Poznan Studies in Contemporary Linguistics manuscript No.
(will be inserted by the editor)

Statistical versus Neural Machine Translation – a Case Study for a Medium Size Domain-Specific Bilingual Corpus

Received: date / Accepted: date

Abstract Neural Machine Translation (NMT) has recently achieved promising results for a number of translation pairs. Although the method requires larger volumes of data and more computational power than Statistical Machine Translation (SMT), it is believed to become dominant in near future. In this paper we evaluate SMT and NMT models learned on a domain-specific English-Polish corpus of a moderate size (1,200,000 segments). The experiment shows that both solutions significantly outperform a general-domain on-line translator. The SMT model achieves a slightly better BLEU score than the NMT model. On the other hand, the process of decoding is noticeably faster in NMT. Human evaluation carried out on a sizeable sample of translations (2,000 pairs) reveals the superiority of the NMT approach, particularly in the aspect of output fluency.

Keywords Machine Translation · Neural Machine Translation · Phrased-Based Statistical Machine Translation

1 Introduction

1.1 Machine Translation for the Polish language

Despite the growing interest in Machine Translation (MT) at the beginning of the 21st century, the Polish language was sporadically involved in international research in the field. This might be explained by the sparsity of bilingual corpora as well as some linguistic features of Polish, such as free word order and rich morphology, which are difficult to handle by Phrase-Based Statistical Machine Translation (PB-SMT). The MT researchers from Poland focused on other language pairs, e.g. Russian-English or German-English. The situation has changed in the last couple of years. New bilingual corpora, such

E-mail:

as *ParaCrawl*[1] and *OpenSubtitles2018*[2][24], have become available. Based on open-access corpora several translation engines with the Polish language have been trained. The solutions developed at the Polish-Japanese Academy of Information Technology participated in the International Workshop for Spoken Language Translation, in 2013, 2014 and 2015 [32] as well as in Workshop for Machine Translation in 2016 and 2017 [33], [34]. The translation direction of the main interest is currently English, however, more exotic pairs such as Japanese-Polish are under study [27].

In the paper we compare the state-of-art solutions in two leading translation approaches (NMT and SMT) applied to the translation from English to Polish, with the training corpus consisting of over a million segments from a specific domain.

The paper is organised as follows. Section 2 contains the description of SMT and NMT models. Section 3 elaborates on the comparison of SMT and NMT models in translation from English to Polish. Section 4 compares our findings from Section 3 with those obtained in related papers. Section 5 describes the manual evaluation of the translation quality – related to similar experiments.

2 Modern Machine Translation

Since 1990s the research on Machine Translation has moved from rule-driven methodology to data-driven one. The latter approach requires large-volume bilingual corpora to train the translation model. In this section we briefly present two types of translation models: computed statistically from pairs of phrases (applied in PB-SMT) and computed by artificial neural networks (applied in NMT)[3]. The following sections aims to provide elementary introduction to SMT, NMT and give an insight into a hybrid combination of the two approaches. More detailed descriptions of those models are presented in [19] and [20].

2.1 Phrase-based Statistical Machine Translation

Let us assume that a sentence f in language F is supposed to be translated into a sentence e in the language E. The fundamental equation of SMT formulates the task in terms of probability:

$$P(e|f) \propto P(f|e) \cdot P(e) \tag{1}$$

The equation may be interpreted as follows: The probability of a translation e of the sentence f is proportional to two factors: the accuracy of translation

[1] `https://paracrawl.eu/`

[2] `http://opus.nlpl.eu/OpenSubtitles2018.php`

[3] NMT also uses statistical methods while building the translation model - it is distinct from SMT by its specific architecture.

$P(f|e)$[4] and the fluency of the output $P(e)$[5]. The former factor is learned statistically from bilingual corpora aligned on the level of word phrases, and the latter – from monolingual texts. Since the actual probabilities are difficult to estimate, they are modelled by so-called *feature functions*. Each feature function focuses on one aspect of translation, such as word order or translation probability of a phrase.

During translation, a source sentence f is split into a number of phrases. Starting with an empty translation, the PB-SMT model chooses source phrases that should be translated first. Each phrase is scored by all feature functions and the best selections are stored in memory. Then, each partial translation is extended by another phrase (using feature functions from the SMT model). The translation process ends when the whole source sentence is covered.

The main drawback of the SMT model is the size of its implementation, which often exceeds tens of gigabytes. The phrase table used in SMT grows proportionally to the amount of training data. The size of the phrase table has a significant effect on its loading time and consequently, on translation efficiency.

The Moses toolkit [21] is an open-source project, which systematically implements up-to-date solutions in the field of PB-SMT. An important contributions to the project was the implementation of *the compact phrase table* [12], which decreased the size of a standard phrase table up to 30 times. The compactness was achieved by applying the Huffman coding and "word ranking"[6].

2.2 Neural Machine Translation

In 2014, two independent groups: from the Google company [30] and the University of Montreal [2] proposed translation systems based on Recurrent Neural Networks (RNNs). To avoid confusion with SMT, the new approach was named Neural Machine Translation (NMT).

Figure 1 presents the scheme of the NMT model.

The first part of the model, called *encoder*, converts the source sentence (e.g. in Polish) to its numerical representation. Firstly, each source word is replaced with its numerical representation (yellow blocks)[7]. Next, an RNN, such

[4] Accuracy is expressed as a value between 0 and 1 – the better the accuracy, the higher the score is.

[5] Fluency is expressed as a value between 0 and 1 – the more fluent the output, the higher the score is.

[6] A phrase table consists of sub-phrases of the input sentence and their potential translations in the target language. The idea of word ranking consists in replacing each target word in the phrase table by a pair: the pointer to the source word it corresponds to and the translation probability rank – the most likely equivalent word is ranked as '1', etc.

[7] A numerical representation of a word is a unique vector of real numbers, calculated by a neural network for each word in a training corpus, which happens to capture the word meaning (close vectors represent similar words). The concept was introduced in [25].

