informative, Doddington (2002) pro-

poses the NIST metric with the information weight added. Furthermore, Doddington (2002) replaces the geometric mean of co-occurrences with the arithmetic average of n-gram counts, extends the n-gram into 5-gram ($N = 5$), and selects the average length of reference translations instead of the nearest length.

ROUGE (Lin and Hovy, 2003) is a recall-oriented evaluation metric, which was initially developed for summaries, and inspired by BLEU and NIST. ROUGE has also been applied in automated TQA in later work (Lin and Och, 2004).

The F-measure is the combination of precision (P) and recall (R), which was firstly employed in information retrieval (IR) and latterly adopted by the information extraction (IE) community, MT evaluations, and others. Turian et al. (2006) carried out experiments to examine how standard measures such as precision, recall and F-measure can be applied to TQA and showed the comparisons of these standard measures with some alternative evaluation methodologies.

Banerjee and Lavie (2005) designed METEOR as a novel evaluation metric. METEOR is based on the general concept of flexible unigram matching, precision and recall, including the match between words that are simple morphological variants of each other with identical word stems and words that are synonyms of each other. To measure how well-ordered the matched words in the candidate translation are in relation to the human reference, METEOR introduces a penalty coefficient, different to what is done in BLEU, by employing the number of matched chunks.

3.1.3 *Revisiting Word Order*

The right word order plays an important role to ensure a high quality translation output. However, language diversity also allows different appearances or structures of a sentence. How to successfully achieve a penalty on really wrong word order, i.e. wrongly structured sentences, instead of on "correct" different order, i.e. the candidate sentence that has different word order to the reference, but is well structured, has attracted a lot of interest from researchers. In fact, the Levenshtein distance (*Section 3.1.1*) and n-gram based measures also contain word order information.

Featuring the explicit assessment of word order and word choice, Wong and yu Kit (2009) developed the evaluation metric ATEC (assessment of text essential characteristics). This is also based

on precision and recall criteria, but with a position difference penalty coefficient attached. The word choice is assessed by matching word forms at various linguistic levels, including surface form, stem, sound and sense, and further by weighing the informativeness of each word.

Partially inspired by this, our work LEPOR (Han et al., 2012) is designed as a combination of augmented evaluation factors including n-gram based *word order penalty* in addition to *precision*, *recall*, and *enhanced sentence-length penalty*. The LEPOR metric (including *h*LEPOR) was reported with top performance on the English-to-other (Spanish, German, French, Czech and Russian) language pairs in ACL-WMT13 metrics shared tasks for *system level* evaluation (Han et al., 2013d). The n-gram based variant *n*LEPOR (Han et al., 2014) was also analysed by MT researchers as one of the three best performing *segment level* automated metrics (together with METEOR and sentBLEU-MOSES) that correlated with human judgement at a level that was not significantly outperformed by any other metrics, on Spanish-to-English, in addition to an aggregated set of overall tested language pairs (Graham et al., 2015).

3.2 Deeper Linguistic Features

Although some of the previously outlined metrics incorporate linguistic information, e.g. synonyms and stemming in METEOR and part of speech (POS) in LEPOR, the simple n-gram word surface matching methods mainly focus on the exact matches of the surface words in the output translation. The advantages of the metrics based on the first category (simple n-gram word matching) are that they perform well in capturing translation fluency (Lo et al., 2012), are very fast to compute and have low cost. On the other hand, there are also some weaknesses, for instance, syntactic information is rarely considered and the underlying assumption that a good translation is one that shares the same word surface lexical choices as the reference translations is not justified semantically. Word surface lexical similarity does not adequately reflect similarity in meaning. Translation evaluation metrics that reflect meaning similarly need to be based on similarity of semantic structure and not merely flat lexical similarity.

3.2.1 *Syntactic Similarity*

Syntactic similarity methods usually employ the features of morphological POS information,

phrase categories, phrase decompositionality or sentence structure generated by linguistic tools such as a language parser or chunker.

In grammar, a **POS** is a linguistic category of words or lexical items, which is generally defined by the syntactic or morphological behaviour of the lexical item. Common linguistic categories of lexical items include noun, verb, adjective, adverb, and preposition. To reflect the syntactic quality of automatically translated sentences, researchers employ POS information into their evaluations. Using the IBM model one, Popović et al. (2011) evaluate translation quality by calculating the similarity scores of source and target (translated) sentences without using a reference translation, based on the morphemes, 4-gram POS and lexicon probabilities. Dahlmeier et al. (2011) developed the TESLA evaluation metrics, combining the synonyms of bilingual phrase tables and POS information in the matching task. Other similar work using POS information include (Giménez and Márquez, 2007; Popovic and Ney, 2007; Han et al., 2014).

