Empirical evaluation of NMT and PBSMT quality for large-scale translation production.

Dimitar Shterionov[α] **Pat Nagle**[α] **Laura Casanellas**[β] **Riccardo Superbo**[β] **Tony O'Dowd**[β]
{dimitars, patn, laurac, riccardos, tonyod}@kantanmt.com
[α] **K**antanLabs, INVENT Building, Dublin City University Campus, Dublin 9, Dublin, IRELAND
[β] **K**antanMT, INVENT Building, Dublin City University Campus, Dublin 9, Dublin, IRELAND

Abstract

Neural Machine Translation (NMT) has recently gained substantial popularity not only in academia, but also in industry. In the present work, we compare the quality of Phrase-Based Statistical Machine Translation (PBSMT) and NMT solutions of a commercial platform for Custom Machine Translation (CMT) that are tailored to accommodate large-scale translation production. In a large-scale translation production line, there is a limited amount of time to train an end-to-end system (NMT or PBSMT). Our work focuses on the comparison between NMT systems trained under a time restriction of 4 days and PBSMT systems. To train both NMT and PBSMT engines for each language pair, we strictly use the same parallel corpora and show that, even if trained within this time limit, NMT quality surpasses substantially that of PBSMT.

Furthermore, we challenge the reliability of automatic quality evaluation metrics (in particular, BLEU) for NMT quality evaluation. We support our hypothesis with both analytical and empirical evidence.

1 Introduction

Recent research in MT based on Artificial Neural Networks – Neural Machine Translation (NMT) (Bahdanau et al., 2014; Cho et al., 2014; Sutskever et al., 2014) – has shown promising results and has gained popularity not only in academia but also in industry. It promises to solve some of the drawbacks that SMT comes upon. Studies like those of Bentivogli et al. (2016), Wu et al. (2016) and Junczys-Dowmunt et al. (2016) indicate that the quality of NMT surpasses that of SMT, and a shift in the state of the art is imminent. Although several MT vendors, such as Google, Microsoft, Systran, KantanMT, offer NMT as part of their services, it is still uncertain to which extent NMT can replace SMT as core technology for large-scale translation projects. The main reasons are the computational (and financial) cost of NMT and the uncertainty in 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 – sufficient for our NMT models to reach high quality without introducing overhead in the production line. We use quality evaluation metrics such as BLEU (Papineni et al., 2002), F-Measure (Melamed, 1995), and TER (Translation Error Rate) (Snover et al., 2006),[1] as well as human evaluation. We challenge the relevance of BLEU for scoring NMT models. Our hypothesis is that BLEU *underestimates* the quality of NMT models. We provide empirical as well as analytical evidence to support our hypothesis.

2 Related work

Since 2015, NMT systems have been clearly outdoing SMT. In the International Workshop on Spoken Language Translation (IWSLT) 2015 competition (Cettolo et al., 2015), an NMT system outper-

[1] BLEU, F-Measure and TER are algorithms for quality evaluation of MT systems, typically used to estimate fluency, adequacy and extent of translation errors.

formed 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 IWSTL 2015 competition (Cettolo et al., 2015). Their results show that NMT is better than all the four different SMT systems on all investigated criteria: (i) higher automatic scores (i.e., BLEU); (ii) lower morphologic, 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 – their NMT system is trained on parallel data of 120,000 tokens, whereas their standard PBSMT system is trained on parallel data of 117,000 tokens and 2.4 billion tokens of monolingual data. Our work compares PBSMT and NMT trained on exactly the same data; we scored our systems and performed side-by-side comparison (i.e., *AB tests*) 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 v 1.0 (Ziemski et al., 2016). Their NMT systems outrank SMT for all but three cases: French→Spanish (the BLEU score for PBSMT is 1.16% higher than NMT), French→English (the BLEU score for the hierarchical system Hiero as implemented in Moses is 1.15% higher than their initial NMT system; after additional training, the BLEU score for NMT is 1.13% higher than Hiero) and Russian→English (the BLEU score for the hierarchical system is respectively 1.32% and 0.75% higher than the initial NMT system and the one with additional training). 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 recently presented their NMT (Google NMT or GNMT) approach and compared it to PBSMT (employing both BLEU scoring and human evaluation) as well as to human translation (Wu et al., 2016). The results they report, although quite disputed, provide once again empirical evidence that the quality of NMT 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., 2014) trained on 96 GPUs[2]. Each model was trained for approximately 6 days, 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.

