
UNIVERZITA
KONŠTANTÍNA
FILOZOFA
V NITRE

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

Matúš Kleštinec

matus.klestinec@ukf.sk

Faculty of Natural Sciences and Informatics, University of Constantine Philosopher in Nitra
7.12.2024

# Data Split - test and validation set

| Number of sentences | % of whole dataset |
|---:|---:|
| 2000 | ≈ 0.33 |
| 4000 | ≈ 0.66 |
| 6000 | ≈ 0.98 |
| 30 000 | ≈ 4.91 |
| 60 000 | ≈ 9.81 |

# Vocabulary in machine translation

- Vocabulary size was set during tokenization using SentencePiece text tokenizer
- Experiment was done with vocabulary sizes: 2000, 4000, 8000, 16 000, 32 000 and 50 000

# Preprocessing steps

| Name of step | Number of rows | Removed or modified lines |
|---|---|---|
| Europarl (ENG - SK) | 640 715 | - |
| Removal of empty rows | 639 954 | 761 |
| Unicode normalization | 639 954 | ENG(17) / SK(13) |
| Removal source = target sentence | 636 704 | 3 250 |
| Removal of duplicate lines | 620 889 | 15 815 |
| Removal of lines that are too long | 620 873 | 16 |
| Additional removal of empty rows | 620 869 | 4 |
| Filtering texts using LangDetect library | 611 479 | 9 390 |
| Reordering of lines | 611 479 | - |

# LangDetect library

| Europarl - english | Europarl - slovak |
| --- | --- |
| the sitting was opened at 9 am | συνεδρίαση αρχίζει στις 9 πμ |
| report elles | informe elles |
| i shall start with a short announcement | začnem krátkym oznámením |
| simple | Je to jednoduché |


# Tokenization of text

- Tokenization with Sentencepiece
- Subword tokenization
- Used tokenization – Byte Pair Encoding

# Training of model

- Training was done using OpenNMT-py toolkit
- Transformer architecture
- Number of training steps = 100 000
- Validation every 10 000 steps


# Evaluation

Evaluation was done using automatic metrics:

- BLEU – library sacrebleu
- METEOR – library spaCy
- COMET – model COMET-22

# Comparison of models with different testing and validation set sizes

| Model | BLEU | METEOR | COMET | Test / Valid size | 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 |

# Comparison of models with different testing and validation set sizes

- Optimal split for 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.


# Comparison of models with different vocabulary sizes

| Model | BLEU | METEOR | COMET | Vocab. Size | Num. 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 |

# Comparison of models with different vocabulary sizes

- **Sentence from source language**: In the current situation the european union should pay particular attention to its 20 million small and medium sized 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
- **Model V9**: V súčasnej situácii by európska únia mala venovať osobitnú pozornosť 20 miliónom malých a stredných podnikov
- **Model V11**: V súčasnej situácii by európska únia mala venovať osobitnú pozornosť 20 miliónom malých a stredných podnikov
- **Model V6**: V súčasnej situácii by európska únia mala venovať osobitnú pozornosť 20 miliónom malých a stredných podnikov
- **Model V7**: V súčasnej situácii by európska únia mala venovať osobitnú pozornosť 20 miliónom malých a stredných podnikov
- **Model V8**: V súčasnej situácii by európska únia mala venovať osobitnú pozornosť 20 miliónom malých a stredných podnikov


# Comparison of models with different vocabulary sizes - Conclusion

- 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.

UNIVERZITA KONŠTANTÍNA FILOZOFA V NITRE

# Thank you for your attention.

MATÚŠ KLEŠTINEC

matus.klestinec@ukf.sk