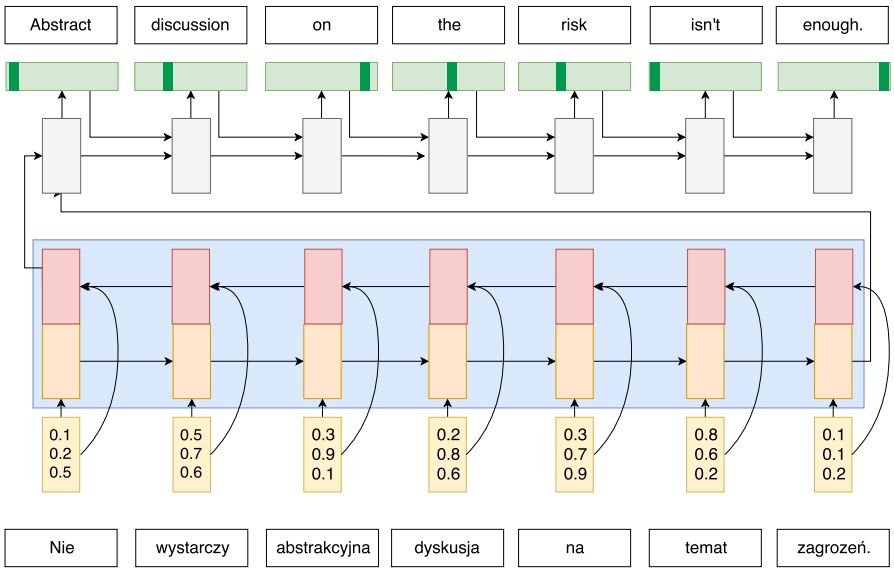

Fig. 1 Scheme of the NMT model. The sentences are taken from the Europarl corpus [18].

as LSTM[8] [11] or GRU[9] [5], builds the sentence representations in both directions: forward (orange) and backward (red)[10]. Finally, both representations are concatenated and make the sentence representation (grey).

The second part of the model, called *decoder*, generates the translation (e.g. in English). At each step, the network estimates the probabilities of a potential next word in the target language (green blocks) and chooses words with highest values. The beam search technique is commonly used to keep several hypotheses (a few best words), or "beams", in memory. Decoding with beam search takes significantly longer than a simple selection of one most probable word at a time. The procedure is repeated until the end of the sentence is reached.

More recently, new model architectures were presented, where RNNs are replaced with other networks. In [8] a researcher from the Facebook company proposed a model, based on convolutional neural networks instead of RNNs. In the same time, the Google research centre published a paper on the attention-based model [31], which selectively focuses on parts of the source sentence during translation.

[8] Long Short-Term Memory (LSTM) was developed to address the problem of vanishing (approaching zero) values that appeared in standard RNNs.

[9] Gated Recurrent Unit (GRU) was invented in order to simplify the LSTM.

[10] The forward representation is formed while traversing the sentence from left to right. The backward representation is formed in the other direction.

The NMT encoders are usually trained using GPU (Graphical Processing Unit), which is much faster than the standard procedure using CPU (Central Processing Unit).

One of the issues to solve when training an NMT model is the handling of *unknown words*, i.e. tokens that do not appear in the training data, such as proper names or numbers. A standard NMT model is capable of finding translations only to words included in the training set. However, the vocabulary size has a huge effect on training time. The practice has shown that the number of word types should be set between 50,000 and 100,000. These numbers are far from sufficient for high-morphological languages, such as Polish. One of the ideas to overcome this problem is to design an NMT model on the level of characters. Such a solution was proposed in [23], where the translation model generates output character by character. Another approach, applied also in the experiment under description, was introduced in [29]. There, words are split into sub-word units by means of a so-called Byte Pair Encoding algorithm (which divides words into the most frequent character sequences)[11]. The goal of the improvement is to decrease the complexity of the calculations in the NMT training by limiting the size of the lexicon. The algorithm boosts the model in a few other ways. First, when an unknown word occurs in a source sentence, it is split into units, which are included in the vocabulary (in the worst case, the unknown word is split into characters). Second, it makes the model capable of producing words in the target sentence that were absent in training data. Finally, the enhancement balances vocabulary sizes between source and target languages.

One of the open-source implementations of NMT is the amuNMT project [13], which is deployable both on CPU and GPU. The decoder supports the most widely used NMT model architectures [10], including the recent attention-based model. The project takes full advantage of the GPU architecture, which makes it possible to perform parallel mathematical operations applied in matrix multiplication.

2.3 Hybrid MT

Hybrid Machine Translation refers to the paradigm, which combines two or more translation methods in one model. The solution, described in [14] is based on the SMT model, with the NMT "plugin" serving as a mechanism that scores the translation candidates (traditionally, the scoring is executed by the language model). In order to speed up the implementation (slowed down by CPU-GPU communication) a so-called Stack Rescoring is applied (see Figure 2).

First, all partial hypotheses in a stack are scored by all feature functions excluding the NMT feature function (on the left). Then, the stack is pruned

[11] A sub-word may or may not correspond to a morpheme, e.g. the word *loved* may be divided into *lov* and *ed*, whereas *code* may be divided into *cod* and *e*.

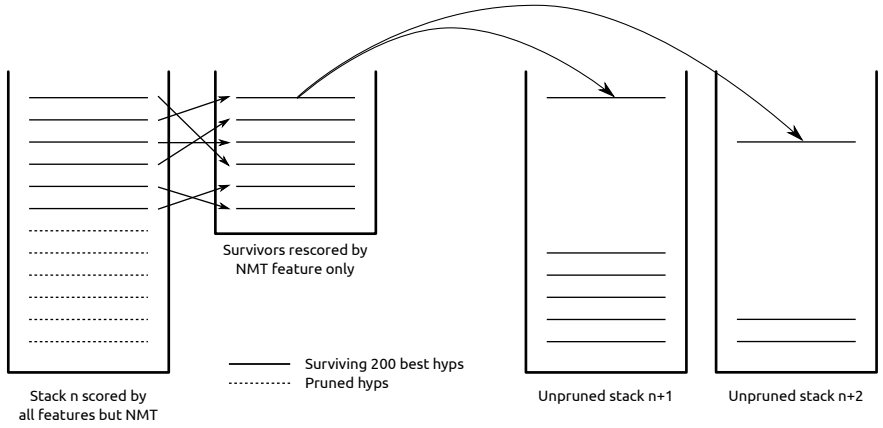

Fig. 2 Stack Rescoring scheme.

and only the best hypotheses are scored by the NMT feature function. The hypotheses that are scored highest by the NMT function are passed to stacks that store translations of extended fragments of text (e.g. stack *n+1* corresponds to a fragment that extends the original one by one next word, etc.).

