SwissText 2021 Task 3: Swiss German Speech to Standard German Text

Michel Plüss **Lukas Neukom** **Manfred Vogel**
Institute for Data Science
University of Applied Sciences and Arts Northwestern Switzerland
Windisch, Switzerland
michel.pluess@fhnw.ch

Abstract

We present the results and findings of SwissText 2021 Task 3 on Swiss German Speech to Standard German Text. Participants were asked to build a system translating Swiss German speech to Standard German text. The objective was to maximize the BLEU score on a new test set covering a large part of the Swiss German dialect landscape. Four teams participated, with the winning contribution achieving a BLEU score of 46.0.

1 Introduction

Swiss German is a family of dialects spoken by around five million people in Switzerland. It is different from Standard German regarding phonetics, vocabulary, morphology, and syntax. Swiss German is mostly a spoken language. While it is also used in writing, particularly in informal text messages, it lacks a standardized writing system. This leads to difficulties for automated text processing such as spelling ambiguities and a huge vocabulary size. Therefore, most use cases for a Swiss German speech-to-text (STT) system require Standard German text as output. This can be viewed as a speech translation problem with similar source and target languages. For example, the Swiss German sentence "Ide Abfahrt hetter de sächsti Platz beleit" can be translated to the Standard German sentence "In der Abfahrt belegte er den sechsten Platz". Here, the sentence structure is very similar, but the past tense changes in Standard German.

Speech-to-text systems for well-resourced languages like English or Standard German work very well. Zhang et al. (2020) set the current state-of-the-art on the popular LibriSpeech test-other benchmark (Panayotov et al., 2015) with a word error rate (WER) of 2.6 %. In comparison, the 2020 shared task on Swiss German STT (Plüss et al., 2020), this task's predecessor, was won by Büchi et al. (2020) with a WER of 40.3 %.

The goal of this task is to spur further progress in the field of Swiss German STT by providing a larger labeled training set, an additional unlabeled training set, and a test set with a dialect distribution similar to the real distribution of Swiss German dialects in Switzerland.

The remainder of this paper is structured as follows: the task, the data, and the evaluation of submissions are described in section 2. An overview of the submissions and results of this task can be found in section 3. Section 4 wraps up the paper and gives directions for future work.

2 Task Description

The objective of the task is to build a sentence-level Swiss German speech to Standard German text speech translation system. The submission with the best BLEU score (Papineni et al., 2002) wins. Participants were encouraged to explore and combine suitable supervised, semi-supervised, and unsupervised learning approaches.

2.1 Data

We provide two training datasets. The first one is the Swiss Parliaments Corpus (Plüss et al., 2021), a labeled 293-hours dataset of Swiss German debates from the Grosser Rat Kanton Bern parliament with corresponding Standard German sentence-level transcriptions[1]. The second one is an unlabeled collection of 1208 hours of Swiss German

[1] https://www.cs.technik.fhnw.ch/i4ds-datasets

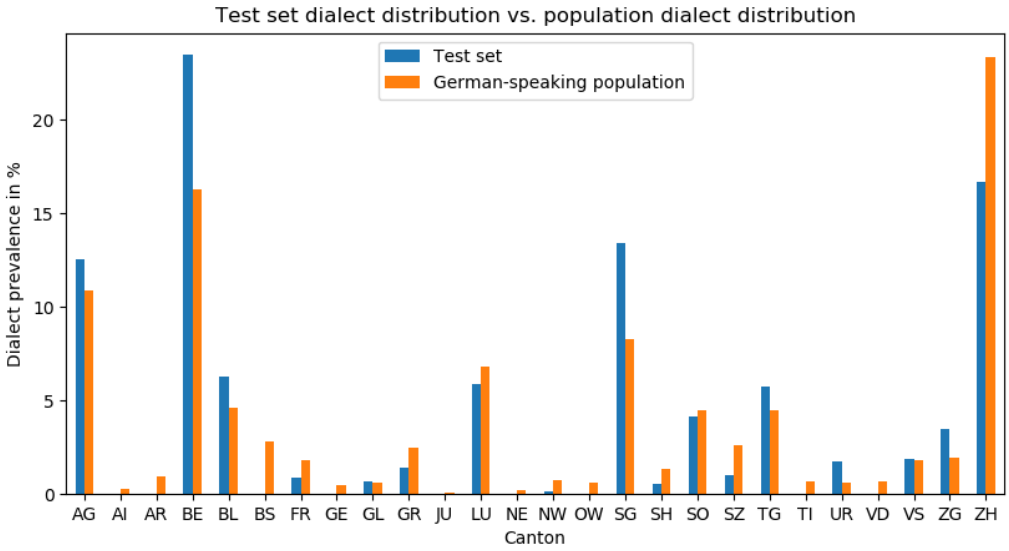

Figure 1: Comparison of the dialect prevalence in Switzerland's German-speaking population with the All Swiss German Dialects Test Set. To make this comparison possible, a dialect is defined as the average dialect spoken in a canton.

debates from the Gemeinderat Zürich parliament[2]. The use of additional datasets is allowed, but has to be declared in the system description.

The test set created for this task, the All Swiss German Dialects Test Set, contains 13 hours of sentence-level Swiss German speech and Standard German text pairs[3]. The set is divided into two equally sized parts, a public part (score on this part was displayed in the public ranking while the task was running) and a private part (final ranking is based on this part, was not available while the task was running). The texts are from the Common Voice project[4] and were spoken by 178 speakers from all over Switzerland. It covers a large part of the Swiss German dialect landscape. Figure 1 compares the test set dialect distribution with the real distribution of Swiss German dialects in Switzerland. The comparison highlights the good match between the test set dialect distribution and the real distribution. There are some exceptions, e.g. there is no data from the cantons AI, AR, and OW due to their small size. Also, BE and SG speakers are overrepresented whereas ZH speakers are underrepresented. There was no distinction made between BL and BS during the collection of the dialect metadata for the test set. BS speakers are therefore included in BL.

