Mach Translat (2018) 32:217–235
https://doi.org/10.1007/s10590-018-9220-z

# Human versus automatic quality evaluation of NMT and PBSMT

**Dimitar Shterionov[1] · Riccardo Superbo[2] · Pat Nagle[2] · Laura Casanellas[2] · Tony O'Dowd[2] · Andy Way[1]**

Received: 4 September 2017 / Accepted: 9 April 2018 / Published online: 8 May 2018

**Abstract** Neural machine translation (NMT) has recently gained substantial popularity not only in academia, but also in industry. For its acceptance in industry it is important to investigate how NMT performs in comparison to the phrase-based statistical MT (PBSMT) model, that until recently was the dominant MT paradigm. In the present work, we compare the quality of the PBSMT and NMT solutions of KantanMT—a commercial platform for custom MT—that are tailored to accommodate large-scale translation production, where there is a limited amount of time to train an end-to-end system (NMT or PBSMT). In order to satisfy the time requirements of our production line, we restrict the NMT training time to 4 days; to train a PBSMT system typically requires no longer than one day with the current training pipeline of KantanMT. To train both NMT and PBSMT engines for each language pair, we strictly use the same parallel corpora and the same pre- and post-processing steps

✉ Dimitar Shterionov
dimitar.shterionov@adaptcentre.ie

Riccardo Superbo
riccardos@kantanmt.com

Pat Nagle
patn@kantanmt.com

Laura Casanellas
laurac@kantanmt.com

Tony O'Dowd
tonyod@kantanmt.com

Andy Way
andy.way@adaptcentre.ie

[1] ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland

[2] KantanMT, Dublin City University, INVENT Building, Dublin, Ireland

(when applicable). Our results show that, even with time-restricted training of 4 days, NMT quality substantially surpasses that of PBSMT. Furthermore, we challenge the reliability of automatic quality evaluation metrics based on $n$-gram comparison (in particular F-measure, BLEU and TER) for NMT quality evaluation. We support our hypothesis with both analytical and empirical evidence. We investigate how suitable these metrics are when comparing the two different paradigms.

**Keywords** Neural machine translation · NMT · Phrase-based statistical machine translation · PBSMT · SMT · Evaluation metrics · Quality evaluation · BLEU · F-measure · F-score · TER · Human evaluation · A/B testing · Ranking · Productivity · Quality comparison

## 1 Introduction

Since the development and the release of the Moses toolkit (Koehn et al. 2007) in 2007, statistical machine translation (SMT) (Koehn 2010) has become the dominant MT paradigm not only in academia but also in industry. Despite the advantages of SMT over previous approaches to MT (e.g. compared to rule-based systems, SMT improves on semantics (Costa-Jussà et al. 2012), training and translation efficiency and requires only a sufficient amount of parallel bilingual data for training), SMT has some limitations such as subject-verb agreement, word reordering, tense modeling, and others (Vanmassenhove et al. 2016).

Recent research in MT based on artificial neural networks—neural machine translation (NMT) (Cho et al. 2014; Sutskever et al. 2014; Bahdanau et al. 2015)—has shown promising results and has gained popularity not only in academia but also in industry. It promises to solve some of the drawbacks of SMT. Studies like those of Bentivogli et al. (2016), Junczys-Dowmunt et al. (2016), Wu et al. (2016), and Castilho et al. (2017) indicate that the quality of NMT can surpass that of PBSMT, and a shift in the-state-of-the-art is imminent. Although several MT vendors, such as Google,[1] Microsoft,[2] Systran,[3] and KantanMT[4] offer NMT as part of their services, it is still uncertain to what extent NMT can replace PBSMT as the core technology for large-scale translation projects. The main reasons are the computational (and financial) cost of NMT and the uncertainty regarding the actual quality; while NMT output is often very fluent, sometimes it lacks adequacy or is even completely wrong.

In this work, we compare phrase-based SMT (PBSMT) and NMT within a translation production line. We set a time limit for training NMT models of 4 days, which is sufficient for our NMT models to reach high quality without introducing delays in the production line. We use quality evaluation metrics such as F-measure (Melamed et al. 2003), BLEU (Papineni et al. 2002) and translation edit rate (TER: Snover et al.

[1] https://translate.google.com/.

[2] https://www.bing.com/translator/.

[3] https://systransoft.com.

[4] https://kantanmt.com/.

(2006)),[5] as well as human evaluation. We challenge the relevance of these metrics for scoring NMT models. Our hypothesis is that they *underestimate* the quality of NMT models. We test this hypothesis by comparing F-measure, BLEU and TER scores with human judgment and provide empirical as well as analytical evidence to support this hypothesis.

The first contribution of this study is the comparison of the quality of PBSMT and NMT in the specific setting of KantanMT's training and translation pipelines. We acknowledge that the variability of the software and hardware settings, the training parameters, the data specifics, the available time for training as well as other settings impact the performance of the MT systems. In order to make a fair comparison, we investigate MT systems as part of the KantanMT translation production line. NMT and PBSMT systems are trained and customised on the same data, using the same general[6] pre- and post-processing steps and optimised towards time and resource allocation. Our view is that such analysis allows us to answer more honestly the question as to whether NMT is suitable for deployment in a large-scale translation production environment

The second contribution is the comparison between human and automatic quality evaluation. This analysis may help to produce a more reliable quality evaluation metric in the future.

Third, we also investigate whether translators would prefer using NMT as part of the translation pipeline given the training (and retraining) time, monetary costs and quality. We thus evaluate the productivity, expressed in terms of post-editing time, of human translators that post-edit PBSMT and NMT outputs.

The remainder of the paper is structured as follows: in Sect. 2, we summarise previous comparative studies of PBSMT and NMT; in Sect. 3 we discuss *n*-gram quality evaluation metrics, focusing on BLEU, and their relevance for NMT. We present our empirical data and its analysis in Sect. 4 and conclude in Sect. 5.

## 2 Related work

Since 2015, NMT systems have been outperforming SMT for many language pairs and translation tasks. In the International Workshop on Spoken Language Translation (IWSLT) 2015 competition (Cettolo et al. 2015),[7] an NMT system outperformed a number of PBSMT systems. Bentivogli et al. (2016) compare and analyse the overall translation quality as well as the translation errors of NMT and PBSMT systems for English → German based on data from the IWSLT 2015 competition (Cettolo et al. 2015). Their results show that NMT is better than all four different SMT systems on all investigated criteria: (i) higher automatic scores (in terms of BLEU); (ii) lower

[5] F-measure, BLEU and TER are algorithms for quality evaluation of MT systems, typically used to estimate fluency, adequacy and extent of translation errors (cf. Way 2018b for more details).

