Observing the Learning Curve of Neural Machine Translation with regard to Linguistic Phenomena

Patrick Stadler, Vivien Macketanz and **Eleftherios Avramidis**
German Research Center for Artificial Intelligence (DFKI), Berlin
firstname.lastname@dfki.de

Abstract

In this paper we present our observations and evaluations by observing the linguistic performance of the system on several steps on the training process of various English-to-German Neural Machine Translation models. The linguistic performance is measured through a semi-automatic process using a test suite. Among several linguistic observations, we find that the translation quality of some linguistic categories decreased within the recorded iterations. Additionally, we notice some drops of the translation quality of certain categories when using a larger corpus.

1 Introduction

During the last years, neural machine translation (NMT) has seen immense progress and achieved high performance. As most machine learning methods, NMT is based on an iterative training process that *learns to translate* given big amounts of parallel corpora. Despite the remarkable achievements of the training process in terms of producing a model able to translate, it is used as a black box. This is also due to the fact that it is training a neural network, one of the least interpretable machine learning algorithms. Thus, little effort has been done in order to investigate how the training process evolves with regards to measurable factors of translation quality, such as the rules of linguistic correctness (grammar, syntax, semantics).

In particular, the training process performs several iterations through which the neural network weights are gradually adjusted to achieve the optimal performance for the training data seen at the moment. After several iterations, the performance of the model, with its current weights, is typically validated against a development set, using some automatic metrics (cross entropy or BLEU score; Papineni et al., 2002), which may also define whether the optimal conditions have been reached and training should stop. Although these automatic metrics have been proven useful for the training process itself, they provide a single number for a generic notion of the translation quality. As specified, we are interested in observing the training process from a more fine-grained perspective and particularly how it proceeds with learning specific linguistic phenomena.

This work is intended to provide NMT researchers and engineers with additional guidance on what to look for when evaluating and designing machine translation systems. This is a preliminary work towards this direction, aiming to investigate how the training process evolves with regards to linguistic performance for several phenomena. We do this by selecting snapshots of particular training epochs and evaluating these snapshots with test suites, which probe the translation of specific linguistic phenomena.

As a result, we can observe the learning curve of those linguistic aspects, along with strengths and weaknesses. We find that as the training ends and the BLEU score reaches the maximum value, some linguistic categories experience a drop in their accuracy. Additionally, we notice further drops of the translation quality of certain categories when using a larger corpus. Finally, we provide further observations on particular linguistic phenomena, by focusing on certain test items. Our experiment is focusing on the language direction English→German.

In the next section (section 2) we review related work. Section 3 presents our used methods, while in section 4 the experiment setup is further discussed. We present our results in section 5 and compare the different models in section 6, followed by a short conclusion and notes on further work in section 7.

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 186–196
August 5–6, 2021. ©2021 Association for Computational Linguistics

2 Related work

2.1 Interpreting NMT with regards to linguistic phenomena

There have been several efforts to interpret the operation of NMT with regards to linguistic phenomena. These works mostly focus on identifying which parts of the neural topology are responsible for learning some particular linguistic aspects. For example they investigate the role of particular neurons (Bau et al., 2019), layers, major components such as the encoder and the decoder (Dalvi et al., 2017; Tang et al., 2019; Belinkov et al., 2020), or different architectures (Tang et al., 2020) with regards to word sense disambiguation and semantics, morphology, long range dependencies and syntax, etc. Contrary to these works, our consideration of the linguistic aspects is not focusing on the elements of the neural network, but on its timely development during the training process.

Recognising the limitations of scoring with cross-entropy or BLEU score, two papers have proposed scoring based on more focused metrics, such as semantic similarity (Wieting et al., 2019) and adequacy (Kong et al., 2018). Here, we are not interested in finding a linguistic metric to improve the training process, but to apply a fine-grained linguistic analysis to the several stages of the training process and make observations.

2.2 Fine-grained evaluation using test suites

Despite the widespread usage of BLEU score, there have been critical voices from the translation community on its role. As stated by Callison-Burch et al. (2006), BLEU sometimes does not reflect improvement in the quality of the produced translations and therefore is not always a reliable metric to rate a system overall. They showed that BLEU score allows for a certain variance and is often unreliable or inconsistent compared to human analysis especially when one is examining linguistic phenomena on a fine grained level (Avramidis et al., 2019).

To overcome the disadvantages and instabilities of the BLEU score, researches have suggested the utilisation of test suites. Such test suites can report scores either through manual (Ahrenberg, 2018; Koh et al., 2001) or semi-automatic evaluation. Semi-automatic evaluation uses certain metrics to be tested against, such as reference translations with specific tokens (Guillou and Hardmeier, 2016; Macketanz et al., 2018a). Another important aspect for using test suites instead of relying solely on automatic evaluation, is the domain-knowledge that only human judges can provide and is required to to assess the translation quality (Vojtěchová et al., 2019).