3 Case Study

Here, we describe an experiment that involves English-to-Polish translation limited to a certain domain. Whenever we refer to domain-specific translation, we assume that the domain of translated documents coincides with the domain of the training corpus (which is not always the case, as shown in the English-to-Croatian experiment reported in section 5.1). In general, the scope of a domain may vary. In some case studies the domain is understood as a wide area of social life (e.g. communication, logistics, law), whereas other implementations are limited to specific documents processed in a given organisation. Assuming a fixed corpus size, the narrower the scope of the domain, the higher quality of the translation may be achieved.

3.1 Corpus

The experiment under description was carried out on a bilingual English–Polish corpus delivered by a Polish shipyard that needs to translate texts characteristic of its narrow field of operation. The corpus field is technical, related to shipbuilding, and contains full of specialist terminology from engineering. Because of the legal restrictions, the corpus has to remain unpublished. Here, we describe how the corpus was created.

In 2012, a Polish shipyard deployed a system for the translation of English orders addressed to the company, into Polish. At the time the company had at

their disposal ca 250 000 bilingual segments resulting from human translation of previous orders. A PB-SMT engine was trained on the corpus, achieving the 29-point BLEU [26] score[12]. The engine was incorporated into a dedicated CAT (Computer Assisted Translation) system that allowed for human post-edition. All sentences from the corpus were stored in the translation memory, which, for any new sentence under translation, was searched for a match before launching the MT engine. The translation memory grew fast, as all sentences, translated automatically, and post-edited by humans were saved into it. In 2018, the memory consisted of over 1 200 000 segments and served as the training corpus for the experiment under description.

Table 1 shows information on the statistics of the corpus. From the corpus, we extracted 2,000 segments for tuning (validation set) and 2,000 segments for testing.

The pre-processing, which takes place before training, was identical for the SMT and NMT models. We used the standard Moses scripts for cleaning, tokenization and truecasing[13] [21]. To avoid the problem of rare words, we applied Byte Pair Encoding [29] for splitting words into sub-word units.

Table 1 The corpus statistics.

Corpus set	segments	Polish tokens (unique)	English tokens (unique)
training set	1,267,862	14,023,717 (346,917)	14,398,261 (280,35)
validation set	2,000	22,181 (7,213)	23,084 (5,668)
test set	2,000	22,429 (7,301)	23,208 (5,739)

3.2 Phrase-Based SMT Model

First, we trained a phrase-based SMT model. We used Fast Align [7] for word alignment[14] and kenLM [9] for building the language model[15]. We applied the Operation Sequence Model (OSM) [6] for the generation of the translation model[16]. We used the validation set to tune the translation parameters set and applied the k-best Mira algorithm [4] for fine-tuning[17]. We used the Moses toolkit [21] for translation.

[12] The BLEU metric estimates, in percentage points, the agreement between an automatically translated text and the reference translation, regarded as perfect.

[13] In Machine Translation, truecasing consists in determining the proper capitalization of words in the target language.

[14] Word alignment consists in automatic matching of words that correspond to each other in the translation pairs.

[15] Language model is a function that assigns probabilities ($P(e)$ in the fundamental SMT equation) to word sequences.

[16] Translation model is a function that assigns probabilities ($P(f|e)$ in the fundamental SMT equation) to potential translations.

[17] Fine-tuning consists in determining optimal weights in linear feature functions.

3.3 NMT Model

The NMT model was trained with the Marian NMT toolkit [15][18], which is the continuation of the amuNMT project, mentioned in Section 2.2 . Marian supports simultaneous training on a number of GPUs (multi-GPU mode) and a number of machines (multi-node mode). We followed the model settings from [28]. We used an encoder–decoder architecture with attention [31]. The model parameters were optimised with the Adam Algorithm[19] [16]. The vocabulary size was set to 50,000, both on source and target sides. We trained the model on a single GPU card NVIDIA GTX 1070.

3.4 Automatic Evaluation

We used the BLEU metric for automatic evaluation of our experiment. The scores are presented in Table 2. We notice that the SMT model performs slightly better than the NMT model. This observation stands in contradiction with the results reported in other recent experiments. We suppose that the reasons are: the very strict, technical style of the training texts, the narrowness of the domain and the size of the corpus – smaller than that used in most experiments with NMT. It is worth mentioning that both solutions significantly outperform the general-purpose on-line Google translator.

Table 2 BLEU scores of compared systems (English to Polish)

Model	validation set	test set
SMT	55.92	55.23
NMT	53.05	51.66
Google	22.63	21.37

In Table 3 we compare the translation speed of the test-set on a medium-class hardware[20]. The translation with the SMT model was executed on 8 threads. We ran the NMT model in two modes: in the single mode one segment was translated at a time. In a batch mode 50 segments were translated simultaneously. The batch-mode NMT proved almost three times faster than the single-mode NMT, which, on the other hand, was more than twice as fast as SMT.

[18] `https://marian-nmt.github.io/`

[19] Adam is a learning algorithm applied in artificial neural networks, which has proved particularly efficient in natural language processing.

[20] (i-7 Intel processor, 16 GB RAM, GPU: NVIDIA® GeForce® GTX 1050 with 4GB GDDR)

Table 3 Comparison of translation speed for the test-set

Model	Duration of translation (seconds)
SMT	296
NMT-1	111
NMT-50	46

3.5 Assessment on Public Data

The privacy of the technical corpus makes the experiment impossible to replicate. To evaluate the translation quality of the environment used in the experiment (described in sections 3.2, 3.3), we performed an experiment on publicly available data from a legal domain. We trained an NMT model with the identical settings. We used English-Polish corpora from the *OPUS* website[21], from which we selected those related to the legal domain. The sizes of the training corpora are included in Table 5. The 4,000-sentence test set was selected at random from a public legal document not used in training[22].

Table 4 shows the BLEU score of the system. Let us notice, that although the training corpus is of a much larger size (ca 6 times), the BLUE score is twice lower than that reported for the shipyard system.

