IESTAC: English-Italian Parallel Corpus for End-to-End Speech-to-Text Machine Translation

Giuseppe Della Corte and **Sara Stymne**
Department of Linguistics and Philology
Uppsala University
giuseppe.dellacorte.1888@student.uu.se
sara.stymne@lingfil.uu.se

Abstract

We discuss a set of methods for the creation of IESTAC: a English-Italian speech and text parallel corpus designed for the training of end-to-end speech-to-text machine translation models and publicly released as part of this work. We first mapped English LibriVox audiobooks and their corresponding English Gutenberg Project e-books to Italian e-books with a set of three complementary methods. Then we aligned the English and the Italian texts using both traditional Gale-Church based alignment methods and a recently proposed tool to perform bilingual sentences alignment computing the cosine similarity of multilingual sentence embeddings. Finally, we forced the alignment between the English audiobooks and the English side of our textual parallel corpus with a text-to-speech and dynamic time warping based forced alignment tool. For each step, we provide the reader with a critical discussion based on detailed evaluation and comparison of the results of the different methods.

1 Introduction

Traditionally, most research on machine translation has been concerned with text-to-text systems. However, there is an increasing interest in speech translation. Speech translation usually refers to the task of translating source language audio signals into a text spoken in a target language. Traditionally, it has been tackled by cascaded ST (speech translation) models that concatenates three technologies: ASR (automatic speech recognition), MT (machine translation), and TTS (text-to-speech). Latency and error propagation are two intrinsic drawbacks of cascaded ST models (Ruiz et al., 2017).

End-to-end speech-to-text machine translation, usually also referred to as direct speech translation, avoid error propagation and reduce latency by directly translating source language audio signals into target language texts. A variety of end-to-end machine and deep learning architectures (Bérard et al., 2016; Weiss et al., 2017; Bérard et al., 2018; Anastasopoulos and Chiang, 2018; Di Gangi et al., 2019b) have been proposed to infer patterns from a first sequence (source language audio utterances) and a second sequence (target language textual translations). Training these models require a large amount of source language audio utterances paired up with their textual translations. Furthermore, since pre-training the encoder on ASR seems to improve the quality of the ST results (Bérard et al., 2018; Di Gangi et al., 2019b), the training data should preferably also include source language audio utterances paired up with their transcriptions.

We discuss the creation of IESTAC (Italian-English Speech and Text Audiobooks Corpus), designed for training English-to-Italian speech-to-text machine translation models. It is publicly available[1] and composed of around 130 hours of English speech aligned with its transcription and Italian textual translation at a sentence level. For a more detailed description of the corpus, see also Della Corte (2020). Our objective is to provide the readers with a methodological contribution for the creation of corpora designed for end-to-end speech-to-text machine translation training. We describe a pipeline to semi-automatically collect audio-textual data from the web and automatically align them. Alignment is performed as a two step process. We first perform bilingual sentences alignment between the English text and its Italian textual translation. Then we force the alignment between the English audio and the English text already aligned with its Italian textual translation. For each step we discuss different possible tools, and evaluate the results, allowing us to give recommendations for tools to use for creating new corpora for other languages.

[1] https://github.com/Giuseppe-Della-Corte/IESTAC

Proceedings of the First International Workshop on Natural Language Processing Beyond Text, pages 41–50

http://www.aclweb.org/anthology/W23-20%2d

2 Related Work

There have been some previous work on creating resources for end-to-end speech-to-text machine translation (Kocabiyikoglu et al., 2018; Di Gangi et al., 2019a; Iranzo-Sánchez et al., 2020; Wang et al., 2020a,b). There are also other related efforts, including computational language documentation for low-resource languages (Godard et al., 2018), multilingual speech corpora creation (Black, 2019), and multi-modal corpora creation (Sanabria et al., 2018). Godard et al. (2018) created a speech-to-text corpus of 5 thousands triplets of Mboshi speech, Mboshi transcription, and French textual translations. Speech elicitation from text was done manually by three qualified speakers. Black (2019) created a large corpus of aligned text, speech, and pronounciation for 700 languages, with texts from the bible. The average duration for each language is 2 hours. Sanabria et al. (2018) created a multi-modal corpus by aligning at a word-level 2000 hours of English instructional YouTube videos with their subtitles. Portuguese textual translations were added by paying bilingual English-Portuguese speakers. Augmented LibriSpeech (Kocabiyikoglu et al., 2018) seem to be the first corpus designed for training end-to-end English-to-French speech-to-text machine translation systems. It was created by collecting public domain audiobooks and e-books from the web and automatically align them. A similar approach was used by Beilharz et al. (2020) to create LibriVoxDeEn, a corpus for German-to-English speech translation. Most recent works have been focused on multilingual corpora creation for speech-to-text machine translation: MuST-C (Di Gangi et al., 2019a), Europarl-ST, CoVoST (Wang et al., 2020a) and CoVoST2 (Wang et al., 2020b)