[6] We consider tokenization and cleaning as general preprocessing steps; word segmentation (e.g. Byte Pair Encoding (BPE): Sennrich et al. 2016) is an NMT-specific pre-processing step.

[7] http://workshop2015.iwslt.org/.

morphological, lexical and reordering (especially, verb reordering) errors; and (iii) reduced post-editing effort.

Despite the thoroughness of their analysis and the significance of their results, Bentivogli et al. (2016) compare systems trained and tuned on different data. In particular, the English → German phrase-based SMT system they analyse (Ha et al. 2015) is built with TED talks, EPPS, news commentary,[8] and Common Crawl data; the NMT system they compare to (Luong and Manning 2015) is a pre-trained NMT system that was further improved with data provided by the IWSLT2015 organizers. Furthermore, the PBSMT system was provided with 2.4 billion tokens of monolingual data to improve the language model. In contrast, our work compares PBSMT and NMT trained on exactly the same data; we scored our systems and performed side-by-side comparison on the same test sets as well.

SMT and NMT systems have also been extensively compared by Junczys-Dowmunt et al. (2016). The authors investigate the BLEU scores of multiple NMT and SMT systems for 10 languages and 30 language directions trained on the United Nations Parallel Corpus v1.0 (Ziemski et al. 2016). Their NMT systems outrank SMT for all but three cases: French → Spanish (the BLEU score for PBSMT is 1.16 points higher than for NMT), French → English (the BLEU score for the hierarchical system Hiero (Chiang 2005) as implemented in Moses is 1.15 points higher than their initial NMT system; after additional training, the BLEU score for NMT is 1.13 points higher than Hiero) and Russian → English (the BLEU score for the hierarchical system is respectively 1.32 and 0.75 points higher than the initial NMT system and the one with additional training).[9] On an NVIDIA GTX 1080, their NMT systems were initially trained for 8 days; for the language pairs that include English, an additional training of 8 days (16 days in total) was performed.

One of the largest providers of MT services (both public and commercial)—Google—has presented their NMT (Google NMT or GNMT) approach and compared it to PBSMT as well as to human translation (Wu et al. 2016). They score their GNMT models using tokenized BLEU and compare them to PBSMT models. They also report results from a three-way side-by-side evaluation by human translators: evaluators are asked to score translations from (i) Google's production PBSMT systems, (ii) the GNMT models, and (iii) human translators fluent in both the source and target languages. The reported results, although quite disputed,[10] provide once again empirical evidence that NMT quality is generally higher than that of PBSMT. The GNMT systems follow a rather optimised implementation of the sequence-to-sequence model (Sutskever et al. 2014) with attention mechanism (Bahdanau et al. 2015) trained on 96 NVIDIA Tesla K80 GPUs. Each model was trained for approximately 6 days, and then refined for approximately 3 days (9 days in total). For training, 36 million parallel sentences for English → German and 5 million parallel sentences for English → French were used.

[8] http://www.casmacat.eu/corpus/news-commentary.html.

[9] BLEU scores are presented in the range of 0–100.

[10] In http://kv-emptypages.blogspot.ie/2016/09/the-google-neural-machine-translation.html the author argues against the generalizability of the results and the appropriateness of the evaluations performed.

Another comparison between NMT and other MT paradigms was presented in Crego et al. (2016). This work investigates the quality (scored in terms of BLEU as well as human evaluation) of NMT systems, PBSMT, rule-based MT and human translation (from both professional and non-professional translators); moreover, an error analysis is presented. Although their NMT systems outperform PBSMT and rule-based MT, they still cannot surpass the quality of translations produced by the human translators employed in the experiment.

In Castilho et al. (2017) the authors investigate the performance of PBSMT and NMT systems for three use-cases and domains: (i) MT for the e-commerce domain, (ii) MT for the patent domain, and (iii) MT for the EU-funded TraMOOC (Translation for Massive Open Online Courses) project.[11] They compare automatic metrics as well as human evaluation. Their work investigates in what scenarios NMT systems cannot outperform other MT solutions. It is interesting to note that in one of their experiments, while the automatic metrics they analysed show that two NMT systems are outperformed by the PBSMT one, the adequacy score, as judged by human evaluators, for one of the NMT systems is on a par with the adequacy for the PBSMT system. This fact indicates a gap between automatic metrics and human evaluation.

The work of Klubička et al. (2017) also performs an in-depth comparative analysis of PBSMT and NMT systems for the English-Croatian language pair. They use the same data (4,786,516 sentence pairs) to train PBSMT, factored PBSMT and NMT systems. They train language models of the PBSMT systems using additional monolingual data. Furthermore, the NMT system is trained for 10 days and an ensemble of the 4 best NMT models (judged according to BLEU) is used for decoding. The presented results show that NMT output is generally judged to be of better quality—not only according to automatic measures, but also having far fewer errors as well as being more fluent and containing more grammatical language—which corroborates other research on the topic of NMT and PBSMT comparison.

The current work extends the work presented in Shterionov et al. (2017) by (i) analysing other automatic metrics than BLEU, i.e. F-measure and TER, and (ii) presenting our findings about post-editing productivity.

## 3 Quality metrics for (N)MT

### 3.1 BLEU

BiLingual Evaluation Understudy (BLEU) (Papineni et al. 2002) is the most widely used quality evaluation metric for MT systems. The correlation between BLEU scores and human judgment of MT translations has been extensively researched. The work of Papineni et al. (2002), Agarwal and Lavie (2008), and Farrús et al. (2012), among others, show that BLEU scores do highly correlate with human judgment; others such as Callison-Burch et al. (2006), Chiang et al. (2008), and Smith et al. (2016) argue about the shortcomings of the metric and its inability to capture the actual quality dimensions of the translation output and present improvements or variations

[11] http://tramooc.eu.

to overcome these shortcomings. However, BLEU dominates other metrics mainly because it is language-independent, very quick and has proven to be the best metric for tuning PBSMT models (Cer et al. 2010).

BLEU measures the precision of an MT system computed through the comparison of the system's output and a set of ideally correct, and usually human-generated reference translations. The BLEU algorithm counts the number of matching $n$-grams (typically $n \in \{1, .., 4\}$) and computes a weighted average. That is, the more $n$-gram matches between a translation and the references, the higher the score. Lower values of $n$ capture lexical coverage of the translation; the higher values of $n$ reflect the word order. The relevant factors for computing BLEU scores are thus: (i) translation length: a correct translation matches the reference in length; (ii) translated words: the words in a correct candidate translation match the words in the reference; (iii) word order: the order of words in a correct candidate translation and in the reference is the same.