The experiment confirms our findings that the translation quality highly depends on the correctness and scope of the training data. The public data that we used had been crawled automatically from the Internet and contained a lot of errors, whereas the enterprise-specific data had been manually edited before training. The thematic scope of the public data (legal texts) was defined much wider than that of the shipyard technical documentation.

Model	test-set
NMT	26.85

Table 4 THE BLEU scores on the legal domain.

Corpora	Number of sentences
DGT	3,109,144
EUbookshop	539,941
Europarl	629,558
JRC-Acquis	1,610,962
ParaCrawl	1,275,162
Total	**7,164,767**

Table 5 The sizes of the public corpora included in the training set.

[21] `http://opus.nlpl.eu/`

[22] `https://www.ifac.org/publications-resources/2018-handbook`

4 Related Experiments: Comparison of NMT to SMT

Since the introduction of the NMT approach, a number of experiments have been carried out in order to compare NMT to SMT. In [13] it was claimed that both methodologies achieved comparable translation quality. The NMT model, however, worked distinctly faster (provided that computations were performed on GPU).

More recently, Koehn [22] revealed main drawbacks of the NMT approach that relate to the length of translated sentence and the size of the training batch. We have generated some statistics to verify Koehn's findings in our experiment.

First, we compare the translation quality (in terms of BLEU) in relation to the length of source sentence. Figure 3 presents BLEU scores of the NMT and SMT models as a function of the length of the source sentence, calculated for the test set (see Section 3.1). The NMT approach seems to better SMT for short sentences (up to 40 words), but is totally outperformed in translation of longer sentences. Actually, the translation quality deteriorates with sentence length for both models but the decrease for the NMT model is far more intensive.

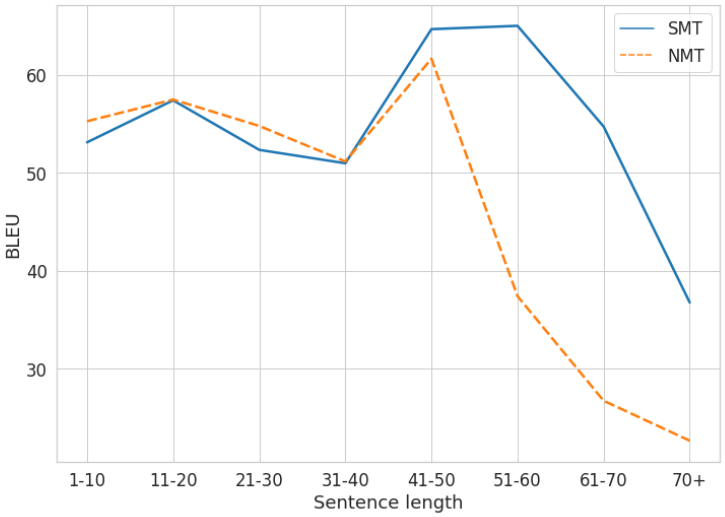

Fig. 3 NMT quality related to sentence length.

Second, we measure the NMT translation quality in relation to the decoding beam size. Contrary to the intuition that the translation quality should boost with a wider space, the Koehn's findings, as well as our experiment demonstrate the opposite. Figure 4 shows the BLEU scores obtained on the

test set with different sizes of the beam during decoding. The translation quality improves with the beam size until it reaches the value of 12. After that, the BLEU scores slowly decline.

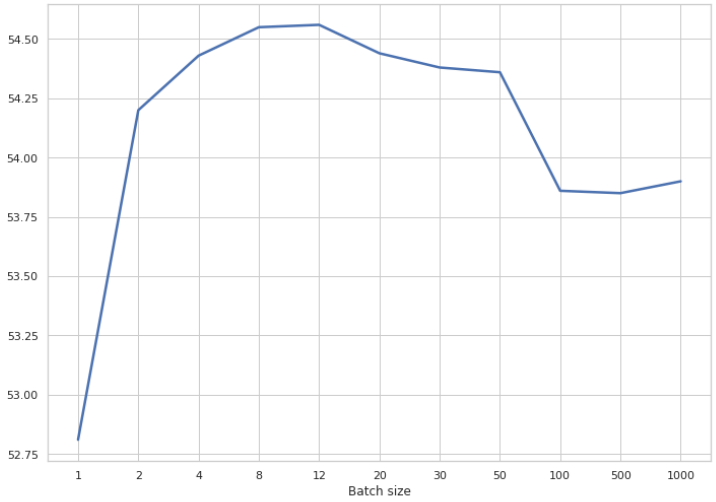

Fig. 4 NMT quality related to beam size. The highest BLEU score is achieved when the beam size equals 16 and rapidly decreases with beam size larger than 50.

5 Human Evaluation

Since the results of automatic comparison had shown similar performance quality of both systems, we made the decision to compare the performance of both solutions manually.

5.1 Similar Experiments

Tilde (https://www.tilde.com/) is a company that applies novel technologies to provide customised translation between English and Eastern-European languages, such as Russian, Polish, Latvian and Estonian. They carried out an experiment to compare the translation quality of SMT and NMT for language pairs: English-Estonian and English-Latvian[23]. Professional translators (5 to 8 depending on the language pair) were enlisted to perform a comparative evaluation and error analysis of MT outputs. The results of the human evaluation

[23] https://www.tilde.com/about/news/316

are contrasted with those calculated by the BLEU metric. Table 6 presents approximate BLEU scores (extracted from a graphical representation) for the English-Latvian translation.

Table 6 BLEU approximate scores for English-Latvian SMT and NMT translations in the Tilde experiment

Direction	SMT score	NMT score
Latvian-English	43	30
English-Latvian	38	25

The evaluation proves a significant advantage of the SMT method in the experiment.

Table 7 presents the results of human comparison of the two systems. For each sentence the translators were expected to blindly (not being aware of the system that generated the translation) point out a better translation, or chose a tie.

Table 7 Human comparison of English-Latvian SMT and NMT translations in the Tilde experiment

Direction	SMT wins (%)	NMT wins (%)	Ties (%)
Latvian-English	39	37	24
English-Latvian	29	45	26

The results from the two experiments show that human evaluators rate the NMT output much higher than it is indicated by an automatic metric. The conclusion stated by the experimenters is the following: "Human comparative evaluation is crucial when comparing MT systems from fundamentally different approaches".