3 Methods

We are interested in observing the learning curve of neural machine translation with regards to linguistic phenomena. Particularly, the aim is to examine how the linguistic performance of a translation model improves along the iterations of the training process. In order to do that, we perform the following steps:

- We train a neural machine translation system.
- We save the state of the translation model after every epoch of the training process.
- We select some epochs of interest (snapshots) based on the BLEU score of the epoch validation on the development set.
- We perform fine-grained evaluation for every snapshot using a linguistically motivated test suite.

By comparing the statistics from the fine-grained evaluation for various snapshots, we intend to get insights with a linguistic perspective in the machine learning process. We can only evaluate particular snapshots, since the functioning of the test suite tool allows semi-automatic error annotation and there is still need to manually evaluate some uncertain decisions and edge cases. To decide which snapshots to pick, we relied on the use of BLEU score as a first indicator, despite its limitations.

Additionally, we build several systems with different architectures and corpus sizes to allow further comparisons. This being a student experiment, the computational and time restrictions allowed a limited number of models trained with an amount of data that is smaller than the state-of-the-art. However, that should serve as proof of concept. Despite the models not being state-of-the-art, our focus remains on the evolution of the linguistic performance, starting from the early steps of the training process. In our experiments we will have three systems: a small RNN model trained on a small amount of corpora, a bigger RNN model with more data than the former, and a transformer model. Technical details for these models are given further in Section 4.3.

3.1 Different neural machine translation models

We trained several models in order to understand the impact of corpus sizes and the architectures to the linguistic performance. A first run using a RNN architecture (Bahdanau et al., 2014) examines the development of the translation quality based on a relatively small corpus (RNN-small). A succeeding run uses the same model type and arguments but utilises a larger corpus (RNN-big). This allows for more direct comparison and helps to understand the impact of the selected data size. To be able to examine the importance of the selected model type and be closer to the state-of-the-art, we trained a transformer system (Vaswani et al., 2017).

3.2 Fine grained evaluation with a test suite

For the fine-grained evaluation of the trained systems performance, we used a test suite similar to Avramidis et al. (2019). As opposed to an outright human evaluation or the sole use of automatic metrics, the test suite relies on automated evaluation based on manually provided rules. Therefore, regular expressions are applied to manually devised test sentences with several linguistic phenomena grouped into categories. Based on the regular expressions, the test suite can then evaluate the linguistic phenomena, strictly by the presence, respectively, absence of certain key terms and phrases, such as false friends or the use of a wrong tense. The score of a system is then presented as the accuracy across the selected phenomena.

The construction of the test suite and the organization of the categories do not follow a specific linguistic theory and we do not claim a full coverage of the whole linguistic spectrum. Other pieces of research may have different categorization, for example unlike other test suites, we include pronouns under the co-reference phenomenon in the category of non-verbal agreement.

4 Experiment setup

4.1 Test suite setup

For the development and application of the test suite we used the tool TQ-AutoTest (Macketanz et al., 2018a). We created 10 sentences per phenomenon, resulting in a total of 585 sentences, examining 49 phenomena organised in 13 categories. The raw test items, as well as the translations evaluated can be found in our repository[1]. The phenomena selected for this experiment are a subset of the ones of German→English MT, as described in Macketanz et al. (2018b) and Avramidis et al. (2020), adapted to the opposite language direction. An extract of the used sentences can be found in table 5.

System name	RNN-small	RNN-big	transf.
Training datasets	europarl	europarl DGT	europarl DGT
Dataset size	1.8M	7M	7M
Vocab size	32000	32000	32000
Mini-Batch-Fit	5000	5000	10000
Learning rate	0.001	0.001	0.003
Encoder depth	1	1	6
Decoder depth	1	1	6
Beam size	6	6	12
Validation freq.	10000	10000	10000
Dropout	0.2	0.2	0.1
Dropout Source	0.1	0.1	
Dropout Target	0.1	0.1	
Transf. heads			8
Early stopping	5	5	10
BLEU min	1.31	5.58	0
BLEU max	14.34	16.02	24.29
Best epoch	39	18	28
Total run time	17 h	56 h	31 h

Table 1: Summary of training settings and development results

4.2 Data

The Europarl corpus ver. 10 (Koehn, 2005) with about 1,8 M sentences and the DGT 2019 corpus (Tiedemann, 2012) with approximately 5,2 M sentences were used, summing up to around 7 M parallel sentences for training. Newstest 2015 (Bojar et al., 2015) was used as a development (validation) set and newstest 2016 (Bojar et al., 2016) as a test set.