In linguistics, a **phrase** may refer to any group of words that form a constituent, and so functions as a single unit in the syntax of a sentence. To measure an MT system's performance in translating new text types, such as in what ways the system itself could be extended to deal with new text types, Povlsen et al. (1998) carried out work focusing on the study of an English-to-Danish MT system. The syntactic constructions are explored with more complex linguistic knowledge, such as the identification of fronted adverbial subordinate clauses and prepositional phrases. Assuming that similar grammatical structures should occur in both source and translations, Avramidis et al. (2011) perform evaluation on source (German) and target (English) sentences employing the features of sentence length ratio, unknown words, phrase numbers including noun phrase, verb phrase and prepositional phrase. Other similar work using phrase similarity includes (Li et al., 2012) that uses noun phrases and verb phrases from chunking, (Echizen-ya and Araki, 2010) that only uses the noun phrase chunking in automatic evaluation, and (Han et al., 2013c) that designs a universal phrase tagset for French to English MT evaluation.

Syntax is the study of the principles and processes by which sentences are constructed in par-

ticular languages. To address the overall goodness of a translated **sentence's structure**, Liu and Gildea (2005) employ constituent labels and head-modifier dependencies from a language parser as syntactic features for MT evaluation. They compute the similarity of dependency trees. Their experiments show that adding syntactic information can improve evaluation performance, especially for predicting the fluency of translation hypotheses. Other works that use syntactic information in evaluation include (Lo and Wu, 2011a) and (Lo et al., 2012) that use an automatic shallow parser and the RED metric (Yu et al., 2014) that applies dependency trees.

3.2.2 *Semantic Similarity*

As a contrast to syntactic information, which captures overall grammaticality or sentence structure similarity, the semantic similarity of automatic translations and the source sentences (or references) can be measured by employing semantic features.

To capture the semantic equivalence of sentences or text fragments, **named entity** knowledge is taken from the literature on named-entity recognition, which aims to identify and classify atomic elements in a text into different entity categories (Marsh and Perzanowski, 1998; Guo et al., 2009). The most commonly used entity categories include the names of persons, locations, organizations and time (Han et al., 2013a). In the MEDAR2011 evaluation campaign, one baseline system based on Moses (Koehn et al., 2007) utilized an Open NLP toolkit to perform named entity detection, in addition to other packages. The low performances from the perspective of named entities causes a drop in fluency and adequacy. In the quality estimation of the MT task in WMT 2012, (Buck, 2012) introduced features including named entity, in addition to discriminative word lexicon, neural networks, back off behavior (Raybaud et al., 2011) and edit distance. Experiments on individual features showed that, from the perspective of the increasing the correlation score with human judgments, the named entity feature contributed the most to the overall performance, in comparisons to the impacts of other features.

Multi-word Expressions (MWEs) set obstacles for MT models due to their complexity in presentation as well as idiomaticity (Sag et al., 2002; Han et al., 2020b,a; Han et al., 2021). To investigate the effect of MWEs in MT evaluation (MTE),

Salehi et al. (2015) focused on the *compositionality* of noun compounds. They identify the **noun compounds** first from the system outputs and reference with Stanford parser. The matching scores of the system outputs and reference sentences are then recalculated, adding up to the Tesla metric, by considering the predicated compositionality of identified noun compound phrases. Our own recent work in this area (Han et al., 2020a) provides an extensive investigation into various MT errors caused by MWEs.

Synonyms are words with the same or close meanings. One of the most widely used synonym databases in the NLP literature is WordNet (Miller et al., 1990), which is an English lexical database grouping English words into sets of synonyms. WordNet classifies words mainly into four kinds of POS categories; Noun, Verb, Adjective, and Adverb, without prepositions, determiners, etc. Synonymous words or phrases are organized using the unit of synsets. Each synset is a hierarchical structure with the words at different levels according to their semantic relations.

Textual entailment is usually used as a directive relation between text fragments. If the truth of one text fragment TA follows another text fragment TB, then there is a directional relation between TA and TB (TB $\Rightarrow$ TA). Instead of the pure logical or mathematical entailment, textual entailment in natural language processing (NLP) is usually performed with a relaxed or loose definition (Dagan et al., 2006). For instance, according to text fragment TB, if it can be inferred that the text fragment TA is *most likely* to be true then the relationship TB $\Rightarrow$ TA is also established. Since the relation is directive, it means that the inverse inference (TA $\Rightarrow$ TB) is not ensured to be true (Dagan and Glickman, 2004). Castillo and Estrella (2012) present a new approach for MT evaluation based on the task of "Semantic Textual Similarity". This problem is addressed using a textual entailment engine based on WordNet semantic features.

Paraphrase is to restate the meaning of a passage of text but utilizing other words, which can be seen as bidirectional textual entailment (Androutsopoulos and Malakasiotis, 2010). Instead of the literal translation, word by word and line by line used by meta-phrases, a paraphrase represents a dynamic equivalent. Further knowledge of paraphrases from the aspect of linguistics is introduced in the works by (McKeown, 1979; Meteer and Shaked, 1988; Barzilay and Lee, 2003). Snover et al. (2006) describe a new evaluation metric TER-Plus (TERp). Sequences of words in the reference are considered to be paraphrases of a sequence of words in the hypothesis if that phrase pair occurs in the TERp phrase table.

