A Large Norwegian Dataset for Weak Supervision ASR

Per Erik Solberg
per.solberg@nb.no

Pierre Beauguitte
pierre.beauguitte@nb.no

Per Egil Kummervold
per.kummervold@nb.no

Freddy Wetjen
freddy.wetjen@nb.no

National Library of Norway, Norway

Abstract

With the advent of weakly supervised ASR systems like Whisper, it is possible to train ASR systems on non-verbatim transcriptions. This paper describes an effort to create a large Norwegian dataset for weakly supervised ASR from parliamentary recordings. Audio from Stortinget, the Norwegian parliament, is segmented and transcribed with an existing ASR system. An algorithm retrieves transcripts of these segments from Stortinget's official proceedings using the Levenshtein edit distance between the ASR output and the proceedings text. In that way, a dataset of more than 5000 hours of transcribed speech is produced with limited human effort. Since parliamentary data is public domain, the dataset can be shared freely without any restrictions.

1 Introduction

Until recently, automatic speech recognition (ASR) models needed to be trained on speech recordings and verbatim transcriptions of such recordings. Transcribing speech manually word-by-word takes time, and is often difficult. Therefore, there was a limited amount of suitable training data for smaller languages. At the time of writing, the ASR datasets with an open license available for Norwegian amount to less than 1000 hours (Solberg and Ortiz, 2022). With the advent of transformer-based ASR systems, the possibility has opened up for using non-dedicated datasets as training data for ASR (Baevski et al., 2020; Radford et al., 2022). In 2022, OpenAI released Whisper, a multilingual ASR model trained on 680 000 hours of audio and subtitles collected from videos online, which performs well across many languages (Radford et al., 2022).

Weakly supervised ASR training, training ASR models on non-verbatim transcriptions such as subtitles, makes it possible to harvest large quantities of pairs of audio and transcription, which is a great advantage for under-resourced languages like Norwegian. As Whisper is not trained on verbatim transcriptions, it often does not produce verbatim transcriptions either: It tends to write easily readable text even when the speaker stutters, repeats herself etc. To some degree one might say that the transcriptions represent what the speaker means rather than what the speaker says. For many downstream tasks such as transcribing videos and interviews, this is the desirable outcome.

This paper describes an ongoing effort to create a dataset for weakly supervised ASR training with data from Stortinget, the Norwegian parliament. Using a method described in (Ljubešić et al., 2022), we have aligned segments of audio from meetings at Stortinget with corresponding text segments in the official proceedings. This results in a 5000+ hours dataset, 1245 of which may be considered near-verbatim. An advantage of using parliamentary data is that there are no copyright or privacy restrictions on the audio data or the textual data. The dataset will therefore be made available with an open license on HuggingFace.

2 Background

Until 2021, the only large, open dataset for Norwegian ASR training was a 540 hour dataset with manuscript-read speech made by the firm Nordisk språkteknologi (NST) at the beginning of the millenium.[1] This is a reasonable dataset for ASR of scripted speech, but it is insufficient as training data for unscripted, spontaneous speech, as it lacks dialectal phenomena as well as hesitations, stuttering, repetitions and other features typical of unplanned speech. Moreover, the dataset has

[1] https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-54/

Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023), pages 48–52, May 22, 2023

almost exclusively transcriptions in Bokmål, the more used of the two written standards of Norwegian. There were no appropriate datasets for ASR training with transcriptions in the other standard, Nynorsk, which is closer in spelling to many rural dialects, particularly on the west coast, and is used by a sizeable minority of the population.

In 2021, the Language Bank at the National Library of Norway released the Norwegian Parliamentary Speech Corpus (NPSC), consisting of 126 hours of transcribed speech from meetings at Stortinget in 2017 and 2018.[2] The NPSC is transcribed verbatim by linguists and philologists, and the transcriptions are tailor-made for ASR training and testing. While the members of parliament often read from a manuscript, a large amount of the NPSC is unplanned speech and dialectal speech, and Solberg and Ortiz (2022) showed that training ASR models on the NPSC has a beneficial effect on the word error rate (WER) across Norwegian dialects compared to models trained on the NST dataset only. Moreover, using the NPSC as training data decreased the differences in WER between dialects.

