Impact of Data Split and Vocabulary Size in Neural Machine Translation for Slovak Language

Matúš Kleštinec

Constantine the Philosopher University Nitra
Trieda Andreja Hlinku 1, 949 01 Nitra, Slovakia
matus.klestinec@ukf.sk

Abstract. In this paper, we focus on machine translation, specifically its process from obtaining texts to evaluating the machine translation model. Our machine translation models were trained from source language English to target language Slovak. We explored the influence of various factors on the machine translation model, such as the size of the test and validation sets, vocabulary size, the impact of tokenization algorithms used, text quality and size of bilingual texts. After training the machine translation model, we also examined the effects of translation parameters on the result text. Evaluation was done using automatic metrics such as BLEU, METEOR, and COMET, as well as manual inspection of sentences. We found that the parameters we investigated had an impact on the machine translation model and optimal settings for our models.

Keywords: Machine Translation, Corpus, Slovak.

1 Introduction

Machine translation is the automated process of translating sentences from one natural language to another using computers [1]. A machine translation model trained with data, in this case, sentences in the desired languages, is capable of evaluating the most suitable words in the target language, thus creating a translation in the desired language. There are not many extensive and high-quality parallel corpora available that include Slovak language, so the optimal preprocessing and training settings have not been thoroughly explored yet. In this work, we present few possible settings for hyperparameters in the preprocessing and training steps for translating, from English to Slovak language.

2 Related Work

Other author [2] have also used the OpenNMT toolkit for training neural machine translation. The author [2] specifically used the OpenNMT-lua variant and trained on the EUR-Lex dataset in the English-Czech language pair, also tokenized text using Byte Pair Encoding. His neural machine translation models

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2024, pp. 77–83, 2024.

achieved BLEU scores 61.07 and 62.02. Authors [3] experimented with vocabulary size of models, which were trained on low-resource languages. For training, they used Akkadian, Manipuri and Lower Sorbian languages. They used automatic metrics such as BLEU and COMET for evaluation and their results showed that, for some languages, its better to have smaller vocabulary size. Another benefit is faster training time, because of smaller vocabulary. In conclusion they concluded, that optimal vocabulary size depends on language.

3 Methodology

For experimentation purposes and to account for available computational resources, we selected a relatively small parallel corpus, Europarl version 7 [4], which contains 640715 sentences in English and Slovak. Europarl corpus is already aligned, so this step was not required. During preprocessing, we applied the following steps:

- Unicode normalization,
- Removal of identical sentences: source sentence = target sentence,
- Removal of duplicate lines,
- Removal of lines that are too long,
- Filtering of text using langdetect library,
- Additional removal of certain characters resulting from previous steps,
- Transformation of text to lowercase,
- Reordering of the lines.

One step requires more detailed explanation: text filtering using the langdetect library. Filtering rows with the Python langdetect library is a useful tool when texts contain words or entire sentences that do not match the desired language or languages. Langdetect processes n-grams of the chosen text and outputs the probability that the text belongs to a particular language [5]. This way, we can filter out sentences that do not belong to English or Slovak.

By applying these preprocessing methods, we adjusted the corpus to the desired form and removed 29236 sentence pairs. Number of sentence pairs left was 611479. We tokenized the corpus into subwords using the Byte Pair Encoding algorithm for most of the models, except one, on which we used Unigram algorithm and split the corpus into training, validation, and testing sets. We trained the models using the OpenNMT-py toolkit [6], which utilizes the Transformer architecture [7]. We evaluated the models using the automatic metrics BLEU [8], METEOR [9], and COMET [10] (specifically COMET-22 [11]). The code for our solution can be found on GitHub [12].

4 The Impact of the Data Split Ratio for Training, Testing, and Validation Sets on Model Quality

We trained models V2 through V5 and model V10. The evaluation results of each model can be seen in Table 1, where the individual models are ordered by

the size of the testing and validation sets. The Test/valid value represents the number of sentences allocated separately for each set. Values highlighted in red represent the worst result within the respective column, while green represents the best result. Our scales for values BLEU, METEOR and COMET is between 0 and 1.

Table 1: Comparison of models with different testing and validation set sizes.

Model	BLEU	METEOR	COMET	Test/Valid	Test/Valid
V10	0.39361	0.6744	0.8978	2000	≈ 0.33
V5	0.39908	0.7038	0.8993	4000	≈ 0.66
V2	0.39889	0.6765	0.9008	6000	≈ 0.98
V3	0.39959	0.6890	0.9009	30000	≈ 4.91
V4	0.40082	0.4942	0.8992	60000	≈ 9.81

As you can see in Table 1, we divided the testing and validation set from approximately 0.33% to 9.81%, with the upper limit representing the standard split of 80/10/10% for training, testing, and validation sets. When comparing models based on BLEU scores, a significant difference appears only between model V10 and the other models, with V10 having a lower score by 0.00528 compared to V2 and by 0.00193 compared to V4. In the case of METEOR scores, the models maintain consistent values, except for model V4, which has a significantly different score from the others. The difference between model V4 and V10 is 0.1802, representing 18%, and is statistically significant. The COMET metric shows smaller differences, with model V10 again having the lowest rating, but it is closer to the other models. The difference between V10 and V3 is only 0.0031, representing 0.31%. The results from all metrics indicate that a small testing and validation set is inadequate, as model V10 achieved the lowest ratings in BLEU and COMET metrics, while METEOR had the second lowest rating.