In [17] the authors conduct a more detailed human analysis of the outputs produced by NMT and SMT systems. They annotate errors found in the translations by their types, according to a pre-defined taxonomy. The evaluation is carried out on English-to-Croatian translations from the news domain. The training corpus was not restricted to any domain – it consisted of close to 5 million sentence pairs from publicly available corpora, such as JRC Acquis, OpenSubtitles or TED talks. Three systems trained on the same data were compared. Two of them apply the SMT approach: Phrased-Based Statistical Machine Translation (PB-SMT) and Factored Phrased-Based Machine Translation[24] (Factored PB-SMT), and the other – the NMT approach. Similar to the Tilde experiment, the authors start from comparing the outcomes by means of the automatic BLEU metric. The results are shown in Table 8.

It is worth noting that here, the NMT approach outscored SMT methods in the automatic metric.

[24] Factored PB-SMT, contrary to "pure" PB-SMT, takes into consideration some linguistic characteristics, such as morphological features.

Table 8 BLEU scores for English-Croatian SMT and NMT translations in the Croatian experiment

System	BLEU
PB-SMT	25.44
Factored PB-SMT	27.00
NMT	30.85

To perform human evaluation, the authors adjust a Multidimensional Quality Metrics (MQM). "MQM provides a framework for describing and defining quality metrics used to assess the quality of translated texts and to identify specific issues in those texts"[25]. MQM defines a taxonomy of translation issues, where the most important nodes are named Accuracy and Fluency. Accuracy is the parent of:

- Addition – "the target text includes text not present in the source",
- Omission – "content is missing from the translation that is present in the source",
- Mistranslation – "the target content does not accurately represent the source content",
- Untranslated – "content that should have been translated has been left untranslated".

Fluency's children are:

- Grammar – "issues related to the grammar or syntax of the text, other than spelling and orthography",
- Grammatical Register – "the content uses the wrong grammatical register, such as using informal pronouns or verb forms when their formal counterparts are required", e.g. using the form "Ty" (Eng. "You") in Polish when the form "Pan" (Eng. "You, Sir") is required,
- Inconsistency – "the text shows internal inconsistency". e.g. inconsistent format of references,
- Spelling – "issues related to spelling of words", e.g. *mięki* instead of *miękki* (Eng. soft),
- Typography – "issues related to the mechanical presentation of text. This category should be used for any typographical errors other than spelling", e.g. *miekki* (missing diacritic in the 'ę' character) instead of *miękki*,
- Unintelligible – "the exact nature of the error cannot be determined. Indicates a major break-down in fluency".

In the experiment reported in [17] the annotators scored 100 sentences randomly selected from 1,000 first sentences of the test set in WMT13[26] translation task (http://www.statmt.org/wmt13/translation-task.html). The

[25] The definition of the MQM taxonomy as well as the description of its nodes originates from http://www.qt21.eu/mqm-definition/definition-2015-12-30.html.

[26] Workshop for Machine Translation is an annual conference that organises contests for MT solutions, which are evaluated on a pre-defined domain-specific set of sentences.

annotators were presented with the source text, a reference translation and the unannotated system outputs at the same time. Altogether 600 sentences were annotated. The task consisted in finding all errors made by the three systems and classifying each of the errors into an appropriate type in the MQM taxonomy.

Table 9 shows the total numbers of errors found by the annotators.

Table 9 Number of errors made by MT system in English-to-Croatian translation

	Annotator 1			Annotator 2		
System	PB-SMT	Factored	NMT	PB-SMT	Factored	NMT
Total errors	317	276	178	264	199	132

The table shows that NMT results in a 42% overall error reduction compared to the factored system, and 54% with respect to pure PB-SMT. The paper gives the statistics on types of errors found by the annotators. Table 10 quotes the aggregated numbers of Accuracy and Fluency errors.

Table 10 Statistics on error types made by MT system in English-to-Croatian translation

	PB-SMT	Factored PB-SMT	NMT
Accuracy	369	291	266
Fluency	641	518	188

The conclusion from the experiment is the following: The NMT approach produces significantly better quality of translation and the main reason for this is fluency rather than accuracy.

5.2 Human Comparison of SMT and NMT for Domain-Specific English-Polish Translation

The evaluation set in our experiment consisted of 2,000 pairs (SMT; NMT) of translated sentences. The order of the translations inside each pair was random and not known to two annotators. Our evaluation followed the Tilde experiment in the method of scoring – the annotators were expected to assign to each pair: either a tie or a win of one of the translations.

Similar to the Croatian experiment, we wanted the annotators to point out the reason for each win (if a win was assigned). Since the annotators were not familiar with the MQM taxonomy, they were asked to tag the reason as one of: Vocabulary, Grammar or Other. The Vocabulary criterion was defined by the question: "Which translation renders the original meaning more accurately?" The win by Vocabulary may be approximately thought of as better accuracy. The Grammar criterion was stated as follows: "Which translation is more correct as far as grammar and language style is concerned?" The win by Grammar corresponds approximately to better fluency.

In the below tables, which show examples of human evaluation, we also deliver the sentence-level BLEU score [3] [27] for each translation.

The annotators were not presented with the reference translation – we did not want them to be influenced by the translations provided by other humans. In some cases this assumption led to different annotations, as shown in Table 11.

Table 11 Different ratings of vocabulary used in translation

Source	Supports for Vapour Return	
Reference	Wsporniki rurociągu powrotu par **ładunku**	
SMT	Wsporniki dla rurociągu zwrotnego pary	
NMT	Wsporniki dla Rurociągu Powrotnego Oparów **Ładunku**	
BLEU	SMT: 30.21	NMT: 26.27
Rating	1: SMT (Vocabulary)	2: NMT (Vocabulary)

Annotator 1 preferred the SMT output probably because in the source sentence there was no equivalent for the word "Ładunku" (Eng. $\text{Load}_{\text{genitive}}$) that appeared only in the NMT output. Annotator 2, however, selected the NMT output as if she were aware of the reference translation, which also contained the word. If the reference translations were presented to the annotators, then probably both of them would be convinced to assign the win to NMT, which, in our opinion, would not be deserved.

