2nd Swiss German Speech to Standard German Text Shared Task at SwissText 2022

Michel Plüss¹, Yanick Schraner^{1,*}, Christian Scheller¹ and Manfred Vogel^{1,*}

¹Institute for Data Science, University of Applied Sciences and Arts Northwestern Switzerland, Windisch, Switzerland

Abstract

We present the results and findings of the 2nd Swiss German speech to Standard German text shared task at SwissText 2022. Participants were asked to build a sentence-level Swiss German speech to Standard German text system specialized on the Grisons dialect. The objective was to maximize the BLEU score on a test set of Grisons speech. 3 teams participated, with the best-performing system achieving a BLEU score of 70.1.

1. Introduction

The topic of this task is automatic speech recognition (ASR) for Swiss German. Swiss German is a family of German dialects spoken in Switzerland, see Plüss et al. [1]. Swiss German ASR is concerned with the transcription of Swiss German Speech to Standard German text and can be viewed as a speech translation task with similar source and target languages, see Plüss et al. [1].

This task has two predecessors. The 2020 task [2] provided a 70-hours labeled training set of automatically aligned Swiss German speech (predominantly Bernese dialect) and Standard German text. The test set also comprised mostly Bernese speech. The winning contribution by Büchi et al. [3] achieved a word error rate (WER) of 40.3 %. The 2021 task [1] provided an improved and extended 293-hours version of the 2020 training set, as well as a 1208-hours unlabeled speech dataset (predominantly Zurich dialect). The test set covered a large part of the Swiss German dialect landscape. The winning contribution by Arabskyy et al. [4] achieved a BLEU score [5] of 46.0.

The goal of this task is to build a system able to translate Swiss German speech to Standard German text and optimize it for the Grisons dialect. To enable this, we provide the Swiss German labeled datasets SDS-200 [6] and SwissDial [7], both including a substantial amount of Grisons speech, as well as the Standard German, French, and Italian labeled datasets of Common Voice 9.0 [8].

2. Task Description

The goal of the task is to build a sentence-level Swiss German speech to Standard German text system specialized on the Grisons dialect. The submission with the best BLEU score on a test set of Grisons dialect speakers wins. Participants were encouraged to explore suitable transfer learning and fine-tuning approaches based on the Swiss German, Standard German, French, and Italian data provided.

2.1. Data

We provide 5 different training datasets to participants, all of which are collections of sentence-level transcribed speech. SDS-200 [6] is a Swiss German dataset with 200 hours of speech from all major Swiss German dialect regions, of which 6 hours are in Grisons dialect. SwissDial [7] is a Swiss German dataset with 34 hours of speech from all major Swiss German dialect regions, of which 11 hours are in Grisons dialect. From version 9.0 of the Common Voice project [8], we provide 1166 hours of Standard German, 926 hours of French, and 340 hours of Italian, all of which are official languages of Switzerland.

The test set was collected in a similar fashion to SDS-200 [6]. It consists of 5 hours of sentence-level transcribed Grisons speech by 11 speakers, of which 8 are female and 3 are male. 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). Two thirds of the texts are from Swiss newspapers and one third is from the minutes of parliament debates in Aarau and Wettingen. Care was taken to avoid any overlap between the Swiss newspaper sentences in this test set and the ones in SDS-200 [6].

2.2. Evaluation

The submissions are evaluated using BLEU score [5]. Our evaluation script, which uses the NLTK [9] BLEU implementation, is open-source¹. The private part of the test

SwissText 2022: Swiss Text Analytics Conference, June 08-10, 2022, Lugano, Switzerland *Corresponding author.

yanick.schraner@fhnw.ch (Y. Schraner);

manfred.vogel@fhnw.ch (M. Vogel)

CONSCIENTING (VI. Voget)
Constant of the second s authors. Use permitted under Creative Commons Licens

¹https://github.com/i4Ds/swisstext-2022-swiss-german-shared-task

Rank	Team	BLEU
1	Baseline	70.1
2	Stucki et al.	68.1
3	Nafisi et al.	55.3

Table 1

Final ranking of the shared task. The BLEU column shows the BLEU score on the private 50 % of the test set.

set is used for the final ranking.

The test set contains the characters a-z, ä, ö, ü, 0-9, and spaces, and the participants' models should support exactly these. Punctuation and casing are ignored for the evaluation. Numbers are not used consistently in the test set, so sometimes they are written as digits and sometimes they are spelled out. We create a second reference by automatically spelling out all numbers and use both the original and this adjusted reference in the BLEU score calculation. Participants were advised to have their models always spell out numbers. 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

3 teams participated in the shared task, including our baseline. Table 1 shows the final ranking.

