Neural *versus* Phrase-Based Machine Translation Quality: a Case Study

Luisa Bentivogli
FBK, Trento
Italy

Arianna Bisazza
University of Amsterdam
The Netherlands

Mauro Cettolo
FBK, Trento
Italy

Marcello Federico
FBK, Trento
Italy

Abstract

Within the field of Statistical Machine Translation (SMT), the neural approach (NMT) has recently emerged as the first technology able to challenge the long-standing dominance of phrase-based approaches (PBMT). In particular, at the IWSLT 2015 evaluation campaign, NMT outperformed well established state-of-the-art PBMT systems on English-German, a language pair known to be particularly hard because of morphology and syntactic differences. To understand in what respects NMT provides better translation quality than PBMT, we perform a detailed analysis of neural versus phrase-based SMT outputs, leveraging high quality post-edits performed by professional translators on the IWSLT data. For the first time, our analysis provides useful insights on what linguistic phenomena are best modeled by neural models – such as the reordering of verbs – while pointing out other aspects that remain to be improved.

1 Introduction

The wave of neural models has eventually reached the field of Statistical Machine Translation (SMT). After a period in which Neural MT (NMT) was too computationally costly and resource demanding to compete with state-of-the-art Phrase-Based MT (PBMT)[1], the situation changed in 2015. For the first time, in the latest edition of IWSLT[2] (Cettolo et al., 2015), the system described in (Luong and Manning, 2015) overtook a variety of PBMT approaches with a large margin (+5.3 BLEU points) on a difficult language pair like English-German – anticipating what, most likely, will be the new NMT era.

This impressive improvement follows the distance reduction previously observed in the WMT 2015 shared translation task (Bojar et al., 2015). Just few months before, the NMT systems described in (Jean et al., 2015b) ranked on par with the best phrase-based models on a couple of language pairs. Such rapid progress stems from the improvement of the recurrent neural network encoder-decoder model, originally proposed in (Sutskever et al., 2014; Cho et al., 2014b), with the use of the attention mechanism (Bahdanau et al., 2015). This evolution has several implications. On one side, NMT represents a simplification with respect to previous paradigms. From a management point of view, similar to PBMT, it allows for a more efficient use of human and data resources with respect to rule-based MT. From the architectural point of view, a large recurrent network trained for end-to-end translation is considerably simpler than traditional MT systems that integrate multiple components and processing steps. On the other side, the NMT process is less transparent than previous paradigms. Indeed, it represents a further step in the evolution from rule-based approaches that explicitly manipulate knowledge, to the statistical/data-driven framework, still comprehensible in its inner workings, to a sub-symbolic framework in which the translation process is totally opaque to the analysis.

What do we know about the strengths of NMT

[1] We use the generic term phrase-based MT to cover standard phrase-based, hierarchical and syntax-based SMT approaches.

[2] International Workshop on Spoken Language Translation (http://workshop2015.iwslt.org/)

and the weaknesses of PBMT? What are the linguistic phenomena that deep learning translation models can handle with such greater effectiveness? To answer these questions and go beyond poorly informative BLEU scores, we perform the very first comparative analysis of the two paradigms in order to shed light on the factors that differentiate them and determine their large quality differences.

We build on evaluation data available for the IWSLT 2015 MT English-German task, and compare the results of the first four top-ranked participants. We choose to focus on one language pair and one task because of the following advantages: *(i)* three state-of-the art PBMT systems compared against the NMT system on the same data and in the very same period (that of the evaluation campaign); *(ii)* a challenging language pair in terms of morphology and word order differences; *(iii)* availability of MT outputs' post-editing done by professional translators, which is very costly and thus rarely available. In general, post-edits have the advantage of allowing for informative and detailed analyses since they directly point to translation errors. In this specific framework, the high quality data created by professional translators guarantees reliable evaluations. For all these reasons we present our study as a solid contribution to the better understanding of this new paradigm shift in MT.

After reviewing previous work (Section 2), we introduce the analyzed data and the systems that produced them (Section 3). We then present three increasingly fine levels of MT quality analysis. We first investigate how MT systems' quality varies with specific characteristics of the input, *i.e.* sentence length and type of content of each talk (Section 4). Then, we focus on differences among MT systems with respect to morphology, lexical, and word order errors (Section 5). Finally, based on the finding that word reordering is the strongest aspect of NMT compared to the other systems, we carry out a fine-grained analysis of word order errors (Section 6).

2 Previous Work

To date, NMT systems have only been evaluated by BLEU in single-reference setups (Bahdanau et al., 2015; Sutskever et al., 2014; Luong et al., 2015; Jean et al., 2015a; Gülçehre et al., 2015). Additionally, the Montreal NMT system submitted to WMT 2015 (Jean et al., 2015b) was part of a manual evaluation experiment where a large number of non-professional annotators were asked to rank the outputs of multiple MT systems (Bojar et al., 2015). Results for the Montreal system were very positive – ranked first in English-German, third in German-English, English-Czech and Czech-English – which confirmed and strengthened the BLEU results published so far. Unfortunately neither BLEU nor manual ranking judgements tell us which translation aspects are better modeled by different MT frameworks. To this end, a detailed and systematic error analysis of NMT versus PBMT output is required.

