*Human Language Technologies – The Baltic Perspective*
*K. Muischnek and K. Müürisep (Eds.)*

*doi:10.3233/978-1-61499-912-6-126*

# Impact of Corpora Quality on Neural Machine Translation

Matīss RIKTERS [1]

*Tilde, Vienības gatve 75A, Rīga, Latvia*

**Abstract.** Large parallel corpora that are automatically obtained from the web, documents or elsewhere often exhibit many corrupted parts that are bound to negatively affect the quality of the systems and models that learn from these corpora. This paper describes frequent problems found in data and such data affects neural machine translation systems, as well as how to identify and deal with them. The solutions are summarised in a set of scripts that remove problematic sentences from input corpora.

**Keywords.** machine translation, parallel corpora, corpora filtering

## 1. Introduction

Machine translation (MT) systems - both, statistical (SMT) and neural (NMT) - rely on large amounts of parallel data for training the models. It is often the case that larger amounts of corpora lead to higher quality models, therefore a common practice is automatic extraction of such corpora from web resources, digitised books and other sources. Such data is prone to be noisy and include all kinds of problematic sentences alongside the high-quality ones. Data quality plays an important role in training of statistical and, especially, neural network based models like NMT, which is quick to memorise bad examples. In the case of training SMT and NMT systems, often the only pre-processing is done using scripts from the Moses Toolkit [1], which is only capable of removing sentences that are longer or shorter than a specified amount or the source-target length ratio is too high.

In this paper, we explore the types of low-quality sentences commonly found in parallel corpora. We also compare the benefits of using additional filters to remove these sentences before training MT systems in contrast to using only the Moses scripts. We introduce a set of corpora cleaning tools [2] that remove sentences that have some of the most common problems found in large corpora. It is published in GitHub with the MIT open-source license.

[1]Corresponding Author: Matīss Rikters; E-mail: name.surname@tilde.lv.
[2]Corpora Cleaning Tools: `https://github.com/M4t1ss/parallel-corpora-tools`

## 2. Related Work

Zipporah [2] is a trainable tool for selecting a high-quality subset of data from a huge amount of noisy data. The authors report that it can improve MT quality by up to 2.1 BLEU, but in order to use it, the tool requires a known high-quality data set for training.

Wolk [3] proposes a method that uses online MT engines to translate source sentences from a parallel corpus and compare them with the given target sentences. It is very expensive to use on real-world parallel corpora, containing tens of millions of parallel sentences. The author reports results on using the method on rather small corpora of only several million words.

Khadivi and Ney [4] introduce a parallel corpora filtering method based on word alignment models. Similar to Zipporah, this method also relies on training using a high-quality corpus.

## 3. Problems in Corpora

This section outlines some often occurring problems in parallel corpora. The specific examples were obtained from the English-Estonian part of the ParaCrawl[3] corpus.

One of the most common defects in parallel corpora is a high mismatch between the non-alphabetic characters between source and target sentences (Figure 1). Also often are sentences that are completely or mostly composed of characters outside the scope of the language in question (Figure 2).

In parallel corpora, we may occasionally see the same sentence of one language aligned to multiple different ones of the other language (Figure 3), but this is not always a bad indication, since they may just be paraphrases of the same concept (Figure 4). It is also wise to check if sentences in specific languages actually consist of text in that language (Figure 5) as there may be citations and other parts of foreign language texts, especially in news domain corpora.