2.1 Augmented LibriSpeech

LibriSpeech (Panayotov et al., 2015) is a corpus for English ASR, created by aligning English audiobooks from the LibriVox project (Kearns, 2014) with their source English e-books from the Gutenberg Project (Stroube, 2003). It was designed to prioritize speaker variety: it contains only a few audio segments per chapter, and a few chapters per book. Augmented LibriSpeech (Kocabiyikoglu et al., 2018) is an augmentation of LibriSpeech with French textual translations. Kocabiyikoglu et al. (2018) used part of the LibriSpeech metadata (around 1500 English e-book titles) to retrieve their corresponding French e-book titles by querying Dbpedia (Auer et al., 2007). They then compared the retrieved French e-book titles against a web index containing public domain French e-books. The collected English and French e-books were aligned with hunalign (Varga et al., 2007), resulting in a textual parallel corpus. Finally, the English side of the parallel corpus was aligned with the LibriSpeech English audio recordings with Gentle[2].

2.2 MuST-C

Data were collected from the English TED website[3]. Di Gangi et al. (2019a) selected those talks that include both a transcription and a textual translation in German, Spanish, French, Italian, Dutch, Portuguese, Romanian or Russian. MuST-C is split in different data-sets for each language direction. Each data-set contains at least 395 hours of English audio utterances aligned with their transcription and their textual translations. Di Gangi et al. (2019a) used the Gargantua sentence alignment tool (Braune and Fraser, 2010) to perform bilingual sentence alignment between transcripts and textual translations. Then, they forced the alignment between the English audio and the English side of the textual parallel corpora with Gentle.

2.3 Europarl-ST

Europarl-ST (Iranzo-Sánchez et al., 2020) is a multilingual corpus for speech-to-text machine translation in 30 language pairs directions from 6 European languages. Data were collected from the LinkedEP database (Van Aggelen et al., 2017), retrieving the European Parliament debates hold between 2008 and 2012 with their transcriptions, time-spans, and translations. The main focus of Iranzo-Sánchez et al. (2020) was to filter out inaccurate labeled EP speeches. To do so, they performed speaker diarization (SD) for each speech and then forced the speech-to-text alignment at a intra-word sentence granularity. Then, they used the character error rate metrics (CER) to further filter out inaccurate transcribed speeches.

2.4 CoVoST and CoVoST2

Facebook AI[4] recently released CoVoST(Wang et al., 2020a) and CoVoST2 (Wang et al., 2020b). Each CoVoST corpus is an augmentation of CoVo

[2] https://github.com/lowerquality/gentle

[3] https://www.ted.com/

[4] https://ai.facebook.com/

(Ardila et al., 2020), a multilingual speech recognition corpus. CoVo already provides pairs of aligned audio and transcription. Wang et al. (2020a) selected 11 languages from Common Voice. Then, they paid professional translators to translate 11 Common Voice data-sets (one for each selected language) into English. In order to ensure the quality of the translations, Wang et al. (2020a) applied different sanity checks to find weak translations and send them back to the professional translators. Interestingly, one of those sanity checks was to compute similarity scores between the sentence embeddings of the source language texts and their translations. By using the same approach, Wang et al. (2020b) released CoVoST2, an extension of CoVoST. It covers training data for end-to-end speech-to-text machine translation for 21 languages to English and for English to 15 languages.

2.5 Corpora and licences

The MuST-C corpus licence (Creative Commons Attribution-NonCommercial-NoDerivs 4.0) does not permit commercial use and prevent derivative works. The Europarl-ST corpus licence (Attribution-NonCommercial 4.0 International license) permits derivative works but it does not allow commercial use. In contrast, the CoVoST corpora have been released under the CC0 licence, while Augmented LibriSpeech has been released under the CC BY 4.0 licence. Both the CC0 and the CC BY 4.0 licences allow commercial use and permit derivative works.