Translation error analysis, as a way to identify systems' weaknesses and define priorities for their improvement, has received a fair amount of attention in the MT community. In this work we opt for the *automatic* detection and classification of translation errors based on *manual* post-edits of the MT output. We believe this choice provides an optimal trade-off between fully manual error analysis (Farrús Cabeceran et al., 2010; Popović et al., 2013; Daems et al., 2014; Federico et al., 2014; Neubig et al., 2015), which is very costly and complex, and fully automatic error analysis (Popović and Ney, 2011; Irvine et al., 2013), which is noisy and biased towards one or few arbitrary reference translations.

Existing tools for translation error detection are either based on Word Error Rate (WER) and Position-independent word Error Rate (PER) (Popović, 2011) or on output-reference alignment (Zeman et al., 2011). Regarding error classification, Hjerson (Popović, 2011) detects five main types of word-level errors as defined in (Vilar et al., 2006): morphological, reordering, missing words, extra words, and lexical choice errors. We follow a similar but simpler error classification (morphological, lexical, and word order errors), but detect the errors differently using TER as this is the most natural choice in our evaluation framework based on post-edits (see also Section 3.4). Irvine et al. (2013) propose another word-level error analysis technique specifically focused on lexical choice and aimed at understanding the effects of domain differences on MT. Their error classification is strictly related to model coverage and insensitive to word order differences. The technique requires access to the sys-

tem's phrase table and is thus not applicable to NMT, which does not rely on a fixed inventory of source-target translation units extracted from the parallel data.

Previous error analyses based on manually post-edited translations were presented in (Bojar, 2011; Koponen, 2012; Popović et al., 2013). We are the first to conduct this kind of study on the output of a neural MT system.

3 Experimental Setting

We perform a number of analyses on data and results of the IWSLT 2015 MT *En-De* task, which consists in translating manual transcripts of English TED talks into German.

Evaluation data are publicly available through the WIT^3 repository (Cettolo et al., 2012).[3]

3.1 Task Data

TED Talks[4] are a collection of rather short speeches (max 18 minutes each, roughly equivalent to 2,500 words) covering a wide variety of topics. All talks have captions, which are translated into many languages by volunteers worldwide. Besides representing a popular benchmark for spoken language technology, TED Talks embed interesting research challenges. Translating TED Talks implies dealing with spoken rather than written language, which is hence expected to be structurally less complex, formal and fluent (Ruiz and Federico, 2014). Moreover, as human translations of the talks are required to follow the structure and rhythm of the English captions, a lower amount of rephrasing and reordering is expected than in ordinary translation of written documents.

As regards the English-German language pair, the two languages are interesting since, while belonging to the same language family, they have marked differences in levels of inflection, morphological variation, and word order, especially long-range reordering of verbs.

3.2 Evaluation Data

Five systems participated in the MT *En-De* task and were manually evaluated on a representative subset of the official 2015 test set. The Human Evaluation (HE) set includes the first half of each of the 12 test talks, for a total of 600 sentences and around 10K words. Five professional translators were asked to post-edit the MT output by applying the minimal edits required to transform it into a fluent sentence with the same meaning as the source sentence. Data were prepared so that all translators equally post-edited the five MT outputs, *i.e.* 120 sentences for each evaluated system.

The resulting evaluation data consist of five new reference translations for each of the sentences in the HE set. Each one of these references represents the *targeted translation* of the system output from which it was derived, but the other four *additional translations* can also be used to evaluate each MT system. We will see in the next sections how we exploited the available post-edits in the more suitable way depending on the kind of analysis carried out.

System	Approach	Data
PBSY (Huck and Birch, 2015)	Combination: Phrase+Syntax-based GHKM string-to-tree; hierarchical + sparse lexicalized reordering models	175M/ 3.1B
HPB (Jehl et al., 2015)	Hierarchical Phrase-based source pre-ordering (dependency tree -based); re-scoring with neural LM	166M/ 854M
SPB (Ha et al., 2015)	Standard Phrase-based source pre-ordering (POS- and tree-based); re-scoring with neural LMs	117M/ 2.4B
NMT (Luong & Manning, 2015)	Recurrent neural network (LSTM) attention-based; source reversing; rare words handling	120M/ –

Table 1: MT systems' overview. Data column: size of parallel/monolingual training data for each system in terms of English and German tokens.

3.3 MT Systems

Our analysis focuses on the first four top-ranking systems, which include NMT (Luong and Manning, 2015) and three different phrase-based approaches: standard phrase-based (Ha et al., 2015), hierarchical (Jehl et al., 2015) and a combination of phrase-based and syntax-based (Huck and Birch, 2015). Table 1 presents an overview of each system, as well as figures about the training data used.[5]

The phrase+syntax-based (PBSY) system com-

[3] wit3.fbk.eu

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

[5] Detailed information about training data was kindly made available by participating teams.

bines the outputs of a string-to-tree decoder, trained with the GHKM algorithm, with those of two standard phrase-based systems featuring, among others, adapted phrase tables and language models enriched with morphological information, hierarchical lexicalized reordering models and different variations of the operational sequence model.