Table 12 gives another example, which shows that the lack of reference translation may have a desired influence on annotations.

Table 12 Scoring vocabulary quality without reference translation

Source	Seat to be cleaned and smooth and new seat packing to be inserted in the valve to be boxed up and tested at owner satisfaction.	
Reference	Siedzenie zaworu należy oczyścić, wygładzić, wstawić nowy pakunek siedzenia, zawór złożyć i wykonać próbę ciśnieniową satysfakcjonującą **Przedstawiciela** Armatora.	
SMT	Siedzenie zaworu należy oczyścić, wygładzić i wstawić nowy pakunek siedzenia, zawór złożyć i wykonać próbę ciśnieniową satysfakcjonującą **Przedstawiciela** Armatora.	
NMT	Siedzenie zaworu należy oczyścić i wygładzić, wstawić nowy pakunek siedzenia i zawór złożyć i wykonać próby w sposób satysfakcjonujący armatora.	
BLEU	SMT: 87.87	NMT: 48.68
Rating	1: Tie	2: NMT (Vocabulary)

The author of the reference translation added the word "Przedstawiciela" (Eng. $\text{Representative}_{\text{genitive}}$), which was absent in the source sentence. The SMT system made exactly the same addition. Annotator 1 scored the trans-

[27] To be precise, we used the second smoothing approximation from the paper, which was already implemented in Moses Toolkit.

lations as a tie, Annotator 2 (correctly in our opinion) downgraded the "over-eagerness", giving the preference to the NMT output. If the annotators were prompted to upgrade the output for its similarity to the reference, then both of them would probably select the SMT output.

Table 13 is the inter-rater agreement matrix of the two annotators. The first row should be read as follows: In 814 cases both annotators agreed to assign a tie, in 21 cases Annotator 2 assigned a tie, when Annotator 1 assigned an SMT win, and in 16 cases Annotator 2 assigned a tie when Annotator 1 assigned an NMT win. The second row gives the number of pairs rated as an SMT win by Annotator 2, and the third row stands for NMT wins (according to Annotator 2).

Table 13 Annotators' scores in English-to-Polish translation

		Annotator 1		
		Tie	SMT wins	NMT wins
Annotator 2	Tie	814	21	16
	SMT wins	136	226	147
	NMT wins	122	73	445

As may be calculated from the Table 13, the annotators agreed in 74% cases and the Cohen's kappa coefficient[28] was equal to 0.4. This indicates fair but not high agreement. We formulate the following conclusions:

- The choice of the better translation was not clear-cut in most cases – this indicates similar quality of outputs generated by the two systems,
- Lack of insight into reference translations might have had negative influence on annotators' inter-rater agreement.

The main goal of our experiment was to compare the quality of the two systems and try to discover the main aspect – accuracy (vocabulary) or fluency (grammar) – which made the human annotators to rate one translation higher than the other.

Table 14 shows a case, when both translators gave preference to the NMT translation in the aspect of vocabulary. NMT correctly translates the term "tachometer" to "obrotomierz". SMT creates a non-existing equivalent of the term: "obroti".

Table 15 shows that NMT is capable of handling typographic errors in the input (here: "Adeck" instead of "A deck"). This feature might be the consequence of dividing source words into sub-words. Note that Annotator 1 classified the win aspect as "vocabulary", whereas Annotator 2 – as "other".

The example in Table 16 confirms that handling the above mentioned type of errors is characteristic of NMT and not of SMT. Here, the word chunk "E5existingserving" is split by the NMT engine into existing English words and translated into Polish.

[28] Cohen's kappa coefficient measures the degree of inter-rater agreement; it differs from the standard percentage in that it subtracts the probability of random agreement [1].

Table 14 Unanimous NMT win by vocabulary

Source	To dismantle filter element tachometer sensor and set tool on both sides.	
Reference	Zdemontować element filtrowy sensor **obrotomierza** i ustawić narzędzia na obu stronach.	
SMT	Zdemontować element filtrowy sensor **obroti** ustawić narzędzia na obu stronach.	
NMT	Zdemontować element filtrowy sensor **obrotomierza** i ustawić narzędzia na obu stronach.	
BLEU	SMT: 67.10	NMT: 100.00
Rating	1: NMT (Vocabulary)	2: NMT (Vocabulary)

Table 15 NMT – error handling (1)

Source	The Adeck shall be used as the roof for the deck store.	
Reference	**Pokład A** powinien stanowić sufit magazynu pokładowego.	
SMT	**Apokład** należy wykorzystać jako stropu dla magazynu pokładowego.	
NMT	**Pokład A** należy wykorzystać jako dach dla magazynu pokładowego.	
BLEU	SMT: 26.27	NMT: 27.90
Rating	1: NMT (Other)	2: NMT (Vocabulary)

Table 16 NMT – error handling (2)

Source	TUnit **E5existingserving** hospital including sanitary space	
Reference	Obecna **jednostka E5 służąca** szpitalowi wraz z przestrzenią sanitarną.	
SMT	Jednostka **E5existingobsługiwania** z uwzględnieniem przestrzeni sanitarnych szpitala	
NMT	**Urządzenie E5 obsługujące** szpitala z uwzględnieniem przestrzeni sanitarnych	
BLEU	SMT: 13.39	NMT: 13.82
Rating	1: NMT (Vocabulary)	2: NMT (Vocabulary)

Table 17 contains a pair of translations, where both annotators gave preference to the SMT translation in the aspect of vocabulary. The difference consists in the translation of the word "communication" (SMT: "korespondencja"; NMT: "przekazanie").

Table 17 Unanimous SMT win by vocabulary

Source	1. Any notification, information, documentation or other communication submitted pursuant to the provisions of this Title shall be supplied in a language acceptable to the competent authorities concerned.	
Reference	1. Wszelkie zgłoszenia, informacje, dokumentacja i inna **korespondencja** przekazywana zgodnie z przepisami niniejszego tytułu przekazywane są w języku akceptowanym przez zainteresowane właściwe organy.	
SMT	1. Wszelkie zgłoszenia, informacje, dokumentacja i inna **korespondencja** zgodnie z przepisami niniejszego tytułu przekazywane są w języku akceptowanym przez zainteresowane właściwe organy.	
NMT	1. Wszelkie zgłoszenia, informacje, dokumentacja lub inna **przekazanie**, złożone zgodnie z przepisami niniejszego tytułu przekazywane są w języku akceptowanym przez właściwe organy.	
BLEU	SMT: 89.86	NMT: 68.16
Rating	1: SMT (Vocabulary)	2: SMT (Vocabulary)

Table 18 shows a pair of translations, for which NMT produces better output in the aspect of word order. The example presents a characteristic feature of the NMT method – it is capable of reversing the order of long components of a sentence. Here, the adverbial part – starting with words "such that" – which ends the English sentence, is correctly moved by the NMT engine to the beginning of the translation ("W celu...").