3 Corpus Creation

Due to the fact that we are interested in releasing a freely available corpus with a permissive licence, we mainly follow the approach proposed by Kocabiyikoglu et al. (2018):

- Text collection: collect English audiobooks, English e-books, and their corresponding Italian e-books
- Bilingual sentence alignment: automatically create a parallel corpus aligned at a sentence level from the English and the Italian e-books
- Forced alignment: force the alignment between the English audio segments and the English side of the parallel corpus

4 Text Collection

Our first challenge was to identify freely available Italian e-books corresponding to available English e-books. As a starting point, we used part of the LibriSpeech metadata, more specifically the list of the Gutenberg Project English e-books titles. Sometimes it might happen that a single book was published several times with different titles. Therefore, we pre-processed the English titles list using regular expressions to increase the number of possible titles. We manually found patterns that indicate the presence of alternative titles, subtitles, or publication specific information. These patterns were used to augment the possible titles. For instance, many Gutenberg Project English e-book titles contained two or more possible titles separated by the sub-string *, or,* (e.g. "Tom Swift and His Sky Racer, or, the Quickest Flight on Record"). At the end of the pre-processing step, the list of English titles was augmented with the inclusion of "Tom Swift and His Sky Racer" and "The Quickest Flight on Record" as two individual list elements.

4.1 Methods

We experimented with three methods for the e-book title translations retrieval task: querying WikiData (Vrandečić and Krötzsch, 2014), querying the WikiMedia endpoint, automatic machine translation with Google Translate WikiData is a knowledge base containing entities (or objects). Each entity is identified by language labels and alternative labels. Each entity belongs to one or more classes and has a set of properties. As a first method, we wrote a SPARQL query search to retrieve all WikiData objects belonging to the class "literary work" and with an English label, an Italian label, plus, if available, the list of alternative Italian and English labels. English and Italian labels and alternative labels correspond to English and Italian book titles. We then compared the results returned by our SPARQL queries against the LibriSpeech English titles list (see Section 4), returning only those WikiData results which English label (or one of the alternative labels) matched one of the items in LibriSpeech English titles list. Our second method was to identify possible Italian e-books titles by querying each element of the LibriSpeech English e-books titles through the WikiMedia endpoint, returning the Italian web page title corresponding to the English e-book title queried and successfully found to match an English web page from one of the WikiMedia Foundation[5] websites. Our third and final method was to use Google Translate for

[5] https://wikimediafoundation.org/

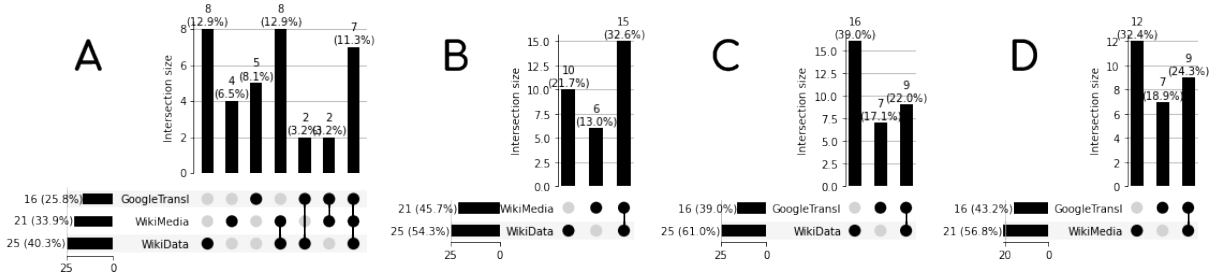

Figure 1: The leftmost columns shows the cardinality and the percentage of unique items in a given set. By unique items we mean the items that are not found in any intersection set. The remaining columns (the rightmost ones) show cardinalities and percentages of the intersection set of exactly two or three sets. A shows the unique terms in the Google Translate, the WikiMedia, and the WikiData sets (first three leftmost columns), the other four columns give information on the intersection sets of two or three sets (black dots linked by a black line). B, C, and D show in more details intersection information between exclusively two sets.

automatically translating the list of English book titles into Italian.

4.2 Evaluation and Discussion

Evaluating these three approaches is challenging, since the nature of the methods' output data is completely different, and we do not have access to a ground truth of which books match. Querying WikiData provides us entities, actual book titles. All retrieved Italian strings represent titles of Italian books that have been published and do have a corresponding English version of the book. We query the knowledge base to retrieve all entities that match with a set of conditions.