Semantic roles are employed by researchers as linguistic features in MT evaluation. To utilize semantic roles, sentences are usually first shallow parsed and entity tagged. Then the semantic roles are used to specify the arguments and adjuncts that occur in both the candidate translation and reference translation. For instance, the semantic roles introduced by Giménez and Márquez (2007); Giméne and Márquez (2008) include causative agent, adverbial adjunct, directional adjunct, negation marker, and predication adjunct, etc. In a further development, Lo and Wu (2011a,b) presented the MEANT metric designed to capture the predicate-argument relations as structural relations in semantic frames, which are not reflected in the flat semantic role label features in the work of Giménez and Márquez (2007). Furthermore, instead of using uniform weights, Lo et al. (2012) weight the different types of semantic roles as empirically determined by their relative importance to the adequate preservation of meaning. Generally, semantic roles account for the semantic structure of a segment and have proved effective in assessing adequacy of translation.

Language models are also utilized by MT evaluation researchers. A statistical language model usually assigns a probability to a sequence of words by means of a probability distribution. Gamon et al. (2005) propose the LM-SVM, language model, and support vector machine methods investigating the possibility of evaluating MT quality and fluency in the absence of reference translations. They evaluate the performance of the system when used as a classifier for identifying highly dis-fluent and ill-formed sentences.

Generally, the linguistic features mentioned above, including both syntactic and semantic features, are combined in two ways, either by following a machine learning approach (Albrecht and Hwa, 2007; Leusch and Ney, 2009), or trying to combine a wide variety of metrics in a more simple and straightforward way, such as (Giméne and Márquez, 2008; Specia and Giménez, 2010; Comelles et al., 2012).

3.3 Neural Networks for TQA

We briefly list some works that have applied deep learning and neural networks for TQA which are promising for further exploration. For instance, Guzmán et al. (2015); Guzmn et al. (2017) use neural networks (NNs) for TQA for pair wise modeling to choose the best hypothetical translation by comparing candidate translations with a reference, integrating syntactic and semantic information into NNs. Gupta et al. (2015b) proposed LSTM networks based on dense vectors to conduct TQA, while Ma et al. (2016) designed a new metric based on bi-directional LSTMs, which is similar to the work of Guzmán et al. (2015) but with less complexity by allowing the evaluation of a single hypothesis with a reference, instead of a pairwise situation.

4 Discussion and Perspectives

In this section, we examine several topics that can be considered for further development of MT evaluation fields.

The first aspect is that development should involve both n-gram word surface matching and the deeper linguistic features. Because natural languages are expressive and ambiguous at different levels (Giménez and Márquez, 2007), simple n-gram word surface similarity based metrics limit their scope to the lexical dimension and are not sufficient to ensure that two sentences convey the same meaning or not. For instance, (Callison-Burch et al., 2006a) and (Koehn and Monz, 2006b) report that simple n-gram matching metrics tend to favor automatic statistical MT systems. If the evaluated systems belong to different types that include rule based, human aided, and statistical systems, then the simple n-gram matching metrics, such as BLEU, give a strong disagreement between these ranking results and those of the human assessors. So deeper linguistic features are very important in the MT evaluation procedure.

However, inappropriate utilization, or abundant or abused utilization, of linguistic features will result in limited popularity of measures incorporating linguistic features. In the future, how to utilize the linguistic features in a more accurate, flexible and simplified way, will be one challenge in MT evaluation. Furthermore, the MT evaluation from the aspects of semantic similarity is more reasonable and reaches closer to the human judgments, so it should receive more attention.

The second major aspect is that MT quality estimation (QE) tasks are different to traditional MT evaluation in several ways, such as extracting reference-independent features from input sentences and their translation, obtaining quality scores based on models produced from training data, predicting the quality of an unseen translated text at system run-time, filtering out sentences which are not good enough for post processing, and selecting the best translation among multiple systems. This means that with so many challenges, the topic will continuously attract many researchers.

Thirdly, some advanced or challenging technologies that can be further applied to MT evaluation include the deep learning models (Gupta et al., 2015a; Zhang and Zong, 2015), semantic logic form, and decipherment model.

5 Conclusions and Future Work

In this paper we have presented a survey of the state-of-the-art in translation quality assessment methodologies from the viewpoints of both manual judgements and automated methods. This work differs from conventional MT evaluation review work by its concise structure and inclusion of some recently published work and references. Due to space limitations, in the main content, we focused on conventional human assessment methods and automated evaluation metrics with reliance on reference translations. However, we also list some interesting and related work in the appendices, such as the quality estimation in MT when the reference translation is not presented during the estimation, and the evaluating methodology for TQA methods themselves. However, this arrangement does not affect the overall understanding of this paper as a self contained overview. We believe this work can help both MT and NLP researchers and practitioners in identifying appropriate quality assessment methods for their work. We also expect this work might shed some light on evaluation methodologies in other NLP tasks, due to the similarities they share, such as text summarization (Mani, 2001; Bhandari et al., 2020), natural language understanding (Ruder et al., 2021), natural language generation (Gehrmann et al., 2021), as well as programming language (code) generation (Liguori et al., 2021).