Finally, a little less common observation for automatically gathered corpora, but somewhat more often in automatically generated (translated) parallel corpora is the repeating of tokens (Figure 6). Sentences like this may not always be incorrect, but they introduce ambiguity when used to train MT systems.

| English | Estonian |
|---|---|
| address: Akariah 3, 8th Floor, Olaya St. | ÇáãæÇÖíÚ ÇáÝÇÆÒÉ : 3 |
| Address by President of the Republic of Hungary Ferenc MÁDL | LENNART MERI LOENG VÄIKERAHVASTE TULEVIKUST |
| Addresses | 1. KOGUD |
| Addresses Speeches of the peoples' representatives Reviews Photos | ETTEKANNE RAHVUSVAHELISEL KONVERENTSIL |
| ADDRESS of BANK: Liivalaia 8, TALLINN, 15040, ESTONIA | Jah, see on võimalik. |
| Add to cart View | Kalorid: 3000 kcal |
| Add to my wishlist | Caffeine 200 Plus |
| Adela Banášová, TV presenter | Adela Banášová, telesaatejuht |

**Figure 1.** An example of a high mismatch in non-alphabetical character counts between source and target.

[3]Large-Scale Parallel Web Crawl: `http://statmt.org/paracrawl`

| | | |
|---|---|---|
| äæÚ: ãÔÇÑßÇÊ; ÚÖæ: Doha Rose | á á ¶â á á ·á á á â á á ¶ – á á ¾á â á á á | संगीत (2) |
| äæÚ: ãÔÇÑßÇÊ; ÚÖæ: Dragonier | à¤ à¤¸ à¤¸à¤ à¤¸à¥ à¤ à¤°à¤£ à¤ à¥ à¤¸à¤¾à | « ఔ ÔÇÑßäÇ ÑÃíß Ýí ÊÞííã ÔÎÕíÇÊ |
| äæÚ: ãÔÇÑßÇÊ; ÚÖæ: drift king | à¤ à¥ à¤ à¤ à¤° à¤ à¤§à¤¿à¤ à¤°à¥ à¤®à¤¾à¤ | 什麼 這是 for? |

**Figure 2.** Examples of sentences with over 50% non-alphabetical symbols.

| English | Estonian |
|---|---|
| I voted in favour. | kirjalikult. – (IT) Hääletasin poolt. |
| I voted in favour. | Ma andsin oma poolthääle. |
| I voted in favour. | Ma hääletasin selle poolt. |

**Figure 3.** An example of an English sentence aligned to multiple different Estonian sentences.

| English | Estonian |
|---|---|
| That is the wrong way to go. | See ei ole õige. |
| This is not true. | See ei ole õige. |
| This is simply wrong. | See ei ole õige. |

**Figure 4.** Multiple English paraphrased sentences aligned to one Estonian sentence.

| English | Estonian |
|---|---|
| Zvér Józsefné Független Nyilvántartásba véve 2010.09.03 | Ötvös Bálint Független Nyilvántartásba véve 2010.09.03 |
| Zvezdnyj bilet : roman. - Moskva, 1961. | – Stockholm : Eesti Raamat, 1948. |
| Zwei 17 | Son Goku |
| ÞÈá : http://i.imgur.com/F42jC6Y.png | ÈÇáàÓÈÉ ááÊØÈíÞ / |
| и ... XXL Booster | Dietary Fibre 300g |

**Figure 5.** Examples of sentences with a different identified language than the one specified.

| English | Estonian |
|---|---|
| Now for , , we get . Now since is bijective and , , and we get . | Nüüd , , Saame . Nüüd kuna on bijective ja , Ja saame . |
| Now if then from (1) we have that and or or or . | Nüüd, kui seejärel (1) meil on, et ja või või või . |
| Now I get it. Thank you very very much Fernando!! obrigado!! | Nüüd ma saan seda. Tänan väga palju Fernando!! aitäh!! |

**Figure 6.** An example repeating tokens (underlined).

## 4. Corpora Filters

The filters described in this section are mainly intended for parallel corpora consisting of two files with identical line-counts where each line of one file is related to the same line of the other file. Several of the filters are applicable to monolingual data as well and can be used to clean data for unsupervised MT training, back-translation, and other use-cases.

**Unique parallel sentence filter** – removes duplicate source-target sentence pairs.

**Equal source-target filter** – removes sentences that are identical in the source side and the target side of the corpus.