On the opposite, automatic translation and scraping the WikiMedia endpoint do not require a set of conditions to be expressed, neither allow to filter out redundant results. Scraping WikiMedia by querying English titles to retrieve the queries' corresponding Italian web page titles do include a great variety of noisy results: web pages referring to movies, theatre and semantic categories that have nothing to do with books. The automatic translation approach is even intrinsically noisier, since all strings in the list of English titles are synthetically translated, without any relationship with the actual book publication.

Due to these reasons, there is no possibility to directly evaluate and compare the accuracy of these methods. Hence we approached the evaluation indirectly, comparing the performance of the three methods on a real-scenario application. First we created a static index containing around 3400 Italian book titles by using web scraping techniques. Then we saved each matching title between each method's list of possible Italian book titles and the index, resulting in three sets of retrieved e-books, one for each method (Google Translate, WikiData, WikiMedia). Finally, we measured, plotted and visualized information regarding the intersection size among the sets using UpSetPlot[6]. Figure 1 shows the information regarding the intersection of the WikiData, the WikiMedia, and the Google Translate sets. We can rank the methods from the most performing one to the least performing one by looking at the percentages of elements that appear exclusively in a set. By following this criterion, WikiData (12.9% of unique items) outperformed both WikiMedia (6.5% of unique items) and Google Translate (8.1% of unique items).

It is also worth noting that the methods are complementary, and that each method found at least four books not identified by any of the other methods. The relatively small amount of retrieved Italian e-books (39) might be due to the rather small size of our index of Italian e-books (around 3400), which size is a fourth of NosLivres[7] (14,845 entries)[8]. NosLivres was the index used by Kocabiyikoglu et al. (2018) for the text collection task, augmented with manual search of French e-books. They collected 315 pairs of English and French e-books.

4.3 Pre-Processing

Once we retrieved the 39 pairs of English and Italian e-books, we first extracted chapters from both the Italian and the English e-books. Secondly, we had to segment each text file (the ones resulting

[6] https://pypi.org/project/UpSetPlot/
[7] https://www.noslivres.net/
[8] NosLivres is updated weekly, so the exact number of entries might change

from step 1) at a sentence level granularity (without performing bitext alignment yet). To extract chapters we used *chapterize*[9], a tool that automatically splits English Gutenberg Project e-books (txt files) in chapters by using a set of regular expressions to recognize chapters headings, retrieve the text between them and finally write in a new folder a series of text files corresponding to each extracted chapter. The same approach was applied to the extracted chapters from the Italian retrieved e-books, using custom regular expressions to recognize possible Italian chapter headings.

Because great variation can be found in chapter names depending on the e-book, we had to manually check if the extracted English and Italian chapters text files really contained comparable chapters. We define comparable chapters as the ones starting and ending with paragraphs strongly semantically correlated between the English and the Italian version. It might happen that due to unseen chapter namings patterns, some chapter headings would have been missed by the regular expression, resulting in the merge of two or more chapters into one or the split of a single chapter in several text files. Therefore manual investigation of possible issues and manual troubleshooting was required to ensure the quality and the parallel property of the extracted chapters. We then stripped all leading and ending spaces from the strings and removed all newlines and tabs. Sentence segmentation was performed using two pre-trained *spaCy*[10] models (*en_core_web_sm* on the English txt files, and *it_core_news_sm* on the Italian txt files).

For our first corpus release, we focused our efforts on a pool of nine books by 8 authors, randomly selected from 39 pairs of English and Italian e-books, with a total amount of 373 pairs of comparable chapters. These 373 pairs of Italian and English texts correspond to around 130 hours of English audio aligned at a sentence granularity with their Italian textual translations and their English source texts. We thought this amount of data was meaningful enough for proceeding to the alignment experiments, prioritizing the comparison and evaluation of different alignment methods instead of using all retrieved Italian e-books (39), which would have resulted in around 500 hours of audio material to be aligned.