Based on these findings in Table 1, we believe that the optimal split for the testing and validation sets should be approximately 0.66% to 4.91%, which we can round to 5%. This range is indicative and may not apply to all machine translation models.

5 The Impact of Vocabulary Size on Model Quality

We experimented with vocabulary sizes ranging from 2000 to 50000. Number of training steps was set to 100000. Table 2 details the models with their respective metrics, vocabulary size, and the number of training steps. The models are sorted by vocabulary size. Since we used the same vocabulary size for both the source and target languages, we only use one column labeled ”Vocabulary Size” to represent this value.

Table 2: Comparison of models with different vocabulary sizes.

Model	BLEU	METEOR	COMET	Vocabulary Size	Number of Training Steps
V12	0.39867	0.7843	0.9024	2000	100000
V9	0.42028	0.6905	0.9058	4000	100000
V11	0.45228	0.6905	0.9131	8000	100000
V6	0.46227	0.6932	0.9147	16000	80000
V7	0.47777	0.7937	0.9147	32000	55000
V8	0.44686	0.7351	0.9194	50000	35000

As the vocabulary size increased, the number of training steps decreased as you can see in Table 2. Models with a vocabulary size exceeding 8000 led to early termination of training through the early stopping condition. Models with a very small vocabulary, such as V12 and V9, had the worst BLEU scores, while models with increasing vocabulary sizes (V11, V6, and V7) achieved better results, with the highest score observed for model V7 with a vocabulary of 32000. On the other hand, the largest model, V8, showed a decrease of 0.03091 in BLEU score, likely due to overfitting. For the COMET metric, scores increased with vocabulary size, but without significant jumps the difference between the worst model, V12, and the best model, V8, was only 0.0170. METEOR scores, however, were inconsistent; model V12, which had the lowest BLEU and COMET scores, achieved the second highest METEOR score, while models V9 and V11 had the lowest METEOR scores. The largest vocabulary size did not yield the best METEOR score either, with a noticeable difference between models V7 and V8 (0.0586).

If we look at how the translated sentences from each model look, we will find that the differences are not that significant. We can demonstrate on one sentence, how different was translation for each model.

Sentence from source language: In the current situation the european union should pay particular attention to its 20 million small and mediumsized enterprises

Reference sentence from target language: V súčasnej situácii by mala európska únia osobitnú pozornosť venovať svojim 20 miliónom malých a stredných podnikov

Translation of model V12: V súčasnej situácii by mala európska únia venovať osobitnú pozornosť svojim 20 miliónom malých a stredných podnikov

Translation of model V9: V súčasnej situácii by európska únia mala venovať osobitnú pozornosť 20 miliónom malých a stredných podnikov

Translation of model V11: V súčasnej situácii by európska únia mala venovať osobitnú pozornosť 20 miliónom malých a stredných podnikov

Translation of model V6: V súčasnej situácii by európska únia mala venovať osobitnú pozornosť 20 miliónom malých a stredných podnikov

Translation of model V7: V súčasnej situácii by európska únia mala venovať osobitnú pozornosť 20 miliónom malých a stredných podnikov

Translation of model V8: V súčasnej situácii by európska únia mala venovať osobitnú pozornosť 20 miliónom malých a stredných podnikov

All sentences maintained their context after translation and are comparable to the reference sentence, with some differences in the order of certain words. Except for model V12, all models produced the same translation for the sentence, but this may not represent the entire set of translated sentences. We selected only one sentence from 4,000 sentences. Manually comparing all sentences would be time consuming.

Based on all metrics, model V12 can be considered the worst and model V7 the best in Table 2. However, the optimal vocabulary size is not universal. It depends on the corpus size and content. The experiment showed that a small vocabulary can negatively impact translation quality, while an excessively large vocabulary may not be beneficial. Further experimentation is necessary to find the optimal value for our specific model.

6 Discussion

Creating a neural machine translation model was also conducted by the author [13]. Besides standard preprocessing steps, such as removing of punctuations, we used langdetect library. This allowed us to filter out 9390 lines of text that were not in either Slovak or English. For training purposes, the author [13] also used OpenNMT and worked with vocabulary sizes of 10000 and 100000. In our approach, we experimented with various vocabulary sizes ranging from 2000 to 50000, which led to significantly higher BLEU scores. For comparison, the BLEU score for the models trained by the author [13] with a vocabulary size of 10000 was 0.0604, and 0.1217 for a vocabulary size of 100000. We converted the BLEU scores presented by the author [13] to our 0-1 scale for consistency. Our model with a vocabulary size of 2000, which was the lowest performer in our comparison in Table 2, achieved a BLEU score of 0.39867. The number of training steps also differed, as most of our models were trained for 100000 steps and validated every 2500 steps, unlike the author's models, which were trained for 20000 steps and validated every 1000 steps. Most notable difference is that author used 2 layer LSTM architecture, while we used 6 layer Transformer architecture.

