IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Bilingual Prosodic Dataset Compilation for Spoken Language Translation

Alp Öktem[1], *Mireia Farrús*[1], *Antonio Bonafonte*[2]

[1]Universitat Pompeu Fabra, Spain
[2]Universitat Politècnica de Catalunya, Spain
[1]{alp.oktem, mireia.farrus}@upf.edu
[2]antonio.bonafonte@upc.edu

Abstract

This paper builds on a previous methodology that exploits dubbed media material to build prosodically annotated bilingual corpora. The almost fully-automatized process serves for building data for training spoken language models without the need for designing and recording bilingual data. The methodology is put into use by compiling an English-Spanish parallel corpus using a recent TV series. The collected corpus contains 7000 parallel utterances totaling to about 10 hours of data annotated with speaker information, word-alignments and word-level acoustic features. Both the extraction scripts and the dataset are distributed open-source for research purposes.

Index Terms: bilingual corpora, spoken machine translation, prosody

1. Introduction

Recent approaches in speech-to-speech translation research gave focus on the transfer of para-linguistic information between the languages involved. These approaches extend on the classic pipeline of S2S translation, which consists of automatic speech recognition (ASR), machine translation (MT) and text-to-speech synthesis (TTS), and introduces models that connect directly the information in input and target speech signal. Prosody, which is the linguistic information encoded in cues such as stress, intonation and rhythm, directly influences the communicative value of the source utterance and thus needs to be carried to the output to achieve a complete translation. For example, the effect of emphasis transfer in S2S translation is shown to influence directly the quality of translation [1].

Text data to train machine translation models are collected from utterances carrying the same linguistic information in different languages. Same way, data-driven models that deal with prosody transfer necessitate audio data that not only carries the same linguistic information, but also the same para-linguistic content, encoded in each language's prosody. For example, Anumanchipalli et. al [2] collected 200 parallel sentences from a flight magazine and recorded using a bilingual speaker to achieve intent transfer in S2S translation. Truong et. al [3] recorded 966 parallel segments with acted emphasis in order to achieve emphasis transfer. A fairly recent project SIWIS [4] that focuses on translation of Swiss languages had 171 prompts with emphasis instructions recorded by many speakers as their training data. These and other similar works [5, 6] rely on small data that is created in laboratory conditions and thus partially reflecting naturalness of conversational language.

Our previous work [7] outlined a methodology for compiling such corpora without the need for preparing bilingual prompts and a recording process. Instead, it exploits professionally created bilingual material that is available through dubbed movies. Dubbing is a carefully designed process where the movie content is first translated and then acted by professionals to reflect original movie lines. A methodology that exploits this parallel nature can be used to create bilingual material for any language pair where dubbing is performed. Using this framework, we present an English-Spanish bilingual corpus of 7000 parallel TV series segments annotated with prosodic features to be used in spoken language translation research.

Contributions of this work can be summarized as follows: (1) We cover the shortcomings of the *movie2parallelDB* framework [8] introduced in [7] by improving word-alignment and parallel segment extraction processes, (2) we expand the methodology to automatically include speaker information from movie scripts, (3) we present an English-Spanish parallel corpus (*Heroes Corpus*) of 10 hours of audio together with transcriptions and prosodic features available through this link[1].

2. Methodology

The methodology introduced by Öktem et al. in [7] consists of three stages: (1) a monolingual step, where audio+text segments are extracted from the movie in both languages using transcripts and cues in subtitles, (2) prosodic feature annotation and (3) alignment of monolingual material to extract the bilingual segments. They discuss some shortcomings of their methodology. Firstly, they report that the word-aligner tool they use is not precise and fast enough. Secondly, they don't address the problem of subtitle/audio mismatch in the dubbed language. This problem occurs because translation of the original movie script for subtitling and dubbing are independent processes that require care in different aspects. While text translation for subtitles can be straightforward, dubbing requires the translated utterances to be in sync with the lip movements of the actors. Because of this reason, many times the dubbed sentences differ substantially from subtitles.