[9] https://github.com/JonathanReeve/chapterize

[10] https://spacy.io/

5 Bilingual Sentence Alignment

Kocabiyikoglu et al. (2018) only use one method for the bilingual sentence alignment task. Instead, we tested and evaluated three different methods: hunalign in conjunction with a small size hand-crafted dictionary, hunalign in conjunction with a larger bilingual dictionary automatically inferred using statistical machine translation techniques, and Vecalign (Thompson and Koehn, 2019). Vecalign is a recently proposed method to perform bilingual text alignment computing cosine similarity of embeddings of consecutive sentences. We did not experiment with Gargantua, the alignment tool used by Di Gangi et al. (2019a), mainly because the tool seems to be more effective on large documents (Abdul-Rauf et al., 2010), while we want to align each pair of English and Italian parallel chapters individually, to maintain parallelism with the English LibriVox chapter audio recordings. Hunalign is not designed for aligning documents with more than 20000 sentences, but is effective on relatively short documents (Abdul-Rauf et al., 2010), as our pairs of comparable chapters.

5.1 Hunalign with LFAligner Dictionary

textitLFAligner[11] is a bilingual sentence alignment tool built upon *hunalign*. It comes with a set of small size accurate and manually evaluated bilingual dictionaries, among which there is also an English-Italian bilingual dictionary. Kocabiyikoglu et al. (2018) used the English-French *LFAligner* dictionary as the starting material for a richer custom bilingual dictionary made of high-quality and manually annotated lexicons, resulting in a final bilingual dictionary of more than 100000 terms. We used hunalign in conjunction with the LFAligner En-It bilingual dictionary (containing around 14500 terms) as our baseline for the bilingual sentence alignment task. We set hunalign to perform two alignments. The first one uses both the lexical and the sentence-length information provided by the LFaligner dictionary and the Gale-Church algorithm. The resulting alignment is used to heuristically increase the size of the bilingual dictionary by looking at the co-occurrences found in the bi-sentences (the output of the first alignment). Finally, a second alignment is performed with the enriched bilingual dictionary (resulting from the first alignment).

[11] https://sourceforge.net/projects/aligner/

5.2 Hunalign with a Bilingual Dictionary Inferred with Moses and Giza++

Since the size of the *LFAligner* Italian-English dictionary was rather small (around 14500 terms) and we did not find other accurate and manually annotated freely available English-Italian lexicons, we investigated if a large automatically created lexicon could be useful. We compiled a large English-Italian corpus (containing 3131200 parallel sentences) by concatenating the *Europarl* (Koehn, 2005), the *Wikipedia* (Wołk and Marasek, 2014), the *GlobalVoices*[12], and the *books*[13] corpora from *OPUS* (Tiedemann, 2012). We used *Giza++* (Och and Ney, 2003) to align the corpus, followed by using Moses SMT (Koehn et al., 2007) to symmetrize the directional alignments, and extract a lexical translation table. The bidirectional tables contain a great amount of extremely low-probability translation terms hypotheses. We inferred a bilingual dictionary containing 692437 bilingual terms by filtering out the terms scoring less than 0.10. This inferred lexicon was used with Hunalign.

5.3 Vecalign

Rather than relying on sentence-length information and bilingual dictionaries, the sentences to be aligned are mapped into their vector representations and the alignment is done by computing the cosine similarity of the sentence embeddings. The underlying theoretical principle is that sentence embeddings seem to capture semantic information. Therefore, the higher the cosine similarity, the higher the probability two sentences in different languages have the same meaning. Vecalign requires multilingual embeddings of consecutive sentences. By concatenation of consecutive sentences we mean all combination of consecutive sentences in a window of size N. If there is a document containing the 3 sentences: *"Hi.", "I'm Jack.", "Nice to meet you."* and the window size is equal to 3 all possible consecutive sentences would be the original three sentences *"Hi.", "I'm Jack.", "Nice to meet you.", Hi. I'm Jack", "I'm Jack. Nice to meet you.", "Hi. I'm Jack. Nice to meet you."*. By embedding consecutive sentences, Vecalign work in scenarios where one sentence in language A should be aligned with multiple sentences in language B (e.g. Italian: "Ciao, sono Jack" - gloss "Hi, I am Jack" - English: "Hi.", "I am Jack") or viceversa. We used Facebook LASER (Schwenk and Douze, 2017; Schwenk and Li, 2018; Schwenk, 2018) to map into vectors all possible English and Italian consecutive sentences in a window of size 10 for each pair of documents to be aligned.

