NMT or SMT: Case Study of a Narrow-domain English-Latvian Post-editing Project

Inguna Skadiņa and **Mārcis Pinnis**
Tilde, Vienibas gatve 75A, Riga, Latvia
{inguna.skadina,marcis.pinnis}@tilde.lv

Abstract

The recent technological shift in machine translation from statistical machine translation (SMT) to neural machine translation (NMT) raises the question of the strengths and weaknesses of NMT. In this paper, we present an analysis of NMT and SMT systems' outputs from narrow domain English-Latvian MT systems that were trained on a rather small amount of data. We analyze post-edits produced by professional translators and manually annotated errors in these outputs. Analysis of post-edits allowed us to conclude that both approaches are comparably successful, allowing for an increase in translators' productivity, with the NMT system showing slightly worse results. Through the analysis of annotated errors, we found that NMT translations are more fluent than SMT translations. However, errors related to accuracy, especially, mistranslation and omission errors, occur more often in NMT outputs. The word form errors, that characterize the morphological richness of Latvian, are frequent for both systems, but slightly fewer in NMT outputs.

1 Introduction

For many years, the central problem in machine translation (MT) has been the quality. MT quality has been recognized as a complicated research question when translation is performed into a morphologically rich (and also under-resourced) language with a relatively free word order, e.g., Bulgarian, Croatian, Estonian, Finnish, Greek or Latvian. Possible solutions for widely used statistical machine translation have been studied for many years (e.g., Koehn and Hoang 2007; Tamchyna and Bojar 2013; Burlot and Yvon 2015).

Today machine translation is experiencing a paradigm shift from (phrase-based) statistical machine translation (SMT) to neural machine translation (NMT). The first results obtained in recent years are promising, as it can be seen from the results of WMT 2016 (Bojar et al., 2016) and WMT 2017 (Bojar et al., 2017).

As NMT becomes more and more popular, the question of what can we expect from NMT in terms of quality becomes very important. Recent analysis of English to German SMT and NMT outputs of manual transcripts of short speeches showed that NMT can decrease the post-editing effort (Bentivogli et al., 2016). A comparison of NMT and SMT systems for nine language directions (English to and from Czech, German, Romanian, Russian, and English to Finnish) on news stories made by Toral and Sánchez-Cartagena (2017) showed that translations produced by NMT systems are more fluent and more accurate in terms of word order compared to translations produced by SMT systems. By analyzing of manually error-annotated outputs of generic English-Croatian MT systems, Klubička et al. (2017) found that NMT handles all types of agreement better than SMT (including factored models).

In this paper, we delve further into analyzing the strengths and weaknesses of NMT from the perspective of translation quality and the needs of the localization industry. We analyze translations of good quality domain-specific (medicine related) English-Latvian SMT and NMT systems that were trained on a rather small (ca. 325K sentences) data set. The target language - Latvian - is a morphologically rich under-resourced language (about 1.5 million speakers). As it is a synthetically inflected language, words change their form according to their grammatical function. In Latvian only half of the word endings are unambiguous, while for

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 373–383,
Taipei, Taiwan, November 27 – December 1, 2017

the rest, multiple base forms may be derived from the inflected form (Skadiņa et al., 2012).

We analyze outputs of NMT and SMT systems in a post-editing (PE) scenario. Data on PE time, keystrokes, and typical operations were collected during the PE process. Analysis of these data allowed us to conclude that both approaches (SMT and NMT) are comparably successful allowing to increase translator productivity, with the NMT system showing slightly worse results. We believe that the reason translations from the SMT system are better in our case, is that from the small amount of data, SMT learns better terminology and phrases which are specific for the particular narrow domain. The situation could be different for broad domain MT systems, as it can be seen from recent WMT 2017 English-Latvian news domain results, where NMT and hybrid approaches were better (Bojar et al., 2017; Pinnis et al., 2017).

In addition, for a small sub-set of the MT system translations, manual error annotation was performed. This allowed us to identify the main error categories for each MT system. Through analysis of annotated errors, we found that NMT translations are more fluent than SMT translations, NMT produces significantly fewer typography errors than SMT. At the same time errors related to accuracy, especially, mistranslation and omission errors, occur more often in NMT outputs. The word form errors, which characterize the morphological richness of Latvian, are slightly fewer in NMT outputs.

2 Related work

Questions on how to evaluate the quality and usefulness of machine translation have been studied for several decades. For localization industry needs, MT quality and PE productivity have been analyzed by Flournoy and Duran (2009); Groves and Schmidtke (2009); Plitt and Masselot (2010); Skadiņš et al. (2011); Pinnis et al. (2016) and others. These studies report significant productivity increase when good quality SMT systems are used. Recently, for English-Spanish Sanchez-Torron and Koehn (2016) reported that "for 1-point increase in BLEU, there is a PE time decrease of 0.16 seconds per word, about 3-4%".