**Multiple sources - one target** and **multiple targets - one source** filters – removes repeating sentence pairs where the same source sentence is aligned to multiple different target sentences and multiple source sentences aligned to the same target sentence.

**Non-alphabetical filters** – remove sentences that contain over 50% non-alphabetical symbols on either the source side or the target and sentence pairs that have significantly more (at least 1:3) non-alphabetical symbols in the source side than in the target side (or vice versa).

**Repeating token filter** – especially useful for filtering back-translated parallel corpora that are created by translating a clean monolingual corpus into another language using NMT. NMT output may sometimes exhibit repeated words in the generated translation, which indicates that the system had problems translating a part of the sentence and it used the repetitions to fill the gap. In such cases the source-target sentence pair is likely to not be a good parallel sentence, therefore the repeating token filter removes them.

**Correct language filter** – uses language identification software [5] to estimate the language of each sentence and removes any sentence that has a different identified language from the one specified.

**Moses Scripts and Subword NMT** – calls Moses scripts for tokenising, cleaning, truecasing, and Subword NMT [6] for splitting into subword units. This process prepares the corpus up to the point where it can be passed on to the NMT system for training.

## 5. Experiments and Results

**Table 1.** Detailed results on filtering English-Estonian/Finnish/Latvian larger common parallel corpora from WMT shared tasks.

| | Paracrawl | | Rapid | | | Europarl | | |
|---|---|---|---|---|---|---|---|---|
| | En-Et | En-Fi | En-Et | En-Fi | En-Lv | En-Et | En-Fi | En-Lv |
| Corpus size | 1298103 | 624058 | 226978 | 583223 | 306588 | 652944 | 1926114 | 638789 |
| Unique | 26 | 37 | 23 | 161463 | 80894 | 23218 | 52686 | 19652 |
| | 0.00% | 0.01% | 0.01% | **27.68%** | **26.39%** | 3.56% | 2.74% | 3.08% |
| src == tgt | 242816 | 41611 | 428 | 3488 | 2929 | 490 | 528 | 707 |
| | **18.71%** | **6.67%** | 0.19% | 0.60% | 0.96% | 0.08% | 0.03% | 0.11% |
| * sources | 267235 | 17239 | 1108 | 1513 | 990 | 1176 | 6631 | 979 |
| 1 target | **20.59%** | 2.76% | 0.49% | 0.26% | 0.32% | 0.18% | 0.34% | 0.15% |
| * targets | 69225 | 9532 | 752 | 1016 | 329 | 462 | 3536 | 435 |
| 1 source | **5.33%** | 1.53% | 0.33% | 0.17% | 0.11% | 0.07% | 0.18% | 0.07% |
| >50% | 200338 | 12919 | 1226 | 5647 | 1699 | 66 | 285 | 72 |
| non-alpha | **15.43%** | 2.07% | 0.54% | 0.97% | 0.55% | 0.01% | 0.01% | 0.01% |
| Non-alpha | 23777 | 12737 | 6674 | 13311 | 6361 | 7211 | 24847 | 4012 |
| mismatch | 1.83% | 2.04% | 2.94% | 2.28% | 2.07% | 1.10% | 1.29% | 0.63% |
| Repeating | 11210 | 1397 | 175 | 396 | 171 | 727 | 2594 | 703 |
| tokens | 0.86% | 0.22% | 0.08% | 0.07% | 0.06% | 0.11% | 0.13% | 0.11% |
| Language | 283152 | 36233 | 14762 | 24854 | 8739 | 8924 | 10932 | 3301 |
| mismatch | **21.81%** | **5.81%** | **6.50%** | **4.26%** | 2.85% | 1.37% | 0.57% | 0.52% |
| $\sum$ removed | 1097779 | 131705 | 25148 | 211688 | 102112 | 42274 | 102039 | 29861 |
| | **85%** | **21%** | **11%** | **36%** | **33%** | 6% | 5% | 5% |