Another comparison between NMT and other MT paradigms was presented by (Crego et al., 2016). Their 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 do not reach human translation quality.

3 BLEU as a quality metric for (N)MT

The most widely used quality evaluation metric for MT systems, i.e., BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002), was one of the first metrics to report high correlation between MT quality and human judgment. 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 compares the n-grams (typically, $n \in \{1, .., 4\}$) of a candidate translation with those of the corresponding reference and counts the number of matches. The more n-gram matches between a translation and the reference, the higher the score.

BLEU scores can be computed either at a document level or at a sentence level (Chen and Cherry, 2014). They range between 0 (or 0% – lowest quality = completely irrelevant to the reference) and 1 (or 100% – highest quality = same as the reference). The relevant factors for computing BLEU scores are: (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.

In PBSMT, phrase-level (n-gram) translations are arranged in a specific order that maximises

[2]The reported GPUs are NVIDIA Tesla K80.

the sentence-level translation likelihood. If an n-gram cannot be translated, usually the original text is transferred. PBSMT translations typically conform with BLEU according to *translation length*, *translated words* and *word order*, as they are both n-gram based.

NMT systems operate differently from PSMT. A typical encoder-decoder system (Sutskever et al., 2014; Cho 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. NMT translations are not bound by the limits of n-grams. As such, NMT output may deviate from the reference according to *sentence length* and *word order* within the n-gram limit specified by the BLEU algorithm. Furthermore, to tackle out-of-vocabulary (OOV) 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. For this reason, NMT output, although representing a correct translation, may deviate significantly from the reference also according to *word choice* (see Example 3.1).

That is why, we believe that BLEU **underestimates** NMT systems. In Section 4, we empirically support our claim. We ought to note that we focus on *sentence-level* BLEU, which has the granularity that suits our sentence-by-sentence comparison.

Example 3.1 An NMT translation with 0% BLEU that is better than a PBSMT one with 58% BLEU.
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 fär Wirtschaft und Wissenschaftspolitik.* ***BLEU:*** 58%
NMT: *Alle Unterlagen müssen von dem für die Volkswirtschaft und die wissenschaftliche Politik zuständigen Ministerium einzeln analysiert werden.* ***BLEU:*** 0% △

4 Comparing NMT to SMT output

4.1 SMT and NMT pipelines

For the present work, we employ KantanMT (`https://kantanmt.com/`) – 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)[3] within a secure environment. Typically, a user creates an engine from scratch; in case their data is not sufficient to train a performant engine, additional data or a pre-built engine can be retrieved from our data banks.

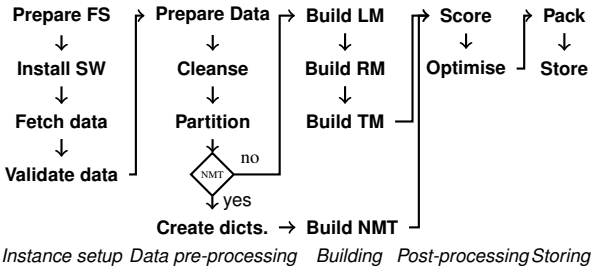

Figure 1: MT training pipeline.

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 suitable format, cleansed and partitioned for training, testing and tuning; in the case of NMT, any duplicate sentence pair that appears in the source and the target sides of the parallel corpus (i.e., the training data) is removed; moreover, the required dictionaries are prepared; 3. *Building of models* – for PBSMT, a translation, a language and a 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, our pipeline uses the Moses toolkit (Koehn et al., 2007) with default settings and lexicalised reordering model with distortion limit of 6 words. We use monolingual data extracted from the target side of the parallel corpus to build a 5-gram language model. For word alignment, we use fast_align (Dyer et al., 2013). Tuning is performed with MERT (Och and Ney, 2003) and a maximum of 25 iterations. For NMT, we employ OpenNMT (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 0.005. Within the scope of this study, we impose the following training limits: minimum number of training epochs is 3; maximum train-

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

ing time is four days; to consider a model fitted for evaluation, its validation perplexity should be below 3 at the end of the 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.