Several studies have recently compared SMT and NMT systems. Bentivogli et al. (2016) conducted a detailed analysis of SMT and NMT output for the English-German language pair on translations of manual transcripts of TED talks [1]. They found that NMT decreases post-editing effort, but degrades faster than SMT for longer sentences. They also found that NMT output contains fewer morphology errors, lexical errors and substantially fewer word order errors. Toral and Sánchez-Cartagena (2017) compared NMT and SMT systems submitted to WMT16 news translation task for nine translation directions (English to and from Czech, German, Romanian, Russian, and English to Finnish). The authors found that the translations produced by NMT systems were more fluent and more accurate in terms of word order compared to translations produced by SMT systems. They observed that NMT systems are also more accurate at producing inflected forms, but they perform poorly when translating very long sentences.

However, when Farajian et al. (2017) compared the performance of generic English-French NMT and SMT systems, that were trained on a generic parallel corpus composed of data from different domains, they found that on such multi-domain data SMT outperforms its neural counterpart. Moreover, Castilho et al. (2017) in their study, in which human evaluators compared NMT and SMT output for a range of language pairs, reported mixed results from the human evaluation. Similarly to the previous authors, they reported an increase in fluency, but inconsistent results for adequacy (the neural model showed a greater number of errors of omission, addition, and mistranslation) for NMT when compared to SMT. They argue that, although "NMT shows significant improvements for some language pairs and specific domains, there is still much room for research and improvement before broad generalizations can be made."

Analysis of NMT and SMT errors was recently made by Klubička et al. (2017) for English-Croatian MT systems. The authors analyzed manual error annotations of SMT and NMT system translations in the news domain and concluded that the NMT system reduces the errors produced by the SMT system by 54%.

3 Data and MT Systems

The SMT and NMT systems were trained on the parallel corpus from the European Medicines Agency (EMEA), which is a part of the OPUS cor-

[1] http://www.ted.com/

Corpus	Sentences before filtering	Sentences after filtering
Parallel	378,869	325,332
Monolingual	378,869	332,652

Table 1: Statistics of the training corpora

pus (Tiedemann, 2009), and the latest documents from the EMEA website (years 2009-2014) [2].

Prior to the training of the MT systems, we preprocessed the training data using tools for corpora cleaning, filtering, non-translatable token (e.g., URL, e-mail address, different code, etc.) identification, tokenization, and true-casing. The statistics of the training corpora before and after preprocessing are given in Table 1.

3.1 Statistical Machine Translation System

The SMT system is a standard phrase-based system that was trained on the Tilde MT platform (Vasiļjevs et al., 2012) with Moses (Koehn et al., 2007). The system features a 7-gram translation model and a 5-gram language model. The language model was trained with KenLM (Heafield, 2011). The system was tuned with MERT (Bertoldi et al., 2009) using a held-out set of 2,000 sentence pairs.

3.2 Neural Machine Translation System

We used the sub-word neural machine translation toolkit Nematus (Sennrich et al., 2017) for training the NMT system. The toolkit allows training attention-based encoder-decoder models with gated recurrent units in the recurrent layers. For word splitting in sub-word units, we use the byte pair encoding tools from the subword-nmt toolkit (Sennrich et al., 2015). The NMT system was trained using a vocabulary of 40,000 word parts (39,500 for byte pair encoding), a projection (embedding) layer of 500 dimensions, recurrent units of 1024 dimensions, a batch size of 20 and dropout enabled. All other parameters were set to the default parameters as used by the developers of Nematus for their WMT 2016 submissions (Sennrich et al., 2016).

3.3 MT System Evaluation

SMT and NMT systems were evaluated on a held-out set of 1000 randomly selected sentence pairs.

[2] http://www.ema.europa.eu/

System	BLEU	NIST	ChrF2
SMT	46.57±1.46	9.45±0.18	0.7586
NMT	38.44±1.62	8.63±0.15	0.7065

Table 2: Automatic evaluation results

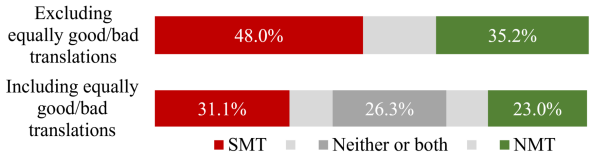

Figure 1: Human comparative evaluation results for SMT and NMT systems

The automatic evaluation results are given in Table 2. The results show that the SMT system achieves better results than the NMT system. This could be explained by the relatively small size of the parallel corpus and a very narrow domain, i.e., from the small amount of data, SMT learns better terminology and phrases which are specific for the particular narrow domain.