### *5.1. Corpora Cleaning*

We used the toolkit to clean parallel corpora provided in the WMT17[4] and WMT18[5] news MT shared tasks for English $\leftrightarrow$ Estonian/Finnish/Latvian. Detailed results of the cleaning process for three of the largest corpora - ParaCrawl, Rapid corpus of EU press

[4]Second Conference on Machine Translation - `http://statmt.org/wmt17`
[5]Third Conference on Machine Translation - `http://statmt.org/wmt18`

releases (Rapid) and European Parliament Proceedings Parallel Corpus (Europarl) - are shown in Table 1.

The results show that ParaCrawl is the most problematic corpus, especially the Estonian part, where 85% had to be removed. The most frequent problems are 1) specified and identified language mismatch; 2) identical sentences appearing on source and target sides; 3) multiple source sentences aligned to the same target sentence; 4) an overwhelming amount of non-alphabetical characters; and 5) multiple target sentences aligned to the same source sentence. All examples of bad sentences in Section 3 were selected from the removed parts of the English-Estonian ParaCrawl corpus.

The Rapid corpus had an overall higher quality with only about 25% of parallel sentences removed. For the three languages it exhibited three main defects - 1) duplicate parallel sentences; 2) specified and identified language mismatch; and 3) mismatch in amounts of non-alphabetical symbols between source and target sentences.

Europarl was by far the cleanest corpus, having only 5-6% of sentences removed by the cleaning toolkit. For all languages, most removed sentences were due to the same two defects as in the Rapid corpus.

We combined and shuffled all three English-Estonian corpora, resulting in 1 012 824 (46.50% of total) sentence parallel corpus for training NMT systems described in the next section. The total amount of English-Finnish parallel sentences was 2 719 104 (82.72% of total) after adding a cleaned version of the Wiki Headlines corpus, and English-Latvian - 1 617 793 (35.85% of total) parallel sentences after adding cleaned versions of LETA translated news, Digital Corpus of European Parliament (DCEP), and Online Books corpora (cleaning details in Table 2). We used the development data sets provided by the WMT shared tasks.

### *5.2. Machine Translation*

To observe the actual benefit of filtering data for NMT, we trained NMT models using filtered and non-filtered data in both translation directions for the three language pairs. We used Sockeye [7] to train transformer architecture models with 6 encoder and decoder layers, 8 transformer attention heads per layer, word embeddings and hidden layers of size 512, dropout of 0.2, shared subword unit vocabulary of 50 000 tokens, maximum sentence length of 128 symbols, and a batch size of 3072 words. All models were trained until they reached convergence on development data.

The final NMT system results in Table 3 show that corpora filtering improves NMT quality for Estonian and Latvian systems, but not Finnish. The lack of improvement for Finnish is mainly due to the Europarl being the largest (about $\frac{3}{5}$ of total) and at the same time the cleanest corpus for this language pair. The biggest corpora for Estonian and Latvian - ParaCrawl (about $\frac{3}{5}$ of total) and DCEP (about $\frac{4}{5}$ of total) respectively were also the most problematic ones with 85% and 78% sentences removed respectively.

Figure 7 shows training progression of all 12 NMT systems. Filtered systems are depicted with solid lines, unfiltered ones - with dotted lines, Estonian systems are in light/dark blue colours, Finnish - orange/yellow, and Latvian are in light/dark red colours. The figure shows that the filtered Estonian and Latvian systems are much quicker to learn than the unfiltered ones, but eventually, they converge close to the unfiltered systems. As for the Finnish systems - there is no significant difference between filtered and unfiltered, as at times one is higher than the other or vice versa.