Our decision to set a limit of four days is guided by economical and practical reasons. Our MT development process has a duration of six weeks. 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 financially inviable.

For data in Chinese, Japanese, Korean or Thai, our 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 according to a byte pair encoding (BPE) (Sennrich et al., 2016) of 40,000 operations. We prepare the dictionaries from normal-cased (i.e., lower- and upper-cased) tokenised data.

4.2 Used data

We built five NMT and five PBSMT engines for the following language pairs: English→German (EN-DE), English→Chinese (EN-ZH-CN),[4] 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 train, test and tune data sets from one engine to another, we can give a more informative comparison of the SMT and NMT engines and their outputs. Details about the data used in our experiments are given in Table 1. The

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

Table 1: Details on the data used for experiments.

data comprises parallel translation memories in the Legal, Medical and Technical domains, acquired from the European Commission (DGT)[5] and from Opus.[6] Prior to training, the data was cleansed and normalised, i.e., duplicates were removed. Untranslated segments and segments constructed of special characters were also removed, as they would not be relevant to the evaluation.

4.3 Evaluation

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.

	PBSMT				NMT				
Lang. Pair	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-ZH-CN	77.16	45.36	46.85	6	71.85	39.39	47.01	2.00	10
EN-JA	80.04	63.27	43.77	9	69.51	40.55	49.46	1.89	68
EN-IT	69.74	56.98	42.54	8	64.88	42.0	48.73	2.70	83
EN-ES	71.53	54.78	41.87	9	69.41	49.24	44.89	2.59	71

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

Side-by-side comparison We set up a side-by-side, or AB Test, project with our online quality evaluation tool. For the test, human evaluators compared 200 segments translated using the aforementioned PBSMT and NMT engines. This exercise was performed by 15 evaluators – three evaluators per language pair – all of whom were native speakers of the (target) language they evaluated. 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. The test was performed online. Each evaluator was instructed on 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. The sentences were presented on the screen as a triplet *(Source, PBSMT Translation, NMT Translation)* – denoted as $(s, t_{NMT}, t_{PBSMT}$. The order of the sentences t_{NMT} and t_{PBSMT} was randomised, i.e., t_{NMT} could be preceding t_{PBSMT} or vice versa. This would ensure that the evaluators do not get used to one style of translation and show preference towards it. The evaluator was instructed to first read the original sentence (s) in English, then the two translation candidates (t_{NMT} or t_{PBSMT}) and then decide which was of better quality or whether they were of equal quality (either good or bad). The test sets did not contain any

[4]By Chinese, we mean Simplified Mandarin Chinese
[5]https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory
[6]http://opus.lingfil.uu.se/

	EN → ZH-CN			EN → JA			EN → DE			EN → IT			EN → ES		
	Same	PBSMT	NMT	Same	PBSMT	NMT	Same	PBSMT	NMT	Same	PBSMT	NMT	Same	PBSMT	NMT
Evaluator 1	41%	20%	39%	21%	19%	60%	19%	27%	54%	25%	19%	56%	12%	28%	60%
Evaluator 2	34%	26%	40%	14%	28%	58%	14%	35%	51%	29%	14%	57%	10%	26%	64%
Evaluator 3	37%	25%	38%	27%	16%	57%	6%	40%	54%	19%	25%	56%	7%	31%	62%
Average	37%	24%	39%	21%	21%	58%	13%	34%	53%	24%	19%	56%	10%	28%	62%

Table 3: Side-by-side PBSMT and NMT evaluation performed by human reviewers.

duplicates – i.e., training, testing and tuning data was normalised beforehand.