Table 18 NMT win by grammar – word order

Source	1. (b) Outlet nozzles shall be dimensioned and fitted such that the extinguishing agent is evenly distributed.	
Reference	(b) W celu równomiernego rozprowadzenia środka gaśniczego dysze wylotowe muszą być odpowiednio zwymiarowane i zamontowane.	
SMT	(B) dysze wylotowe muszą być odpowiednio zwymiarowane i zamontowane równomiernego rozprowadzenia środka gaśniczego.	
NMT	(B) W celu równomiernego rozprowadzenia środka gaśniczego dysze wylotowe muszą być odpowiednio zwymiarowane i zamontowane.	
BLEU	SMT: 63.66	NMT: 100.00
Rating	1: NMT (Grammar)	2: NMT (Grammar)

Table 19 shows how the two engines deal with sentences out of domain. The NMT system produces better output in terms of grammar.

Table 19 NMT win by grammar – text out of domain

Source	1. We are not prepared simply to accept that more and more steps are being taken to dismantle democracy in Russia.	
Reference	Po prostu nie jesteśmy przygotowani, aby zaakceptować fakt, ze podejmuje się coraz więcej kroków mających na celu likwidację demokracji w Rosji.	
SMT	Jesteśmy nie została przygotowana po prostu zaakceptuje coraz więcej stopnie są wykonywane w celu rozmontowania demokracji w Rosji.	
NMT	Nie jesteśmy w prostu zaakceptować, by przyjąć więcej i więcej kroków w celu rozmontowania demokracji w Rosji.	
BLEU	SMT: 18.98	NMT: 19.36
Rating	1: NMT (Grammar)	2: NMT (Grammar)

The example given in Table 20 is one of infrequent cases of SMT winning in the aspect of grammar. Let us notice that both translations may be regarded as correct – the SMT output, however, looks more fluent.

Table 20 SMT win by grammar

Source	Aux. Boiler No.1 complete survey	
Reference	Kompletna inspekcja kotła pomocniczego # 1	
SMT	Kompletna inspekcja klasyfikacyjna kotła pomocniczego Nr. 1	
NMT	Pomocniczy kocioł nr 1 - kompletna inspekcja klasyfikacyjna	
BLEU	SMT: 32.17	NMT: 22.68
Rating	1: SMT (Grammar)	2: SMT (Grammar)

A frequent reason for the NMT wins was the truecasing. This is illustrated by Table 21. In the source translation all words begin with a capital letter. The case is preserved in the reference and in the NMT output. SMT produces output with two words that begin with a capital letter and three – with a small letter.

Table 21 NMT win by truecasing

Source	0902.12 Main Engine - Unit Overhaul	
Reference	0902.12 **Przegląd Układów Cylindrowych** Silnika Głównego	
SMT	0902.12 - **przegląd układów cylindrowych** Silnika Głównego	
NMT	0902.12 **Przegląd Układów Cylindrowych** Silnika Głównego	
BLEU	SMT: 70.71	NMT: 100.00
Rating	1: NMT (Other)	2: NMT (Other)

The ties were naturally assigned to all sentences translated in the identical way by both systems. When an annotator decided to choose a tie to different outcomes, that meant that in her/his opinion the translations were equally correct (rather than she/he could not decide which translation was of better quality). Table 22 presents a pair of translations, which was classified by one of the annotators as equally correct. (The other annotator regarded the SMT output as more accurate). The SMT translation changes the infinitive form of the sentence into a passive form (equivalent to the English expression: "Action must be done"), whereas the NMT engine produces a nominal, gerund form (equivalent to: "Doing something").

Table 22 Tie – both translations are correct

Source	Remove, clean, and replace nozzle ring as directed.	
Reference	Należy zdemontować, oczyścić i należy wymienić pierścień dyszy na nowy według wskazań.	
SMT	**Należy** zdemontować, oczyścić i wymienić, pierścień dyszy, według wskazań.	
NMT	**Zdemontowanie**, oczyszczenie i wymiana pierścienia dyszy zgodnie z zaleceniami.	
BLEU	SMT: 45.82	NMT: 10.90
Rating	1: Tie	2: SMT (Grammar)

Table 23 shows a pair of translations, for which both systems made one error. SMT correctly translated "sets" into "kompletów" but made a typographic error in the word "Sszakle" (unnecessary doubled "s"). NMT avoided the typographic error but incorrectly translated "set" into "szt." (Eng. "piece"). Annotator 1 scored it as a tie, Annotator 2 assigned a win to SMT (probably because its output required fewer post-edit operations on characters).

The overall statistics on the evaluation are given in Tables 24, 25 and 26. Table 24 presents the scores assigned by Annotator 1.

Table 25 presents the scores assigned by Annotator 2.

Table 23 Tie – both translations are incorrect

Source	Sackles.. 35 sets..	
Reference	**Szakle**.. 35 **kompletów**	
SMT	**Sszakle**.. 35 **komplety**..	
NMT	**Szakle**..: 35 **szt.**.	
BLEU	SMT: 35.93	NMT: 31.24
Rating	1: Tie	2: SMT (Vocabulary)

Table 24 First annotator's scores in English-to-Polish translation

Winner	Annotations	Percentage
SMT wins total	**320**	**16.0%**
SMT wins by grammar	67	3,35%
SMT wins by vocabulary	215	10,75%
SMT other wins	38	1.90%
NMT wins total	**608**	**30.40%**
NMT wins by grammar	302	15.10%
NMT wins by vocabulary	279	13.95%
NMT other wins	27	1.35%
Ties	**1072**	**53.60**

Table 25 Second annotator's scores in English-to-Polish translation

Winner	Annotations	Percentage
SMT wins total	**509**	**25.45%**
SMT wins by grammar	114	5.70
SMT wins by vocabulary	329	16.45%
SMT other wins	66	3.30%
NMT wins total	**640**	**32.00%**
NMT wins by grammar	268	13.40%
NMT wins by vocabulary	318	15.90%
NMT other wins	54	2.70%
Ties	**851**	**42.55%**

Table 26 presents the aggregated scores of both annotators.

