LibriVoxDeEn: A Corpus for German-to-English Speech Translation and German Speech Recognition

Abstract

We present a corpus of sentence-aligned triples of German audio, German text, and English translation, based on German audio books. The corpus consists of over 100 hours of audio material and over 50k parallel sentences. The audio data is read speech and thus low in disfluencies. The quality of audio and sentence alignments has been checked by a manual evaluation, showing that speech alignment quality is in general very high. The sentence alignment quality is comparable to well-used parallel translation data and can be adjusted by cutoffs on the automatic alignment score. To our knowledge, this corpus is to date the largest resource for end-to-end speech translation for German.

Introduction

Direct speech translation has recently been shown to be feasible using a single sequence-to-sequence neural model, trained on parallel data consisting of source audio, source text and target text. The crucial advantage of such end-to-end approaches is the avoidance of error propagation as in a pipeline approaches of speech recognition and text translation. While cascaded approaches have an advantage in that they can straightforwardly use large independent datasets for speech recognition and text translation, clever sharing of sub-networks via multi-task learning and two-stage modeling BIBREF0, BIBREF1, BIBREF2 has closed the performance gap between end-to-end and pipeline approaches. However, end-to-end neural speech translation is very data hungry while available datatsets must be considered large if they exceed 100 hours of audio. For example, the widely used Fisher and Call-home Spanish-English corpus BIBREF3 comprises 162 hours of audio and $138,819$ parallel sentences. Larger corpora for end-to-end speech translation have only recently become available for speech translation from English sources. For example, 236 hours of audio and $131,395$ parallel sentences are

available for English-French speech translation based on audio books BIBREF4, BIBREF5. For speech translation of English TED talks, 400-500 hours of audio aligned to around $250,000$ parallel sentences depending on the language pair have been provided for eight target languages by DiGangiETAL:19. Pure speech recognition data are available in amounts of $1,000$ hours of read English speech and their transcriptions in the LibriSpeech corpus provided by PanayotovETAL:15.

When it comes to German sources, the situation regarding corpora for end-to-end speech translation as well as for speech recognition is dire. To our knowledge, the largest freely available corpora for German-English speech translation comprise triples for 37 hours of German audio, German transcription, and English translation BIBREF6. Pure speech recognition data are available from 36 hours BIBREF7 to around 200 hours BIBREF8.

We present a corpus of sentence-aligned triples of German audio, German text, and English translation, based on German audio books. The corpus consists of over 100 hours of audio material aligned to over 50k parallel sentences. Our approach mirrors that of KocabiyikogluETAL:18 in that we start from freely available audio books. The fact that the audio data is read speech keeps the number of disfluencies low. Furthermore, we use state-of-the art tools for audio-text and text-text alignment, and show in a manual evaluation that the speech alignment quality is in general very high, while the sentence alignment quality is comparable to widely used corpora such as that of KocabiyikogluETAL:18 and can be adjusted by cutoffs on the automatic alignment score. To our knowledge, the presented corpus is to data the largest resource for end-to-end speech translation for German.

Overview

In the following, we will give an overview over our corpus creation methodology. More details will be given in the following sections.

Creation of German corpus (see Section sourcecorpus.)

Data download

Download German audio books from LibriVox web platform

Collect corresponding text files by crawling public domain web pages

Audio preprocessing

Manual filtering of audio pre- and postfixes

Text preprocessing

Noise removal, e.g. special symbols, advertisements, hyperlinks

Sentence segmentation using spaCy

Speech-to-text alignments

Manual chapter segmentation of audio files

Audio-to-text alignments using forced aligner aeneas

Split audio according to obtained timestamps using SoX

Creation of German-English Speech Translation Corpus (see Sections targetcorpus. and corpusfiltering.)

Download English translations for German texts

Text preprocessing (same procedure as for German texts)

Bilingual text-to-text alignments

Manual text-to-text alignments of chapters

Dictionary creation using parallel DE-EN WikiMatrix corpus BIBREF9

German-English sentence alignments using hunalign BIBREF10

Data filtering based on hunalign alignment scores

Source Corpus Creation ::: Data Collection

We acquired pairs of German books and their corresponding audio files starting from LibriVox, an open source platform for people to publish their audio recordings of them reading books which are available open source on the platform Project Gutenberg. Source data were gathered in a semi-automatic way: The URL links were collected manually by using queries containing metadata descriptions to find German books with LibriVox audio and possible German transcripts. These were later automatically scraped using BeautifulSoup4 and Scrapy, and saved for further processing and cleaning. Public domain web pages crawled include https://gutenberg.spiegel.de, http://www.zeno.org, and https://archive.org.