About 13% of the transcriptions in the NPSC are in Norwegian Nynorsk. The AI lab at the National Library of Norway fine-tuned separate, unsupervised wav2vec2 models on the Bokmål and Nynorsk transcriptions of the NPSC (De la Rosa et al., 2023). The 1B parameter Bokmål model[3] obtains a WER of 6.4% on the Bokmål part of the NPSC test set, and the 300M parameter Nynorsk model[4] obtains a WER of 12.7% on the Nynorsk part of the test set. In informal tests, the models also give decent results on non-parliamentary data, although not quite as good as on recordings from Stortinget.

While the NPSC is a valuable resource for speech recognition of Norwegian, more data is needed. However, manual, verbatim transcription is very time-consuming, expensive and difficult to do in a systematic way. There are often both recordings and official non-verbatim transcripts of meetings of parliaments, and there have been several successful efforts of extracting transcription from parliamentary data (Helgadóttir et al., 2017; Iranzo-Sánchez et al., 2020; Kirkedal et al., 2020; Virkkunen et al., 2022). As the parliamentary transcripts are not verbatim and older ASR systems need verbatim transcriptions as training data, the extraction process often involves significant data cleaning to retain only fragments which are verbatim or close to verbatim (see e.g. Kirkedal et al., 2020).

Ljubešić et al. (2022) describe the creation of a 1860 hour ASR dataset from Croatian parliamentary data, ParlaSpeech-HR. The transcriptions are automatically extracted and are not necessarily verbatim, but a score indicates how closely the ASR output and the extracted transcription match. They fine-tuned multilingual and Slavic wav2vec2 models on a 300 hour subset of the data. The best-performing model has a WER of 4.3% on the test set of ParlaSpeech-HR.

Radford et al. (2022) show that it is possible to obtain good ASR results with weak supervision, using subtitles as training data. Audio segments of parliamentary speech aligned with official proceedings may be a good data source for weakly supervised ASR: There is a large amount of accessible speech data, and there are transcriptions created by humans. Furthermore, like subtitles, the parliament proceedings transcribe what the speakers mean instead of transcribing word by word what the speakers say, and they correspond more closely to the desirable output in many downstream tasks.[5]

3 Producing the dataset

3.1 Retrieving and segmenting the data

The audio for the dataset was obtained by retrieving the links to all the videos in the Stortinget video archive[6], which contains videos of most plenary meetings from 2010 to 2022. We then ran the audio files through voice activity detection (VAD) using Silero VAD.[7] The tool often produces segments of just a few seconds. These segments are too short for the subsequent extraction task. We therefore used code from the ParlaSpeech-HR

[2] https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-58/

[3] https://huggingface.co/NbAiLab/nb-wav2vec2-1b-bokmaal

[4] https://huggingface.co/NbAiLab/nb-wav2vec2-300m-nynorsk

[5] There are situations where transcriptions should be verbatim, such as investigative police interviews and linguistic research. For such use cases, wav2vec2 models trained on dedicated ASR datasets, e.g. the NPSC, may be a better option.

[6] https://www.stortinget.no/no/Hva-skjer-pa-Stortinget/Videoarkiv/Videoarkiv/

[7] https://github.com/snakers4/silero-vad

project which merges shorter segments together to larger segments up to 30 seconds.[8] Using the timestamps from this segmentation, we split the audio extracted from the Stortinget video files into smaller mp3 files with a sampling rate of 16 kHz. The total duration of the identified speech segments is 6182 hours.

The textual data from the official proceedings are taken from the ParlaMint-NO corpus, which contains the proceedings from Stortinget from 1998 to 2022.[9] ParlaMint-NO is part of the European ParlaMint project and contains rich metadata on a standardized format, e.g. on the gender and date of birth of the speakers and the written standard (Bokmål or Nynorsk) of the paragraphs (Erjavec et al., 2022). We have not yet used the metadata from ParlaMint-NO in this project.