We applied standard preprocessing including normalization, sentence filtering, tokenization and byte-pair encoding by using the default Marian setting (Junczys-Dowmunt et al., 2018) with embedded SentencePiece (Kudo and Richardson, 2018). Concerning the length of the individual sentences, we followed the general practice and limited the sentences to a maximum length of 100.

4.3 Training setup

The NMT systems were trained using Marian ver. 1.9.0 (Junczys-Dowmunt et al., 2018). In order to follow the learning curve of the training process,

[1] https://github.com/pstadler1990/nmt_paper21_appendix

we kept one checkpoint every 10,000 iterations. To do so, we disabled the `overwrite` option from the CLI call of Marian. As per default, cross entropy was used as a validation metric, whereas the training processes were run on a computational server Quadro RTX 6000 (4608 cores, 96 ROPs and a 24 GB memory size) using 2 out of its 8 GPUs.

The validation iterations in the results are labeled as following: $iter_{val} = \frac{iter_{tr}}{f_{val}}$ where $iter_{tr}$ is the reported training iteration number (up.) and f_{val} is the specified validation frequency. For our trained systems, we set this to 10,000. So, a validation iteration of 10,000 training iterations is labeled as 1.

An overview of the settings of the three systems can be seen in Table 1. In particular, the following three systems were trained:

Small RNN model This system was built with Europarl with a final size after pre-processing of 1,828,521 sentences, using an RNN with single-layer encoder and decoder and a minibatch size of 10,000.

Big RNN model In order to build a bigger RNN model, we used the larger dataset consisting of both, Europarl and the DGT corpora, following the same settings as for the small RNN model.

Transformer We used the same training, dev and test sets as in the big RNN model, and the example configuration for a transformer model from Marian[2] adapted to our needs as shown in Table 1. This configuration utilises a six-layer deep encoder and decoder, learning rate warm-up and tied embeddings for source, target and output layer. As suggested by Karita et al. (2019), we increased the minibatch size for the transformer model from 5,000 to 10,000.

5 Results

5.1 Evaluation of the small RNN model

The small RNN model was trained for 17 hours and achieved a BLEU score of 14.34.

5.1.1 Snapshot selection

The best reported BLEU score was reached in epoch 39, out of total 46 epochs, having started with 1.31 in epoch 1. Figure 1 shows the BLEU score evolution, with the black triangle marks indicating the snapshots that we chose to examine, based on the following criteria:

- Epoch 1 (iteration 1): Start of training
- Epoch 4 (iteration 4): BLEU score > 10
- Epoch 11 (iteration 12): Sudden BLEU drop
- Epoch 23 (iteration 26): Mid-high
- Epoch 39 (iteration 45): Highest BLEU score

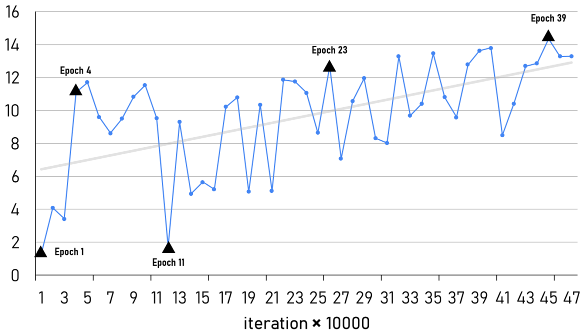

Figure 1: Progress of BLEU score during the training of the small RNN model

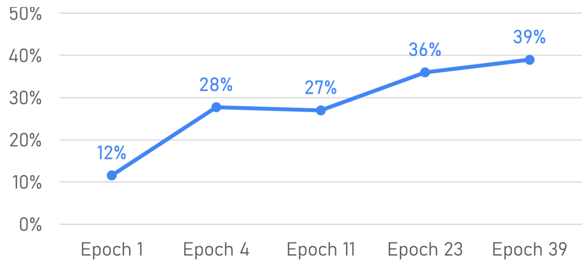

Figure 2: Progress of the average test suite accuracy for the chosen snapshots while training the small RNN model

The complete dataset can be found online in the repository[3].

5.1.2 Evaluation of linguistic categories over time

There was an unsteady but visible rise in the BLEU score over time and also a positive development in the average test suite accuracy (see figure 2), achieving the best accuracy in epoch 39 after a more or less constant improvement.