While BLEU was initially proposed as a document-level metric, later adaptations present sentence-level BLEU (Chen and Cherry 2014). BLEU scores range between 0 (or 0% – lowest quality = completely unrelated to the reference) and 1 (or 100% – highest quality = same as the reference).[12]

## 3.2 TER and F-measure

We briefly outline the mechanics of F-measure (Melamed et al. 2003) and TER (Snover et al. 2006). Similar to BLEU, we look into the factors that lead to high scores.

F-measure is the harmonic mean of the precision and the recall of a system. Similar to BLEU, it is concerned with the comparison of candidate translations to a set of reference translations at the $n$-gram level. Precision is the fraction of the number of correctly translated $n$-grams to the total number of translated $n$-grams; recall is the fraction of the number of correctly translated $n$-grams to the total number of reference $n$-grams. Typically, F-measure is computed based on unigrams, i.e. words or tokens.[13] That is, F-measure is concerned with the correctly translated words regardless of their order.

TER estimates the effort needed to edit a translation to match a reference. The TER score is computed as the minimum number of complete-word edits (such as insertions, deletions and shifts) normalized by the average length of the sentences. Factors for a high F-measure score as well as for TER are the length of the translated sentence and the correct word choice.

We present our empirical results and analysis using these metrics in Sect. 4.

## 3.3 Conformity with $n$-gram evaluation metrics

In PBSMT, phrase-level ($n$-gram) translations are arranged in a specific order that maximises the sentence-level translation likelihood. The phrase-level translations originate

[12] Few translations will attain a score of 1 unless they are identical to a reference translation. Even translations by professional translators will not necessarily obtain a BLEU score of 1.

[13] Character-based F-measure was also shown to correlate well with human judgment (Popović 2015).

from the translation model represented by the phrase-table; the language models, typically operating also at phrase-level, determine the word-order and affect the word choice in order to maximise the translation likelihood. If an $n$-gram cannot be translated, usually the original text is transferred. The phrase length $n$ for both the translation and language models is selected to optimise the translation quality/efficiency ratio and is usually 7 and 5, respectively.

NMT systems operate differently from PBSMT. A typical encoder–decoder system (Cho et al. 2014; Sutskever et al. 2014) would generate a sentence translation based on the complete sequence of tokens from the source sentence, as well as all preceding translated tokens from the current sentence. That is, NMT translations are not bound by the limits of $n$-grams and can substantially deviate from a reference according to translation length and word order. Furthermore, to tackle out-of-vocabulary issues and reduce vocabulary size, it is customary to build NMT systems on subword units (Sennrich et al. 2016) or even characters (Chung et al. 2016). This would provide the network with greater flexibility and allow it to extend beyond exact words or phrases from the training data. However, the generated words may not exist in the training data (although the subword units they are composed of do) and may even even be (linguistically) incorrect. Word embeddings[14] used to encode (semantically) similar words as vectors with smaller distance and dissimilar words as vectors with larger distance may also affect the word choice in the translation. For these reasons, although representing a correct translation, NMT output may also deviate significantly from the reference according to *word choice*. In Example 1, we illustrate how NMT generates a correct sentence that deviates significantly from the reference.

In sum, while in PBSMT $n$-gram (with $1 < n < 7$) translations are combined in a sentence that is optimised according to *translation length*, *translated words* and *word order* with respect to the training data, NMT output may deviate from any reference when it comes to the same factors. As such, PBSMT translations conform with the factors for F-measure, BLEU and TER better than NMT translations do.

*Example 1* An NMT translation with low BLEU score that is better (as judged by human evaluators) than a PBSMT one with a higher BLEU score.
*Source (EN)* All dossiers must be individually analysed by the ministry responsible for the economy and scientific policy.
*Reference (DE)* Jeder Antrag wird von den Dienststellen des zuständigen Ministers für Wirtschaft und Wissenschaftspolitik individuell geprüft.
*PBSMT* Alle Unterlagen müssen einzeln analysiert werden von den Dienststellen des zuständigen Ministers für Wirtschaft und Wissenschaftspolitik. BLEU: 55.82%
*NMT* Alle Unterlagen müssen von dem für die Volkswirtschaft und die wissenschaftliche Politik zuständigen Ministerium einzeln analysiert werden. BLEU: 3.21% □

The BLEU scoring mechanism relies on sentence length, word order and word matching between translation output and reference, as noted earlier. In Example 1, the PBSMT translation has a sentence length of 16, closer to the length of the reference (15) than the NMT translation (17). Furthermore, the PBSMT translation has more

[14] See e.g. https://deeplearning4j.org/word2vec.html.

*n*-grams in common with the reference, e.g. *von den Dienststellen des zuständigen Ministers für Wirtschaft und Wissenschaftspolitik.*; the NMT translation, in contrast, has *no* 4-gram, 3-gram nor 2-gram phrases in common with the reference; common unigrams are *von*, *für*, *und* and *zuständigen* leading to the aforementioned score of just 3.21% BLEU. Nonetheless, during the evaluation, the NMT translation of Example 1 was preferred over the PBSMT output, i.e. it was judged to be of higher quality by all three of our evaluators.

Our hypothesis is that F-measure, BLEU and TER underestimate the quality of NMT systems. That is, F-measure, BLEU and TER correlate better with human judgement of the quality of PBSMT than of NMT systems. In Sect. 4, we empirically support this hypothesis. We ought to note that we focus on *sentence-level* BLEU, F-measure and TER which has the granularity that suits our sentence-by-sentence comparison. In our examples and experiments we present BLEU scores computed with the script provided together with the Moses toolkit.[15]

Previous research has challenged the reliability of BLEU also for SMT systems. A complete discussion of the metrics' shortcomings is given in Callison-Burch et al. (2006) and Smith et al. (2016). However, to the best of our knowledge, our research is the first to question the reliability of BLEU for NMT systems (cf. also Way 2018a).

## 4 Comparing NMT to PBSMT output

### 4.1 PBSMT and NMT pipelines

For the present work, we employ KantanMT, a cloud-based MT platform which delivers MT services individually to each user. A user can create, customise and exploit their own MT engine(s)[16] within a secure environment.[17] Typically, a user creates an engine from scratch; in case their data is not sufficient to train an engine with good performance, additional data can be added, or a pre-built engine can be retrieved from KantanMT's data banks.

The training pipeline for both NMT and PBSMT engines follows the same architecture: 1. *Instance setup* hardware is allocated, software is set up: and data is downloaded; 2. *Data pre-processing*: data is converted to a suitable format, cleaned and partitioned for training, testing and tuning; for NMT the required dictionaries are prepared; 3. *Building of models*: for PBSMT, translation, language and recasing models are built; for NMT an encoder–decoder model is built; 4. *Engine post-processing*: the engine is evaluated, optimised and stored for future use. Figure 1 illustrates these steps.