The hierarchical phrase-based MT (HPB) system leverages thousands of lexicalised features, data-driven source pre-ordering (dependency tree-based), word-based and class-based LMs, and n-best re-scoring models based on syntactic and neural LMs.

The standard phrase-based MT (SPB) system features an adapted phrase-table combining in-domain and out-domain data, discriminative word lexicon models, multiple language models (word-, POS- and class-based), data-driven source pre-ordering (POS- and constituency syntax-based), n-best re-scoring models based on neural lexicons and neural LMs.

Finally, the neural MT (NMT) system is a 4-layer long short-term memory (LSTM) network featuring 1,000-dimension word embeddings, attention mechanism, source reversing, 50K source and target vocabularies, and out-of-vocabulary word handling. Training with TED data was performed on top of models trained with large out-domain parallel data.

With respect to the use of training data, it is worth noticing that NMT is the only system not employing monolingual data in addition to parallel data. Moreover, NMT and SPB were trained with smaller amounts of parallel data with respect to PBSY and HPB (see Table 1).

3.4 Translation Edit Rate Measures

The *Translation Edit Rate* (TER) (Snover et al., 2006) naturally fits our evaluation framework, where it traces the edits done by post-editors. Also, TER *shift* operations are reliable indicators of reordering errors, in which we are particularly interested. We exploit the available post-edits in two different ways: *(i)* for *Human-targeted TER* (HTER) we compute TER between the machine translation and its manually post-edited version (targeted reference), *(ii)* for *Multi-reference TER* (mTER), we compute TER against the closest translation among all available post-edits (*i.e.* targeted and additional references) for each sentence.

Throughout sections 4 and 5, we mark a score achieved by NMT with the symbol * if this is better than the score of its best competitor at statistical significance level 0.01. Significance tests for HTER and mTER are computed by bootstrap re-sampling, while differences among proportions are assessed via one-tailed z-score tests.

system	BLEU	HTER	mTER
PBSY	25.3	28.0	21.8
HPB	24.6	29.9	23.4
SPB	25.8	29.0	22.7
NMT	31.1*	21.1*	16.2*

Table 2: Overall results on the HE Set: BLEU, computed against the original reference translation, and TER, computed with respect to the targeted post-edit (HTER) and multiple post-edits (mTER).

4 Overall Translation Quality

Table 2 presents overall system results according to HTER and mTER, as well as BLEU computed against the original TED Talks reference translation. We can see that NMT clearly outperforms all other approaches both in terms of BLEU and TER scores. Focusing on mTER results, the gain obtained by NMT over the second best system (PBSY) amounts to 26%. It is also worth noticing that mTER is considerably lower than HTER for each system. This reduction shows that exploiting all the available post-edits as references for TER is a viable way to control and overcome post-editors variability, thus ensuring a more reliable and informative evaluation about the real overall performance of MT systems. For this reason, the two following analyses rely on mTER. In particular, we investigate how specific characteristics of input documents affect the system's overall translation quality, focusing on *(i)* sentence length and *(ii)* the different talks composing the dataset.

4.1 Translation quality by sentence length

Long sentences are known to be difficult to translate by the NMT approach. Following previous work (Cho et al., 2014a; Pouget-Abadie et al., 2014; Bahdanau et al., 2015; Luong et al., 2015), we investigate how sentence length affects overall translation quality. Figure 1 plots mTER scores against source sentence length. NMT clearly outperforms every PBMT system in any length bin, with statistically significant differences. As a general tendency, the

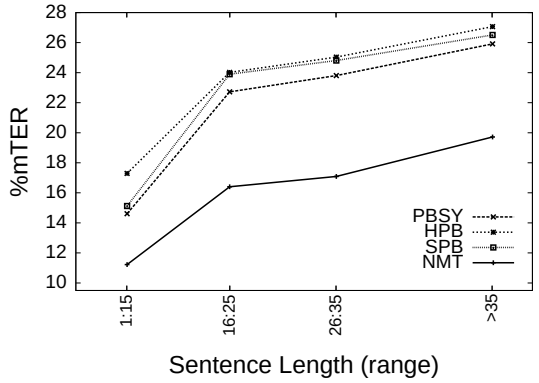

Figure 1: mTER scores on bins of sentences of different length. Points represent the average mTER of the MT outputs for the sentences in each given bin.