In this work, we introduced machine translation and its significance, the steps necessary to achieve a functional machine translation model, and addressed the settings for training and translation parameters.

Based on experiments with the size of the test and validation set distribution, we found that the most suitable division ranges from approximately 0.66% to 4.91%. These values are indicative, representing an approximate distribution that yielded the best results. It is important to note that for a larger model, this range could differ. We worked with a relatively small corpus, making it impossible to experiment with a larger number of sentences with the same quality.

Experiments with vocabulary size demonstrated that vocabulary size can significantly influence the model. An excessively small vocabulary can negatively impact the final model, but the same holds true for an excessively large vocabulary. Based on our findings, we concluded that it is always necessary to consider the ideal vocabulary size, which may vary for each solution. We only experimented with English and Slovak, using one corpus size, so our values apply only to this pair of languages and for an approximate corpus size. The languages have different lexical richness, which could also affect the appropriate vocabulary size for the machine translation model. A larger corpus could contain a richer vocabulary, which would also influence the suitable vocabulary size.

By comparing models with different tokenization algorithms, we reached the conclusion that Byte Pair Encoding yields better results than the Unigram algorithm. The experiment with tokenization algorithms was conducted on a small sample, as we compared only two models. While the sample is small, all metrics, especially METEOR, indicate a decline in translation quality.

The numerous possibilities make it both time consuming and computationally demanding to test everything, which can be considered the greatest limitation of this work. Testing all possible combinations of settings would be impractical, so our results cover only a small portion of potential solutions for improving machine translation. Another limitation is the relatively small amount of quality bilingual texts in English and Slovak. While Europarl is a high quality corpus, its quite small.

Acknowledgements. This work was supported by the Slovak Research and Development Agency under the Contract no. APVV-23-0554.

Disclosure of Interests. The author have no competing interests to declare that are relevant to the content of this article.

References

1. Tan, Z. Wang, S. Yang, Z. Chen, G. Huang, X. Sun, M. Liu, Y.: Neural machine translation: A review of methods, resources, and tools. In: AI Open, Volume 1, pp. 5-21. (2020) https://doi.org/10.1016/j.aiopen.2020.11.001
2. Wörgötter, B.: Domain-specific English-Czech Neural Machine Translation, `https://is.muni.cz/th/k8nt8/` (2018)
3. Signoroni, E. Rychlý, P.: Better Low-Resource Machine Translation with Smaller Vocabularies. In: Nöth, E., Horák, A., Sojka, P. (eds) Text, Speech, and Dialogue. TSD 2024. Lecture Notes in Computer Science(), vol 15048. Springer, Cham. (2024) `https://doi.org/10.1007/978-3-031-70563-2_15`
4. Tiedemann, J.: Parallel Data, Tools and Interfaces in OPUS. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pp. 2214–2218. European Language Resources Association (ELRA), Istanbul, Turkey. (2012)
5. Langdetect, `https://github.com/fedelopez77/langdetect?tab=readme-ov-file`, last accessed 2024/02/29

6. Klein, G. KIM, Y. DENG, Y. SENELLART, J. RUSH, A.: OpenNMT: Open-Source Toolkit for Neural Machine Translation. In: Proceedings of ACL 2017, System Demonstrations, pp. 67-72. Association for Computational Linguistics, Vancouver, Canada (2017)
7. Vaswani, A. Shazeer, N. Parmar, N. Uszkoreit, J. Jones, L. Gomez, A. Kaiser, L. Polosukhin, I.: Attention Is All You Need. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000-6010. Curran Associates Inc, 57 Morehouse LaneRed HookNYUnited States (2017)
8. Papineni, K. Roukos, S. Ward, T. Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 311-318. Association for Computational Linguistics, Philadelphia, USA (2002)
9. BANERJEE, S. LAVIE, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp 65-72. Association for Computational Linguistics, Ann Arbor, Michigan (2005)
10. REI, R. STEWARD, C. FARINHA, A. LAVIE, A.: COMET: A Neural Framework for MT Evaluation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2685-2702. Association for Computational Linguistics (2020)
11. REI, R. SOUZA, J. ALVES, D. ZERVA, C. FARINHA, A. GLUSHKOVA, T. LAVIE, A. COHEUR, L. MARTINS, A.: COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In: Proceedings of the Seventh Conference on Machine Translation (WMT), pp. 578-585. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (2022)
12. GitHub repository, `https://github.com/ukf-matusklestinec/Strojovy-preklad-DP`
13. Pavlišin, P.: Slovenský neurónový strojový preklad pomocou knižnice OpenNMT, `https://opac.crzp.sk/?fn=detailBiblioForm&sid=2F8FD603D177BA679037F0795407` (2022)