To train PBSMT models, the KantanMT pipeline uses Moses 2.1 with default settings and a lexicalised reordering model with a distortion limit of 6 words. We use monolingual data extracted from the target side of the parallel corpus to build a

[15] https://github.com/moses-smt/mosesdecoder/scripts/generic/multi-bleu.perl.

[16] An MT engine refers to the package of models (translation, language and recasing models for PBSMT, and an encoder–decoder model for NMT) as well as to the required rules and dictionaries for pre- and post-processing.

[17] KantanMT provides both cloud-based and on-premise solutions.

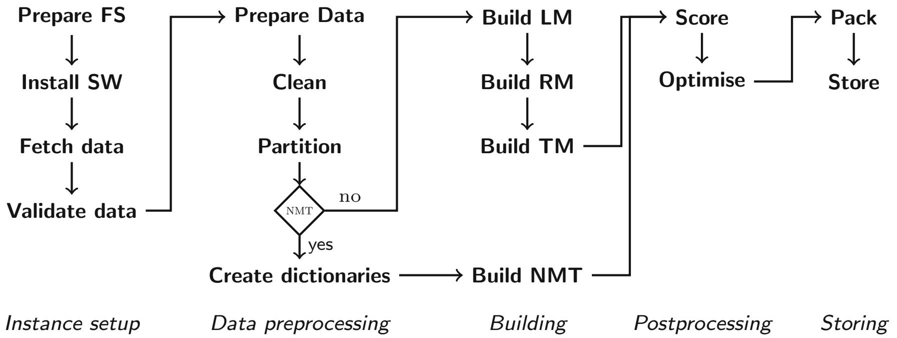

**Fig. 1** KantanMT training pipeline for PBSMT and NMT engines. *FS* file system, *SW* software, *NMT* neural machine translation, *LM* language model, *RM* recasing model, *TM* translation model

5-gram language model.[18] For word alignment, we use fast_align (Dyer et al. 2013), following experimentation described in Shterionov et al. (2016). Tuning is performed using MERT (Och and Ney 2003) with a maximum of 25 iterations.

For NMT, we employ the OpenNMT toolkit (Klein et al. 2017). A single NMT model is trained on one NVIDIA G520 GPU with 4GB RAM. As a learning optimiser, we use Adam (Kingma and Ba 2014) with a learning ratio of 0.0005. Within the scope of this study, we impose the following training limits: minimum number of training epochs is 3; maximum training time is 4 days; we consider a model to be fit for evaluation if its perplexity is below 3 at the end of training. One exception, English → German, has a perplexity of 3.02 at the end of the fourth day; we ought to note also that the English → Chinese engine achieved perplexity of 2 on the first day after only 2 training epochs; due to the low validation perplexity score the training was terminated at that point.

For data in Chinese, Japanese, Korean or Thai, KantanMT's pipeline uses dictionaries based on character-by-character segmentation (Chung et al. 2016). For other languages, we use dictionaries built from word-subunits. These subunits are generated from the training data using byte-pair encoding (BPE) with 40,000 operations. We prepare the dictionaries from true-cased (i.e. lower- and upper-case) tokenised data, which revokes the requirement for a recasing model.

The setup mentioned above was determined to optimise the trade-off between quality of NMT engines and training time via a number of experiments during the implementation and initial evaluation of KantanMT's NMT pipeline. In particular, we observed that engines that reach a validation perplexity of (approximately) 3 within the first 3 epochs would improve insignificantly if left to train for longer.[19]

[18] Typically additional monolingual data is employed to train the language model of a PBSMT engine. While we acknowledge that monolingual data, among other optimisations, may improve a PBSMT engine, within the scope of this work we keep the same-data assumption. That is, we do not employ any other data except for the parallel corpus provided. This requirement is vital when trying to answer the question: "When a user has at their disposal a set of parallel data, which of the two paradigms is preferable to train: PBSMT or NMT?".

[19] Expressed in terms of increases in BLEU and F-measure and decreases in TER, as well as according to internal human evaluation.

**Table 1** Details about the data used for experiments

| Lang. pair | Sent. count | Word count | Dict. size | Domain |
|---|---|---|---|---|
| EN–DE | 8,820,562 | 110,150,238 | 859,167 | Legal/medical |
| EN–ES | 3,681,332 | 44,917,583 | 752,089 | Legal |
| EN–IT | 2,756,185 | 35,295,535 | 765,930 | Medical |
| EN–JA | 8,545,366 | 87,252,129 | 676,244 | Legal/technical |
| EN–ZH | 6,522,064 | 84,426,931 | 956,864 | Legal/technical |

Our decision to set a training limit of four days is also guided by economic and practical reasons related to KantanMT's MT development process time restrictions. In particular, the MT development process has a (maximum) duration of six weeks and includes: (i) data preparation/cleaning, (ii) engine creation and training with data augmentation (3 iterations), (iii) translation and linguistic evaluation with engine retraining, and (iv) engine deployment. Training an engine for more than four days would disrupt the structure of this process and may impose further delays in a large-scale translation project. Furthermore, it is also economically inviable.

## 4.2 Data used

We built five NMT and five PBSMT engines for the following language pairs: English → German (EN–DE), English → Chinese (EN–ZH),[20] English → Japanese (EN–JA), English → Italian (EN–IT), and English → Spanish (EN–ES). For each language pair, both the PBSMT and the NMT engines were built using strictly the same data set. By keeping identical training, test and tuning data sets from one engine to another, we can give a more informative comparison of the PBSMT and NMT engines and their outputs. Details about the data used in our experiments are given in Table 1. The data comprises parallel translation memories in the legal, medical and technical domains, acquired from the European Commission (DGT)[21] and from Opus.[22] Prior to training, the data was cleaned and normalised, i.e. duplicates were removed. Untranslated segments and segments made of special characters were also removed. Segments of low linguistic quality may introduce noise to the engine and decrease its translation ability. Examples include segments entirely composed of product numbers, markup, file path names, and others. For such cases it is preferable to ignore them during engine training, but create a terminology table that would provide desired 'translations' if required.

We selected these language pairs and domains for the following reasons:

- EN–DE is a language pair that is hard for PBSMT to deal with. The major issue is word reordering.

[20] By Chinese, we mean Simplified Mandarin Chinese.

[21] https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory.

[22] http://opus.lingfil.uu.se/.