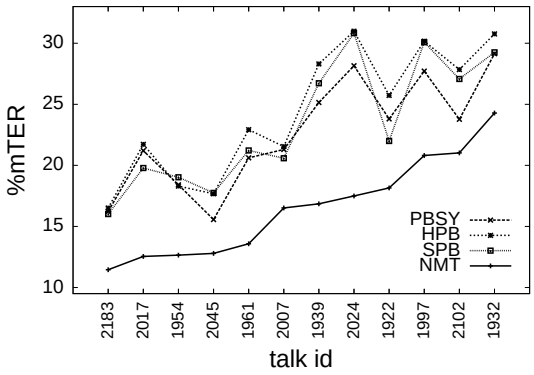

Figure 2: mTER scores per talk, sorted in ascending order of NMT scores.

performance of all approaches worsens as sentence length increases. However, for sentences longer than 35 words we see that NMT quality degrades more markedly than in PBMT systems. Considering the percentage decrease with respect to the preceding length bin (26-35), we see that the $\%\Delta$ for NMT (-15.4) is much larger than the average $\%\Delta$ for the three PBMT systems (-7.9). Hence, this still seems an issue to be addressed for further improving NMT.

4.2 Translation quality by talk

As we saw in Section 3.1, the TED dataset is very heterogeneous since it consists of talks covering different topics and given by speakers with different styles. It is therefore interesting to evaluate translation quality also at the talk level.

Figure 2 plots the mTER scores for each of the twelve talks included in the HE set, sorted in ascending order of NMT scores. In all talks, the NMT system outperforms the PBMT systems in a statistically significant way.

We analysed different factors which could impact translation quality in order to understand if they correlate with such performance differences. We studied three features which are typically considered as indicators of complexity (see (François and Fairon, 2012) for an overview), namely *(i)* the length of the talk, *(ii)* its average sentence length, and *(iii)* the type-token ratio[6] (TTR) which – measuring lexical diversity – reflects the size of a speaker's vocabulary and the variety of subject matter in a text.

For the first two features we did not find any correlation; on the contrary, we found a moderate Pearson correlation (R=0.7332) between TTR and the mTER gains of NMT over its closest competitor in each talk. This result suggests that NMT is able to cope with lexical diversity better than any other considered approach.

5 Analysis of Translation Errors

We now turn to analyze which types of linguistic errors characterize NMT *vs.* PBMT. In the literature, various error taxonomies covering different levels of granularity have been developed (Flanagan, 1994; Vilar et al., 2006; Farrús Cabeceran et al., 2010; Stymne and Ahrenberg, 2012; Lommel et al., 2014). We focus on three error categories, namely *(i)* morphology errors, *(ii)* lexical errors, and *(iii)* word order errors. As for lexical errors, a number of existing taxonomies further distinguish among translation errors due to missing words, extra words, or wrong lexical choice. However, given the proved difficulty of disambiguating between these three subclasses (Popović and Ney, 2011; Fishel et al., 2012), we prefer to rely on a more coarse-grained linguistic error classification where lexical errors include all of them (Farrús Cabeceran et al., 2010).

For error analysis we rely on HTER results under the assumption that, since the targeted translation is generated by post-editing the given MT output, this method is particularly informative to spot MT errors.

[6]The type-token-ratio of a text is calculated dividing the number of word types (vocabulary) by the total number of word tokens (occurrences).

We are aware that translators' subjectivity is still an issue (see Section 4), however in this more fine-grained analysis we prefer to focus on what a human implicitly annotated as a translation error. This particularly holds in our specific evaluation framework, where the goal is not to measure the absolute number of errors made by each system, but to compare systems among each other. Moreover, the post-edits collected for each MT output within IWSLT allow for a fair and reliable comparison since systems were equally post-edited by all translators (see Section 3.2), making all analyses uniformly affected by such variability.

5.1 Morphology errors

A morphology error occurs when a generated word form is wrong but its corresponding base form (lemma) is correct. Thus, we assess the ability of systems to deal with morphology by comparing the HTER score computed on the surface forms (*i.e.* morphologically inflected words) with the HTER score obtained on the corresponding lemmas. The additional matches counted on lemmas with respect to word forms indicate morphology errors. Thus, the closer the two HTER scores, the more accurate the system in handling morphology.

To carry out this analysis, the lemmatized (and POS tagged) version of both MT outputs and corresponding post-edits was produced with the German parser ParZu (Sennrich et al., 2013). Then, the HTER-based evaluation was slightly adapted in order to be better suited to an accurate detection of morphology errors. First, punctuation was removed since – not being subject to morphological inflection – it could smooth the results. Second, *shift* errors were not considered. A word form or a lemma that matches a corresponding word or lemma in the post-edit, but is in the wrong position with respect to it, is counted as a *shift* error in TER. Instead – when focusing on morphology – exact matches are not errors, regardless their position in the text.[7]

Table 3 presents HTER scores on word forms and lemmas, as well as their percentage difference which gives an indication of morphology errors. We can see that NMT generates translations which are morphologically more correct than the other systems. In particular, the %Δ for NMT (-13.7) is lower than that of the second best system (PBSY, -16.9) by 3.2% absolute points, leading to a percentage gain of around 19%. We can thus say that NMT makes at least 19% less morphology errors than any other PBMT system.

system	HTERnoShft		
	word	lemma	%Δ
PBSY	27.1	22.5	-16.9
HPB	28.7	23.5	-18.4
SPB	28.3	23.2	-18.0
NMT	21.7*	18.7*	-13.7

Table 3: HTER ignoring shift operations computed on words and corresponding lemmas, and their % difference.