Acknowledgments

References

J. Albrecht and R. Hwa. 2007. A re-examination of machine learning approaches for sentence-level mt evaluation. In *Proceedings of the 45th Annual Meeting of the ACL, Prague, Czech Republic*.

Jon Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

D. Arnold. 2003. *Computers and Translation: A translator's guide-Chap8 Why translation is difficult for computers*. Benjamins Translation Library.

Eleftherios Avramidis, Maja Popovic, David Vilar, and Aljoscha Burchardt. 2011. Evaluate with confidence estimation: Machine ranking of translation outputs using grammatical features. In *Proceedings of WMT 2011*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL 2005*.

Srinivas Bangalore, Owen Rambow, and Steven Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of INLG 2000*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019.

Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of NAACL 2003*.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 workshop on statistical machine translation. In *Proceedings of WMT 2013*.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 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.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp,

Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 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. Association for Computational Linguistics.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Christian Buck. 2012. Black box features for the wmt 2012 quality estimation shared task. In *Proceedings of WMT 2012*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007a. (meta-) evaluation of machine translation. In *Proceedings of WMT 2007*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007b. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 64–71. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of WMT 2008*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaridan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the WMT 2010*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of WMT 2012*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the 4th WMT 2009*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F. Zaridan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of WMT 2011*.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006a. Improved statistical machine translation using paraphrases. In *Proceedings of HLT-NAACL 2006*.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006b. Re-evaluating the role of bleu in machine translation research. In *Proceedings of EACL 2006*, volume 2006, pages 249–256.

John B. Carroll. 1966. An experiment in evaluating the quality of translation. *Mechanical Translation and Computational Linguistics*, 9(3-4):67–75.

Julio Castillo and Paula Estrella. 2012. Semantic textual similarity for MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 52–58, Montréal, Canada. Association for Computational Linguistics.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan. Association for Computational Linguistics.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.

Kenneth Church and Eduard Hovy. 1991. Good applications for crummy machine translation. In *Proceedings of the Natural Language Processing Systems Evaluation Workshop*.

Jasob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):3746.

Elisabet Comelles, Jordi Atserias, Victoria Arranz, and Irene Castellón. 2012. Verta: Linguistic features in mt evaluation. In *LREC*, pages 3944–3950.

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. In *Learning Methods for Text Understanding and Mining workshop*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. *Machine Learning Challenges:LNCS*, 3944:177–190.

Daniel Dahlmeier, Chang Liu, and Hwee Tou Ng. 2011. Tesla at wmt2011: Translation evaluation and tunable metric. In *Proceedings of WMT 2011*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *HLT Proceedings*.

Bonnie Dorr, Matt Snover, and etc. Nitin Madnani. 2009. Part 5: Machine translation evaluation. In *Bonnie Dorr edited DARPA GALE program report*.

Jennifer B. Doyon, John S. White, and Kathryn B. Taylor. 1999. Task-based evaluation for machine translation. In *Proceedings of MT Summit 7*.

H. Echizen-ya and K. Araki. 2010. Automatic evaluation method for machine translation using noun-phrase chunking. In *Proceedings of the ACL 2010*.

Matthias Eck and Chiori Hori. 2005. Overview of the iwslt 2005 evaluation campaign. In *In proceeding of International Workshop on Spoken Language Translation (IWSLT)*.

Project EuroMatrix. 2007. 1.3: Survey of machine translation evaluation. In *EuroMatrix Project Report, Statistical and Hybrid MT between All European Languages, co-ordinator: Prof. Hans Uszkoreit*.

Marcello Federico, Luisa Bentivogli, Michael Paul, and Sebastian Stüker. 2011. Overview of the iwslt 2011 evaluation campaign. In *In proceeding of International Workshop on Spoken Language Translation (IWSLT)*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-level mt evaluation without reference translations beyond language modelling. In *Proceedings of EAMT*, pages 103–112.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Rubungo Andre Niyongabo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics. *arXiv e-prints*, page arXiv:2102.01672.

Jesús Giméne and Lluís Márquez. 2008. A smorgasbord of features for automatic mt evaluation. In *Proceedings of WMT 2008*, pages 195–198.

Jesús Giménez and Lluís Márquez. 2007. Linguistic features for automatic evaluation of heterogenous mt systems. In *Proceedings of WMT 2007*.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1183–1191.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2016. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Yvette Graham and Qun Liu. 2016. Achieving accurate conclusions in evaluation of automatic machine translation metrics. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1–10.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceeding of SIGIR*.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015a. Machine translation evaluation using recurrent neural networks. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 380–384, Lisbon, Portugal. Association for Computational Linguistics.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015b. Reval: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 Conference on Emperical Methods in Natural Language Processing*, pages 1066–1072. Association for Computational Linguistics, o.A.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2015. Pairwise neural machine translation evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint*

Conference of the Asian Federation of Natural Language Processing (ACL'15), pages 805–814, Beijing, China. Association for Computational Linguistics.

Francisco Guzmn, Shafiq Joty, Llus Mrquez, and Preslav Nakov. 2017. Machine translation evaluation with neural networks. *Comput. Speech Lang.*, 45(C):180–200.