Another problem in their approach is in the bilingual

[1] *http://hdl.handle.net/10230/35572*

10.21437/IberSPEECH.2018-5

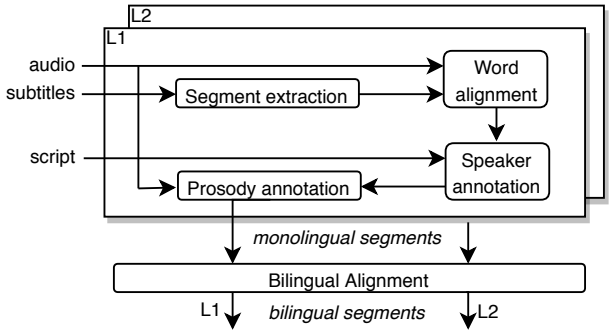

Figure 1: *Overall corpus extraction pipeline. Audio excerpts are first processed in each language and then aligned to obtain bilingual segments.*

segment alignment stage. In order to extract the parallel sentences from two languages, they first translate the sentences in one language to the other using a machine translation system. Then, sentences ordered in both languages are matched by selecting most similar pairs of sentences in terms of a translation scoring metric. As sentence structures can differ between two scripts, this approach employing an unreliable MT system can lead to mismatched segments.

This section explains how we addressed these shortcomings and also ensured the following requirements for the segments to be extracted: (1) They should contain the utterance of only one speaker, (2) transcriptions should be exactly what is being spoken in the audio, (3) it should have f0/intensity and speech rate information labeled on word level.

The overall scheme of the corpus extraction methodology is illustrated in Figure 1.

2.1. Audio segment mining using subtitles

Subtitles are the source for obtaining both (1) audio transcriptions, (2) timing information related to utterances in a movie. These information are contained in a standard *srt* subtitle file, entry by entry like the structure below:

```
1
00:00:09,980 --> 00:00:12,256
Please, tell me who I am,
2
00:00:12,540 --> 00:00:13,974
and what the future holds.
Where are we?
3
00:00:14,740 --> 00:00:16,572
-We're in New York.
-Where is everyone?
```

Each subtitle entry is represented by an index, time cues and the script being spoken at that time in the movie. The script portion can consist of multiple (#2), complete (#2,3) or incomplete sentences (#1) and from single (#1,2) or multiple speakers (#3). Using only the time cues for extracting audio segments with complete sentences of a single speaker does not suffice.

For both the objective of obtaining word-level prosodic features and for segmenting multi-speaker portions, we have used a speech-to-text aligner software. Multi-speaker segments were split from the words following speech-dashes. For merging incomplete segments, punctuation information was used.

2.2. Speech-to-text alignment

For speech-to-text alignment, we used the open-source tool *Montreal Forced Aligner* (MFA) [9]. Forced alignment process is built on an automatic speech recognition system and requires its own acoustic models and a pronunciation dictionary. Although pre-trained models for both English and Spanish is provided through the tool's website[2], Spanish pronunciation dictionary isn't openly available. For this reason, we have created a Spanish pronunciation dictionary[3] that uses the same phoneme set as MFA using word list from the open-source spell checker tool *ISpell*[4] and obtaining their phonetic transcriptions using *TransDic*[5].

2.3. Speaker annotation through scripts

Movie scripts, which contain dialogue and scene information, are valuable pieces of information for determining the segment speaker labels. Scripts follows approximately the same format: Actor/actress name is followed by the line they say. And in between, there might be non-spoken information in brackets. An example excerpt from a movie script of TV series Heroes is given below:

```
Claire: What did you do? What the hell is going
    on?
[Caption: Manhattan 16 years ago]
Noah: (in Japanese) We think she died in the
    fire.
Claire: Dad? (Hiro covers her mouth to be quiet
    .)
Kaito: (in Japanese) Once again. Not a request.
    (Kaito hands baby Claire to Noah.)
```

Unlike subtitles, scripts don't have timing information. In order to map subtitle segments with the speaker information we followed an automatic procedure. We first removed all non-spoken text, which is included in brackets. Then, speaker tags and corresponding lines are extracted with regular expressions depending on the format of the script. Next, segments coming from subtitles are mapped one by one to lines in the script. If 70% of the words in a subtitle segment is included in a script turn, then the segment is labeled with the speaker of that turn. We found this metric to successfully label 95% of the segments.