**Table 2** Evaluation scores (in %), training time ($T$) in hours and perplexity ($P$) (only for NMT)

| Lang. pair | PBSMT | | | | NMT | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | F-measure↑ | BLEU↑ | TER↓ | $T$ | F-measure↑ | BLEU↑ | TER↓ | $P$ | $T$ |
| EN–DE | 62.00 | 53.08 | 54.31 | 18 | 62.53 | 47.53 | 53.41 | 3.02 | 92 |
| EN–ES | 71.53 | 54.78 | 41.87 | 9 | 69.41 | 49.24 | 44.89 | 2.59 | 71 |
| EN–IT | 69.74 | 56.98 | 42.54 | 8 | 64.88 | 42.00 | 48.73 | 2.70 | 83 |
| EN–JA | 80.04 | 63.27 | 43.77 | 9 | 69.51 | 40.55 | 49.46 | 1.89 | 68 |
| EN–ZH | 77.16 | 45.36 | 46.85 | 6 | 71.85 | 39.39 | 47.01 | 2.00 | 10 |

- EN–JA and EN–ZH are two other language pairs that are hard for PBSMT, but they also require different word segmentation when training the corresponding NMT engines: BPE for the English side of the corpus and SCN (single character *n*-gram, cf. Chung et al. 2016) for the Chinese/Japanese side.
- EN–ES and EN–IT are two language pairs for which PBSMT performs typically very well.
- For all engines we have used legal, medical or technical data which typically contains specific terminology.

## 4.3 Evaluation

### *4.3.1 Quality evaluation metrics*

Table 2 shows the scores of the quality evaluation metrics we use (F-measure, BLEU and TER) for both PBSMT and NMT engines. We also show the training time in hours; for the NMT engines, each model's perplexity on the test set is also given.

### *4.3.2 Side-by-side comparison*

We set up a side-by-side (or A/B test) project with KantanLQR$^{\text{TM}}$, an online quality evaluation tool with ranking capabilities. For the test, human evaluators compared 200 segments translated using the PBSMT and NMT engines described above. The test sets did not contain any duplicates.[23] The evaluation was performed by three evaluators per language pair, all of whom were native speakers of the language they evaluated, i.e. the target language. All evaluators were translation studies students recruited from five different universities in Europe, holding certificates of English proficiency or attending courses taught in English. All evaluators of one language pair had to compare the same segments translated by the two engines (PBSMT and NMT). Evaluators had no communication with other evaluators.

The test was performed online via the interface of KantanLQR. Each evaluator was instructed as to how to access the platform and how to perform the test. Each evaluator was requested to evaluate all test sentences without taking any significant break. Three

[23] Training, testing and tuning data was normalised prior to building the MT engines.

sentences at a time were presented on the screen: *Source, PBSMT Translation, NMT Translation*. We denote these sentences as $s$, $t_{\mathrm{PBSMT}}$, $t_{\mathrm{NMT}}$, respectively. The order of the sentences $t_{\mathrm{NMT}}$ and $t_{\mathrm{PBSMT}}$ was randomised according to provenance of MT output, i.e. $t_{\mathrm{NMT}}$ could precede $t_{\mathrm{PBSMT}}$ or vice versa. This ensured that the evaluators did not get used to one style of translation and show a preference towards it. An evaluator was instructed to first read the original sentence ($s$) in English, then the two translation candidates ($t_{\mathrm{NMT}}$ or $t_{\mathrm{PBSMT}}$) and then decide which was of better quality or whether they were of equal quality (either good or bad).

According to the A/B test, each evaluator assigned each translation triplet $(s, t_{\mathrm{NMT}}, t_{\mathrm{PBSMT}})$ to one of three classes: PBSMT = NMT if the quality of the two translations was the same; PBSMT > NMT if the quality of the PBSMT translation was better than the NMT translation; PBSMT < NMT if the quality of the NMT translation was better than the PBSMT translation. For each evaluator and class, we count the translation triplets assigned to that class and compute the ratio towards the total number of entries, i.e. 200. These results together with an average over all evaluators (per language pair) are presented in Table 3 and visualised in Fig. 2.

To assess the agreement between each three evaluators we compute Fleiss' kappa coefficient (Fleiss 1971) (also presented in Table 3 and denoted by $\kappa$), which indicates to what extent evaluators agree with each other above chance. According to Landis and Koch (1977), the values of $\kappa$ should be interpreted as follows: $\kappa < 0\%$—poor agreement; $0\% < \kappa \leq 20\%$—slight agreement; $20\% < \kappa \leq 40\%$—fair agreement; $40\% < \kappa \leq 60\%$—moderate agreement; $60\% < \kappa \leq 80\%$—substantial agreement; $80\% < \kappa \leq 100\%$—almost perfect agreement.

We observe that all evaluators scored more of the translations that originate from an NMT engine better (i.e. being translations of higher linguistic quality and/or expressing more accurately the meaning of the source sentences) than their PBSMT alternatives. Averaging over all evaluators, the amount of better NMT translations are 18%, 34%, 27%, 38% and 15% more than those PBSMT translations denoted as better for the EN–DE, EN–ES, EN–IT, EN–JA and EN–ZH language pairs, respectively. This shows that NMT is always adjudged to be better under the conditions specified in Sect. 4.1. Moreover, it shows a discrepancy between evaluation metrics and human judgement of NMT quality. In Sect. 4.3.4 we present our comparative analysis of the quality evaluation metrics and human judgement of PBSMT and NMT quality.

We also note that for EN–JA and EN–ZH the inter-annotator agreement (expressed by the kappa coefficient, 62.65% and 62.68%, respectively) is substantial while for the other language pairs it is fair to moderate. That is, we can be strongly confident in the evaluation results for EN–JA and EN–ZH, and still confident in the results for the other language pairs.

It is also interesting to observe (see Table 3) that for the EN–ZH data, on average 37% of the translations are scored the same and the difference between the scores for the PBSMT > NMT and PBSMT < NMT is the smallest (average among all evaluators 15%); in general, for this language pair, the NMT engine is not evaluated as high as the others. As noted earlier, this engine was trained quite quickly as it reached a low training perplexity that allowed the training process to terminate at an early stage. While further investigation regarding whether additional training would lead to

**Table 3** Side-by-side evaluation of PBSMT and NMT output performed by human evaluators

| Lang. Pair | PBSMT = NMT | | | | PBSMT > NMT | | | | PBSMT < NMT | | | | $\kappa$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | H1 (%) | H2 (%) | H3 (%) | Avg. (%) | H1 (%) | H2 (%) | H3 (%) | Avg. (%) | H1 (%) | H2 (%) | H3 (%) | Avg. (%) | Coef. (%) |
| EN–DE | 14 | 6 | 14 | 11 | 35 | 40 | 24 | 33 | 51 | 54 | 47 | 51 | 33.14 |
| EN–ES | 12 | 10 | 7 | 10 | 28 | 26 | 31 | 28 | 60 | 64 | 63 | 62 | 36.82 |
| EN–IT | 35 | 34 | 45 | 38 | 23 | 26 | 14 | 21 | 42 | 40 | 42 | 41 | 28.99 |
| EN–JA | 21 | 28 | 14 | 21 | 19 | 15 | 28 | 21 | 60 | 57 | 58 | 59 | 62.65 |
| EN–ZH | 34 | 37 | 42 | 38 | 26 | 25 | 20 | 24 | 40 | 38 | 39 | 39 | 62.68 |