5.4 Evaluation Methodology

There are two main challenges in evaluating the sentence aligners. First, the lack of a gold standard file for computing the F1 score. Second, the fact that in many cases sentence alignments are not ambiguous and are easy to spot. For instance, Varga et al. (2007) reports precision and recall of over 0.97 on several texts. Lacking a gold standard, we have to resort to evaluating a small sample manually. If we sample sentences randomly, there will probably be very little difference between the aligners, due to the high number of easy-to-align sentences. Instead, we decided to focus our evaluation effort on a set of difficult alignment scenarios. Therefore, we focused only on the cases in which the results of the three aligners differed: a subset of 2030 likely difficult alignments from all pairs of 373 aligned chapters, with a total of 70204 possible alignments.[14] We then sampled 200 cases from this pool. For each sentence we compared the results of the three methods by assigning three possible values: correct alignment, wrong alignment, or partial alignment. By correct alignment we mean a perfect one to one, one to many, or many to one alignment: e.g. the English sentence *"It was delightful once upon a time"* aligned with the Italian sentence *"Era un piacere allora!"*. By wrong alignment we mean the cases in which the alignment was totally wrong: e.g. the English sentence *'Yours affectionately'* aligned with the Italian sentence *'Barkis ha intenzione di andare'* (gloss:'*Barkis plans to go*'). Finally, by partial alignment we mean the cases in which the alignment was neither totally wrong, nor totally correct, but only partially correct. This usually happens when a different number of sentences is used in the two texts to express the same meaning: e.g. the English sentence *'I see it now'* is aligned with the Italian sentence *'Mi sembra di rivederla: una lunga sala, con tre lunghe file di*

[12] http://casmacat.eu/corpus/global-voices.html

[13] http://opus.nlpl.eu/Books-v1.php

[14] The number of overlapping alignments between the systems could give an idea of the overall performance of the sentence aligners. While not all identical alignments can be expected to be correct, a high proportion of them are likely to be. Hunalign with and without the inferred dictionary have a 93.6% agreement. Hunalign and Vecalign have a 79.2% agreement if we exclude zero alignments not given by Hunalign.

	Hunalign	Hunalign+Inf	Vecalign
Correct	42	43	169
Wrong	108	89	5
Partial	50	68	26

Table 1: Evaluation of the bilingual text alignment task. For each method the number of correct, wrong, and partial alignments is given out of a selected pool of 200 sentences. *Inf* stands for the inferred dictionary.

piccoli scrittoi' (gloss: *I see it now. A long room with three long rows of desks'*).

5.5 Results and Discussion

Table 1 shows the results of the comparison of the 200 aligned sentences. For each one of them, we compared the alignments provided by our baseline, hunalign in conjunction with the inferred dictionary, and Vecalign. Vecalign outperforms the two variants of hunalign on this sample, with 169 correct alignments and only 5 wrong alignments, compared to hunalign which had just over 40 correct alignments for either variant. We explain the outperformance of *Vecalign* over the two hunalign methods to be strictly correlated to the use of multilingual embeddings of possible consecutive sentences. Computing the cosine similarity of consecutive sentences embeddings allows an easier spot of one to many, or many to one alignments. For instance, Vecalign managed to align correctly the two English sentences *'I see it now.'* and *'A long room with three long rows of desks.* with the single Italian sentence: *'Mi sembra di rivederla: una lunga sala, con tre lunghe file di piccoli scrittoi'* (gloss: *I see it now. A long room with three long rows of desks'*). Furthermore, *Vecalign* also gave as output a zero to one or zero to many alignments, which might mean that it filtered out the cases where the two e-books differed drastically in terms of paragraphs and sentences. For these reasons, we decided to use Vecalign for the final corpus. Approaching the sentence alignment problem using sentence embeddings has the advantage of obliterating bilingual lexical resources. The results of the two hunalign methods gives interesting insights over the quality and the size of the bilingual dictionary supporting the Gale-Church alignment algorithm. The use of the inferred dictionary (roughly six times the size of the manually annotated dictionary) led to a reduction of the amount of wrong alignments (19 less errors), without increasing the amount of the correct alignments.

Correct	Mild Errors	Severe Error
204	2	4

Table 2: Evaluation table for the forced alignment task with *Aeneas*

6 Forced Alignment

Forced alignment is the process to return time intervals matching word or sentence utterances in a audio file with their corresponding strings in a parallel text file. We forced the alignment between 373 LibriVox wav files and the English side of our textual parallel corpus, using Aeneas[15] rather than Gentle (as in Kocabiyikoglu et al. (2018)).