Scripts are usually only available in the original language. The dubbed language segments are labeled the same as their matches after the alignment step.

[2] https://montreal-forced-aligner.readthedocs.io/

[3] Resource available in: https://github.com/TalnUPF/phonetic_lexica

[4] https://www.gnu.org/software/ispell/

[5] https://sites.google.com/site/juanmariagarrido/research/resources/tools/transdic

2.4. Word-level acoustic feature annotation

Each word in the extracted segments is automatically annotated with the following acoustic features: mean fundamental frequency (f0), mean intensity, speech rate and silence intervals (pauses) before and after. The first two features are extracted with the *ProsodyTagger* toolkit [10] built on *Praat* [11]. Pause information is calculated from word-boundary information and speech rate is calculated using:

$$word\ speech\ rate = \frac{\#syllables\ in\ word}{word\ duration} \quad (1)$$

To represent speaker independent, perceptual acoustic variations in the segments, both f0 and intensity values are converted into logarithmic semitone scale relative to the speaker norm value. Thus, speaker mean values were represented by zero values in both cases. Semitone values are calculated with the corresponding formula:

$$semitone(x, norm) = 12 * \log(\frac{x}{norm}) \quad (2)$$

2.5. Cross-lingual segment alignment based on subtitle cues

The first three methodologies presented in this section dealt with extraction of segments in each language. This subsection explains how segments extracted for each language are aligned to create the bilingual segment pairs.

We have developed an aligning process based on timing information of the extracted segments. Note that the segment alignments can be one-to-one, one-to-many, many-to-one or many-to-many depending on the sentencing structure in the subtitles. To create our own alignment algorithm based on time cues, we first defined a metric that measures the correlation percentage between two sets of ordered segments S=$\langle s_1, ..., s_N \rangle$ and E=$\langle e_1, ..., e_N \rangle$:

$$segments\ correlation = max(0, \frac{correlating}{span} \times 100) \quad (3)$$

$$correlating = min(e^s_N, s^e_N) - max(s^s_1, s^s_1) \quad (4)$$

$$span = max(e^e_N, s^e_N) - min(e^s_1, s^s_1) \quad (5)$$

where e^s_x and e^e_x denote the starting and ending time of the x^{th} segment in set E, s^s_x and s^e_x denote the starting and ending time of the x^{th} segment in set S.

The alignment procedure is as follows: Two indexes i_E, i_S are kept which slide through the segments of each language. First, segments corresponding to each index are checked if they correlate more than the T_{Sure} threshold. If they do, they are assigned as a one-to-one matched pair. If not, the possibilities of one-to-many, many-to-one or many-to-many matches are considered. This is done through computing the correlations between combinations of the current and two following segments and selecting the most correlating segment set pair. While considering combinations of the segments it is made sure that two merged segments belong to the same speaker and are not more than 10 seconds far from each other. If the combined segment set pair with highest correlation has a correlation of more than T_{Merged} threshold, then the combinations are merged into one segment and paired with each other.

Although the T_{Sure} threshold catches most of the one-to-one mapping segments, we realized that many of them fall below this threshold even if they map. So, we added another decision step that if one-to-one mapping correlation scores higher than merged pairings and it scores above a T_{OK} threshold, then it is preferred as a matched pair.

2.6. Output format

We needed to store the corpus segments in a convenient way to use with machine learning based applications. We used the *Proscript* library [12] for storing the enhanced transcripts. This library makes it possible to store and manipulate speech transcript related data. The segments are stored in *csv* files that keep the information listed in Table 1. A csv file containing all the segments is created for each episode as well.