When translation is performed into a morphologically rich language, such as Latvian, automatic metrics (e.g. BLEU score) are not always good indicators of translation quality. Table 3 illustrates a case, where both translations have the same quality, but because of different word order the SMT translation received 41.38 BLEU points, while the NMT translation - only 24.42 points. To validate the automatic evaluation results, we performed a small blind comparative evaluation task. The task was performed by 5 professional translators who evaluated 198 segments in total. The results of the comparative evaluation show that the translations of the SMT system are preferred more often by evaluators than the translations of the NMT system (see Figure 1). However, the difference is not statistically significant according to the methodology by Skadiņš et al. (2010). Therefore, both systems were further used in the post-editing and error annotation experiments.

4 What Can Be Learned from Post-edits?

4.1 Post-editing process

For post-editing, we compiled a list of 22,500 segments (360,000 words) from EMEA documents. Then, we split the list into documents consisting of 100 segments so that the original sequence of sentences is preserved, and translated the documents

Sentence	BLEU	Text
Source	-	Seek medical advice straight away if you develop a severe rash, itching or shortness of breath or difficulty breathing.
Human	100.00	Nekavējoties meklējiet medicīnisku palīdzību , ja Jums parādās izsitumi , rodas nieze vai elpas trūkums , vai apgrūtināta elpošana .
SMT	41.38	Nekavējoties meklējiet medicīnisko palīdzību , ja Jums rodas smagi izsitumi , nieze vai elpas trūkums vai apgrūtināta elpošana .
NMT	24.42	Ja Jums rodas smagi izsitumi, nieze vai elpas trūkums vai apgrūtināta elpošana , nekavējoties meklējiet medicīnisko palīdzību .

Table 3: Influence of word order on BLEU score for similar translations by SMT and NMT systems

with both MT systems.

At first, translators were asked to post-edit SMT translations. Then, three months later, they were asked to post-edit NMT translations. For the NMT post-editing task, the documents were redistributed to translators, to ensure that each translator has different set of documents in SMT and NMT post-editing tasks.

We asked translators to post-edit translated segments with the post-editing tool PET (Aziz et al. 2012). It allowed us to track the time spent on each segment and to log all keystrokes that the translator performed while post-editing each segment. Translators were asked not to spend excessive amounts of time on each segment because the quality expectations were not "human translation quality", but rather "post-editing quality".

To assist post-editing, translators were provided with an automatically extracted in-domain term collection that was integrated into PET and provided translation suggestions for known terms.

After post-editing each segment, translators were asked to evaluate the quality of the MT translation, marking it as one of the following: "*near perfect*", "*very good*", "*poor*", and "*very poor*". If the translator did not apply any changes, the system automatically assigned the highest quality rating - "*Unchanged*".

Five professional translators were involved in the SMT post-editing task and seven in the NMT post-editing task. Finally, we asked the translators who participated in both tasks (4 in total) to translate two documents without pre-translated segments in order to measure each translator's pure translation productivity.

4.2 Post-editing Results

Most of the translators involved in this experiment post-edited 20 documents (in each post-editing task). To perform a fair comparison between SMT and NMT post-editing tasks, we limit our analysis to the first 20 documents post-edited by each translator participating in both post-editing tasks. We perform the analysis only on segments that were not found in the MT system training data (approximately 36% of segments were discarded). The statistics of the post-edited data that are used for the further analysis is given in Table 4.

	Doc.	Segments	Tokens
Translation	8	797	14,924
SMT	80	5,280	99,375
NMT	80	4,688	86,651
Total	**168**	**10,765**	**200,950**

Table 4: Statistics of post-edited data

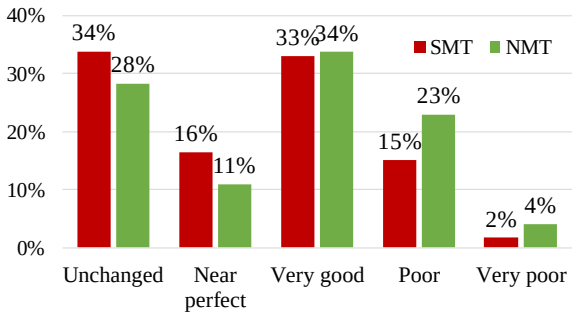

Figure 2: Distribution of rankings for MT segments

We start the analysis by examining the MT quality assessments produced by translators during post-editing. The Figure 2 summarizes the distribution of rankings showing that the SMT system produced a larger proportion of near perfect and perfect translations than the NMT system - 50.2% compared to just 39.3%.