H1, H2 and H3 denote the different human evaluators

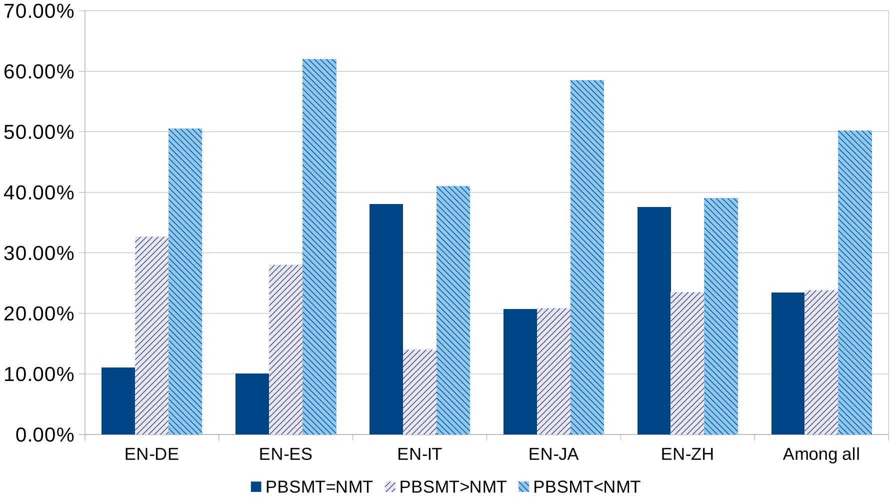

**Fig. 2** Average scores of the side-by-side human evaluation for each language pair and total average for the complete set of experimental engines

improving these scores is required, we stress the importance of training iterations or epochs for an NMT engine.

### *4.3.3 Productivity*

To further assess the comparative quality of our NMT and PBSMT engines we conducted a productivity experiment. Similar to the side-by-side evaluation, we employed three translators for each language pair with the same background as for the A/B tests. Each translator was requested to: (i) manually translate 50 sentences into English, (ii) to post-edit PBSMT output, and (iii) to post-edit NMT output. The test set for (i) was extracted from data used to train our engines so that it is in the same domain and style as the other test sets. The test set for (ii) is composed of the PBSMT translations judged by the human evaluators in our AB tests to be better than the NMT ones together with randomly selected PBSMT translations from the PBSMT $=$ NMT category (where both PBSMT and NMT quality was judged the same). Similarly, the test set for (iii) is composed of the NMT translations judged in the AB test to be better than the PBSMT ones together with the translations from the PBSMT $=$ NMT category that were not yet selected for (ii). We then measured the time to perform the requested task: ($h_{\text{trans}}$—time for translation; $h_{\text{pe}--\text{PBSMT}}$—time for post-editing PBSMT output; and $h_{\text{pe}-\text{NMT}}$—time for post-editing NMT output) and counted the number of words that were processed: either produced by translating the English sentences, or reviewed in the post-editing tasks ($w_{\text{trans}}$, $w_{\text{pe}-\text{PBSMT}}$ and $w_{\text{pe}-\text{NMT}}$, respectively). We then calculated the productivity rate as words per hour: $\frac{w_{\text{Y}}}{h_{\text{Y}}}$, for $Y \in$ {trans, pe-PBSMT, pe-NMT}. Our results are presented in Table 4.

From Table 4 we first notice that both post-editing PBSMT and post-editing NMT are more productive than human translation per se, which conforms with previous research. Second, the majority of translators are more productive when post-editing

**Table 4** Words per hour for translating (trans), post-editing PBSMT (pe-PBSMT) or post-editing NMT (pe-NMT) output performed by human translators

| Lang. pair | Trans | | | | pe-PBSMT | | | | pe-NMT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | H1 | H2 | H3 | Avg. | *H*1 | H2 | H3 | Avg. | H1 | H2 | H3 | Avg. |
| EN–DE | 522 | 622 | 807 | 650 | 1641 | 1331 | 2016 | 1663 | **1989** | **2192** | **2923** | **2368** |
| EN–ES | 410 | 741 | 766 | 576 | 1264 | 2589 | **1754** | **1869** | **1425** | **2849** | 1097 | 1790 |
| EN–IT | 559 | 493 | 432 | 495 | 956 | 1046 | 560 | 854 | **1338** | **1173** | **682** | **1064** |
| EN–JA | 304 | 166 | 261 | 243 | 538 | 203 | 569 | 437 | **644** | **235** | **812** | **564** |
| EN–ZH | 368 | 129 | 245 | 247 | **1100** | **434** | 550 | **695** | 886 | 302 | **605** | 597 |

In bold font are the highest rates for each translator and language pair. H1, H2 and H3 denote the human evaluators

NMT output, with the exception of the EN–ZH NMT engine. We already noted that the training of this engine was terminated earlier when it reached low perplexity, but possibly not the highest possible quality. However, we exploit this shortcoming to express a particular observation. Consider the entries for the EN–ZH entries in Table 3 and Fig. 2. While all three evaluators scored NMT output better than PBSMT, Table 4 shows that translators' productivity is higher when post-editing PBSMT. This indicates that NMT leads to errors that are harder to correct while the output may seem of higher quality. While the work presented in Bentivogli et al. (2016), Crego et al. (2016), Castilho et al. (2017), and Klubička et al. (2017)—as well as the papers in this volume—conducts an in-depth analysis of the translation errors, in the future we aim to further look into this phenomenon and, similar to Daems et al. (2017), analyse the correlation between errors from NMT translations and productivity. Third, the fact that the majority of translators are more productive when post-editing NMT output supports our hypothesis that automatic scores underestimate NMT quality. We address this in more detail in the next section.

### *4.3.4 Underestimation of NMT output quality by evaluation metrics*

We focus on the data from the A/B test and use it to analyse to what extent F-measure, BLEU and TER underestimate NMT quality as compared to human judgement. This analysis aids in supporting our hypothesis that current automatic evaluation metrics underestimate NMT quality.