2.2 Evaluation

The submissions are evaluated using BLEU score (Papineni et al., 2002). Our evaluation script, which uses the NLTK (Bird et al., 2009) BLEU implementation, is open-source[5]. The private part of the test set is used for the final ranking. The test set contains the characters a-z, ä, ö, ü, and spaces, and the participants' models should support exactly these. Punctuation and casing are ignored for the evaluation. Numbers are spelled out. All other characters are removed from the submission (see evaluation script for details). Participants were therefore advised to replace each additional character in their training set with a sensible replacement.

3 Results

Four teams participated in the shared task. Table 1 shows the final ranking.

The team in first place, Arabskyy et al. (2021), achieved a BLEU score of 46.0. They use a hybrid system with a lexicon that incorporates translations, a first pass language model that deals with Swiss German particularities, an acoustic model transfer-learned from a large Standard German dataset, and

[2] https://www.cs.technik.fhnw.ch/i4ds-datasets

[3] https://www.cs.technik.fhnw.ch/i4ds-datasets

[4] https://github.com/common-voice/common-voice/tree/main/server/data/de

[5] https://github.com/i4Ds/swisstext-2021-task-3

Rank	Team	BLEU
1	Arabskyy et al.	46.0
2	Plüss et al.	41.0
3	Ulasik et al.	39.4
4	DeJa	17.1

Table 1: Final ranking of the shared task. The BLEU column shows the BLEU score on the private 50 % of the All Swiss German Dialects Test Set.

a strong neural language model for second pass rescoring.

Our baseline ranks second with 41.0 BLEU. The system is described in (Plüss et al., 2021) (section 5). We train an end-to-end Conformer (Gulati et al., 2020) model using a hybrid CTC / attention encoder-decoder framework. The training data consists of the Swiss Parliaments Corpus (Plüss et al., 2021), an additional 250 hours corpus of automatically aligned Swiss German parliament debates, and the Standard German Common Voice corpus (Ardila et al., 2019).

The team in third place, Ulasik et al. (2021), achieved a BLEU score of 39.4. Their approach combines three models trained on multilingual, Standard German, and Swiss German data using ensembling.

The team called DeJa ranked fourth and achieved a BLEU score of 17.1. We have not received a system description for this submission.

4 Conclusion

We have described SwissText 2021 Task 3 on Swiss German Speech to Standard German Text. Submissions were evaluated on the All Swiss German Dialects Test Set, which we introduced in this work. It covers a large part of the Swiss German dialect landscape. Four teams participated in the task, with the winning team reaching a BLEU score of 46.0. The results are hard to compare to the results of this task's predecessor, GermEval 2020 Task 4 (Plüss et al., 2020), due to the different test set and metric. Last year's winning contribution achieved a WER of 40.3 %. In our experiments in (Plüss et al., 2021), ranking second in this year's task, we achieved a WER of 27.8 % on a test set comparable to GermEval 2020 Task 4. The relative improvement of 31 % indicates that a lot of progress has been made in the field of Swiss German STT over the past year.

Despite recent advances in semi-supervised and unsupervised learning for STT, see e.g. (Park et al., 2020) and (Baevski et al., 2020), none of the participants made use of the provided unlabeled training set. This seems to be a promising direction for further improvements of Swiss German STT given that the amount of available labeled training data is still comparatively small.

Acknowledgments

We thank our participants for their interest in the shared task, for their participation, and for their timely feedback, which have helped us make this task a success.

We also thank Elias Schorr for his great work on the submission and evaluation website.

References

Yuriy Arabskyy, Aashish Agarwal, Subhadeep Dey, and Oscar Koller. 2021. Dialectal speech recognition and translation of swiss german speech to standard german text: Microsoft's submission to swisstext 2021. In preparation.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

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

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Matthias Büchi, Malgorzata Anna Ulasik, Manuela Hürlimann, Fernando Benites, Pius von Däniken, and Mark Cieliebak. 2020. Zhaw-init at germeval 2020 task 4: Low-resource speech-to-text. In *SWISSTEXT & KONVENS 2020*, Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS).

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proceedings of Interspeech*, pages 5036–5040.

V. Panayotov, G. Chen, D. Povey, and S. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Daniel S. Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V. Le. 2020. Improved noisy student training for automatic speech recognition. *Interspeech 2020*.

Michel Plüss, Lukas Neukom, Christian Scheller, and Manfred Vogel. 2021. Swiss parliaments corpus, an automatically aligned swiss german speech to standard german text corpus.

Michel Plüss, Lukas Neukom, and Manfred Vogel. 2020. Germeval 2020 task 4: Low-resource speech-to-text. In *SWISSTEXT & KONVENS 2020*, Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS).

Malgorzata Anna Ulasik, Manuela Hurlimann, Bogumila Dubel, Yves Kaufmann, Silas Rudolf, Jan Deriu, Katsiaryna Mlynchyk, Hans-Peter Hutter, and Mark Cieliebak. 2021. Zhaw-cai: Ensemble method for swiss german speech to standard german text. In preparation.

Yu Zhang, James Qin, Daniel S. Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V. Le, and Yonghui Wu. 2020. Pushing the limits of semi-supervised learning for automatic speech recognition.