The detailed logs of each translators work allowed to measure the time spent on post-editing

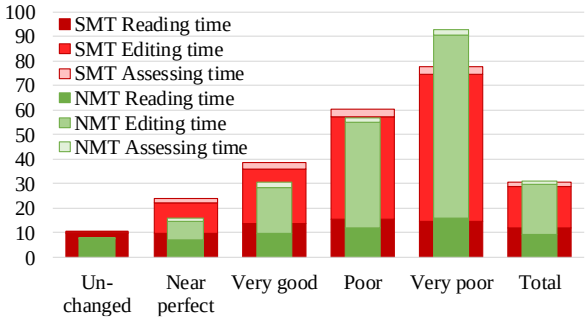

Figure 3: Average time in seconds spent on a segment

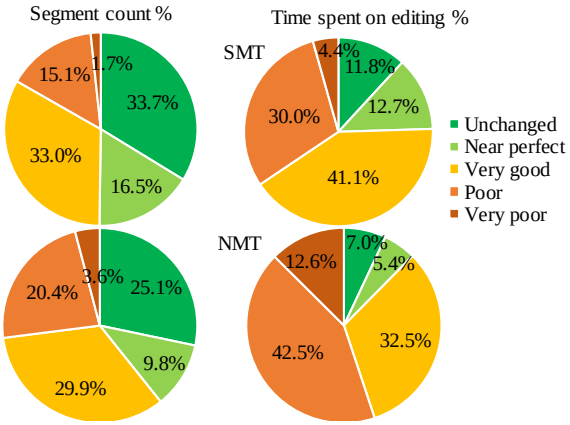

Figure 4: Segment count and editing time distribution for different quality MT segments

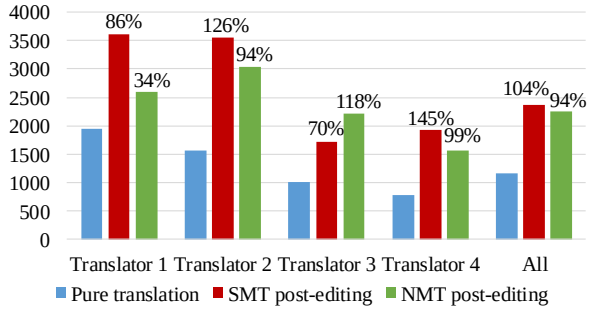

Figure 5: Individual translator productivity (tokens translated/post-edited per hour) based on actually measured numbers

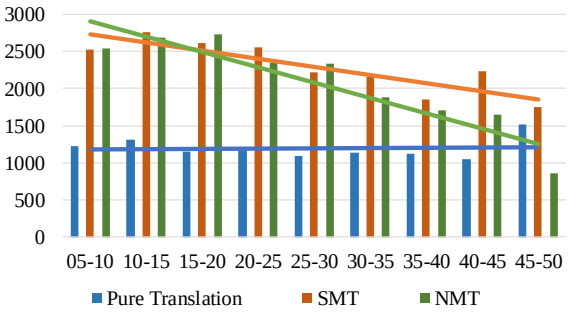

Figure 6: Translation and post-editing productivity (tokens translated/post-edited per hour) for segments with different length with linear trendlines

in three distinct intervals: the amount of time that elapsed between the appearance of an MT segment and the first click, or "*reading time*"; the amount of time between the first edit and approval of the segment, or "*editing time*"; and the amount of time spent between approval of the segment and completion of the quality assessment, referred to as "*assessment time*". The results of the log data analysis in Figure 3 show that on average it takes 30% more time for translators to start editing SMT translations. It is also obvious that editing of good, very good and near perfect SMT translations requires 16-62% more time than for NMT translations. However, the situation is opposite for poor and very poor translations - it requires 3-25% more time to post-edit NMT translations. This difference is more noticeable in Figure 4, which shows that post-editing poor and very poor NMT translations (24% of all post-edited NMT translations) required more than half of the editing time (55.1%). In comparison post-editing of poor SMT translations (16.8% of all post-edited SMT translations) required just 34.4% of time.

In terms of productivity (see Figure 5), it is evident that both tasks (SMT and NMT post-editing) obtain higher productivity than pure translation. However, the productivity is higher for post-editing SMT translations (104% compared to 94%).

When analyzing the effect of the length of segments on productivity (tokens translated/post-edited per hour), the results in Figure 6 showed that there is an obvious decrease in post-editing productivity for longer segments, with the NMT post-editing productivity decreasing faster than for SMT post-editing. It is interesting that there is almost no change in productivity when translating without MT support.