5.2 Lexical errors

Another important feature of MT systems is their ability to choose lexically appropriate words. In order to compare systems under this aspect, we consider HTER results at the lemma level as a way to abstract from morphology errors and focus only on actual lexical choice problems. The evaluation on the lemmatised version of the data performed to identify morphology errors fits to this purpose, since its driving assumptions (*i.e.* punctuation can be excluded and lemmas in the wrong order are not errors) hold for lexical errors too.

The lemma column of Table 3 shows that NMT outperforms the other systems. More precisely, the NMT score (18.7) is better than the second best (PBSY, 22.5) by 3.8% absolute points. This corresponds to a relative gain of about 17%, meaning that NMT makes at least 17% less lexical errors than any PBMT system. Similarly to what observed for morphology errors, this can be considered a remarkable improvement over the state of the art.

5.3 Word order errors

To analyse reordering errors, we start by focusing on *shift* operations identified by the HTER metrics. The first three columns of Table 4 show, respectively: *(i)* the number of words generated by each system *(ii)* the number of shifts required to align each system output to the corresponding post-edit; and *(iii)* the corresponding percentage of shift errors. Notice that the *shift* error percentages are incorporated in

[7] Note that the TER score calculated by setting to 0 the cost of shifts approximates the Position-independent Error Rate (Tillmann et al., 1997).

system	#words	#shifts	%shifts	KRS
PBSY	11,517	354	3.1	84.6
HPB	11,417	415	3.6	84.3
SPB	11,420	398	3.5	84.5
NMT	11,284	173	1.5*	88.3*

Table 4: Word reordering evaluation in terms of shift operations in HTER calculation and of KRS. For each system, the number of generated words, the number of shift errors and their corresponding percentages are reported.

the HTER scores reported in Table 2.

We can see in the table that the percentage of *shift* errors in NMT translations is definitely lower than for the other systems. The gap between NMT and the second best system (PBSY) is about 50%.

It should be recalled that these numbers only refer to *shifts* detected by HTER, that is (groups of) words of the MT output and post-edits that are identical but occurring in different positions. Words that had to be moved and modified at the same time (for instance replaced by a synonym or a morphological variant) are not counted in HTER *shift* figures, but are detected as *substitution*, *insertion* or *deletion* operations. To ensure that our reordering evaluation is not biased towards the alignment between the MT output and the post-edit performed by HTER, we run an additional assessment using KRS, or Kendall Reordering Score (Birch et al., 2010), which measures the similarity between the source-reference reorderings and the source-MT output reorderings.[8] Being based on bilingual word alignment via the source sentence, KRS detects reordering errors also when post-edit and MT words are not identical. Also unlike TER, KRS is sensitive to the *distance* between the position of a word in the MT output and that in the reference.

Looking at the last column of Table 4, we can say that our observations on HTER are confirmed by the KRS results: the reorderings performed by NMT are much more accurate than those performed by any PBMT system.[9] Moreover, according to the approximate randomization test, KRS differences are statistically significant between NMT and all other systems, but not among the three PBMT systems.

Given the concordant results of our two quantitative analyses, we conclude that one of the major strengths of the NMT approach is its ability to place German words in the right position even when this requires considerable reordering. This outcome calls for a deeper investigation, which is carried out in the following section.

6 Fine-grained Word Order Error Analysis

We have observed that word reordering is a very strong aspect of NMT compared to PBMT, according to both HTER and KRS. To better understand this finding, we investigate whether reordering errors concentrate on specific linguistic constructions across our systems. Using the POS tagging and dependency parsing of the post-edits produced by ParZu, we classify the *shift* operations detected by HTER and count how often a word with a given POS label was misplaced by each of the systems (alone or as part of a shifted block). For each word class, we also compute the percentage order error reduction of NMT with respect to the PBMT system that has highest reordering accuracy overall, that is PBSY. Results are presented in Table 5, ranked by NMT-vs-PBSY gain. We omit punctuation as well as word classes that were shifted less than 10 times by all systems. Examples of salient word order error types are presented in Table 6.

The upper part of Table 5 shows that verbs are by far the most often misplaced word category in all PBMT systems – an issue already known to affect standard phrase-based SMT between German and English (Bisazza and Federico, 2013). Reordering is particularly difficult when translating *into* German, since the position of verbs in this language varies according to the clause type (*e.g.* main versus subordinate). Our results show that even syntax-informed PBMT does not solve this issue. Using syntax at decoding time, as done by one of the systems combined within PBSY, appears to be a better strategy than using it for source pre-ordering, as done by the HPB and SPB systems. However this only results in a moderate reduction of verb reordering errors (-12% and -25% versus HPB and SPB

[8] To compute the word alignments required by KRS, we used the FastAlign tool (Dyer et al., 2013).

[9] To put our results into perspective, note that Birch (2011) reports a difference of 5 KRS points between the translations of a PBMT system and those produced by four *human* translators tested against each other, in a Chinese-English experiment.