While looking at the evolution of the accuracy on particular linguistic categories (Table 2), a positive trend is observed when a constant improvement for a specific category has been encountered, a negative trend when there is either a constant decrease

[2] https://github.com/marian-nmt/marian-examples/blob/master/transformer

[3] https://github.com/pstadler1990/nmt_paper21_appendix

category\epoch	1	4	11	22	39
Ambiguity	10%	10%	10%	11%	20%
Coordination & ellipsis	0%	20%	10%	20%	30%
False friends	50%	50%	56%	50%	50%
Function word	30%	50%	30%	50%	60%
Long distance dependency & interrogative	30%	40%	40%	40%	40%
MWE	0%	10%	0%	22%	22%
Named entity & terminology	10%	30%	11%	20%	20%
Negation	20%	60%	60%	60%	60%
Non-verbal agreement	0%	20%	20%	40%	20%
Punctuation	0%	20%	33%	50%	60%
Subordination	0%	40%	40%	60%	70%
Verb tense/ aspect/mood	0%	10%	10%	20%	20%
Verb valancy	0%	10%	30%	20%	30%

Table 2: Progress of accuracy for linguistic categories (small RNN model) over selected epochs

in translation quality for the category or there is a decrease after a peak, whereas any other trend is considered neutral, meaning a positive overall trend characterised by peaks and valleys, which indicate a shift in quality over time or a trend without any development. From the 13 examined categories we found a positive trend in nine categories (70%), two are to be considered neutral (15%) and there was a negative trend in two categories (15%, non-verbal agreement and NER and terminology). Further we provide details on 3 particular categories:

Ambiguity For this category, 10 sentences from a single phenomenon (lexical ambiguity) were examined. Until epoch 39, only one sentence was correctly translated (`Beijing is the capital of China.`). In epoch 39, another sentence was translated in the right way (`What is today's date?`). In epoch 1, 4 and 11, the regular expression provided by the test suite reported a valid translation, because it focused on the ambiguity for the word `china` (wrong translation would be `Porzellan(geschirr)`). However, the translation `Kapital` for the English word `capital` (as in capital city) is wrong. In epoch 22, this is corrected.

Non-verbal agreement A total of 10 sentences from three distinct phenomena were examined. In the first epoch, no sentence was correctly translated (four of them were not translated at all). In the epoch four, two sentences were correctly translated according to the test suite. In epoch 22, four sentences were correctly translated, while interestingly the accuracy decreased to 20% in epoch 39; two formerly correct sentences were mistranslated in this epoch.

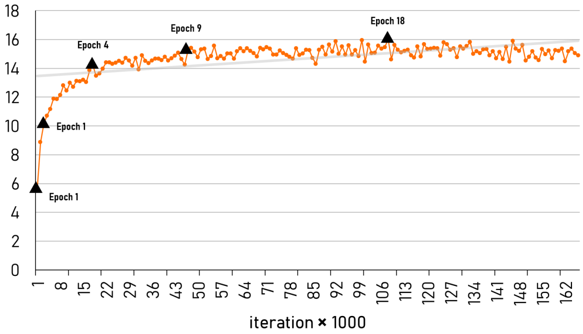

Figure 3: Progress of BLEU score during the training of the big RNN model

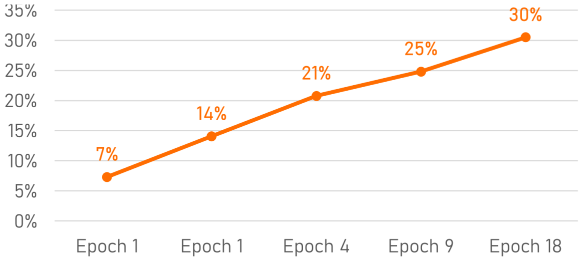

Figure 4: Progress of the average test suite accuracy for the chosen snapshots while training the big RNN model

Subordination For this category, 10 sentences from eight different phenomena were evaluated. We found a constant increase in the translation quality over the selected epochs. Starting with zero correctly translated sentences in the first epoch, the system already reached 40% in epoch 4. The translation quality was quite decent, even when regarding the remaining words that were not part of the examined phenomenon.

5.2 Evaluation of the big RNN model

The big RNN model was trained for 56 hours and achieved a BLEU score of 16.

5.2.1 Snapshot selection

Figure 3 shows the BLEU score evolution over all 164 iterations (28 epochs). We chose the five snapshots for further evaluation based on the following criteria:

category\epoch	1	1	4	9	18
Ambiguity	20%	20%	10%	10%	10%
Coordination & ellipsis	0%	0%	20%	10%	30%
False friends	22%	50%	40%	40%	60%
Function word	10%	30%	30%	30%	44%
Long distance dependency & interrogative	10%	20%	30%	30%	50%
MWE	0%	0%	10%	0%	0%
Named entity & terminology	22%	40%	40%	40%	40%
Negation	0%	0%	40%	40%	50%
Non-verbal agreement	11%	22%	20%	30%	30%
Punctuation	0%	0%	10%	20%	20%
Subordination	0%	0%	20%	40%	30%
Verb tense/ aspect/mood	0%	0%	0%	10%	10%
Verb valancy	0%	0%	0%	20%	20%

Table 3: Progress of accuracy (big RNN model) for linguistic categories over selected epochs