Table 26 Aggregated annotators' scores in English-to-Polish translation

Winner	Annotations	Percentage
SMT wins total	**829**	**20.73%**
SMT wins by grammar	181	4.53%
SMT wins by vocabulary	544	13.60%
SMT other wins	104	2.60%
NMT wins total	**1248**	**31.20%**
NMT wins by grammar	570	14.25%
NMT wins by vocabulary	597	14.93%
NMT other wins	81	2.03%
Ties	**1923**	**48.08%**

Table 26 reveals that almost half of the translations produced by the two systems were rated as being of equal quality. Of the remaining half a significant

majority (31 percentage points against 21 percentage points) was rated in favour of NMT. The wins by vocabulary (approximately corresponding to accuracy) were split almost evenly (595 against 545 in favour of NMT). The difference was made by the grammar (fluency) of the output. Here, the NMT method outscored SMT significantly: 571 against 179.

Our human evaluation of English-to-Polish domain-specific translation systems trained on a medium- size corpus confirms the findings formulated in [17]: Human evaluation contrasted with automatic evaluation favours the NMT approach over the SMT approach. The main feature of the NMT methodology, which causes humans to rate it higher than SMT, is better fluency of the output.

6 Conclusions

We simulated a situation when an organisation demands a domain-adaptive MT system, having at their disposal a bilingual technical corpus of a medium size (over a million segments). The experiment shows that training the system on enterprise-specific texts may result in the translation quality over twice as good as that of a general system and almost twice as good as that trained for a specific domain on publicly available data. Surprisingly, Statistical Machine Translation, which requires fewer data and less computational power than Neural Machine Translation, reports better results in the BLEU metric. On the other hand, decoding carried out on a standard office machine is performed much faster within NMT than SMT. We carried out an evaluation experiment, which consisted in human comparing the output quality of 2,000 sentences. The annotators were not aware of reference translation in order not to be affected by human-created translation. This assumption might have been the reason for the medium inter-rater agreement correlation, as measured by the Cohen's kappa co-efficient. The experiment confirmed previous findings for similar pairs of languages: human evaluators favour NMT over SMT, particularly in the aspect of output fluency.

References

1. Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.
2. Dymitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*, 2014.
3. Boxing Chen and Colin Cherry. A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, 2014.
4. Colin Cherry and George Foster. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 427–436, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

5. Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.
6. Nadir Durrani, Helmut Schmid, Alexander M. Fraser, Philipp Koehn, and Hinrich Schütze. The Operation Sequence Model - Combining N-Gram-Based and Phrase-Based Statistical Machine Translation. *Computational Linguistics*, 41:185–214, 2015.
7. Chris Dyer, Victor Chahuneau, and Noah A. Smith. 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*, pages 644–648. Association for Computational Linguistics, 2013.
8. Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional Sequence to Sequence Learning. *ArXiv e-prints*, May 2017.
9. Kenneth Heafield. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 187–197, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
10. Hieu Hoang, Tomasz Dwojak, Rihards Krislauks, Daniel Torregrosa, and Kenneth Heafield. Fast Neural Machine Translation Implementation. In *Proceedings of the NMT 2018*. Association for Computational Linguistics, September 2018.
11. Sepp Hochreiter and Jürgen Schmidhuber. Long Short-term Memory. *Neural computation*, 9:1735–80, 12 1997.
12. Marcin Junczys-Dowmunt. Phrasal rank-encoding: Exploiting phrase redundancy and translational relations for phrase table compression. *The Prague Bulletin of Mathematical Linguistics*, 98:63–74, 2012.
13. Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. Is Neural Machine Translation ready for Deployment? A Case Study on 30 Translation Directions. *arXiv preprint arXiv:1610.01108*, 2016.
14. Marcin Junczys-Dowmunt, Tomasz Dwojak, and Rico Sennrich. The AMU-UEDIN Submission to the WMT16 News Translation Task: Attention-based NMT Models as Feature Functions in Phrase-based SMT. In *Proceedings of the First Conference on Machine Translation*, pages 319–325, Berlin, Germany, August 2016. Association for Computational Linguistics.
15. Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast Neural Machine Translation in C++. *arXiv preprint arXiv:1804.00344*, 2018.
16. Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, 12 2014.
17. Filip Klubicka, Antonio Toral, and Víctor M. Sánchez-Cartagena. Fine-grained human evaluation of neural versus phrase-based machine translation. *CoRR*, abs/1706.04389, 2017.
18. Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.
19. Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2009.
20. Philipp Koehn. Neural machine translation. *CoRR*, abs/1709.07809, 2017.
21. Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.
22. Philipp Koehn and Rebecca Knowles. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, 2017.
23. Jason Lee, Kyunghyun Cho, and Thomas Hofmann. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378, 2017.

24. Pierre Lison and Jörg Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *LREC*, 2016.
25. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.
26. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
27. Świeczkowska Patrycja. Towards a direct Japanese-Polish machine translation system . In *Proceedings of the 8th Language & Technology Conference*, Poznan, Poland, November 2017.
28. Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 389–399, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
29. Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics, 2016.
30. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.
31. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.
32. Krzysztof Wołk and Krzysztof Marasek. PJAIT Systems for the IWSLT 2015 Evaluation Campaign Enhanced by Comparable Corpora. In *Proceedings of the International Workshop on Spoken Language Translation, December 3-4, 2015 Da Nang, Vietnam*, 2015.
33. Krzysztof Wołk and Krzysztof Marasek. PJAIT Systems for the WMT 2016. In *Proceedings of the First Conference on Machine Translation*, 2016.
34. Krzysztof Wołk and Krzysztof Marasek. PJAIT's Systems for WMT 2017 Conference. In *Proceedings of the Second Conference on Machine Translation*, 2017.