Auxiliary-main verb construction [aux:V]:			
	SRC	in this experiment , individuals **were shown** hundreds of hours of YouTube videos	
(a)	HPB	in diesem Experiment , Individuen **gezeigt wurden** Hunderte von Stunden YouTube-Videos	✗
	PE	in diesem Experiment **wurden** Individuen Hunderte von Stunden Youtube-Videos **gezeigt**	
	NMT	in diesem Experiment **wurden** Individuen hunderte Stunden YouTube Videos **gezeigt**	✓
	PE	in diesem Experiment **wurden** Individuen hunderte Stunden YouTube Videos **gezeigt**	
Verb in subordinate (adjunct) clause [neb:V]:			
	SRC	... when coaches and managers and owners **look** at this information streaming ...	
(b)	PBSY	... wenn Trainer und Manager und Eigentümer **betrachten** diese Information Streaming ...	✗
	PE	... wenn Trainer und Manager und Eigentümer dieses Informations-Streaming **betrachten** ...	
	NMT	... wenn Trainer und Manager und Besitzer sich diese Informationen **anschauen** ...	✓
	PE	... wenn Trainer und Manager und Besitzer sich diese Informationen **anschauen** ...	
Prepositional phrase [pp:PREP det:ART pn:N] acting as temporal adjunct:			
	SRC	so like many of us , I 've lived in a few closets **in my life**	
(c)	SPB	so wie viele von uns , ich habe in ein paar Schränke **in meinem Leben** gelebt	✗
	PE	so habe ich wie viele von uns **während meines Lebens** in einigen Verstecken gelebt	
	NMT	wie viele von uns habe ich in ein paar Schränke **in meinem Leben** gelebt	✗
	PE	wie viele von uns habe ich **in meinem Leben** in ein paar Schränken gelebt	
Negation particle [adv:PTKNEG]:			
	SRC	but I eventually came to the conclusion that that just did **not** work for systematic reasons	
(d)	HPB	aber ich kam schlielich zu dem Schluss , dass nur aus systematischen Gründen **nicht** funktionieren	✓
	PE	aber ich kam schlielich zu dem Schluss , dass es einfach aus systematischen Gründen **nicht** funktioniert	
	NMT	aber letztendlich kam ich zu dem Schluss , dass das einfach **nicht** aus systematischen Gründen funktionierte	✗
	PE	ich musste aber einsehen , dass das aus systematischen Gründen **nicht** funktioniert	

Table 6: MT output and post-edit examples showing common types of reordering errors.

respectively). On the contrary, NMT reduces verb order errors by an impressive -70% with respect to PBSY (-74% and -77% versus HPB and SPB respectively) despite being trained on raw parallel data without any syntactic annotation, nor explicit modeling of word reordering. This result shows that the recurrent neural language model at the core of the NMT architecture is very successful at generating well-formed sentences even in languages with less predictable word order, like German (see examples in Table 6(a,b)). NMT, though, gains notably less on nouns (-47%), which is the second most often misplaced word category in PBSY. More insight on this is provided by the lower part of the table, where reordering errors are divided by their dependency label as well as POS tag. Here we see that order errors on nouns are notably reduced by NMT when they act as syntactic objects (-65% obja:N) but less when they act as preposition complements (-36% pn:N) or subjects (-33% subj:N).

The smallest NMT-vs-PBSY gains are observed on prepositions (-18% pp:PREP), negation particles (-17% PTKNEG) and articles (-4% det:ART). Manual inspection of a data sample reveals that misplaced prepositions are often part of misplaced prepositional phrases acting, for instance, as temporal or instrumental adjuncts (e.g. *'in my life'*, *'with this video'*). In these cases, the original MT output is overall understandable and grammatical, but does not conform to the order of German semantic arguments that is consistently preferred by post-editors (see example in Table 6(c)). Articles, due to their commonness, are often misaligned by HTER and marked as *shift* errors instead of being marked as two unrelated substitutions. Finally, negation particles account for less than 1% of the target tokens but play a key role in determining the sentence meaning. Looking closely at some error examples, we found that the correct placement of the German particle *nicht* was determined by the focus of negation in the source sentence, which is difficult to detect in English. For instance in Table 6(d) two interpretations are possible (*'that did not work'* or *'that worked, but not for systematic reasons'*), each resulting in a dif-

Class	NMT-vs-PBSY	NMT	PBSY	HPB	SPB
V	-70%	35	116	133	155
PRO	-57%	22	51	53	62
PTKZU	-54%	6	13	4	11
ADV	-50%	14	28	44	36
N	-47%	37	70	99	56
KON	-33%	6	9	8	12
PREP	-18%	18	22	27	28
PTKNEG	-17%	10	12	10	7
ART	-4%	26	27	38	35
aux:V	-87%	3	23	17	18
neb:V	-83%	2	12	7	19
objc:V	-79%	3	14	21	24
subj:PRO	-70%	12	40	34	46
root:V	-68%	6	19	28	27
adv:ADV	-67%	8	24	33	28
obja:N	-65%	6	17	28	12
cj:V	-59%	7	17	21	22
part:PTKZU	-54%	6	13	4	11
obja:PRO	-38%	5	8	14	7
mroot:V	-36%	7	11	26	20
pn:N	-36%	16	25	33	19
subj:N	-33%	6	9	10	7
pp:PREP	-30%	14	20	19	23
adv:PTKNEG	-17%	10	12	10	7
det:ART	-4%	26	27	38	34
all	-48%	222	429	493	488