Table 1: *Segment information kept in a Proscript format csv file.*

Information	Details
word	tokenized
id	unique word id
timing	start and end times
pause	coming before and after
punctuation	attached to beginning and end
f0	in Hertz and log-scale (semitones)
intensity	in Decibels and log-scale
speech rate	relative to syllables

3. Compiling the Heroes corpus

We put our methodology into practice by compiling a corpus from the science fiction TV series *Heroes*[6]. Originating from United States, Heroes ran in TV channels worldwide between the years 2006 and 2010. The whole series consists of 4 seasons and 77 episodes and is dubbed into many languages including Spanish, Portuguese, French and Catalan. Each episode runs for a length of 42 minutes.

We chose this series as we had access to the DVD's with Spanish dubbing. Also, we found it to have the Spanish subtitles closest to the Spanish dubbing scripts among other series.

3.1. Raw data acquisition

The DVD's of the series were obtained from the Pompeu Fabra University Library. Episodes were extracted using the Handbrake software and were saved as Matroska format (mkv) files. Mkv files can hold multiple channels of audios and subtitles embedded in it like DVDs. In order to run our scripts we first needed to extract the audio and subtitle pairs for both languages. Audio is extracted using the *mkvextract* command line tool[7]. As subtitles were embedded as bitmap images in the DVD, we had to

[6] Produced by Tailwind Productions, NBC Universal Television Studio (2006-2007) and Universal Media Studios (2007-2010)

[7] `https://mkvtoolnix.download/`

run optical character recognition (OCR) in order to get srt format subtitles. As OCR is an error-prone process, the resulting srt files needed to be spell checked.

We collected English and Spanish audio of 21 episodes totaling to 25 hours of raw audio and their corresponding subtitles. The episode scripts were obtained from a fan-site in the Internet[8].

3.2. Manual subtitle correction work

A speech corpus necessitates properly transcripted speech segments. Our method is based on obtaining transcriptions from subtitles. Although subtitles are highly reliable sources for obtaining proper transcriptions in the original language of the movie, this is not the case in the dubbed languages. This is due to the fact that dubbing transcript needs to satisfy visual alignment such as lip movements, whereas subtitles do not. Also, subtitles are often done in a more concise way to facilitate easy reading. In our case, we observed that the Spanish subtitles were matching the Spanish audio in approximately 80% of the cases. To accommodate this issue, we manually corrected the Spanish subtitles to match with the Spanish audio. Both subtitle transcripts and timestamps had to be corrected. This process was done using a subtitle editing program *Aegisub*[9].

An advantage the manual correction process gives is the opportunity to filter out unwanted audio portions that would end up in the corpus. Subtitle segments that contained noise and music, overlapping or unintelligible speech and speech in other languages (e.g. Japanese) were removed during this process. The spell checking and timestamps and script correction of 21 episodes was done by two annotators and took 60 hours in total.

3.3. Heroes corpus in numbers

We present the statistics of the first preparation sprint of *The Heroes Corpus*. 21 episodes from season 2 and season 3 were processed. The totaled audio durations of 7000 parallel segments approaches 10 hours (see Table 2). Counts of several linguistic units in the final parallel corpus are presented in Table 3. A summary of how much of the content in one episode ended up in the dataset in average is presented in Table 4.

	English	**Spanish**
Total duration	4:45:36	4:43:20
Avg. duration/segment	0:00:02.44	0:00:02.42

Table 2: *Corpus duration information.*

	English	**Spanish**
# words	56320	48593
# tokens	72565	63014
# sentences	9892	9397
Avg. # words/sentence	5.69	5.17
Avg. # words/segment	8.04	6.94
Avg. # sentences/segment	1.41	1.34

Table 3: *Word, token, sentence counts and average word count for parallel English and Spanish segments.*

[8] `https://heroes-transcripts.blogspot.com/`
[9] `http://www.aegisub.org/`

	English	**Spanish**
Avg. # sentences (subtitles)	647	554
Avg. # sentences (extracted)	628	513
Avg. # segments	526	459
Avg. # parallel segments	334	

Table 4: *Averages numbers for each episode.*

4. Discussion

The first version of the Heroes corpus shows that our automated method for bilingual corpus building is successful in terms of the quality of the segments extracted. Our manual inspections show that the segments are correctly aligned and the transcriptions are correctly stored with the audio segments.

