The PASSAGE project : Standard German Subtitling of Swiss German TV content

Pierrette Bouillon, Johanna Gerlach, Jonathan Mutal, Marianne Starlander
Faculty of translation and interpreting, University of Geneva
40, bd du Pont d'Arve, 1211 Geneva, Switzerland
{pierrette.bouillon, johanna.gerlach,
jonathan.mutal, marianne.starlander}@unige.ch

Abstract

We present the PASSAGE project, which aims at automatic Standard German subtitling of Swiss German TV content. This is achieved in a two step process, beginning with ASR to produce a normalised transcription, followed by translation into Standard German. We focus on the second step, for which we explore different approaches and contribute aligned corpora for future research.

1 Introduction

Swiss German, a primarily spoken language with many regional dialects and no standardised written form (Honnet et al., 2018), is spoken by two thirds of the population of Switzerland. It is widely used on Swiss TV, e.g. in news reports or interviews, which are subtitled in Standard German to make them accessible to people who cannot understand spoken Swiss German. Producing these subtitles automatically would be advantageous in terms of time and cost. This task is the focus of the PASSAGE project (Nov. 2020- Dec. 2022) "Sous-titrage automatique du suisse allemand en allemand standard", which is a collaboration between Geneva University, SRF (Schweizer Radio und Fernsehen) and recapp.[1]

In this project a first automatic speech recognition (ASR) step is used to produce a normalised transcription of spoken Swiss German, keeping the original syntax and expressions but using Standard German words. In a second step, different approaches are explored to transform this normalised transcription into correct written Standard German (see Figure 1). To achieve this, multiple issues must be dealt with: ASR errors, incorrect detection of sentence boundaries, features related to spontaneous spoken language, such as dysfluencies or informal language, and finally the syntactic divergences between Swiss German and Standard German (Glaser and Bart, 2021). The three goals of the project are 1) to create data sets for Swiss German, 2) to build systems for the translation of ASR output into Standard German, and 3) to evaluate the usability of the system output.

2 Data

The following data were provided by SRF:

- Normalised transcriptions of TV shows: originally created to train the Swiss German speech recogniser, these human transcriptions keep the original syntax and expressions but use Standard German words. (98,126 segments)
- Original Standard German subtitles of the TV shows (DE): batches of subtitles, not aligned with the transcriptions. (101,150 segments)

Based on these data, we have so far created several aligned corpora, which were used to train and specialise the first systems:

- Normalised transcriptions - Standard German: this corpus was produced by manual post-editing of the transcriptions. (20,634 segments)
- Normalised transcriptions - original subtitles: this corpus was aligned automatically. (70,374 segments)

[1] https://www.media-initiative.ch/project/subtitling-of-swiss-german-into-standard-german-automatic-post-editing/

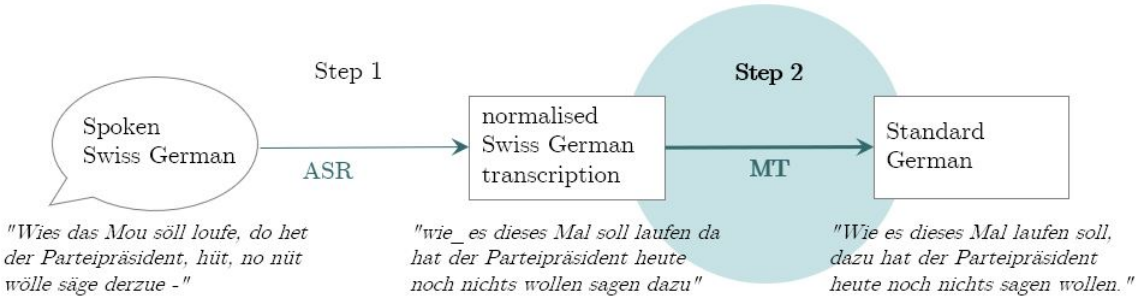

Figure 1: Overview of the subtitling pipeline

- Synthetic parallel data for some of the syntactic divergences between Swiss and Standard German: these corpora were generated by applying hand-crafted transformation rules to the post-edited transcriptions and the original subtitles. Rules were created using the SpaCy toolkit's Matcher[2] to change word order or verb forms to artificially produce Swiss German syntax (e.g. Man *verschliesst* sich solchen Fragen sicher nicht → Man *tut* sich solchen Fragen sicher nicht *verschliessen*). (4,418 and 13,896 segments generated from post-edited transcriptions and subtitles).

Finally, our project partner recapp[3] provided real ASR output, which was used for evaluation, and is currently being aligned with the original subtitles to create another parallel resource.

3 Systems

The project aims at investigating different approaches suitable for tasks where only few changes are needed and applicable to low-resource languages. We will also explore different settings, such as the impact of different training data, e.g. transcriptions vs ASR output, automatic vs manual alignment.

4 First results

The first two approaches tested for this task are 1) a neural machine translation (NMT) transformer architecture with copy attention (Gehrmann et al., 2018) and 2) an edit-based model (Ed) with a task-specific attention mechanism that predicts types of edits (Berard et al., 2017).

Our first system evaluations were carried out on normalised transcriptions, which simulate a perfect ASR result. An automatic evaluation using the post-edited version as reference showed that overall the NMT system was slightly better than the Ed system (BLEU 64.91 vs 61.49), and that NMT makes more edits than Ed (HTER 22.59 vs. 12.69).

Another round of evaluations was performed on real ASR output, with a focus on the systems' ability to transform Swiss German syntactic phenomena into their Standard German counterparts. Here the NMT system outperforms the Ed system. For NMT, the addition of targeted synthetic training data improves the results, in terms of transformed phenomena and precision.

Next steps include an evaluation with end-users to assess the impact on satisfaction.

[2]https://spacy.io/api/matcher
[3]https://recapp.ch/

Acknowledgements

This project has received funding from the Initiative for Media Innovation based at Media Center, EPFL, Lausanne, Switzerland.

References

Bérard, Alexandre, Laurent Besacier, and Olivier Pietquin. 2017. Lig-cristal submission for the WMT2017 automatic post-editing task. *Proceedings of the Second Conference on Machine Translation*. ACL. pp. 623–629.

Gehrmann, Sebastian, Yuntian Deng, and Alexander M. Rush. 2018. Bottom-up abstractive summarization. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. ACL. pp. 4098–4109.

Glaser, Elvira and Gabriela Bart. 2021. *Syntaktischer Atlas der deutschen Schweiz (SADS)*. A. Francke Verlag

Honnet, Pierre-Edouard, Andrei Popescu-Belis, Claudiu Musat, and Michael Baeriswyl. 2018. Machine translation of low-resource spoken dialects: Strategies for normalizing Swiss German. *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. pp. 3781–3788.