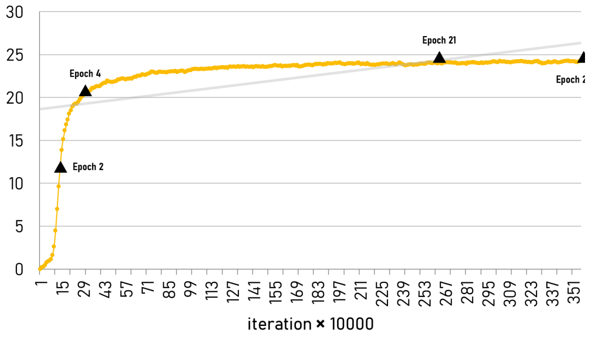

Figure 5: Progress of BLEU score during the training of the transformer model

- Epoch 1 (iteration 1): start of training
- Epoch 1 (iteration 3): BLEU score < 10
- Epoch 4 (iteration 18): BLEU score > 14
- Epoch 9 (iteration 44): BLEU > 15
- Epoch 18 (iteration 108): highest BLEU score

5.2.2 Evaluation of linguistic categories over time

While studying the accuracy progress for particular linguistic categories, we observe that three of them have a negative thread, ending with a lower accuracy than the one achieved during some earlier epochs (ambiguity, multi-word expressions and subordination). Additionally, we observe the following particular issues:

Named entities and terminology Four out of ten sentences from five different phenomena were translated correctly in this category: Proper name (1 out of 1), Date (0 out of 2), Measuring unit (2 out of 3), Location (1 out of 2) and Domain specific term (0 out of 1). Dates were not properly converted into the German format (`dd.mm.yyyy`), however the named entities were kept in their original spelling (`Marilyn Monroe` , `Pearl Harbor`) in both cases. In our final recorded snapshot, the system was able to translate 2 out of 3 measuring units accordingly: `The human brain has a volume of about 600 to 800 cubic centimetres.` and `The room was 17 feet long.`. The system struggled with the sentence `Stella had her hair cut six inches last week.`, no matter the progress. The locations `Saarland` (Saarland) and `Palatinate` (Pfalz) were only correctly translated in iteration 3 and 18 and mistranslated in lower and higher iterations. Regarding the domain-specific term `neurotransmitter serotonin`, the system was not able to get the capitalisation right in most cases and randomly got it either correct or wrong from iteration to iteration.

False Friends False friends were translated correctly in 60% (6 out of 10 sentences). Three sentences contained the word `Genie` and were all translated wrong over all recorded snapshots. Four sentences examined the different meanings of `serious` and were all translated correct in all recorded snapshots but the first (epoch 1). The system struggled with the sentence `For the Christmas party, the chef sculpted an angel out of chocolate.`; in no case the translation was correct. It seems to be obvious that words like `Genie` were not part of the two used corpora or at least not used in the given meaning and thus unable to translate correctly. Overall though, the system performed well with false friends.

5.3 Evaluation of the transformer model

The transformer was trained for 31 hours and achieved a BLEU score of 24.29. Our trained model is comparable to the one trained by Sennrich et al. (2015) that achieved a BLEU score of 22.7 to 25.7 for English→German with a similar dev and test set (newstest14 and newstest15).

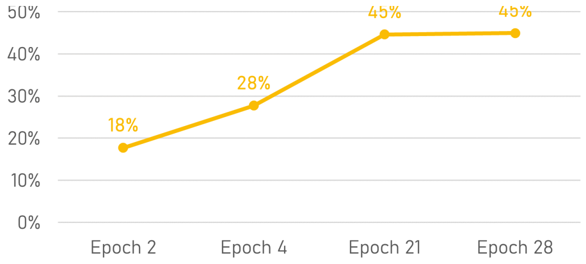

Figure 6: Progress of the average test suite accuracy for the chosen snapshots while training the transformer model

category\epoch	2	4	21	28
Ambiguity	20%	20%	20%	20%
Coordination & ellipsis	0%	10%	20%	10%
False friends	50%	50%	50%	50%
Function word	20%	40%	50%	60%
Long distance dependency & interrogative	20%	40%	70%	70%
MWE	10%	10%	10%	10%
Named entity & terminology	20%	40%	70%	80%
Negation	40%	50%	60%	60%
Non-verbal agreement	20%	40%	40%	50%
Punctuation	20%	10%	30%	50%
Subordination	0%	20%	80%	80%
Verb tense/ aspect/mood	10%	10%	30%	40%
Verb valancy	0%	20%	60%	50%

Table 4: Progress of accuracy for linguistic categories (transformer) over selected epochs

5.3.1 Snapshot selection

Figure 5 shows the BLEU score evolution over all 351 iterations (28 epochs). For the transformer model, we picked only four snapshots for further examination, as there were no big changes after certain epochs:

- Epoch 2 (iteration 15): BLEU score 10
- Epoch 4 (iteration 39): BLEU score >20
- Epoch 21 (iteration 255): BLEU score 24 (no great changes from now on)
- Epoch 28 (iteration 355): Final epoch, BLEU score 24

5.3.2 Evaluation of linguistic categories over time

A total of 49 phenomena from 13 categories were examined for the transformer-based system within the test suite. There was a steady and visible rise in the BLEU score development over time and a positive development in the average score as reported by the test suite. The highest recorded BLEU score 24,28 was achieved in epoch 28 (iteration 348). However, there is only a small difference between epoch 21 and the final epoch 28 – this is also perceptible from the BLEU score (figure 5); the system became satisfactory around epoch 20 to 21. Regarding the test suite accuracy, there was a notable increase from the first epoch to epoch 21 (see figure 6). Here, one observes that two linguistic categories, verb valency and coordination and ellipsis, end up with 10% less accuracy than the one achieved during the previous snapshot. Another three categories (ambiguity, MWE, and false friend) have a flat trend, maintaining the same accuracy as the one achieved in epoch 2, whereas negation is also very close with a relatively mild increase. A steady increase was achieved for NER and terminology, whereas the steepest trend is shown by subordination, which starts with 0% and ends with 80%. Looking on particular items, we can observe the following:

Ambiguity The system struggled with ambiguity – only 2 out of 10 test sentences (20%) were correctly translated, and this was stable from the first snapshot until the final system. The system didn't make a correct lexical choice for any of the three sentences focusing on the ambiguity of the word `bat`: `The player hit the ball with the bat.`, `The woman hit the burglar with the bat.` and `Bats sleep upside-down.` The two sentences containing the words `date` respectively `date palm`, were both translated incorrectly.

Function words Question tags were mistranslated in nearly all cases within the recorded snapshots. In the first snapshots, the question tags were completely ignored in the translation, however, the system understood the sentences contained a question and therefore ended the sentences with a question mark; yet, the important words were skipped. In epoch 4, the system began to translate some parts of the subordinate clauses (the question tags), but was not able to translate them in an appropriate way (No one still goes voluntarily in one of these old-style libraries, right? → Niemand geht noch immer freiwillig in einer dieser alten Bibliotheken, Recht?). In epoch 21, one question tag was translated accurately (You saw her last week, didn't you? → Sie haben sie letzte Woche gesehen, nicht wahr?). Focus particles such as even, only or also were translated almost without any errors (9 out of 10 in epoch 26). However, the word even in the sentence He didn't even drink a single glass of wine. was never translated correctly.

Named entities The translations for dates were highly accurate within the latest recorded snapshots (epoch 21 and epoch 28); Two dates have been correctly translated from the American / English format to the dd.mm.yyyy format commonly used in Germany. Measuring units were not converted (as intended) and correctly translated (3 out of 3 sentences in epoch 28). Location information was not translated well enough; especially well-known proper names, such as the names of the German federal states still caused difficulties for the system

However, a slight improvement towards the end could be recognized here. An interesting transition in quality can be found for the sentence The Saarland and the Palatinate enjoy a fierce regional rivalry. where the translation quality actually dropped in the last two recorded epochs 21 and 28; it seemed the system had been overfitted to some specific word combinations, resulting in the use of Flughafen Pfalz (airport Pfalz) for the English word Palatine (German: Pfalz or pfälzisch) instead of Pfalz (epochs 2 and 4).

Coordination and ellipsis The system had difficulty translating sentences from this category. An accurate evaluation of the phenomena is difficult because many of the necessary vocabularies were not correctly translated, making the sentences incomplete or partially meaningless. However, two sentences were translated correctly: Goethe wrote Faust, not Schiller. and Jackie likes the doctor but she doesn't like the nurse. were both translated correctly in epoch 21, but not in epoch 28 and 4. In epoch 2, no sentence was translated correctly.

Verb valency There was an increasing development until epoch 21 (best score for this category) - in the following epoch 28 the translation quality dropped from 60% back to 50% due to a mistranslated sentence in the last epoch (I want to talk to your neighbors.).

6 Comparison between iterations and models

As figure 7 shows, there is a clear difference between the two RNN trainings regarding the resilience of the BLEU score over time. While there is a lot of jittering in the RNN model with a small amount of data, a nearly constant increase is given in the model with a bigger amount of data, showing no huge peaks or valleys. Regarding BLEU scores, the system with the larger corpus performed a little bit better (~16) than the one smaller one (~14), but this was not reflected in the test suite comparison, where there was no big difference in terms of test suite accuracy, even though the used corpus has more than doubled in the big RNN model. Additionally, it can be observed that some categories in the bigger RNN perform worse than what was achieved in the smaller one. The inability of the bigger model to take advantage of the additional data may be addressed to the rather shallow architecture of the encoder and the decoder. With the current range of experiments, there are some open questions regarding further comparisons between RNN-small and RNN-big models. Future experiments could investigate the reasons for the fact that RNN-small and RNN-big systems perform comparably on the test suite, e.g. whether this can be attributed to the shallow architecture, to a subtle domain mismatch between Europarl and DGT or to the domain mismatch between the training data and the test suite.