The results we gathered, summarised in Table 3, clearly contradict the scores presented in Table 2. 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. This (i) shows that NMT is better under the conditions specified in Section 4.1, and (ii) supports our claim that quality evaluation metrics are not reliable for NMT. It is, however, interesting to observe that for the EN-ZH-CN data, 37% of the translations are scored the same; in general, for this language pair, the NMT engine is not evaluated as high as the others. A closer investigation shows that this engine was trained quite quickly reaching a low perplexity that allowed the training process to terminate at an early stage. While further investigation for whether additional training will lead to improving these scores is required, we ought to stress the importance of how much time is devoted to training an NMT engine.

BLEU underestimation of NMT output quality We use the data from our AB Test to analyse to what extent BLEU underestimates NMT quality as compared to human judgement.

For each language pair, we selected the set of triplets (s, t_{NMT}, t_{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^{NMT}. Then, from this set we counted the number of translations with a BLEU score lower than their PBSMT counterparts. Let us denote this number as d^{NMT}_{PBSMT}. We then computed the fraction $\frac{d^{NMT}_{PBSMT}}{d^{NMT}}$. 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 $\frac{d^{PBSMT}_{NMT}}{d^{PBSMT}}$. We present these scores as percentages in Table 4. We observe that the percentage of underestimated sentences for NMT is significantly higher then for PBSMT. It is interesting to highlight that two of the language pairs, EN-JA and EN-ES, do not have any underestimated scores for PBSMT, but they are respectively the highest and the third highest underestimated language pairs in the NMT case. On average, the underestimation of BLEU for our NMT engines and our test sentences amounts to 48%. That is, we can say that *on average,* 48% *of the NMT translations with BLEU scores worse than for their PBSMT counterparts are judged by the human evaluators as better.* We should also mention that, for the other quality evaluation metrics (i.e., F-Measure and TER), the results are rather similar. As it extends beyond our current research (which focuses on BLEU), further analysis will be addressed in future work.

	EN-ZH-CN	EN-JP	EN-DE	EN-IT	EN-ES	**Average**
NMT	40	59	55	34	53	**48**
SMT	12	0	9	9	0	**6**

Table 4: Underestimation of BLEU scores ($\%$).

5 Conclusions and future work

In this work, we analysed the NMT and PBSMT systems of a commercial MT platform. We trained five NMT and five 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. In all cases, the human reviewers, all native speakers of the evaluated language pairs, ranked the quality of the NMT engines higher than that of PBSMT. While these results are in agreement with previous research, we show that BLEU scores do not always conform with NMT quality. Rather, they underestimate NMT quality.

In the future, we plan to perform quality ranking of other language pairs, including challenging ones, e.g., Baltic languages. Furthermore, we intend to measure the quality of the NMT output in comparison to the quality of the PBSMT output to observe if the difference is significant and if it varies depending on the language pairs. Given the current differences in terms of setup and cost be-

tween 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 Beltr, as well as University College London, Dublin City University, KU Leuven, University of Strasbourg, and University of Stuttgart.

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR, Accepted for oral presentation at the International Conference on Learning Representations (ICLR) 2015*, abs/1409.0473.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*.

Cettolo, Mauro, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The IWSLT 2015 evaluation campaign. *Proc. of IWSLT, Da Nang, Vietnam*.

Chen, Boxing and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*.

Cho, Kyunghyun, Bart van Merriënboer, Çalar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of EMNLP 2014*, Doha, Qatar, October. Association for Computational Linguistics.

Chung, Junyoung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the ACL*, Berlin, Germany, August.

Crego, Josep Maria, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, and Peter Zoldan. 2016. Systran's pure neural machine translation systems. *CoRR*, abs/1610.05540.

Dyer, Chris, Victor Chahuneau, and Noah A. Smith. 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, June.

Junczys-Dowmunt, Marcin, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? A case study on 30 translation directions. *CoRR*, abs/1610.01108.

Kingma, Diederik P. and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Klein, G., Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Koehn, Philipp, 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 ACL 2007, demonstration session*, Prague, Czech Republic, June.

Melamed, I. Dan. 1995. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Proceedings of the third Workshop on Very Large Corpora*, Cambridge, Massachusetts, USA.

Och, Franz Josef and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1).

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, July.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 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, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Ziemski, Michal, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *Proceedings of LREC 2016, Portorož, Slovenia, May 23-28, 2016.*