Table 5: Main POS tags and dependency labels of words occurring in shifted blocks detected by HTER. NMT-vs-PBSY denotes the reduction of reordering errors in NMT versus PBSY system. Only word classes that were shifted 10 or more times in at least one system output are shown.

ferent, but equally grammatical, location of *nicht*. In fact, negation-focus detection calls for a deep understanding of the sentence semantics, often requiring extra-sentential context (Blanco and Moldovan, 2011). When faced with this kind of translation decisions, NMT performs as bad as its competitors.

In summary, our fine-grained analysis confirms that NMT concentrates its word order improvements on important linguistic constituents and, specifically in English-German, is very close to solving the infamous problem of long-range verb reordering which so many PBMT approaches have only poorly managed to handle. On the other hand, NMT still struggles with more subtle translation decisions depending, for instance, on the semantic ordering of adjunct prepositional phrases or on the focus of negation.

7 Conclusions

We have analysed the output of four state-of-the-art MT systems that participated in the English-to-German task of the IWSLT 2015 evaluation campaign. Our selected runs were produced by three phrase-based MT systems and a neural MT system. The analysis leveraged high quality post-edits of the MT outputs, which allowed us to profile systems with respect to reliable measures of post-editing effort and translation error types.

The outcomes of the analysis confirm that NMT has significantly pushed ahead the state of the art, especially in a language pair involving rich morphology prediction and significant word reordering. To summarize our findings: *(i)* NMT generates outputs that considerably lower the overall post-edit effort with respect to the best PBMT system (-26%); *(ii)* NMT outperforms PBMT systems on all sentence lengths, although its performance degrades faster with the input length than its competitors; *(iii)* NMT seems to have an edge especially on lexically rich texts; *(iv)* NMT output contains less morphology errors (-19%), less lexical errors (-17%), and substantially less word order errors (-50%) than its closest competitor for each error type; *(v)* concerning word order, NMT shows an impressive improvement in the placement of verbs (-70% errors).

While NMT has proved superior to PBMT with respect to all error types that we have investigated, our analysis has also pointed out some aspects of NMT that deserve further work, such as the handling of long sentences and the reordering of particular linguistic constituents requiring a deep semantic understanding of text. Machine translation is definitely not a solved problem, but the time is finally ripe to tackle its most intricate aspects.

Acknowledgments

FBK authors were supported by the CRACKER and QT21 projects, which received funding from the European Unions Horizon 2020 research and innovation programme under grant no. 645357 and no. 645452, respectively. AB's work was funded in part by the Netherlands Organisation for Scientific Research (NWO) under project numbers 639.022.213 and 612.001.218.

References

[Bahdanau et al.2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR*, San Diego, US-CA.

[Birch et al.2010] Alexandra Birch, Miles Osborne, and Phil Blunsom. 2010. Metrics for MT evaluation: evaluating reordering. *Machine Translation*, 24(1):15–26.

[Birch2011] Alexandra Birch. 2011. *Reordering Metrics for Statistical Machine Translation*. Ph.D. thesis, School of Informatics, University of Edinburgh, UK.

[Bisazza and Federico2013] Arianna Bisazza and Marcello Federico. 2013. Efficient solutions for word reordering in German-English phrase-based statistical machine translation. In *Proc. of WMT*, Sofia, Bulgaria.

[Blanco and Moldovan2011] Eduardo Blanco and Dan Moldovan. 2011. Semantic representation of negation using focus detection. In *Proc. of ACL-HLT*, Portland, US-OR.

[Bojar et al.2015] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proc. of WMT*, Lisbon, Portugal.

[Bojar2011] Ondrej Bojar. 2011. Analyzing error types in English-Czech machine translation. *The Prague Bulletin of Mathematical Linguistic*, (95):63–76.

[Cettolo et al.2012] Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proc. of EAMT*, Trento, Italy.

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

[Cho et al.2014a] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014a. On the properties of neural machine translation: encoder–decoder approaches. In *Proc. of SSST-8*, Doha, Qatar.

[Cho et al.2014b] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014b. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proc. of EMNLP*, Doha, Qatar.

[Daems et al.2014] Joke Daems, Lieve Macken, and Sonia Vandepitte. 2014. On the origin of errors: a fine-grained analysis of MT and PE errors and their relationship. In *Proc. of LREC*, Reykjavik, Iceland.

[Dyer et al.2013] Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proc. of NACL-HLT*, Atlanta, US-GA.

[Farrús Cabeceran et al.2010] Mireia Farrús Cabeceran, Marta Ruiz Costa-Jussà, José Bernardo Mariño Acebal, and José Adrián Rodríguez Fonollosa. 2010. Linguistic-based evaluation criteria to identify statistical machine translation errors. In *Proc. of EAMT*, Saint-Raphaël, France.