Source Corpus Creation ::: Data Preprocessing

We processed the audio data in a semi-automatic manner which included manual splitting and alignment of audio files into chapters, while also saving timestamps for start and end of chapters. We removed boilerplate intros and outros and as well as noise at the beginning and end of the recordings.

Preprocessing the text included removal of several items, including special symbols like *, advertisements, hyperlinks in [], <>, empty lines, quotes, - preceding sentences, indentations, and noisy OCR output.

German sentence segmentation was done using spaCy based on a medium sized German corpus that contains the TIGER corpus and the WikiNER dataset dataset. Furthermore we added rules to adjust the segmenting behavior for direct speech and for semicolon-separated sentences.

Source Corpus Creation ::: Text-to-Speech Alignment

To align sentences to onsets and endings of corresponding audio segments we made use of aeneas – a tool for an automatic synchronization of text and audio. In contrast to most forced aligners, aeneas does not use automatic speech recognition (ASR) to compare an obtained transcript with the original text. Instead, it works in the opposite direction by using dynamic time warping to align the mel-frequency cepstral coefficients extracted from the real audio to the audio representation synthesized from the text, thus aligning the text file to a time interval in the real audio.

Furthermore, we used the maps pointing to the beginning and the end of each text row in the audio file produced with SoX to split the audio into sentence level chunks. The timestamps were also used to filter boilerplate information about the book, author, speaker at the beginning and end of the audio file.

Statistics on the resulting corpus are given in Table TABREF36.

Target Corpus Creation ::: Data Collection and Preprocessing

In collecting and preprocessing the English texts we followed the same procedure as for the source language corpus, i.e., we manually created queries containing metadata descriptions of English books (e.g. author names) corresponding to German books which then were scraped. The spaCy model for sentence segmentation used a large English web corpus. See Section sourcecorpus. for more information.

Target Corpus Creation ::: Text-to-Text Alignment

To produce text-to-text alignments we used hunalign with a custom dictionary of parallel sentences, generated from the WikiMatrix corpus. Using this additional dictionary improved our alignment scores. Furthermore we availed ourselves of a realign option enabling to save a dictionary generated in a first pass and profiting from it in a second pass. The final dictionary we used for the alignments consisted of a combination of entries of our corpora as well as the parallel corpus WikiMatrix. For further completeness we reversed the arguments in hunalign to not only obtain German to English alignments, but also English to German. These tables were merged to build the union by dropping duplicate entries and keeping those with a higher confidence score, while also appending alignments that may only have been produced when aligning in a specific direction.

Statistics on the resulting text alignments are given in Table TABREF37.

Data Filtering and Corpus Structure ::: Corpus Filtering

A last step in our corpus creation procedure consisted out filtering out empty and incomplete alignments, i.e., alignments that did not consist of a DE-EN sentence pair. This was achieved by dropping all entries with a hunalign score of -0.3 or below. Table TABREF38 shows the resulting corpus after this filtering step.

Moreover, many-to-many alignments by hunalign were re-segmented to source-audio sentence level for German, while keeping the merged English sentence to provide a complete audio lookup. The corresponding English sentences were duplicated and tagged with <MERGE> to mark that the German sentence was involved into a many-to-many alignment.

The size of our final cleaned and filtered corpus is thus comparable to the cleaned Augmented LibriSpeech corpus that has been used in speech translation experiments by BerardETAL:18.

Statistics on the resulting filtered text alignments are given in Table TABREF38.

Data Filtering and Corpus Structure ::: Corpus Structure

Our corpus is structured in following folders:

contains German text files for each book

contains English text files for each book

alignment maps produced by aeneas

sentence level audio files

text2speech, a lookup table for speech alignments

text2text, a lookup table for text-to-text alignments

Further information about the corpus and a download link can be found here:

https://www.cl.uni-heidelberg.de/statnlpgroup/librivoxdeen/.

Corpus Evaluation ::: Human Evaluation

For a manual evaluation of our dataset, we split the corpus into three bins according to ranges $(-0.3,0.3]$, $(0.3,0.8]$ and $(0.8,\infty)$ of the hunalign confidence score (see Table TABREF56).