**Table 2.** Detailed results on filtering English-Finnish/Latvian smaller parallel corpora from WMT shared tasks.

| | En-Fi | | En-Lv | |
|---|---|---|---|---|
| | Wiki | DCEP | Leta | Books |
| Corpus size | 153728 | 3542280 | 15671 | 9577 |
| Unique | 0 | 2277397 | 454 | 434 |
| | 0.00% | **64.29%** | 2.90% | 4.53% |
| src == tgt | 42438 | 339861 | 2 | 4 |
| | **27.61%** | **9.59%** | 0.01% | 0.04% |
| * sources | 161 | 12474 | 2 | 35 |
| 1 target | 0.10% | 0.35% | 0.01% | 0.37% |
| * targets | 339 | 9450 | 15 | 12 |
| 1 source | 0.22% | 0.27% | 0.10% | 0.13% |
| >50% | 488 | 31842 | 0 | 13 |
| non-alpha | 0.32% | 0.90% | 0.00% | 0.14% |
| Non-alpha | 4616 | 38838 | 946 | 20 |
| mismatch | 3.00% | 1.10% | 6.04% | 0.21% |
| Repeating | 38 | 1242 | 47 | 8 |
| tokens | 0.02% | 0.04% | 0.30% | 0.08% |
| Language | 74507 | 48910 | 59 | 1074 |
| mismatch | **48.47%** | 1.38% | 0.38% | **11.21%** |
| $\sum$ removed | 122587 | 2760014 | 1525 | 1600 |
| | **80%** | **78%** | **10%** | **17%** |

It is generally visible that in both translation directions the filtered systems achieve higher BLEU scores and reach higher quality quicker. For both English-Estonian systems, the unfiltered version catches up to the filtered one later on in the training, but never quite reaches or surpasses it.

**Table 3.** Translation quality results (BLEU scores) for all translation directions on development data. The best results are marked in bold. The second row shows how much of the initial parallel corpora remained after filtering for each language pair.

| | En-Et | Et-En | En-Fi | Fi-En | En-Lv | Lv-En |
|---|---|---|---|---|---|---|
| Unfiltered | 15.45 | 21.55 | **20.07** | **25.25** | 21.29 | 24.12 |
| Corpus after filtering | 46.50% | | 82.72% | | 35.85% | |
| Filtered | **15.80** | **21.62** | 19.64 | 25.04 | **22.89** | **24.37** |
| Difference | **+0.35** | **+0.07** | -0.43 | -0.21 | **+1.60** | **+0.25** |

## 6. Conclusion

This paper introduced several types of problematic sentences that can be found in large text corpora and a set of filters that help to remove them in order to train higher quality neural machine translation models using the remaining clean part of the corpora. Results show that in cases where the majority of given parallel corpora are very noisy and there is a small fraction of high-quality corpora, cleaning boosts NMT performance. This is

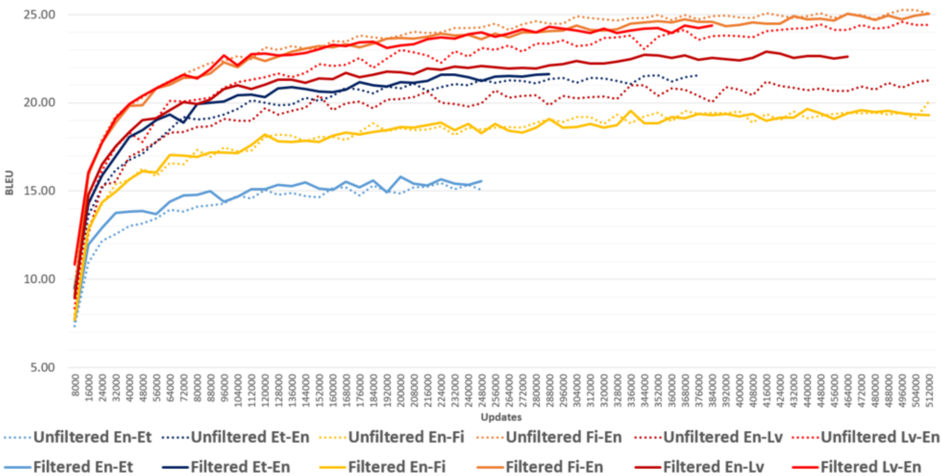


**Figure 7.** Training progress of English ↔ Estonian/Finnish/Latvian NMT systems.

especially evident for translation into morphologically rich languages like Estonian and Latvian.