[Federico et al.2014] Marcello Federico, Matteo Negri, Luisa Bentivogli, and Marco Turchi. 2014. Assessing the impact of translation errors on machine translation quality with mixed-effects models. In *Proc. of EMNLP*, Doha, Qatar.

[Fishel et al.2012] Mark Fishel, Ondrej Bojar, and Maja Popović. 2012. Terra: a collection of translation error-annotated corpora. In *Proc. of LREC*, Istanbul, Turkey.

[Flanagan1994] Mary Flanagan. 1994. Error classification for MT evaluation. In *Proc. of AMTA*, Columbia, US-MD.

[François and Fairon2012] Thomas François and Cédrick Fairon. 2012. An "AI readability" formula for French as a foreign language. In *Proc. of EMNLP-CoNLL*, Jeju Island, Korea.

[Gülçehre et al.2015] Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535.

[Ha et al.2015] Thanh-Le Ha, Jan Niehues, Eunah Cho, Mohammed Mediani, and Alex Waibel. 2015. The KIT translation systems for IWSLT 2015. In *Proc. of IWSLT*, Da Nang, Vietnam.

[Huck and Birch2015] Matthias Huck and Alexandra Birch. 2015. The Edinburgh machine translation systems for IWSLT 2015. In *Proc. of IWSLT*, Da Nang, Vietnam.

[Irvine et al.2013] Ann Irvine, John Morgan, Marine Carpuat, Hal Daumé III, and Dragos Munteanu. 2013. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics*, 1:429–440.

[Jean et al.2015a] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015a. On using very large target vocabulary for neural machine translation. In *Proc. of ACL-IJCNLP*, Beijing, China.

[Jean et al.2015b] Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015b. Montreal neural machine translation systems for WMT15. In *Proc. of WMT*, Lisbon, Portugal.

[Jehl et al.2015] Laura Jehl, Patrick Simianer, Julian Hitschler, and Stefan Riezler. 2015. The Heidelberg university English-German translation system for IWSLT 2015. In *Proc. of IWSLT*, Da Nang, Vietnam.

[Koponen2012] Maarit Koponen. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In *Proc. of WMT*, Montréal, Canada.

[Lommel et al.2014] Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proc. of EAMT*, Dubrovnik, Croatia.

[Luong and Manning2015] Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proc. of IWSLT*, Da Nang, Vietnam.

[Luong et al.2015] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proc. of EMNLP*, Lisbon, Portugal.

[Neubig et al.2015] Graham Neubig, Makoto Morishita, and Satoshi Nakamura. 2015. Neural Reranking Improves Subjective Quality of Machine Translation: NAIST at WAT2015. In *Proc. of WAT2015*, Kyoto, Japan.

[Popović and Ney2011] Maja Popović and Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688.

[Popović et al.2013] Maja Popović, Eleftherios Avramidis, Aljoscha Burchardt, Sabine Hunsicker, Sven Schmeier, Cindy Tscherwinka, David Vilar, and Hans Uszkoreit. 2013. Learning from human judgments of machine translation output. In *Proc. of MT Summit*, Nice, France.

[Popović2011] Maja Popović. 2011. Hjerson: an open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistic*, (96):59–68.

[Pouget-Abadie et al.2014] Jean Pouget-Abadie, Dzmitry Bahdanau, Bart van Merrienboer, Kyunghyun Cho, and Yoshua Bengio. 2014. Overcoming the curse of sentence length for neural machine translation using automatic segmentation. In *Proc. of SSST-8*, Doha, Qatar.

[Ruiz and Federico2014] Nicholas Ruiz and Marcello Federico. 2014. Complexity of spoken versus written language for machine translation. In *Proc. of EAMT*, Dubrovnik, Croatia.

[Sennrich et al.2013] Rico Sennrich, Martin Volk, and Gerold Schneider. 2013. Exploiting synergies between open resources for German dependency parsing, POS-tagging, and morphological analysis. In *Proc. of RANLP*, Hissar, Bulgaria.

[Snover et al.2006] Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of AMTA*, Boston, US-MA.

[Stymne and Ahrenberg2012] Sara Stymne and Lars Ahrenberg. 2012. On the practice of error analysis for machine translation evaluation. In *Proc. of LREC*, Istanbul, Turkey.

[Sutskever et al.2014] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proc. of NIPS*, Montréal, Canada.

[Tillmann et al.1997] Christoph Tillmann, Stephan Vogel, Hermann Ney, Alexander Zubiaga, and Hassan Sawaf. 1997. Accelerated DP based search for statistical translation. In *Proc. of Eurospeech*, Rhodes, Greece.

[Vilar et al.2006] David Vilar, Jia Xu, Luis Fernando d'Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proc. of LREC*, Genoa, Italy.

[Zeman et al.2011] Daniel Zeman, Mark Fishel, Jan Berka, and Ondrej Bojar. 2011. Addicter: what is wrong with my translations? *The Prague Bulletin of Mathematical Linguistic*, (96):79–88.