The information on the time spent on each segment allows us to analyze the relationship between the post-editing productivity and the post-editing effort that is expressed with the help of the Human-targeted Translation Edit Rate (HTER; Snover et al. 2006). Figure 7 depicts the aver-

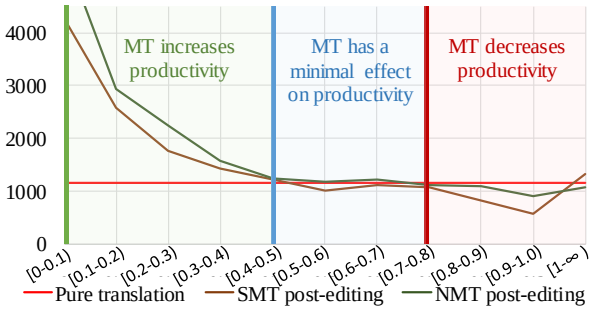

Figure 7: Average productivity (tokens translated/post-edited per hour; *y* axis) at different MT suggestion quality thresholds (HTER; *x* axis)

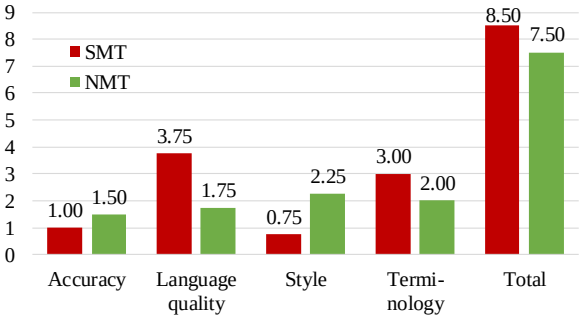

Figure 8: Average error score (per 1000 words)

age productivity for different MT translation quality intervals. It shows that we can identify average MT system quality thresholds, at which post-editing becomes productive (HTER of 0.4 or less) and at which it stops being productive (HTER of 0.7 or higher). The average HTER scores of the SMT and NMT systems are 0.22 and 0.31 respectively. The figure also shows that there is little difference between SMT and NMT post-editing, with the NMT post-editing being faster at individual quality levels. Still, because the NMT system produced more poor translations, the overall post-editing productivity is higher for the SMT post-editing task.

To validate, whether the post-edits are of good quality, we performed quality assessment of the post-edits according to the LISA Quality Assurance model[3]. The quality assessment was performed by professional editors from our localization department. The results in Figure 8 show that even though the task for translators was to perform light post-editing, the quality of the post-edited translations is rated as excellent (i.e., the average error score for both SMT and NMT post-edits is below 10 per 1000 words).

5 MT Error Annotation

The aim of the error annotation task was to identify common and specific errors for both MT architectures and their influence on the overall quality of MT output.

5.1 Error Annotation Task

For error annotation (EA), 1800 English segments and their translations into Latvian by SMT and NMT systems were selected. Only translations that were marked as "*Very good*" during post-editing for both MT systems were included. The main reason for including only segments that have good translations was the necessity to avoid wrong annotations due to very bad input.

The error classification used, in this task, is based on Multidimensional Quality Metrics (MQM; Lommel et al. 2014). More specifically, the subset that is defined by Burchardt and Lommel (2014) was used. In this classification, errors are divided into three top categories: accuracy, fluency, and terminology. These top level categories then include more detailed categories from the MQM issue type hierarchy.

The EA was performed four months after finishing both post-editing tasks. Two translators, who participated in both post-editing tasks, were involved to ensure consistency between post-editing and error annotation tasks and to avoid a situation when translators annotate errors, which were not requested to be corrected during post-editing.

The error annotation was performed in the Translate5[4] platform. Before translators started the error annotation, they were introduced to a video tutorial, written guidelines, and the decision process. During annotation, translators saw the source segment, MT output, and post-edited MT output.

Each translator annotated 1000 segments translated by the SMT system and the same 1000 segments translated by the NMT system. Although inter-annotator agreement was not our main interest, 200 translations from each system were annotated by both translators.

[3]LISA QA model: http://web.archive.org/web/20080124014404/http://www.lisa.org/products/qamodel/

Error type	SMT error annotation			NMT error annotation		
	Count	Total	%	Count	Total	%
Accuracy	**39**	**1078**	**28%**	**50**	**1634**	**44%**
Addition	282	282	7%	271	271	7%
Mistranslation	275	275	7%	683	683	19%
Omission	402	402	10%	568	568	15%
Untranslated	80	80	2%	62	62	2%
Fluency	**234**	**2734**	**71%**	**213**	**2023**	**55%**
Grammar	11	1329	35%	2	1006	27%
Function words	0	171	4%	0	136	4%
Extraneous	49	49	1%	49	49	1%
Incorrect	56	56	1%	55	55	1%
Missing	66	66	2%	32	32	1%
Word form	282	809	21%	266	714	19%
Part of speech	38	38	1%	35	35	1%
Agreement	429	429	11%	367	367	10%
Tense/aspect/mood	60	60	2%	46	46	1%
Word order	338	338	9%	154	154	4%
Spelling	326	326	8%	394	394	11%
Typography	835	835	22%	396	396	11%
Unintelligible	10	10	0%	14	14	0%
Terminology	**35**	**35**	**1%**	**31**	**31**	**1%**
All types		**3847**	**100%**		**3688**	**100%**