Anders Hald. 1998. *A History of Mathematical Statistics from 1750 to 1930*. ISBN-10: 0471179124. Wiley-Interscience; 1 edition.

Aaron L-F Han, Derek F Wong, and Lidia S Chao. 2013a. Chinese named entity recognition with conditional random fields in the light of chinese characteristics. In *Language Processing and Intelligent Information Systems*, pages 57–68. Springer.

Lifeng Han. 2014. *LEPOR: An Augmented Machine Translation Evaluation Metric*. University of Macau, Macao.

Lifeng Han. 2016. Machine Translation Evaluation Resources and Methods: A Survey. *arXiv e-prints*, page arXiv:1605.04515.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020a. AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020b. MultiMWE: Building a multi-lingual multi-word expression (MWE) parallel corpora. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2970–2979, Marseille, France. European Language Resources Association.

Lifeng Han, Gareth J. F. Jones, Alan F. Smeaton, and Paolo Bolzoni. 2021. Chinese Character Decomposition for Neural MT with Multi-Word Expressions. *arXiv e-prints*, page arXiv:2104.04497.

Lifeng Han, Derek F. Wong, Lidia S. Chao, Liangye He, Yi Lu, Junwen Xing, and Xiaodong Zeng. 2013b. Language-independent model for machine translation evaluation with reinforced factors. In *Machine Translation Summit XIV*, pages 215–222. International Association for Machine Translation.

Lifeng Han, Derek Fai Wong, and Lidia Sam Chao. 2012. A robust evaluation metric for machine translation with augmented factors. In *Proceedings of COLING*.

Lifeng Han, Derek Fai Wong, Lidia Sam Chao, Liangeye He, Shuo Li, and Ling Zhu. 2013c. Phrase tagset mapping for french and english treebanks and its application in machine translation evaluation. In *International Conference of the German Society for Computational Linguistics and Language Technology, LNAI Vol. 8105*, pages 119–131.

Lifeng Han, Derek Fai Wong, Lidia Sam Chao, Liangeye He, and Yi Lu. 2014. Unsupervised quality estimation model for english to german translation and its application in extensive supervised evaluation. In *The Scientific World Journal. Issue: Recent Advances in Information Technology*, pages 1–12.

Lifeng Han, Derek Fai Wong, Lidia Sam Chao, Yi Lu, Liangye He, Yiming Wang, and Jiaji Zhou. 2013d. A description of tunable machine translation evaluation systems in wmt13 metrics task. In *Proceedings of WMT 2013*, pages 414–421.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30:81–93.

Maurice G. Kendall and Jean Dickinson Gibbons. 1990. *Rank Correlation Methods*. Oxford University Press, New York.

Marrgaret King, Andrei Popescu-Belis, and Eduard Hovy. 2003. Femti: Creating and using a framework for mt evaluation. In *Proceedings of the Machine Translation Summit IX*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of Conference on Association of Computational Linguistics*.

Philipp Koehn and Christof Monz. 2006a. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. 2006b. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of WMT 2006*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

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

Jamal Laoudi, Ra R. Tate, and Clare R. Voss. 2006. Task-based mt evaluation: From who/when/where extraction to event understanding. In *in Proceedings of LREC-06*, pages 2048–2053.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL 2003*.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Alon Lavie. 2013. Automated metrics for mt evaluation. *Machine Translation*, 11:731.

Guy Lebanon and John Lafferty. 2002. Combining rankings using conditional probability models on permutations. In *Proceeding of the ICML*.

Gregor Leusch and Hermann Ney. 2009. Edit distances with block movements and error rate confidence estimates. *Machine Translation*, 23(2-3).

A. LI. 2005. Results of the 2005 nist machine translation evaluation. In *Proceedings of WMT 2005*.

Liang You Li, Zheng Xian Gong, and Guo Dong Zhou. 2012. Phrase-based evaluation for machine translation. In *Proceedings of COLING*, pages 663–672.

Pietro Liguori, Erfan Al-Hossami, Domenico Cotroneo, Roberto Natella, Bojan Cukic, and Samira Shaikh. 2021. Shellcode_IA32: A Dataset for Automatic Shellcode Generation. *arXiv e-prints*, page arXiv:2104.13100.

Chin-Yew Lin and E. H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of NAACL 2003*.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of ACL 2004*.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedingsof theACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Chi Kiu Lo, Anand Karthik Turmuluru, and Dekai Wu. 2012. Fully automatic semantic mt evaluation. In *Proceedings of WMT 2012*.

Chi Kiu Lo and Dekai Wu. 2011a. Meant: An inexpensive, high- accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of ACL 2011*.

Chi Kiu Lo and Dekai Wu. 2011b. Structured vs. flat semantic role representations for machine translation evaluation. In *Proceedings of the 5th Workshop on Syntax and Structure in StatisticalTranslation (SSST-5)*.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67.