Our baseline achieves a BLEU score of 70.1. We use the model *Transformer Baseline* described in Plüss et al. [6]. We train the model from scratch on SDS-200, SwissDial, and the Standard German part of Common Voice. Contrary to Plüss et al. [6], we employ a Transformer-based language model (LM) with 12 decoder layers, 16 attention heads, an embedding dimension of 512, and a fully connected layer with 1024 units. The LM is trained on 67M Standard German sentences. We use a beam width of 60 during decoding. The same model achieves 65.3 BLEU on the 2021 task test set [1].

Stucki et al. achieve a BLEU score of 68.1. They use an XLS-R 1B model [10], pre-trained on 436K hours of unlabeled speech in 128 languages, not including Swiss German. They fine-tune the model on SDS-200 and Swiss-Dial. A KenLM 5-gram LM [11] trained on the German Wikipedia is employed.

Nafisi et al. achieve a BLEU score of 55.3. They use an XLS-R 1B model [10], pre-trained on 436K hours of unlabeled speech in 128 languages, not including Swiss German. They fine-tune the model on SDS-200. No LM is employed.

4. Conclusion

We have described the 2nd Swiss German speech to Standard German text shared task at SwissText 2022. The best-performing system on the Grisons speech test set is our baseline with a BLEU score of 70.1. The same system achieves a BLEU score of 65.3 on the 2021 task test set [1], a relative improvement of 42 % over the highest score of the 2021 task. This highlights the large progress in the field over the last year. The main drivers for this progress seem to be the new dataset SDS-200 [6] as well as the use of models pre-trained on large amounts of unlabeled speech as demonstrated by the teams Stucki et al. and Nafisi et al., who employed XLS-R models [10]. The addition of an LM seems to be especially important for XLS-R models. The main difference between Nafisi et al. and Stucki et al. is that the latter add an LM, leading to a relative improvement of 23 % BLEU.

On the other hand, none of the 3 participating teams made a significant effort to optimize their system for the Grisons dialect. The best approach to create an ASR system optimized for a specific dialect remains to be found in future work. Incorporating the provided French and Italian data for training is another possible direction for future research.

References

- M. Plüss, L. Neukom, M. Vogel, Swisstext 2021 task 3: Swiss german speech to standard german text, in: Proceedings of the Swiss Text Analytics Conference 2021, 2021.
- [2] M. Plüss, L. Neukom, M. Vogel, Germeval 2020 task 4: Low-resource speech-to-text, in: Proceedings of the 5th Swiss Text Analytics Conference (Swiss-Text) & 16th Conference on Natural Language Processing (KONVENS), 2020.
- [3] M. Büchi, M. A. Ulasik, M. Hürlimann, F. Benites, P. von Däniken, M. Cieliebak, Zhaw-init at germeval 2020 task 4: Low-resource speech-to-text, in: Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS), 2020.
- [4] Y. Arabskyy, A. Agarwal, S. Dey, O. Koller, Dialectal speech recognition and translation of swiss german speech to standard german text: Microsoft's submission to swisstext 2021, in: Proceedings of the Swiss Text Analytics Conference 2021, 2021.
- [5] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002.
- [6] M. Plüss, M. Hürlimann, M. Cuny, A. Stöckli,

N. Kapotis, J. Hartmann, M. A. Ulasik, C. Scheller, Y. Schraner, A. Jain, J. Deriu, M. Cieliebak, M. Vogel, Sds-200: A swiss german speech to standard german text corpus, in: Proceedings of the Language Resources and Evaluation Conference, 2022.

- [7] P. Dogan-Schönberger, J. Mäder, T. Hofmann, Swissdial: Parallel multidialectal corpus of spoken swiss german, CoRR abs/2103.11401 (2021). URL: https: //arxiv.org/abs/2103.11401. arXiv:2103.11401.
- [8] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, G. Weber, Common Voice: A Massively-Multilingual Speech Corpus, in: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 2020.
- [9] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, "O'Reilly Media, Inc.", 2009.
- [10] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, M. Auli, Xls-r: Selfsupervised cross-lingual speech representation learning at scale, arXiv abs/2111.09296 (2021).
- [11] K. Heafield, Kenlm: Faster and smaller language model queries, in: Proceedings of the sixth workshop on statistical machine translation, 2011.