6.1 Gentle and Aeneas

Both tools are used for forcing speech to text alignment, but they differ in the way forced alignment is reached. Gentle is based on a pre-trained Kaldi (Povey et al., 2011) English ASR model: a hidden Markov model (HMM) determines the location of phonemes and words in the audio. For this reason, Gentle returns the time intervals for each word. Sentence time intervals can be obtained by a post-processing step on the Gentle output. The advantage of this method is that it can handle large portions of spurious text or audio. Aeneas is based on a TTS-DTW (text-to-speech and dynamic time warping) algorithm: the text transcript is first read by a text-to-speech software, then the dynamic time warping algorithm (Sakoe and Chiba, 1978) is deployed to compare the two audio sequences and return a synchronisation map with time intervals. The advantage of this approach is that it directly gives the sentence time intervals. The disadvantage is that it cannot handle large portions of spurious text: the audio has to match the text. As we did not encounter large portions of spurious text and audio in our corpus[16], we chose to use Aeneas for the forced alignment task.

6.2 Evaluation

We sampled 210 pairs of audio segments and their corresponding English texts. Manual evaluation was carried out by listening to each audio segment while reading its corresponding text file. Each

[15] https://github.com/readbeyond/aeneas

[16] the only discrepancies between the 373 LibriVox chapter audio recordings and our corpus were the LibriVox disclaimers at the beginning or at the end of each LibriVox recording, which we manually cut

	Number		
Speakers	98		
Hours	131.23		**Avg**
Chapters	373	**Per Speaker**	3.80
Segments	60561	**Duration**	7.80 s

Table 3: Corpus statistics, including the total number of chapters, speakers, segments and hours (rounded to decimals). In addition, the average (Avg) segment duration and the average number of chapters read by each speaker is given

audio-to-text alignment was given one of the following labels: *correct*, *mild*, and *severe*. The label *correct* indicates that text and audio match and all words are pronounced entirely and clearly, without brutal and abrupt cuts. The label *mild error* describes a scenario where a letter was missing from the starting or ending of the audio or was not clearly pronounced. The *severe error* label indicates that the audio and its corresponding text are severely off-sync or completely wrong. Due to the fact that *severe errors* can cause error propagation, every time we ran across a severe error, we also checked the preceding and following audio-to-text alignments. Table 2 shows the results of our manual evaluation. Out of 210 manually checked alignments, we encountered only 4 severe errors and none of them caused error propagation. The few severe errors were due to actual mismatches between the audio and the text (e.g. the speaker decided to read a footnote from the original e-book).

7 Corpus Statistics and Structure

IESTAC[17] contains 60561 triplets of English audio, English source text, and Italian textual translation. Statistics are given in Table 3. The corpus is available as a zipped folder containing two parallel raw text files[18] and nine folders, one for each book. Each folder is named after a Gutenberg Project Ebook ID. Each one of these folders contains several sub-folders, one for each aligned chapter. The chapter folders are named as increasing integers. In each chapter folder, there are several triplets of files. The alignment is preserved by the base-name notation: each element of the triplet has a base-name composed of three concatenated integers (E-bookID, ChapterID, SegmentID). The file-name extension is used to disambiguate between the audio, the Italian textual translation, and the English source text (e.g. 83_07_33.wav, 83_07_33.it, 83_07_33.en). We also provide the users with a SQL database to allow them to query the corpus according to their needs.

8 Future Work

In future work, we want first of all to perform an extrinsic evaluation of our corpus on both automatic speech recognition and end-to-end speech-to-text machine translation tasks. A possible tool for that might be FBK-Fairseq-ST[19]. We also want to further improve the quality of the alignments by filtering out low-quality ones. For such task we might investigate the use of alignment scores, character error rate, language model perplexity, sentence embeddings similarity scores, and length ratio heuristic. Another direction of interest is to work towards a full taxonomy for collecting speech-to-text datasets for machine translation.