Qingsong Ma, Fandong Meng, Daqi Zheng, Mingxuan Wang, Yvette Graham, Wenbin Jiang, and Qun Liu. 2016. Maxsd: A neural machine translation evaluation metric optimized by maximizing similarity distance. In *Natural Language Understanding and Intelligent Applications - 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, IC-CPOL 2016, Kunming, China, December 2-6, 2016, Proceedings*, pages 153–161.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

I. Mani. 2001. Summarization evaluation: An overview. In *NTCIR*.

Elaine Marsh and Dennis Perzanowski. 1998. Muc-7 evaluation of ie technology: Overview of results. In *Proceedingsof Message Understanding Conference (MUC-7)*.

Kathleen R. McKeown. 1979. Paraphrasing using given and new information in a question-answer system. In *Proceedings of ACL 1979*.

Marie Meteer and Varda Shaked. 1988. Microsoft research treelet translation system: Naacl 2006 europarl evaluation. In *Proceedings of COLING*.

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

Douglas C. Montgomery and George C. Runger. 2003. *Applied statistics and probability for engineers*, third edition. John Wiley and Sons, New York.

Erwan Moreau and Carl Vogel. 2013. Weakly supervised approaches for quality estimation. *Machine Translation*, 27(3–4):257–280.

Erwan Moreau and Carl Vogel. 2014. Limitations of MT quality estimation supervised systems: The tails prediction problem. In *Proceedings of COLING*

2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 2205–2216, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*.

M. Paul. 2009. Overview of the iwslt 2009 evaluation campaign. In *Proceeding of IWSLT*.

Michael Paul, Marcello Federico, and Sebastian Stüker. 2010. Overview of the iwslt 2010 evaluation campaign. In *Proceeding of IWSLT*.

Karl Pearson. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(5):157–175.

Maja Popović, David Vilar, Eleftherios Avramidis, and Aljoscha Burchardt. 2011. Evaluation without references: Ibm1 scores as evaluation metrics. In *Proceedings of WMT 2011*.

M. Popovic and Hermann Ney. 2007. Word error rates: Decomposition over pos classes and applications for error analysis. In *Proceedings of WMT 2007*.

Maja Popović. 2020a. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5059–5069, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maja Popović. 2020b. Relations between comprehensibility and adequacy errors in machine translation output. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 256–264, Online. Association for Computational Linguistics.

Claus Povlsen, Nancy Underwood, Bradley Music, and Anne Neville. 1998. Evaluating text-type suitability for machine translation a case study on an english-danish system. In *Proceeding LREC*.

Sylvain Raybaud, David Langlois, and Kamel Smaïli. 2011. "this sentence is wrong." detecting errors in machine-translated sentences. *Machine Translation*, 25(1):1–34.

Florence Reeder. 2004. Investigation of intelligibility judgments. In *Machine Translation: From Real Users to Research*, pages 227–235, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation. *arXiv e-prints*, page arXiv:2104.07412.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bahar Salehi, Nitika Mathur, Paul Cook, and Timothy Baldwin. 2015. The impact of multiword expression compositionality on machine translation evaluation. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 54–59, Denver, Colorado. Association for Computational Linguistics.

Mattthew Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceeding of AMTA*.

L. Specia and J. Giménez. 2010. Combining confidence estimation and reference-based metrics for segment-level mt evaluation. In *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA)*.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, 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, Naheh Hajlaoui, Catalina Hallett, and Wilker Aziz. 2011. Predicting machine translation adequacy. In *Machine Translation Summit XIII*.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*.

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.

Keh-Yih Su, Wu Ming-Wen, and Chang Jing-Shin. 1992. A new quantitative quality measure for machine translation systems. In *Proceeding of COLING*.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. Accelerated dp based search for statistical translation. In *Proceeding of EUROSPEECH*.

Joseph P Turian, Luke Shea, and I Dan Melamed. 2006. Evaluation of machine translation and its evaluation. Technical report, DTIC Document.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Conference on Neural Information Processing System*, pages 6000–6010.

Clare R. Voss and Ra R. Tate. 2006. Task-based evaluation of machine translation (mt) engines: Measuring how well people extract who, when, where-type elements in mt output. In *In Proceedings of 11th Annual Conference of the European Association for Machine Translation (EAMT-2006*, pages 203–212.

Warren Weaver. 1955. Translation. *Machine Translation of Languages: Fourteen Essays*.

John S. White, Theresa O' Connell, and Francis O' Mara. 1994. The arpa mt evaluation methodologies: Evolution, lessons, and future approaches. In *Proceeding of AMTA*.

John S. White and Kathryn B. Taylor. 1998. A task-oriented evaluation metric for machine translation. In *Proceeding LREC*.

Billy Wong and Chun yu Kit. 2009. Atec: automatic evaluation of machine translation via word choice and word order. *Machine Translation*, 23(2-3):141–155.

Hui Yu, Xiaofeng Wu, Jun Xie, Wenbin Jiang, Qun Liu, and Shouxun Lin. 2014. RED: A reference dependency based MT evaluation metric. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 2042–2051.

Jiajun Zhang and Chengqing Zong. 2015. Deep neural networks in machine translation: An overview. *IEEE Intelligent Systems*, (5):16–25.