The evaluation of the text alignment quality was conducted according to the 5-point scale used in KocabiyikogluETAL:18:

Wrong alignment

Partial alignment with slightly compositional translational equivalence

Partial alignment with compositional translation and additional or missing information

Correct alignment with compositional translation and few additional or missing information

Correct alignment and fully compositional translation

The evaluation of the audio-text alignment quality was conducted according to the following 3-point scale:

Wrong alignment

Partial alignment, some words or sentences may be missing

Correct alignment, allowing non-spoken syllables at start or end.

The evaluation experiment was performed by two annotators who each rated 30 items from each bin, where 10 items were the same for both annotators in order to calculate inter-annotator reliability.

Corpus Evaluation ::: Evaluation Results

Table TABREF54 shows the results of our manual evaluation. The audio-text alignment was rated as in general as high quality. The text-text alignment rating increases corresponding to increasing hunalign confidence score which shows that the latter can be safely used to find a threshold for corpus filtering. Overall, the audio-text and text-text alignment scores are very similar to those reported by KocabiyikogluETAL:18.

The inter-annotator agreement between two raters was measured by Krippendorff's $\alpha $-reliability score BIBREF11 for ordinal ratings. The inter-annotator reliability for text-to-text alignment quality ratings scored 0.77, while for audio-text alignment quality ratings it scored 1.00.

Corpus Evaluation ::: Examples

In the following, we present selected examples for text-text alignments for each bin. A closer inspection reveals properties and shortcomings of hunalign scores which are based on a combination of dictionary-based alignments and sentence-length information.

Shorter sentence pairs are in general aligned correctly, irrespective of the score (compare examples with score 0.30. 0.78 and 1.57, 2.44 below). Longer sentences can include exact matches of longer substrings, however, they are scored based on a bag-of-words overlap (see the examples with scores 0.41 and 0.84 below).

Schigolch Yes, yes; und mir träumte von einem Stück Christmas Pudding.

She only does that to revive old memories. LULU.

Und hätten dreißigtausend Helfer sich ersehn.

And feardefying Folker shall our companion be; He shall bear our banner; better none than he.

Kakambo verlor nie den Kopf.

Cacambo never lost his head.

Es befindet sich gar keine junge Dame an Bord, versetzte der Proviantmeister.

He is a tall gentleman, quiet, and not very talkative, and has with him a young lady — There is no young lady on board, interrupted the AROUND THE WORLD IN EIGPITY DAYS. purser..

Ottilie, getragen durch das Gefühl ihrer Unschuld, auf dem Wege zu dem erwünschtesten Glück, lebt nur für Eduard.

Ottilie, led by the sense of her own innocence along the road to the happiness for which she longed, only

lived for Edward.

Was ist geschehen? fragte er.

What has happened ? he asked.

Es sind nun drei Monate verflossen, daß wir Charleston auf dem Chancellor verlassen, und zwanzig Tage, die wir schon auf dem Flosse, von der Gnade der Winde und Strömungen abhängig, verbracht haben!

JANUARY st to th.More than three months had elapsed since we left Charleston in the Chancellor, and for no less than twenty days had we now been borne along on our raft at the mercy of the wind and waves.

Charlotte stieg weiter, und Ottilie trug das Kind.

Charlotte went on up the cliff, and Ottilie carried the child.

Fin de siecle, murmelte Lord Henry.

Fin de siecle, murmured Lord Henry.

Conclusion

We presented a corpus of aligned triples of German audio, German text, and English translations for speech translation from German to English. The audio data in our corpus are read speech, based on

German audio books, ensuring a low amount of speech disfluencies. The audio-text alignment and text-to-text sentence alignment was done with state-of-the-art alignment tools and checked to be of high quality in a manual evaluation. The audio-text alignment was generally rated very high. The text-text sentence alignment quality is comparable to widely used corpora such as that of KocabiyikogluETAL:18. A cutoff on a sentence alignment quality score allows to filter the text alignments further for speech translation, resulting in a clean corpus of $50,427$ German-English sentence pairs aligned to 110 hours of German speech. A larger version of the corpus, comprising 133 hours of German speech and high-quality alignments to German transcriptions is available for speech recognition.

Acknowledgments

The research reported in this paper was supported in part by the German research foundation (DFG) under grant RI-2221/4-1.