Table 5: Summary of error annotation task (count - number of errors for particular category; total - sum of errors, including subcategories)

5.2 Observations from the Error Annotation Task

The overall results of the error annotation task are summarized in Table 5. Results show that although the segments were ranked as good, most of them contain more than one error per segment. The total number of errors is higher for SMT. There are twice as many errors related to fluency (77%) as to accuracy (28%) for SMT, while for NMT the fluency errors comprise 55% of errors, but accuracy errors - 44%.

The complexity of Latvian morphology is a reason why more than 1/4 of errors are grammar errors (35% for SMT and 27% for NMT), from which almost 1/5 of errors are word form errors (SMT 21%, NMT - 19%). For instance, both MT systems generate the wrong form for the word "*aerosols (spray)*" when translating the sentence "*How to use the nasal spray*": the SMT system generates the singular nominative form *aerosols (spray)*, while the NMT system generates singular genitive form *aerosola (spray)*.

A significant difference between SMT and NMT outputs has been observed for three error subcategories - typography (the subcategory of fluency), mistranslation (the subcategory of accuracy) and omission (the subcategory of accuracy).

Typography errors are much more widespread in SMT (21.70%) than in NMT (11%). Usually these are cases where spaces are wrongly used (e.g. "*beta - 2 - agonisti*" instead of "*beta-2-agonisti*" (beta-2-agonists), or wrong separators appear in numbers (e.g. "*3,644*" instead of "*3644*", or "*0.5*" instead of "*0,5*"). These errors, especially wrong separators, are not frequent in NMT translations.

The Latvian language has a very rich, morphology-based word-building potential (words are usually built by adding affixes to the stem). This feature resulted in a high number (19%) of mistranslations from the NMT system. Typical cases of mistranslation from the NMT system include the incorrect translation of numbers (e.g., *30 July **2012*** is translated as ***2008.** gada 30. jūlijs*), terms (e.g., *drop (piliens)* is translated as *injekcija (injection)*) and named entities (e.g., *Naglazyme (Naglazyme)* is translated

[4]http://translate5-metashare.dfki.de

as *MabCampath*).

Latvian also has a relatively free word order. In the case of a formal, narrow domain, where usually the word order is strict, it has a rather small influence even for the SMT system (9% of errors), while in the case of more general systems this could have much greater impact.

Errors of omission are much more frequent for NMT (15%) than for SMT outputs (10%). NMT also produces fewer (4%) word order errors than SMT (9%), while SMT has fewer (8%) spelling errors than NMT (11%).

5.3 Inter-annotator Agreement

Although the aim of this research was not to study consistency between annotations, but to identify and analyze the main error categories, 200 segments translated by SMT and NMT systems were annotated by two translators. The reason for having only two annotators was seriously debated in the consortium of the QT21 project [5] by a number of leading MT researchers. It was agreed that, to show inconsistencies/issues, common understanding of the annotation task, it is enough to have two annotators. The inter-annotator agreement is more like a sanity check for the fine-grained annotation levels (whether annotators have common understanding or not). Table 6 presents the summary on errors annotated in these segments.

Similarly to the whole error annotation task, slightly more errors are found in the SMT system's output. Table 6 also confirms the finding from the overall error annotation task, that NMT produces less typography and word order errors than SMT, but it produces more mistranslation and omission errors.

There are several error categories where translators have different opinions about the applicability of the particular categories. The table clearly demonstrates that the most complicated case was the identification of a correct subcategory for wrong word form errors. The annotator A1 mostly assigned the top category "word form" for such errors, while the annotator A2 marked them as agreement errors.

Another case of significant disagreement between annotators can be observed for fluency errors in the NMT post-editing task. As there was no consistent correspondence between an error category assigned by annotator A2 for cases where annotator A1 marked fluency errors, we asked annotator A1 to explain her reasoning. She told us that she marked fluency errors where a post-editor during post-editing applied just stylistic corrections. After inspecting these cases, we agreed with her explanation.