The Spanish subtitle correction task was the only time-consuming part of the whole process. However, it showed that it is also useful for obtaining clean parallel segments. Subtitle segments that were removed during the correction process ensured the elimination of unwanted audio portions.

We can interpret Table 4 to show us the amount of information loss at various stages. The first one being the word-alignment process where in average 5% of the sentences are lost due to the failure of the word aligner in segmenting words. We found out that many of the segments that are lost this way were either noisy or erroneous. The biggest loss happens at the stage of alignment where in average 30% of the segments in each language are left unaligned. This percentage is directly affected by the alignment parameters explained in Section 2.5. For example, selecting a lower T_{Sure} leads to detecting more aligned segments but also to more mismatches. A similar logic applies to T_{OK}. Also, choosing a lower T_{Merged} leads to more coverage of the sentences but more as combinations with others, leading to fewer and longer segments. After experimenting with a handful of parameter combinations, we decided on this configuration for obtaining the corpus presented in this paper: $T_{Sure} = 70\%$, $T_{Merged} = 80\%$ and $T_{OK} = 30\%$.

5. Conclusions

We have presented an English-Spanish bilingual corpus of dubbed TV series content. The corpus that consists of 7000 parallel audio segments with transcriptions and annotated prosodic features is made openly available. We hope both our methodology and the corpus we compiled be useful for the speech-to-speech translation research community.

6. Acknowledgements

Special thanks to annotators Sandra Marcos Bonet and Laura Gómez Fisas for their work during the Spanish subtitle correction process. The annotation work carried by the annotators was financed with the 2018 Maria de Maeztu Reproducibility Award from Department of Information and Communication Technologies of Universitat Pompeu Fabra received by the first author. The second author is funded by the Spanish Ministry through the *Ramón y Cajal* program.

7. References

[1] A. Tsiartas, P. G. Georgiou, and S. S. Narayanan, "A study on the effect of prosodic emphasis transfer on overall speech translation quality," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 8396–8400.

[2] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black, "Intent transfer in speech-to-speech machine translation," in *Spoken Language Technology (SLT) Workshop*. IEEE, 2012, pp. 153–158.

[3] Q. T. Do, S. Sakti, and S. Nakamura, "Toward expressive speech translation: A unified sequence-to-sequence lstms approach for translating words and emphasis," in *INTERSPEECH*, 2017.

[4] J.-P. Goldman, P.-E. Honnet, R. Clark, P. N. Garner, M. Ivanova, A. Lazaridis, H. Liang, T. Macedo, B. Pfister, M. S. Ribeiro, E. Wehrli, and J. Yamagishi, "The siwis database: a multilingual speech database with acted emphasis," 2016.

[5] P. D. Agüero, J. Adell, and A. Bonafonte, "Prosody generation for speech-to-speech translation," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. IEEE, 2006, pp. 557–560.

[6] T. Kano, S. Takamichi, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "An end-to-end model for cross-lingual transformation of paralinguistic information," *Machine Translation*, Apr 2018. [Online]. Available: https://doi.org/10.1007/s10590-018-9217-7

[7] A. Öktem, M. Farrús, and L. Wanner, "Automatic extraction of parallel speech corpora from dubbed movies," in *Proceedings of the 10th Workshop on Building and Using Comparable Corpora (BUCC)*, Vancouver, Canada, 2017, pp. 31–35.

[8] A. Öktem, "movie2parallelDB: Automatic parallel speech database extraction from dubbed movies," 2018. [Online]. Available: https://github.com/alpoktem/movie2parallelDB

[9] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proc. Interspeech*, 2017, pp. 498–502.

[10] M. Dominguez, M. Farrús, and L. Wanner, "An automatic prosody tagger for spontaneous speech," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, 2016, pp. 377–386. [Online]. Available: http://www.aclweb.org/anthology/C16-1037

[11] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [Computer software], retrieved from http://www.praat.org/," 2017.

[12] A. Öktem, "Proscript: Python library for prosodic annotation of speech segments," 2018. [Online]. Available: https://github.com/alpoktem/proscript