Appendices

Appendix A: Evaluating TQA

A.1: Statistical Significance

If different MT systems produce translations with different qualities on a dataset, how can we ensure that they indeed own different system quality? To explore this problem, Koehn (2004) presents an investigation statistical significance testing for MT evaluation. The bootstrap re-sampling method is used to compute the statistical significance intervals for evaluation metrics on small test sets. Statistical significance usually refers to two separate notions, one of which is the p-value, the probability that the observed data will occur by chance in a given single null hypothesis. The other one is the "Type I" error rate of a statistical hypothesis test, which is also called "false positive" and measured by the probability of incorrectly rejecting a given null hypothesis in favour of a second alternative hypothesis (Hald, 1998).

A.2: Evaluating Human Judgment

Since human judgments are usually trusted as the gold standards that automatic MT evaluation metrics should try to approach, the reliability and coherence of human judgments is very important. Cohen's kappa agreement coefficient is one of the most commonly used evaluation methods (Cohen, 1960). For the problem of nominal scale agreement between two judges, there are two relevant quantities p_0 and p_c. p_0 is the proportion of units in which the judges agreed and p_c is the proportion of units for which agreement is expected by chance. The coefficient k is simply the proportion of chance-expected disagreements which do not occur, or alternatively, it is the proportion of agreement after chance agreement is removed from consideration:

$$k = \frac{p_0 - p_c}{1 - p_c} \qquad (2)$$

where $p_0 - p_c$ represents the proportion of the cases in which beyond-chance agreement occurs and is the numerator of the coefficient (Landis and Koch, 1977).

A.3: Correlating Manual and Automatic Score

In this section, we introduce three correlation coefficient algorithms that have been widely used at recent WMT workshops to measure the closeness of automatic evaluation and manual judgments. The choice of correlation algorithm depends on whether scores or ranks schemes are utilized.

Pearson Correlation

Pearson's correlation coefficient (Pearson, 1900) is commonly represented by the Greek letter ρ. The correlation between random variables X and

Y denoted as ρ_{XY} is measured as follows (Montgomery and Runger, 2003).

$$\rho_{XY} = \frac{cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (3)$$

Because the standard deviations of variable X and Y are higher than 0 ($\sigma_X > 0$ and $\sigma_Y > 0$), if the covariance σ_{XY} between X and Y is positive, negative or zero, the correlation score between X and Y will correspondingly result in positive, negative or zero, respectively. Based on a sample of paired data (X,Y) as $(x_i, y_i), i = 1\,to\,n$, the Pearson correlation coefficient is calculated as:

$$\rho_{XY} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}} \quad (4)$$

where μ_x and μ_y specify the means of discrete random variable X and Y respectively.

Spearman rank Correlation

Spearman rank correlation coefficient, a simplified version of Pearson correlation coefficient, is another algorithm to measure the correlations of automatic evaluation and manual judges, e.g. in WMT metrics task (Callison-Burch et al., 2008, 2009, 2010, 2011). When there are no ties, Spearman rank correlation coefficient, which is sometimes specified as (rs) is calculated as:

$$rs_{\varphi(XY)} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \quad (5)$$

where d_i is the difference-value (D-value) between the two corresponding rank variables $(x_i - y_i)$ in $\vec{X} = \{x_1, x_2, ..., x_n\}$ and $\vec{Y} = \{y_1, y_2, ..., y_n\}$ describing the system φ.

Kendall's τ

Kendall's τ (Kendall, 1938) has been used in recent years for the correlation between automatic order and reference order (Callison-Burch et al., 2010, 2011, 2012). It is defined as:

$$\tau = \frac{\text{num concordant pairs} - \text{num discordant pairs}}{\text{total pairs}} \quad (6)$$

The latest version of Kendall's τ is introduced in (Kendall and Gibbons, 1990). Lebanon and Lafferty (2002) give an overview work for Kendall's τ showing its application in calculating how much the system orders differ from the reference order. More concretely, Lapata (2003) proposed the use of Kendall's τ, a measure of rank correlation, to estimate the distance between a system-generated and a human-generated gold-standard order.

A.4: Metrics Comparison

There are researchers who did some work about the comparisons of different types of metrics. For example, Callison-Burch et al. (2006b, 2007b); Lavie (2013) mentioned that, through some qualitative analysis on some standard data set, BLEU cannot reflect MT system performance well in many situations, i.e. higher BLEU score cannot ensure better translation outputs. There are some recently developed metrics that can perform much better than the traditional ones especially on challenging sentence-level evaluation, though they are not popular yet such as nLEPOR and SentBLEU-Moses (Graham et al., 2015; Graham and Liu, 2016). Such comparison will help MT researchers to select th appropriate metrics to use for specialist tasks.