9 Conclusion

We have explored methods for creating a bilingual corpus with English speech and text, and Italian text. The corpus collection is based on available English speech and text in the LibriVox and the Project Gutenberg collection of audiobooks and e-books, which we mapped to Italian texts in this work. We explored and evaluated a number of methods for the different steps needed in the corpus creation, which might guide future work in creating corpora for other language pairs. For the text collection, we needed to map English books to their equivalent Italian translations. We proposed three methods based on WikiData matching, Wikipedia matching, and MT of book titles. We found that the three methods were complementary, each contributing some unique titles. The next step was bilingual sentence alignment, for which we found that Vecalign, a method based on computing the cosine similarity of consecutive sentence embeddings, outperformed the traditional Gale-Church method when dealing with difficult alignments. Finally we argued for the use of a TTS dynamic time warping system for forcing the alignment between English speech and text, and showed that the results were of high quality. As part of this work we release IESTAC

[17] https://github.com/Giuseppe-Della-Corte/IESTAC

[18] *parallel.it* and *parallel.en* contain respectively: total tokens (1425072 - 1577118), unique tokens (63,774 - 38,325)

[19] https://github.com/mattiadg/FBK-Fairseq-ST

References

Sadaf Abdul-Rauf, Mark Fishel, Patrik Lambert, Sandra Noubours, and Rico Sennrich. 2010. Evaluation of sentence alignment systems. Technical report, Fifth MT Marathon.

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 82–91, New Orleans, Louisiana. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *ISWC 2007, ASWC 2007: The semantic web*, Lecture Notes in Computer Science, vol 4825, pages 722–735. Springer, Busan, Korea.

Benjamin Beilharz, Xin Sun, Sariya Karimova, and Stefan Riezler. 2020. LibriVoxDeEn: A corpus for German-to-English speech translation and German speech recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3590–3594, Marseille, France. European Language Resources Association.

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-end automatic speech translation of audiobooks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6224–6228, Calgary, Canada. IEEE.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*, Barcelona, Spain.

Alan W Black. 2019. CMU wilderness multilingual speech dataset. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975, Brighton, UK. IEEE.

Fabienne Braune and Alexander Fraser. 2010. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Coling 2010: Posters*, pages 81–89, Beijing, China. Coling 2010 Organizing Committee.

Giuseppe Della Corte. 2020. Text and Speech Alignment Methods for Speech Translation Corpora Creation: Augmenting English LibriVox Recordings with Italian Textual Translations. Master's thesis, Uppsala University, Sweden.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019a. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Mattia Antonino Di Gangi, Matteo Negri, Roldano Cattoni, Dessi Roberto, and Marco Turchi. 2019b. Enhancing transformer for end-to-end speech-to-text translation. In *Machine Translation Summit XVII*, pages 21–31, Dublin, Ireland.

Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noel Kouarata, Lori Lamel, Hélène Maynard, Markus Mueller, Annie Rialland, Sebastian Stueker, François Yvon, and Marcely Zanon-Boito. 2018. A very low resource language speech corpus for computational language documentation experiments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-ST: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233, Barcelona, Spain. IEEE.

Jodi Kearns. 2014. Librivox: Free public domain audiobooks. *Reference Reviews*, 28(1):7–8.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting librispeech with French translations: A multimodal corpus for direct speech translation evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit X*, volume 5, pages 79–86, Phuket, Thailand.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open

source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev 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, Brisbane, Australia. IEEE.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, location = Hilton Waikoloa Village, Hawaii, US,. IEEE Signal Processing Society.

Nicholas Ruiz, Mattia Antonino Di Gangi, Nicola Bertoldi, and Marcello Federico. 2017. Assessing the tolerance of neural machine translation systems against speech recognition errors. In *Proceedings of Interspeech 2017*, page 2635–2639, Stockholm, Sweden.

Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. In *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2018 Workshop*, Montreal, Canada.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Bryan Stroube. 2003. Literary freedom: Project Gutenberg. *XRDS: Crossroads, The ACM Magazine for Students*, 10(1).

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Astrid Van Aggelen, Laura Hollink, Max Kemman, Martijn Kleppe, and Henri Beunders. 2017. The debates of the European parliament as linked open data. *Semantic Web*, 8(2):271–281.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing (RANLP)*, pages 247–258, Borovets, Bulgaria.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020a. CoVoST: A diverse multilingual speech-to-text translation corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4197–4203, Marseille, France. European Language Resources Association.

Changhan Wang, Anne Wu, and Juan Pino. 2020b. CoVoST 2 and massively multilingual speech-to-text translation. *arXiv e-prints arXiv:2007.10310v2*.

Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. In *Proceedings of Interspeech 2017*, pages 2625–2629, Stockholm, Sweden.

Krzysztof Wołk and Krzysztof Marasek. 2014. Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs. *Procedia Technology*, 18:126–132.