3.2 Producing Automatic Transcriptions

All audio segments were transcribed using the NBAiLab wav2vec2 models, as described in (De la Rosa et al., 2023). These models exist both as a large (1B) Norwegian Bokmål version and a base (300M) Norwegian Nynorsk version. Both models were used in the processing of the audio segments, producing a transcription in Bokmål and a transcription in Nynorsk for each segment. The models were obtained from HuggingFace, and the Transformers Pipeline was employed for automatic batching and transcription of the segments, which were executed on a NVIDIA A6000 GPU.

3.3 Matching ASR and proceedings

In order to associate an audio excerpt with its corresponding portion of the proceedings, we are left with the task of searching for a short text (ASR output) within a longer text, while allowing for small differences. There exists many different solutions to this problem, but the method used in (Ljubešić et al., 2022) fits our purpose.

In short, the method starts by finding occurrences of the first word of a segment in the proceedings, then computes a score based on the Levenshtein edit distance[10] between the segment and the portion of the proceedings starting at one such position. A match is found when the score is high enough, and the search continues for the next segment on the remainder of the proceedings. If the first word of the segment is not found, it looks for the second, then the third, etc. This method is not guaranteed to find the best match for a segment, as it can put too much emphasis on the start of the segment, but computing a full substring edit distance for each segment would be much more time-consuming. We only retained matches with a score larger than 0.5.

A popular method for similar tasks is described in (Panayotov et al., 2015). It requires running the Smith-Waterman local alignment algorithm with the whole reference text and the sequence of segments. The method we used is much less time-consuming, as it computes edit distances only between individual segments and small excerpts of the reference text, and as it only looks for the starting words of the next segment in the remainder of the reference text. Panayotov et al. (2015) also aim at finding near-exact matches, whereas we are also interested in non-verbatim transcriptions.

We used the code[11] released alongside (Ljubešić et al., 2022) as our starting point, and modified it to suit the specifics of our project, with the following normalization steps:

- Our ASR model was trained to transcribe vocal and nasal hesitations and unintelligible sounds as *eee*, *mmm* and *qqq* respectively. These were simply removed from the segment before performing the search.
- The output from our model contains no punctuation or upper case letters. To accommodate this, the tokenized proceedings were stripped of all punctuation and lower cased before matching. The original tokens were also preserved so we could reconstruct the original text from the positions of a match.
- Numbers are typically written with digits in the proceedings, and spelled out in the ASR output. An inverse-normalization filter developed by the Language Bank is applied to transform them to digits.[12]

Our code is publicly available.[13] As mentioned above, we actually run one ASR model for each

[8] https://github.com/danijel3/CroatianSpeech/blob/main/Croatian.ipynb

[9] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-77/

[10] the `ratio` method of the Levenshtein python package is used: https://maxbachmann.github.io/Levenshtein/levenshtein.html#ratio

[11] https://github.com/clarinsi/parlaspeech

[12] https://github.com/Sprakbanken/sprakbanken_normalizer

[13] https://github.com/Sprakbanken/transcription_matching

Score	Bokm.	Nyn.	Total	Coverage
> 0.5	4566h	624h	5190h	84%
> 0.8	3168h	176h	3345h	54%
> 0.9	1229h	16h	1245h	20%

Table 1: Duration and percentage of the total speech audio for subsets with a match score larger than 0.5 (i.e. the whole matched dataset), 0.8 and 0.9.

written standard of Norwegian. Each audio segment then results in two ASR outputs. The matching algorithm is run with both outputs, and we retain the one getting the highest score.

4 Results

By using this method, we extracted transcriptions for 724 783 segments, which amounts to 5190 hours, i.e. 84% of the duration of the identified speech segments (6182 hours, cf. section 3.1). The average segment word length for the extracted transcriptions is 59.15.