Error type	SMT		NMT	
	A1	A2	A1	A2
Accuracy	2	0	0	10
Addition	42	37	32	23
Mistranslation	11	16	17	24
Omission	32	26	37	22
Untranslated	8	10	8	10
Fluency	3	0	33	4
Grammar	6	0	0	0
Function words	0	0	0	0
Extraneous	1	2	0	7
Incorrect	0	3	0	4
Missing	1	3	0	12
Word form	43	0	41	1
Part of speech	0	5	0	2
Agreement	4	41	8	46
Tense/aspect/mood	3	8	0	8
Word order	18	16	6	4
Spelling	43	44	58	56
Typography	84	71	43	42
Unintelligible	3	1	1	1
Terminology	3	0	0	5
All categories	307	283	284	281

Table 6: Error annotation summary for 200 segments annotated by 2 translators (A1 and A2)

For inter-annotator agreement, we calculated free-marginal kappa under three different conditions (see Table 7): perfect match analysis (i.e., by taking the precise positions and (sub)categories of errors into account), error count analysis (i.e., by ignoring error positions), and error presence analysis (i.e., by just looking at whether both annotators identified that a segment contains a certain (sub)category of errors)[6]. The results show that when taking positions into account, there is just slight agreement between the annotators. This is explained by the different understanding of where errors need to be marked: one translator annotated errors at the character level, while the other - at the token level. For instance, in the case of wrong

[5] http://www.qt21.eu/

[6] Free-marginal kappa is interpreted as: 0.01-0.20 = slight agreement, 0.21-0.40 = fair agreement, 0.41-0.60 = moderate agreement, 0.61-0.80 = substantial agreement, 0.81-1.00 = almost perfect agreement (Landis and Koch, 1977)

	SMT	NMT	Both
Perfect match analysis			
Instances	493	446	939
Agreed inst.	54	54	108
Kappa	0.065	0.077	0.071
Agreement %	11%	12%	12%
Error count analysis			
Instances	401	418	819
Agreed inst.	189	147	336
Kappa	0.445	0.319	0.381
Agreement %	47%	35%	41%
Error presence analysis			
Instances	355	388	743
Agreed inst.	172	133	305
Kappa	0.459	0.310	0.381
Agreement %	48%	34%	41%

Table 7: Inter-annotator agreement (free-marginal kappa) on the 200 segment data sets

separators in numbers (e.g. 7.5), one annotator marked only the punctuation mark, while the other - the whole number. If we analyze the agreement on just error count and error presence levels, we see that the annotators reached moderate agreement for the annotation of errors for the SMT system's translations, but only fair agreement for the NMT system's translations. This is mainly due to the disagreement on how to annotate fluency errors.

The inter-annotator agreement scores highlight the necessity for improvements in the general guidelines to mitigate the potential for disagreement. That being said, the inter-annotator agreement in the higher error levels (i.e., if we do not split errors up in 4 levels of sub-categories, but analyze only the top 2 levels) is good (over 0.6) for SMT and moderate (over 0.4) for NMT.

6 Conclusion

In this paper, we presented an analysis of narrow domain English-Latvian SMT and NMT systems, that were trained on a rather small in-domain corpus.

Translations of both systems were post-edited by professional translators and ranked depending on the complexity of editing. 83% of SMT translations and 73% of NMT translations were ranked as perfect, near perfect or very good, thus confirming the fact that in-domain MT systems can produce good quality translations even when the amount of training data is limited. The analysis of post-edited data allowed us to conclude that both approaches allow for an increase in translator productivity, with the NMT system showing slightly worse results in general, but better for good quality MT output. We believe that the lower results for the NMT system are linked to the relatively small size of the parallel corpus and the narrow domain.

By analysis of the manually annotated errors, we found that the SMT system produced twice as many errors related to fluency (77%) in comparison to those related to accuracy (28%), while for the NMT system the fluency errors comprise 55% of all errors, but accuracy errors - 44%. In terms of error subcategories, widespread errors for both systems are grammar errors (35% for SMT and 27% for NMT), especially wrong word form errors (21% for SMT and 19% for NMT), indicating that morphologically rich languages, e.g., Latvian, are problematic for both MT systems, while improving with NMT. A significant difference between SMT and NMT outputs has been observed for three error subcategories - typography (22% for SMT and 11% for NMT), mistranslation (7% for SMT and 19% for NMT) and omission (10% for SMT and 15% for NMT).

The obtained results show that in the case of a narrow domain, if MT systems are trained on a small amount of data, the SMT system performs better than the NMT system. The reason why the SMT system in our case is better, is that from the small amount of data, SMT learns better terminology and phrases which are specific for the particular narrow domain. The situation differs for broad domain MT systems, as it has been demonstrated by recent WMT 2017 English-Latvian news domain results, where NMT and hybrid approaches were better.