Appendix B: MT QE

In past years, some MT evaluation methods that do not use manually created gold reference translations were proposed. These are referred to as "Quality Estimation (QE)". Some of the related works have already been introduced in previous sections. The most recent quality estimation tasks can be found at WMT12 to WMT20 (Callison-Burch et al., 2012; Bojar et al., 2013, 2014, 2015; Specia et al., 2018; Fonseca et al., 2019; Specia et al., 2020). These defined a novel evaluation metric that provides some advantages over the traditional ranking metrics. The DeltaAvg metric assumes that the reference test set has a number associated with each entry that represents its extrinsic value. Given these values, their metric does not need an explicit reference ranking, the way that Spearman ranking correlation does. The goal of the DeltaAvg metric is to measure how valuable a proposed ranking is according to the extrinsic values associated with the test entries.

$$DeltaAvg_v[n] = \frac{\sum_{k=1}^{n-1} V(S_{1,k})}{n - 1} - V(S) \quad (7)$$

For scoring, two task evaluation metrics were used that have traditionally been used for measur-

ing performance in regression tasks: Mean Absolute Error (MAE) as a primary metric, and Root of Mean Squared Error (RMSE) as a secondary metric. For a given test set S with entries $s_i, 1 \leqslant i \leqslant |S|$, $H(s_i)$ is the proposed score for entry s_i (hypothesis), and $V(s_i)$ is the reference value for entry s_i (gold-standard value).

$$\text{MAE} = \frac{\sum_{i-1}^{N}|H(s_i) - V(s_i)|}{N} \quad (8)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i-1}^{N}(H(s_i) - V(s_i))^2}{N}} \quad (9)$$

where $N = |S|$. Both these metrics are non-parametric, automatic and deterministic (and therefore consistent), and extrinsically interpretable.

Some further readings on MT QE are the comparison between MT evaluation and QE Specia et al. (2010) and the QE framework model QuEst (Specia et al., 2013); the weakly supervised approaches for quality estimation and the limitations analysis of QE Supervised Systems (Moreau and Vogel, 2013, 2014), and unsupervised QE models (Fomicheva et al., 2020); the recent shared tasks on QE (Fonseca et al., 2019; Specia et al., 2020).

In very recent years, the two shared tasks, i.e. MT quality estimation and traditional MT evaluation metrics, have tried to integrate into each other and benefit from both knowledge. For instance, in WMT2019 shared task, there were 10 reference-less evaluation metrics which were used for the QE task, "QE as a Metric", as well (Ma et al., 2019).

Appendix C: Mathematical Formulas

Some mathematical formulas that are related to aforementioned metrics:

Section 2.1.2 - Fluency / Adequacy / Comprehension:

$$\text{Comprehension} = \frac{\#\text{Cottect}}{6} \quad (10)$$

$$\text{Fluency} = \frac{\frac{\text{Judgment point}_{-1}}{\text{S}_{-1}}}{\#\text{Sentences in passage}} \quad (11)$$

$$\text{Adequacy} = \frac{\frac{\text{Judgment point}_{-1}}{\text{S}_{-1}}}{\#\text{Fragments in passage}} \quad (12)$$

Section 3.1.1 - Editing Distance:

$$\text{WER} = \frac{\text{substitution+insertion+deletion}}{\text{reference}_{\text{length}}}. \quad (13)$$

$$\text{PER} = 1 - \frac{\text{correct} - \max(0, \text{output}_{\text{length}} - \text{reference}_{\text{length}})}{\text{reference}_{\text{length}}}. \quad (14)$$

$$\text{TER} = \frac{\#\text{of edit}}{\#\text{of average reference words}} \quad (15)$$

Section 3.1.2 - Precision and Recall:

$$\text{BLEU} = \text{BP} \times \exp \sum_{n=1}^{N} \lambda_n \log \text{Precision}_n, \quad (16)$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r, \\ e^{1-\frac{r}{c}} & \text{if } c <= r. \end{cases} \quad (17)$$

where c is the total length of candidate translation, and r refers to the sum of effective reference sentence length in the corpus. Bellow is from NIST metric, then F-measure, METEOR and LEPOR:

$$\text{Info} = \log_2 \left(\frac{\#\text{occurrence of } w_1, \cdots, w_{n-1}}{\#\text{occurrence of } w_1, \cdots, w_n} \right) \quad (18)$$

$$F_\beta = (1 + \beta^2) \frac{PR}{R + \beta^2 P} \quad (19)$$

$$\text{Penalty} = LP0.5 \times \left(\frac{\#\text{chunks}}{\#\text{matched unigrams}} \right)^3 \quad (20)$$

$$\text{MEREOR} = \frac{10PR}{R + 9P} \times (1 - \text{Penalty}) \quad (21)$$

$$\text{LEPOR} = LP \times NPosPenal \times Harmonic(\alpha R, \beta P) \quad (22)$$

$$h\text{LEPOR} = Harmonic(w_{LP}LP, w_{NPosPenal}NPosPenal, w_{HPR}HPR)$$

$$n\text{LEPOR} = LP \times NPosPenal \times exp(\sum_{n=1}^{N} w_n log HPR)$$

where, in our own metric LEPOR and its variations, *n*LEPOR (*n*-gram *precision* and *recall* LEPOR) and *h*LEPOR (*harmonic* LEPOR), *P* and *R* are for precision and recall, *LP* for length penalty, *NPosPenal* for n-gram position difference penalty, and *HPR* for harmonic mean of precision and recall, respectively (Han et al., 2012, 2013b; Han, 2014; Han et al., 2014).