For each language pair, we selected the set of triplets $(s, t_{\text{NMT}}, t_{\text{PBSMT}})$ for which the translation produced by the NMT engine was considered of better quality by all three evaluators. Let us denote their count as $d_{\text{PBSMT}<\text{NMT}}$. Then, from this set we counted the number of translations with F-measure and BLEU scores lower and a TER score higher than their PBSMT counterparts (i.e. the PBSMT translation is scored as better). Let us denote these numbers as $d^{\text{FM}}_{\text{PBSMT}>\text{NMT}}$, $d^{\text{BLEU}}_{\text{PBSMT}>\text{NMT}}$ and $d^{\text{TER}}_{\text{PBSMT}>\text{NMT}}$. For each of these we then computed the fraction $\frac{d^{\text{X}}_{\text{PBSMT}>\text{NMT}}}{d_{\text{PBSMT}<\text{NMT}}}$ (for $X \in \{\text{FM}, \text{BLEU}, \text{TER}\}$). We performed the same check for the PBSMT candidates that were considered of better quality by the three evaluators, i.e. we computed the fraction $\frac{d^{X}_{\text{PBSMT}<\text{NMT}}}{d_{\text{PBSMT}>\text{NMT}}}$ (for $X \in \{\text{FM}, \text{BLEU}, \text{TER}\}$). We present the underestimation ratios as percentages in Table 5.

**Table 5** Underestimation of quality evaluation scores

| Lang. pair | F-measure | | BLEU | | TER | |
|---|---|---|---|---|---|---|
| | PBSMT (%) | NMT (%) | PBSMT (%) | NMT (%) | PBSMT (%) | NMT (%) |
| EN–DE | 4 | 52 | 8 | 52 | 8 | 41 |
| EN–ES | 6 | 52 | 12 | 53 | 6 | 37 |
| EN–IT | 13 | 40 | 20 | 40 | 13 | 28 |
| EN–JA | 0 | 61 | 0 | 65 | 0 | 44 |
| EN–ZH | 20 | 38 | 20 | 35 | 20 | 28 |
| Average | 9 | 49 | 12 | 49 | 9 | 38 |

The lower the underestimation score, the more accurate (judged according to human evaluators) the evaluation metric expresses the quality of an MT engine. From Table 5, we observe that the percentage of underestimated sentences for NMT is higher than for PBSMT for all three metrics. Most significantly, this phenomenon is expressed for the BLEU and F-measure metrics, and furthermore for all language pairs (49% for each). That is, we can say that BLEU and F-measure are more suitable to predict the human judgements of quality in an AB test for PBSMT engines rather than for NMT engines. The TER metric has the lowest underestimation for NMT engines (only 38% on average). As such, we can consider TER to express the quality of an NMT engine the closest to human judgment. The closer TER underestimation scores for PBSMT and NMT arise from the fact that TER is not strictly based on the *n*-gram correlation between reference and translation sentences, but it considers the number of edits between the reference and translated sentences (shifts, insertions or deletions). As such TER scores of NMT quality correlate better with human judgement than BLEU and F-measure.

It is also interesting to highlight that EN–JA does not have any underestimated scores for PBSMT, but it is the highest underestimated language pair (according to all three metrics) in the NMT case. On average, the underestimation of F-measure, BLEU and TER for our NMT engines and our test sentences amounts to 49%, 49% and 38%. That is, we can say that on average, 49%, 49% and 38% of the NMT translations with F-measure, BLEU and TER scores worse than for their PBSMT counterparts are actually judged by the human evaluators as better.

This analysis shows that quality evaluation metrics based on *n*-grams do behave differently for PBSMT and NMT engines. It supports our hypothesis stated in Sect. 3 that F-measure, BLEU and TER underestimate NMT quality. Furthermore, it shows that these metrics correlate better with human judgement of the quality of PBSMT than of NMT systems.

## 5 Conclusions and future work

In this work, we analysed the NMT and PBSMT systems of a commercial MT platform—KantanMT. We trained four NMT and four PBSMT engines on the same data and under a time limitation that would allow for a large-scale translation development with no delays. We then compared the quality evaluation scores (F-measure, TER and BLEU) of these engines with human evaluation. While the quality evaluation scores indicate that the PBSMT engines perform better, human reviewers show the opposite results, i.e. that NMT outperforms PBSMT. The human reviewers, all native speakers of the evaluated language pairs, ranked the quality of the NMT engines higher than that of PBSMT in all cases. While these results are in agreement with previous research, we show that F-measure, BLEU and TER scores do not always conform with NMT quality, as determined by human experts. Rather, they underestimate NMT quality. We performed an extensive empirical evaluation on the three quality evaluation metrics and confirmed our hypothesis. We also conducted a productivity test, measuring the time for translating, post-editing PBSMT and post-editing NMT outputs. This experiment shows that the majority of translators are most productive when post-editing NMT output. These results further support our hypothesis that *n*-gram-based metrics correlate better with human judgement of the translation quality of PBSMT systems than of NMT ones.

In the future, we plan to perform quality ranking of other language pairs, including more challenging ones, e.g. Baltic or Indian languages. Furthermore, we intend to measure the quality of the NMT output in comparison to the quality of the PBSMT output to observe whether the difference is significant and whether it varies depending on the language pairs. Given the current differences in terms of setup and cost between PBSMT and NMT, this information is essential for MT users in a commercial environment.

**Acknowledgements** We would like to thank our external evaluators: Xiyi Fan, Ruopu Wang, Wan Nie, Ayumi Tanaka, Maki Iwamoto, Risako Hayakawa, Silvia Doehner, Daniela Naumann, Moritz Philipp, Annabella Ferola, Anna Ricciardelli, Paola Gentile, Celia Ruiz Arca, Clara Beltrá, Giulia Mattoni, as well as University College London, Dublin City University, KU Leuven, University of Strasbourg, and University of Stuttgart.

**Funding** The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

## References

Agarwal A, Lavie A (2008) METEOR, M-BLEU and M-TER: evaluation metrics for high-correlation with human rankings of machine translation output. In: Proceedings of the third workshop on statistical machine translation, Columbus, Ohio, pp 115–118

Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Proceedings of the 6th international conference on learning representations (ICLR 2015), San Diego, CA, USA

Bentivogli L, Bisazza A, Cettolo M, Federico M (2016) Neural versus phrase-based machine translation quality: a case study. In: Proceedings of the 2016 conference on empirical methods in natural language processing, Austin, Texas, pp 257–267

Callison-Burch C, Osborne M, Koehn P (2006) Re-evaluating the role of BLEU in machine translation research. In: Proceedings of the eleventh conference of the European chapter of the association for computational linguistics, Trento, Italy, pp 249–256

Castilho S, Moorkens J, Gaspari F, Calixto I, Tinsley J, Way A (2017) Is neural machine translation the new state of the art? Prague Bull Math Linguist 108(1):109–120

Cer D, Manning CD, Jurafsky D (2010) The best lexical metric for phrase-based statistical MT system optimization. In: Human language technologies: the 2010 annual conference of the North American chapter of the association for computational linguistics, Los Angeles, California, pp 555–563

