Generating coherent text from noisy speech transcripts

Leif Førland Schill, Pranjal Garg and Virgile Blanchet-Møhl

1 Department of Mathematics and Computer Science, Technical University of Denmark; 2 DTU Compute, Technical University of Denmark

Introduction

Automatic speech recognition models are capable of producing accurate transcripts, but these transcripts will often contain a variety of errors. This project aims to remedy this issue through a computationally inexpensive model that takes in noisy transcripts produced by aforementioned models and produces accurate coherent speech.
Our model uses the Multilingual Lexical Normalization model[1] which is based on the ByT5[2] model, which we have fine tuned on new data for the Danish language to improve the lexical normalization produced by the multilexnorm for Danish

The Multilex Normalization Model

1. Based on ByT5 foundation model which is a byte-level generative sequence to sequence model which processes a sequence of bytes of UTF-8 encoding on input and produces a sequence of UTF-8 encoding bytes on output.
2. Pretrained on synthetic data and fine tuned on authentic data pulled from social media websites mainly twitter.
3. Uses masking for span of around 20 bytes using special sentinel tokens and reconstructs the masked span.
4. Best performance in multilingual Lexical Normalization for W-NUT 2021: Multilingual Lexical Normalization shared task.

Dataset

The dataset used in this project consists of noisy danish speech transcripts produced from audiobooks, the text of the audiobook being used as a reference. Each data sample contains the transcription of a sentence, the reference text for said sentence, the audiobook of origin and the Word error rate (WER) over the sentence. The WER is used to determine which sentences contain noise and need to be corrected. The data samples from these transcriptions are saved as individual sentences in which the words are mapped into singular strings, thus retaining the context of the sentence when evaluating each individual word.

Data pre-processing

- Excluding sentences with wer =0 : Sentences with no errors were filtered out of the dataset to just focus on noisy texts.
- Alignment of the data- The input and output sentences could possible have different number of words which was not accepted by the model. So a word to word mapping had to be done for input to output where some words were mapped to an empty string if they were missing in either the input or output.

Examples

Unnormalized:
han ku sgu osse bare la være med at være så urimelig.
Normalized:
han kunne sgu også bare lade være med at være så urimelig.

Input:
han ku sgu <extra_id_0 > osse <extra_id_1 > bare la være med at være så urimelig.
Output:
også

Error measures

- Word error rate / Character error rate:

$$\text{WER / CER} = \frac{S + D + I}{N}$$

- Error Reduction rate

$$\text{ERR} = \frac{TP - FP}{TP + FN}$$

- TN = Annotator did not normalize, system did not normalize.
- FP = Annotator did not normalize, system normalized.
- FN = Annotator normalized, system normalized incorrectly or did not normalize at all.
- TP = Annotator normalized, system normalized correctly.

Baselines

- Leave-As-Is (LAI): Do not not normalize anything, use input as output.
- Most-frequent-Replacement (MFR): Use most frequent replacement based on the training data.
- Unfinetuned byt5 small multilexnorm2021 danish model.

Results

Table 1: Comparison of generalization performances between the Untrained MLN and Trained MLN.

Model	Test sentences number	Word error rate	Correct predictions
Untrained MLN	3601	1.00444	25
Trained MLN	3601	0.34296	2405

As it can be seen from the table above, just finetuning on a new dataset has increased the performance of multilex norm model by over 10 times

Key Learnings

- Our system achieves high degree of improvement when trained on a new formalized Danish dataset compared to the original multilexnorm model.
  - Applying transfer learning on a pretrained transformer is the key to create a computationally inexpensive language processing model.
  - Byte type models which produce sequence of bytes perform remarkably well on noisy text data.
- Training a pretrained model without adding any extra layers is also enough when using transfer learning to significantly improve model performance

Future Work

- Currently for the model input sentences have to be encoded separately for each word which can be improved through separating the input words through different sentinel tokens.
- Adding another LSTM layer over the top of existing model can capture structural language indicators especially in Danish with lower flexibility in its structure .

References

HappyTransformer: https://www.vennify.ai/fine-tune-grammar-correction/ T5 blogpost:
https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html
Word Error Rate (WER): "https://en.wikipedia.org/wiki/Word_error_rate"
Bi-Lingual Evaluation Understudy: "https://en.wikipedia.org/wiki/BLEU"
GLEU: https://huggingface.co/metrics/gleu
(https://aclanthology.org/P15-2097.pdf)

[1] D. Samuel and M. Straka. UFAL at MultiLexNorm 2021: Improving multilingual lexical normalization by fine-tuning ByT5. In *Proceedings of the 7th Workshop on Noisy User-generated Text (W-NUT 2021)*, Punta Cana, Dominican Republic