Acknowledgments

We would like to thank Tilde's Localization Department for the hard work they did to prepare material for the analysis presented in this paper. The work within the QT21 project has received funding from the European Union under grant agreement n° 645452. The research has been supported by the ICT Competence Centre (www.itkc.lv) within the project "2.2. Prototype of a Software and Hardware Platform for Integration of Machine Translation in Corporate Infrastructure" of EU Structural funds, ID n° 1.2.1.1/16/A/007.

References

Wilker Aziz, Sheila CM De Sousa, and Lucia Specia. 2012. Pet: a tool for post-editing and assessing machine translation. In *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC12)*, pages 3982–3987.

Luisa Bentivogli, 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*, pages 257–267.

Nicola Bertoldi, Barry Haddow, and Jean-Baptiste Fouet. 2009. Improved Minimum Error Rate Training in Moses. *The Prague Bulletin of Mathematical Linguistics*, 91(1):7–16.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation (wmt16). *Proceedings of WMT*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Aljoscha Burchardt and Arle Lommel. 2014. Practical guidelines for the use of mqm in scientific research on translation quality. *Preparation and Launch of a Large-scale Action for Quality Translation Technology, report*, page 19.

Franck Burlot and François Yvon. 2015. Morphology-aware alignments for translation to and from a synthetic language. In *Proc. IWSLT*, pages 188–195.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108(1):109–120.

M Amin Farajian, Marco Turchi, Matteo Negri, Nicola Bertoldi, and Marcello Federico. 2017. Neural vs. phrase-based machine translation in a multi-domain scenario. *EACL 2017*, page 280.

Raymond Flournoy and Christine Duran. 2009. Machine translation and document localization at adobe: From pilot to production. *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, pages 425–428.

Declan Groves and Dag Schmidtke. 2009. Identification and analysis of post-editing patterns for mt. In *Proceedings of MT Summit*, volume 12, pages 429–436.

Kenneth Heafield. 2011. KenLM : Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 2009, pages 187–197. Association for Computational Linguistics.

Filip Klubička, Antonio Toral, and Víctor M Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):121–132.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *EMNLP-CoNLL*, pages 868–876.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ond\vrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. 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. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumtica: tecnologies de la traducci*, 0(12):455–463.

Mārcis Pinnis, Rihards Kalniņš, Raivis Skadiņš, and Inguna Skadiņa. 2016. What Can We Really Learn from Post-editing? In *Proceedings of the 12th Conference of the Association for Machine Translation in the Americas (AMTA 2016), vol. 2: MT Users*, pages 86–91, Austin, USA. Association for Machine Translation in the Americas.

Mārcis Pinnis, Rihards Krišlauks, Toms Miks, Daiga Deksne, and Valters Šics. 2017. Tilde's machine translation systems for wmt 2017. In *Proceedings of the Second Conference on Machine Translation*, pages 374–381.

Mirko Plitt and François Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague bulletin of mathematical linguistics*, 93:7–16.

Marina Sanchez-Torron and Philipp Koehn. 2016. Machine translation quality and post-editor productivity. *AMTA 2016, Vol.*, page 16.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, et al. 2017. Nematus: a toolkit for neural machine translation. *arXiv preprint arXiv:1703.04357*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation (WMT 2016), Volume 2: Shared Task Papers*.

Raivis Skadiņš, Kārlis Goba, and Valters Šics. 2010. Improving SMT for Baltic Languages with Factored Models. In *Human Language Technologies: The Baltic Perspective: Proceedings of the Fourth International Conference, Baltic HLT 2010*, volume 219, pages 125–132. IOS Press.

Inguna Skadiņa, Andrejs Veisbergs, Andrejs Vasiļjevs, Tatjana Gornostaja, Iveta Keiša, and Alda Rudzīte. 2012. *The Latvian language in the digital age*. Springer.

Raivis Skadiņš, Māris Puriņš, Inguna Skadiņa, and Andrejs Vasiļjevs. 2011. Evaluation of SMT in Localization to Under-Resourced Inflected Language. In *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT 2011)*, May, pages 35–40, Leuven, Belgium. European Association for Machine Translation.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th biennial conference of the Association for Machine Translation in the Americas*, August, pages 223–231, Cambridge, MA, USA.

Aleš Tamchyna and Ondřej Bojar. 2013. No Free Lunch in Factored Phrase-Based Machine Translation. In *Proc. of CICLing 2013*, volume 7817 of *LNCS*, pages 210–223, Samos, Greece. Springer-Verlag.

Jörg Tiedemann. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.

Antonio Toral and Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1063–1073, Valencia, Spain. Association for Computational Linguistics.

Andrejs Vasiļjevs, Raivis Skadiņš, and Jörg Tiedemann. 2012. Letsmt!: a cloud-based platform for do-it-yourself machine translation. In *Proceedings of the ACL 2012 System Demonstrations*, pages 43–48. Association for Computational Linguistics.