The transformer model development takes some more iterations until it reaches a competitive BLEU score but then clearly outperforms both RNN systems by more than 60%, although the comparison with the RNN models is not direct, since the transformer is built with more layers and a not directly comparable architecture.

There is no generalizable development over all examined categories; some performed better than others, while some of the categories had no development at all. Scoring with the test suite was difficult for many sentences, because of insufficient vocabulary and wrong lexical choices. The system

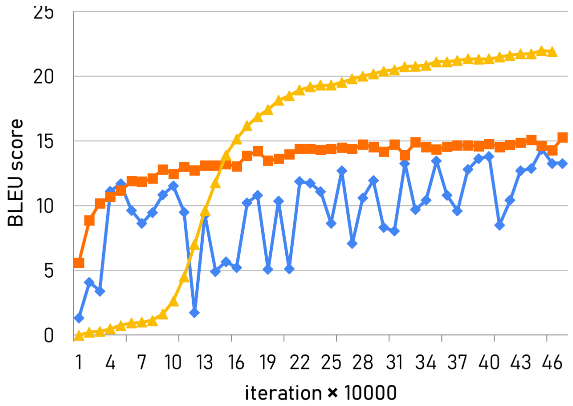

Figure 7: BLEU scores for the models small RNN (◆), big RNN (■) and transformer (△)

had trouble with punctuation, such as quotation marks. Names were often translated with fragments or as mixtures of different fragments, clearly coming from the Europarl proceedings.

7 Conclusions and further work

We performed a fine-grained evaluation on several training stages of three different NMT models. The most interesting observation is that although the training process stops when the best scores of the automatic metrics are achieved (*early stopping*), the accuracy of some linguistic phenomena is dropping, as compared to previous epochs. For this reason, the contribution of the scoring metric and the stopping criterion should be further investigated, while it might be also depend on whether the development sets contain these phenomena.

The fact that some linguistic categories have a steeper curve than the others may also signalise the difficulty of these categories from a machine learning perspective.

Since this is a preliminary study, the amount of items per linguistic category is small and does not allow for statistically significant conclusions. This could be improved in the future with further annotation effort. Finally, the systems examined are taken as random samples in terms of settings and parameters. We should repeat the measurements on state of the art systems, allowing fair comparisons among different architectures and design decisions.

Acknowledgments

This work has been accomplished as a semester project, part of the MSc program of Media Informatics of the Technical University of Berlin, hosted by the Quality and Usability Lab and organized by Neslihan Iskender under the chair of Prof. Sebastian Möller. Supervision from the side of DFKI has been supported by the projects TextQ (funded by the German Research Foundation; DFG) and SocialWear (funded by the German Ministry of Education; BMBF), with the support of Dr. Aljoscha Burchardt, Jana Majella Hirschberg, Malte Ostendorff and Christian Schulze.

References

Lars Ahrenberg. 2018. A challenge set for english-swedish machine translation. *Thanks to our sponsors: Gold: Stora Skuggans Värdshus Silver: TT Nyhetsbyrån and Lingsoft Bronze: Convertus, Digital Grammars, IQVIA and Voice Provider*, page 27.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, Aljoscha Burchardt, and Sebastian Möller. 2020. Fine-grained linguistic evaluation for state-of-the-art Machine Translation. In *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. 2019. Linguistic evaluation of german-english machine translation using a test suite. *arXiv preprint arXiv:1910.07457*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *Computer Research Repository*, abs/1409.0.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and James Glass. 2019. Identifying and Controlling Important Neurons in Neural Machine Translation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2020. On the linguistic representational power of neural machine translation models. *Computational Linguistics*, 46(1):1–52.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

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 *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, and Stephan Vogel. 2017. Understanding and Improving Morphological Learning in the Neural Machine Translation Decoder. *undefined*.

Liane Guillou and Christian Hardmeier. 2016. Protest: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. 2019. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the tenth Machine Translation Summit*, volume 5, pages 79–86, Phuket, Thailand.

Sungryong Koh, Jinee Maeng, Ji-Young Lee, Young-Sook Chae, and Key-Sun Choi. 2001. A test suite for evaluation of english-to-korean machine translation systems. In *MT Summit'conference, Santiago de Compostela*. Citeseer.