In this paper, we mainly focused on cleaning parallel corpora, but the toolkit is also capable of cleaning monolingual corpora separately. In the MT system training workflow, cleaning monolingual data is useful before performing back-translation of an in-domain corpus, so that only filtered sentences get translated.

We release the corpora cleaning toolkit on GitHub under the MIT open-source license. The toolkit was used as an integral part of the runner-up English-Estonian NMT system submission [8] in the WMT18 news translation task for cleaning parallel and back-translatable monolingual data, as well as synthetic parallel data produced via back-translation.

## 7. Acknowledgements

The research has been supported by the European Regional Development Fund within the research project "Neural Network Modelling for Inflected Natural Languages" No. 1.1.1.1/16/A/215.

## References

[1] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst, Moses: open source toolkit for statistical machine translation, Association for Computational Linguistics, 2007, pp. 177–180. `https://dl.acm.org/citation.cfm?id=1557821`.

[2] H. Xu and P. Koehn, Zipporah: a Fast and Scalable Data Cleaning System for Noisy Web-Crawled Parallel Corpora, *Emnlp* (2017), 2935–2940. `http://www.aclweb.org/anthology/D17-1319%0Ahttp://aclweb.org/anthology/D17-1318`.

[3] K. Wolk, Noisy-parallel and comparable corpora filtering methodology for the extraction of bi-lingual equivalent data at sentence level, *Computer Science @BULLET Computer Science* **16**(162) (2015), 169–184. ISBN ISBN 9788361182085. doi:10.7494/csci.2015.16.2.169.

https://arxiv.org/ftp/arxiv/papers/1510/1510.04500.pdfhttp://dx.doi.org/10.7494/csci.2015.16.2.169.

[4] S. Khadivi and H. Ney, Automatic Filtering of Bilingual Corpora for Statistical Machine Translation, in: *Natural Language Processing and Information Systems, 10th International Conference on Applications of Natural Language to Information Systems*, Vol. 3513, Springer, Berlin, Heidelberg, 2005, pp. 263–274, ISSN 03029743. ISBN ISBN 3-540-26031-5. doi:10.1007/11428817_24. http://link.springer.com/10.1007/11428817_24.

[5] M. Lui and T. Baldwin, langid.py: An off-the-shelf language identification tool, in: *Proceedings of the ACL 2012 System Demonstrations*, Association for Computational Linguistics, 2012, pp. 25–30. https://dl.acm.org/citation.cfm?id=2390475http://dl.acm.org/citation.cfm?id=2390475.

[6] R. Sennrich, B. Haddow and A. Birch, Neural Machine Translation of Rare Words with Subword Units, in: *In Proceedings of the 54th An- nual Meeting of the Association for Computational Linguistics (ACL 2016)*, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1715–1725. ISBN ISBN 9781510827585. http://www.research.ed.ac.uk/portal/files/25478429/subword_1.pdfhttp://arxiv.org/abs/1508.07909.

[7] F. Hieber, T. Domhan, M. Denkowski, D. Vilar, A. Sokolov, A. Clifton and M. Post, Sockeye: A Toolkit for Neural Machine Translation, *ArXiv e-prints* (2017). https://arxiv.org/abs/1712.05690.

[8] M. Pinnis, M. Rikters and R. Krišlauks, Tilde's Machine Translation Systems for WMT 2018, in: *Proceedings of the Third Conference on Machine Translation (WMT 2018), Volume 2: Shared Task Papers*, Association for Computational Linguistics, Brussels, Belgium, 2018.