Table 1 reports the duration in hours with different match scores.[14] As the score is a calculation of the distance between the ASR output and the matched segment, there are two requirements for the score to be high: Firstly, the proceedings text must report what was said quite faithfully. Secondly, the ASR output must be an accurate representation of the speech in the audio file. The second requirement entails that the segments with high scores are already handled well by the ASR systems used in this project. We suspect that segments with a lower match score may be useful in weakly supervised ASR training, both because they are less verbatim and because they may be of a kind that current models have been less exposed to and therefore handle less well.

5 Future work

For now, we have only extracted texts from the ParlaMint-NO corpus. However, ParlaMint-NO contains rich metadata. The complete release of the dataset will include ParlaMint speaker identifiers as well as identifiers of the relevant sections in ParlaMint-NO. With this information, it is possible to couple the segments of the dataset with the metadata in ParlaMint-NO and enrich them with, e.g. the gender and age of the speakers. We can also extract language codes directly from the metadata, which is likely a more accurate method for identifying Bokmål and Nynorsk segments than the ASR-based method used here.

The best way to evaluate the quality of this dataset is to use it to train a weakly supervised ASR system. We plan to use it to train Whisper models for Norwegian and compare the performance of these models to existing Whisper models and other Norwegian ASR models. Once this training is complete, we will release the scores, which will give more insights into how useful this dataset may be.

6 Conclusion

This paper describes the creation of a large, transcribed audio dataset for ASR training with relatively small human effort. When it is released, the dataset will be several times larger than all openly available, transcribed audio data in Norwegian today. We believe that the dataset is the largest open dataset of its kind that can be made for Norwegian, as there do not exist, to our knowledge, any other sources of speech recordings and corresponding transcriptions that are free of copyright and privacy restrictions. Parliamentary speech is also of a quite different genre from most subtitled videos, which currently constitute the primary source of data for weakly supervised ASR, so we expect the dataset to add valuable variation. Moreover, the dataset contains 623 hours of speech transcribed to Nynorsk, 176 of which have a match ratio above 0.8. We expect this data to be a particularly valuable addition, as there is currently very limited training data for Nynorsk ASR and few systems that support Nynorsk.

Acknowledgments

This work has been partially supported by the Research Council of Norway through the IKTPLUSS grant for the SCRIBE project[15] (KSP21PD).

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

[14] In rare cases where the score is exactly the same for Bokmål and Nynorsk, the segment is assumed to be in Bokmål.

[15] https://scribe-project.github.io/

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, et al. 2022. The ParlaMint corpora of parliamentary proceedings. *Language resources and evaluation*, pages 1–34.

Inga Rún Helgadóttir, Róbert Kjaran, Anna Björk Nikulásdóttir, and Jón Guðnason. 2017. Building an asr corpus using Althingi's parliamentary speeches. In *Proceedings of INTERSPEECH 2017*, pages 2163–2167.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerda, Javier Jorge, Nahuel Roselló, Adria Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Andreas Kirkedal, Marija Stepanovic, and Barbara Plank. 2020. FT SPEECH: Danish parliament speech corpus. In *Proceedings of INTERSPEECH 2020*, pages 111–116.

Nikola Ljubešić, Danijel Koržinek, Peter Rupnik, and Ivo-Pavao Jazbec. 2022. ParlaSpeech-HR: a freely available ASR dataset for Croatian bootstrapped from the ParlaMint Corpus. In *Proceedings of the ParlaCLARIN III Workshop*, pages 111–116.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Javier De la Rosa, Braaten Rolv-Arild, Per E Kummervold, and Freddy Wetjen. 2023. Boosting Norwegian automatic speech recognition. In *Proceedings of the 24rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, Faroe Islands. Linköping University Electronic Press, Sweden.

Per Erik Solberg and Pablo Ortiz. 2022. The Norwegian Parliamentary Speech Corpus. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 1003–1008.

Anja Virkkunen, Aku Rouhe, Nhan Phan, and Mikko Kurimo. 2022. Finnish parliament ASR corpus - analysis, benchmarks and statistics.