Xiang Kong, Zhaopeng Tu, Shuming Shi, Eduard Hovy, and Tong Zhang. 2018. Neural machine translation with adequacy-oriented learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, page 19. arXiv.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP 2018 - Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Proceedings*, pages 66–71. Association for Computational Linguistics (ACL).

Vivien Macketanz, Renlong Ai, Aljoscha Burchardt, and Hans Uszkoreit. 2018a. Tq-autotest–an automated test suite for (machine) translation quality. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, and Hans Uszkoreit. 2018b. Fine-grained evaluation of German-English Machine Translation based on a Test Suite. In *Proceedings of the Third Conference on Machine Translation (WMT18)*, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Why self-attention? A targeted evaluation of neural machine translation architectures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 4263–4272. Association for Computational Linguistics.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2019. Encoders Help You Disambiguate Word Senses in Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1429–1435, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC12)*, pages 2214–2218, Istanbul, Turkey.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Tereza Vojtěchová, Michal Novák, Miloš Klouček, and Ondřej Bojar. 2019. Sao wmt19 test suite: Machine translation of audit reports. *arXiv preprint arXiv:1909.01701*.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU: Training Neural Machine Translation with Semantic Similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355.

A Appendix

Example sentence	Category	Phenomenon
Beijing is the capital of China.	Ambiguity	Lexical ambiguity
The manager suspects the president of theft.	Verb valency	Case government
I stopped reading the poster.	Verb valency	Catenative verb
John sang the baby to sleep.	Verb valency	Resultative
Goethe wrote Faust, not Schiller.	Coordination & ellipsis	Stripping
Hand me a Kleenex, please.	Named entitiy & terminology	Proper name
Marilyn Monroe was born as Norma Jeane Mortenson on June 1, 1926.	Named entitiy & terminology	Date
The room was 17 feet long.	Named entitiy & terminology	Measuring unit
John is studying at the Technical University of Vienna.	Named entitiy & terminology	Location
In the latter case, this would be the neurotransmitter serotonin.	Named entitiy & terminology	Domainspecific term
For the Christmas party, the chef sculpted an angel out of chocolate.	False friends	False friends
No one still goes voluntarily in one of these old-style libraries, right?	Function word	Question tag
I saw him only once.	Function word	Focus particle
You will have passed John the ball.	Verb tense/aspect/mood	Ditransitive - future II simple
She had been baking Tim a cake.	Verb tense/aspect/mood	Ditransitive - past perfect progressive
Neither John nor Mary could do anything about the problem.	Long distance dependency & interrogative	Multiple connectors
Never again will he eat raw spaghetti.	Long distance dependency & interrogative	Negative inversion
To whom should the documents be sent?	Long distance dependency & interrogative	Pied piping
No walking on the grass!	Negation	Negation
Susan dropped the plate and it shattered loudly.	Non-verbal agreement	Coreference
The man who you mentioned is my friend.	Subordination	Relative clause
What do you think they did that upset everyone?	Long distance dependency & interrogative	Extraposition
I'd like to have a round of applause for our next guest!	MWE	Collocation
John can play the guitar, and Mary can too.	Coordination & ellipsis	VP-ellipsis
Jackie likes the doctor but she doesn't the nurse.	Coordination & ellipsis	Pseudogapping
She likes the car more than her husband does.	Subordination	Adverbial clause
Oh, what a beautiful morning! Jim said to himself.	Punctuation	Quotation marks
They are well-behaved children.	MWE	Compound
Don't put all your eggs in one basket.	MWE	Idiom
Rebecca said she would be in Munich next week.	Subordination	Indirect speech
We didn't realize she was so ill.	Subordination	Object clause
We are determined to completely solve the problem.	Long distance dependency & interrogative	Split infinitive
Are you going to the beach today?	Long distance dependency & interrogative	Polar question
They may not know it.	Verb tense/aspect/mood	Modal negated
They are teaching themselves Spanish.	Verb tense/aspect/mood	Reflexive - present progressive
I would be kicking Tim.	Verb tense/aspect/mood	Transitive - conditional I progressive
You would have been eating the potatoes.	Verb tense/aspect/mood	Transitive - conditional II progressive
She will have been painting the house.	Verb tense/aspect/mood	Transitive - future II progressive
I have been painting the house.	Verb tense/aspect/mood	Transitive - present perfect progressive
He looks up to his older brother.	MWE	Verbal MWE
She has lost her shoes.	Non-verbal agreement	Possession
Before leaving, John has been at home.	MWE	Prepositional MWE
What was the man looking for in the fridge?	Long distance dependency & interrogative	Wh-movement
Mandy's brother John plays football.	Non-verbal agreement	Genitive
It was Lena who had baked the cake.	Subordination	Cleft sentence
What I did in the end was to go home.	Subordination	Pseudo-cleft sentence

Table 5: Extracted example sentences for each examined category and phenomenon