Cettolo M, Niehues J, Stüker S, Bentivogli L, Cattoni R, Federico M (2015) The IWSLT 2015 evaluation campaign. In: Proceedings of the 12th international workshop on spoken language translation, Da Nang, Vietnam, pp 2–14

Chen B, Cherry C (2014) A systematic comparison of smoothing techniques for sentence-level BLEU. In: Proceedings of the ninth workshop on statistical machine translation (WMT@ACL 2014), Baltimore, Maryland, USA, pp 362–367

Chiang D (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), Ann Arbor, Michigan, pp 263–270

Chiang D, DeNeefe S, Chan YS, Ng HT (2008) Decomposability of translation metrics for improved evaluation and efficient algorithms. In: Proceedings of the conference on empirical methods in natural language processing, Honolulu, Hawaii, USA, pp 610–619

Cho K, van Merriënboer B, Gülçehre Ç, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing, Doha, Qatar, pp 1724–1734

Chung J, Cho K, Bengio Y (2016) A character-level decoder without explicit segmentation for neural machine translation. In: Proceedings of the 54th annual meeting of the association for computational linguistics, ACL 2016, vol 1, long papers, Berlin, Germany, pp 1693–1703

Costa-Jussà MR, Farrús M, Mariño JB, Fonollosa JAR (2012) Study and comparison of rule-based and statistical Catalan-Spanish machine translation systems. Comput Inform 31(2):245–270

Crego JM, Kim J, Klein G, Rebollo A, Yang K, Senellart J, Akhanov E, Brunelle P, Coquard A, Deng Y, Enoue S, Geiss C, Johanson J, Khalsa A, Khiari R, Ko B, Kobus C, Lorieux J, Martins L, Nguyen D, Priori A, Riccardi T, Segal N, Servan C, Tiquet C, Wang B, Yang J, Zhang D, Zhou J, Zoldan P (2016) Systran's pure neural machine translation systems. CoRR arXiv:1610.05540

Daems J, Vandepitte S, Hartsuiker RJ, Macken L (2017) Identifying the machine translation error types with the greatest impact on post-editing effort. Front Psychol 8:1282

Dyer C, Chahuneau V, Smith NA (2013) A simple, fast, and effective reparameterization of IBM model 2. In: Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies (NAACL), Atlanta, USA, pp 644–649

Farrús M, Costa-jussà MR, Popović M (2012) Study and correlation analysis of linguistic, perceptual, and automatic machine translation evaluations. J Assoc Inf Sci Technol 63(1):174–184

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378–382

Ha TL, Niehues J, Eunah C, Mediani M, Waibel A (2015) The KIT translation systems for IWSLT 2015. In: Proceedings of the 12th international workshop on spoken language translation, Da Nang, Vietnam, pp 62–69

Junczys-Dowmunt M, Dwojak T, Hoang H (2016) Is neural machine translation ready for deployment? A case study on 30 translation directions. In: Proceedings of the 9th international workshop on spoken language translation, Seattle, WA

Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. CoRR arXiv:1412.6980

Klein G, Kim Y, Deng Y, Senellart J, Rush AM (2017) Opennmt: open-source toolkit for neural machine translation. In: Proceedings of the 55th annual meeting of the association for computational linguistics, ACL 2017, System Demonstrations, Vancouver, Canada, pp 67–72

Klubička F, Toral A, Sánchez-Cartagena VM (2017) Fine-grained human evaluation of neural versus phrase-based machine translation. The Prague Bulletin of Mathematical Linguistics, pp 121–132

Koehn P (2010) Statistical machine translation, 1st edn. Cambridge University Press, New York, NY

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (2007) Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, Prague, Czech Republic, pp 177–180

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174

Luong MT, Manning CD (2015) Stanford neural machine translation systems for spoken language domains. In: Proceedings of the 12th international workshop on spoken language translation (IWSLT), Da Nang, Vietnam, pp 76–79

Melamed ID, Green R, Turian JP (2003) Precision and recall of machine translation. In: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, pp 61–63

Och F, Ney H (2003) A systematic comparison of various statistical alignment models. Comput Linguist 29(1):19–51

Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, Philadelphia, Pennsylvania, USA, pp 311–318

Popović M (2015) chrF: character n-gram f-score for automatic MT evaluation. In: Proceedings of the tenth workshop on statistical machine translation (WMT@EMNLP 2015), Lisbon, Portugal, pp 392–395

Sennrich R, Haddow B, Birch A (2016) Neural machine translation of rare words with subword units. In: Proceedings of the 54th annual meeting of the association for computational linguistics, ACL 2016, vol 1, Long Papers, Berlin, Germany, pp 1715–1725

Shterionov D, Du J, Palminteri MA, Casanellas L, O'Dowd T, Way A (2016) Improving KantanMT training efficiency with FastAlign. In: Proceedings of AMTA 2016, the twelfth conference of the Association for Machine Translation in the Americas, vol 2, MT Users' Track, Austin, TX, USA, pp 222–231

Shterionov D, Nagle P, Casanellas L, Superbo R, ODowd T (2017) Empirical evaluation of NMT and PBSMT quality for large-scale translation production. In: Proceedings of the user track of the 20th annual conference of the European Association for Machine Translation (EAMT), Prague, Czech Republic, pp 74–79

Smith A, Hardmeier C, Tiedemann J (2016) Climbing Mont BLEU: the strange world of reachable high-BLEU translations. In: Proceedings of the 19th annual conference of the European Association for Machine Translation, EAMT 2017, Riga, Latvia, pp 269–281

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: AMTA 2006. Proceedings of the 7th conference of the association for machine translation of the Americas. Visions for the future of machine translation, Cambridge, Massachusetts, USA, pp 223–231

Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Proceedings of advances in neural information processing systems 27: annual conference on neural information processing systems, Montreal, Quebec, Canada, pp 3104–3112

Vanmassenhove E, Du J, Way A (2016) Improving subject-verb agreement in SMT. In: Proceedings of the fifth workshop on hybrid approaches to translation, Riga, Latvia

Way A (2018a) Machine translation: where are we at today? In: Angelone E, Massey G, Ehrensberger-Dow M (eds) The Bloomsbury Companion to language industry studies. Bloomsbury, London

Way A (2018b) Quality expectations of machine translation. In: Moorkens J, Castilho S, Gaspari F, Doherty S (eds) Translation quality assessment: from principles to practice. Springer, Berlin

Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser L, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado G, Hughes M, Dean J (2016) Google's neural machine translation system: bridging the gap between human and machine translation. CoRR arXiv:1609.08144

Ziemski M, Junczys-Dowmunt M, Pouliquen B (2016) The United Nations Parallel Corpus v1.0. In: Proceedings of the tenth international conference on language resources and evaluation, Portorož, Slovenia, pp 3530–3534