The Norwegian Parliamentary Speech Corpus

Per Erik Solberg [1] Pablo Ortiz [2]

[1]National Library of Norway [2]Telenor Research

Abstract

The Norwegian Parliamentary Speech Corpus (NPSC) is a speech dataset with recordings of meetings from Stortinget, the Norwegian parliament. It is the first, publicly available dataset containing unscripted, Norwegian speech designed for training of automatic speech recognition (ASR) systems. The recordings are manually transcribed and annotated with language codes and speakers, and there are detailed metadata about the speakers. The transcriptions exist in both normalized and non-normalized form, and non-standardized words are explicitly marked and annotated with standardized equivalents. To test the usefulness of this dataset, we have compared an ASR system trained on the NPSC with a baseline system trained on only manuscript-read speech. These systems were tested on an independent dataset containing spontaneous, dialectal speech. The NPSC-trained system performed significantly better, with a 22.9% relative improvement in word error rate (WER). Moreover, training on the NPSC is shown to have a "democratizing" effect in terms of dialects, as improvements are generally larger for dialects with higher WER from the baseline system.

1. Introduction

- The Norwegian Parliamentary Speech Corpus (**NPSC**): first publicly available dataset for training ASR of Norwegian unscripted speech
- Transcribed speech from Stortinget, the Norwegian parliament
- Distributed by the Language Bank at the National Library of Norway

`https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-58/`

2. The Content of the NPSC

Audio files: sentence-segmented audio files of 41 days of parliamentary debates from 2017 and 2018

Transcriptions: manual transcriptions in **Bokmål and Nynorsk**, the two written standards of Norwegian

- **Non-normalized and normalized** transcriptions
- **Sentence-tokenized and word-tokenized** transcriptions
- **Training, test and evaluation** sets (80/10/10%)

Speaker metadata: Name, gender, place and date of birth, dialect etc.

	Bokmål	Nynorsk	Total
Duration pauses incl.	-	-	140.3h
Duration pauses excl.	110.5h	15.7h	126.2h
Word count	1.054M	144K	1.198M
Sentences	56.2K	8.3K	64.5K
Language distribution	87.2%	12.8%	100%
Gender	F: 39.2%	F: 32.6%	F: 38.3%
distribution	M: 60.8%	M: 67.4%	M: 61.7%

Table 1. Corpus statistics.

3. Making the NPSC

Why use parliamentary recordings?

- The audio files are public domain
- The speakers are public figures, so there is a lot of metadata about them
- There is a lot of dialect variation, since MPs tend to speak their own dialect in parliament
- Official proceedings exist, which can be used in the preprocessing

Preprocessing steps:

1. The audio files of the parliamentary meetings are transcribed with Google Cloud StT
2. A script replaces words in the transcriptions with words from the official proceedings

Manual transcription: The ASR transcriptions are corrected by trained linguists/philologists. Each sentence is annotated with the name of the speaker. The transcriptions are proofread by a fellow transcriber.

Postprocessing of the transcriptions: Metadata about the speakers is extracted from Wikidata. A dialect is added manually for each speaker. Normalization grammars produce normalized versions of the transcriptions.

4. Evaluating the dataset

ASR model. We use DeepSpeech 2 [Amodei et al. 2015] as acoustic model (AM), combined with n-gram language model (LM).

Datasets. We train and test models using the following data:

- Nordic Speech Technology (**NST**). Clean, manuscript-read speech, recorded on silent environments. 300 hours used for training acoustic and language models.
- **NPSC**. Relatively noisy and spontaneous. Dialectal speech with moderate use of dialect-specific and non-standard words. 100 hours used for training acoustic and language models.
- NB Tale module 3 (**NB**). Free, spontaneous, dialectal speech with frequent use of dialect-specific and non-standard words. Classified in 12 dialect groups. 6.4 hours used for testing.

All of the above datasets are distributed by the Language Bank at the National Library of Norway: `https://www.nb.no/sprakbanken/en/resource-catalogue/`.

Experiments and results. We test NST and NPSC acoustic models with different language models on an independent dataset with highly spontaneous and dialectal speech, namely NB:

Model	AM	LM	NST_{test}	$\text{NPSC}_{\text{test}}$	NB ($\bar{\sigma}_{\text{dial}}/\bar{\sigma}_N$)
M_1	NST	LM_{base}	2.9	40.6	48.4 (1.092)
M_2	NPSC_{NST}	$\text{LM}_{\text{NST+NPSC}}$	-	15.9	39.6 (0.897)
M_3	NPSC_{NST}	LM_{base}	-	17.8	37.3 (0.984)
M_4	NPSC_{NST}	LM_{NPSC}	-	17.1	41.5 (-)

Table 2. Performance measured in terms of word error rate (WER) percentage. For the NB data, we also measure the standard deviation across dialects normalized by the average WER ($\bar{\sigma}_{\text{dial}}$) relative to the normalized standard deviation of the number of samples for each dialect ($\bar{\sigma}_N$).

NPSC_{NST} is an acoustic model pre-trained on the NST data and fine-tuned on the NPSC data.
LM_{base} is a 5-gram language model trained with $\approx$ 13 million sentences from newspapers.
$\text{LM}_{\text{NST+NPSC}}$ is a 3-gram language model trained with $\approx$ 300,000 sentences from NST and NPSC.
LM_{NPSC} is a 3-gram language model trained with $\approx$ 50,000 sentences from NPSC.

Analysis of results

- The model trained on NST data (M_1) performs very well on clean data (WER = 2.9%), but performance notably decreases on more realistic datasets (WER > 40%). **Applying ASR to realistic situations requires data resources such as NPSC**.
- The models fine-tuned on NPSC data (M_2, M_3, M_4) naturally perform much better on NPSC. We also observe relative improvements on NB data up to **22.9%**, meaning that **the NPSC data improves the adaptation of ASR models to spontaneous, dialectal speech**.
- The differences across dialects for M_1 on NB data is larger than what one would expect from statistical fluctuations due to the number of samples per dialect ($\bar{\sigma}_{\text{dial}} > \bar{\sigma}_N$). Using the NPSC data we reduce dialect biases up to a relative **17.94%**. Thus **the NPSC data makes ASR models more democratic towards the whole spectrum of dialects**.

5. Future work and Conclusions

- Standardized output and written standards. Modelling the two written standards of Norwegian (Bokmål and Nynorsk) requires either two separate sets of data and models or a single two-step model with output standardization. Further research will explore the different scenarios to find out what is best.
- Dialectal forms, optimization and metrics. Dialect-specific and non-standard words include inflections and variations which usually are not part of the dictionary. Loss functions and metrics that do not have a notion of semantics identify these variants as a full error, which strongly disagrees with human judgement. Implementing semantic-aware losses and metrics is the subject of future work, however the flexibility towards multiple written variants works against a standard output, so this trade-off has to be considered as well.
- Non-verbal noises. Fillers and hesitations are ubiquitous in spontaneous speech. Here we model them as "blanks", but future work will aim to exploit the information implicitly carried by non-verbal noises.

Take-home message

The NPSC is a highly documented and annotated public resource. With its more than 126 hours of transcribed speech, it is incredibly valuable for a low-resource language as Norwegian. In addition, it enables further research of ASR methods for any language, in particular aspects of spontaneous speech, dialects and written variants, and their relationship with optimization, metrics and output normalization. Using the NPSC substantially improves ASR quality on spontaneous, dialectal speech.

References

Amodei, D. et al. (2015). "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin.". In: *CoRR* abs/1512.02595. URL: `http://arxiv.org/abs/1512.02595`.