EUROPARL-ST: A MULTILINGUAL CORPUS FOR SPEECH TRANSLATION OF PARLIAMENTARY DEBATES

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló,
Adrià Giménez, Albert Sanchis, Jorge Civera, Alfons Juan

Machine Learning and Language Processing (MLLP) research group
Valencian Research Institute for Artificial Intelligence (VRAIN)
Universitat Politècnica de València, Spain

ABSTRACT

Current research into spoken language translation (SLT), or speech-to-text translation, is often hampered by the lack of specific data resources for this task, as currently available SLT datasets are restricted to a limited set of language pairs. In this paper we present *Europarl-ST*, a novel multilingual SLT corpus containing paired audio-text samples for SLT from and into 6 European languages, for a total of 30 different translation directions. This corpus has been compiled using the debates held in the European Parliament in the period between 2008 and 2012. This paper describes the corpus creation process and presents a series of automatic speech recognition, machine translation and spoken language translation experiments that highlight the potential of this new resource. The corpus is released under a Creative Commons license and is freely accessible and downloadable.

Index Terms— speech translation, spoken language translation, automatic speech recognition, machine translation, multilingual corpus

1. INTRODUCTION

The significant developments in the automatic speech recognition (ASR) and machine translation (MT) fields in the last five years, which have been mainly driven by advances in deep learning models and greater data availability, have picked up interest in spoken language translation (SLT) as the natural convergence of the two previous fields.

However, SLT is far from solved. Two approaches are currently used: cascade [1, 2, 3] and end-to-end models [4, 5, 6], without one being clearly adopted by the community. The latest IWSLT 2018 evaluation campaign showed that the cascade approach outperforms end-to-end models [7], but recent developments in the area are shrinking that gap [8]. The performance of SLT, and especially end-to-end SLT models, is limited by the lack of SLT corpora when compared with the more resource-rich ASR and MT fields. Furthermore, most of the existing SLT corpora are limited to only English speech data paired with translations into other languages, such as the recently released MuST-C corpus [9]. This fact limits the SLT research than could be carried out in language pairs other than English. Moreover, recent studies report their main results using either the paid Fisher/Callhome corpora [1, 4, 5, 6, 10], or private propietary datasets [8], which limits reproducibility for the research community.

In order to alleviate these problems, we have created the *Europarl-ST* corpus out of European Parliament (EP) debates and their official transcriptions and translations. To our knowledge, Europarl-ST is the first fully self-contained, publicly available corpus with both, multiple (speech) source and target languages, which will also enable further research into multilingual SLT (cf. [11]). The Europarl-ST corpus is released under a Creative Commons Attribution-NonCommercial 4.0 International license (CC BY-NC 4.0), and can be freely accessed and downloaded at `www.mllp.upv.es/europarl-st`.

2. DATA COLLECTION AND PROCESSING

The corpus has been created using the publicly available videos from European Parliament debates[1]. In order to ease the access to the different attributes of each debate the LinkedEP database is used [12]. The basic unit of this corpus is a *speech*, an intervention made by a single speaker at the Parliament.

The EP debates suffer from missing videos, inaccurate timestamps and, as of 2011, many translations into languages other than English are missing. Indeed, after 2012, the translation of EP debates is not available. Additional data is discarded when constructing the Europarl-ST corpus, since in order to build a corpus of audio-transcription-translation triples,

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 761758 (X5gon); MCIU/AEI/FEDER,UE under the Multisub (RTI2018-094879-B-I00) research project and the Government of Spain's FPU scholarship FPU18/04135.

[1] http://www.europarl.europa.eu/plenary/en/debates-video.html

Table 1. Number of speech hours after each step of the data filtering pipeline, and CER of the filtered data sets.

	Initial	Step 1	Step 2	CER
De	207	149	44	10.7
En	346	252	120	12.9
Es	80	59	34	9.1
Fr	183	132	47	10.7

it is necessary to properly define forced audio-text and text-text sentence alignments, and intra-sentence word-alignment.

For this initial release of the corpus, experiments are reported from and into English (En), German (De), French (Fr) and Spanish (Es), since these languages accumulate a larger number of speech hours. Additional languages, such as Italian and Portuguese, will also be included in the initial release, but experimental results are not reported due to time constraints.

2.1. Audio-to-text alignment and data filtering

One of the challenges processing this corpus is that timestamps provided for the EP speeches can be wildly inaccurate, and as a side-effect, they often contain fragments from both the preceding and following speeches. In order to ameliorate this, first we carried out a Speaker Diarization (SD) step for each speech using the *LIUM SpkDiarization* [13] toolkit. Second, for each speech, the longest sequence of audio segments belonging to the same speaker was clipped, making the assumption that it does correspond to the actual intervention of the speaker of this speech. Finally, a forced alignment of the clipped audio segments was carried out against their corresponding transcriptions to obtain correct word timestamps. Forced alignments were carried out using the TLK toolkit's decoder [14] and the FF-DNN acoustic models (AM) described in Section 3.1, restricting the search graph of the decoder to the provided transcription. As a result of the procedure describe above (Step 1), around 28% of the original audio data was discarded (see Table 1 for language-based statistics).

Next, in order to produce a reliable corpus than could be used to both train and evaluate models, a second data filtering step was carried out based on character error rate (CER) computed at the speech level. First, we apply ASR over all speeches, using the ASR system described in Section 3.1. Second, we measure how much the recognition outputs differ from the provided reference transcriptions by computing CER values. Our aim is to eliminate speeches that exhibit significant amounts of non-verbatim transcriptions, as well as non-transcribed speech or unuttered transcripts that could be present either due to mistakes of the SD process or to annotation errors in the original data. In comparison with the well-known word error rate (WER) metric, the CER is more convenient for our purposes, as it better gauges the phonetic similarity between the recognised speech and the candidate reference transcripts, and alleviates the effect of ASR out-of-vocabulary words.

Table 2. Statistics of the preprocessed Europarl-ST corpus.

Src	Trg	Speeches	Sent.	Hours	Src w.	Trg w.
De	En	1521	18.1K	42	345K	409K
	Es	863	10.2K	24	196K	242K
	Fr	839	9.6K	24	191K	265K
En	De	3233	35.5K	89	811K	793K
	Es	3184	34.4K	87	796K	865K
	Fr	3174	34.5K	87	794K	974K
Es	De	694	7.0K	20	193K	186K
	En	1131	11.2K	32	305K	307K
	Fr	684	6.9K	20	190K	225K
Fr	De	832	9.6K	25	263K	227K
	En	1306	15.1K	38	394K	371K
	Es	817	9.4K	25	260K	246K

Finally, language-dependent CER thresholds were defined, 15% for French, German and Spanish and 20% for English, in order to exclude those speeches whose CER exceeded these thresholds. Thresholds were defined based on previous experience filtering crawled speech data. As a result of this filtering step (Step 2), around 40-70% of the audio data selected in the previous step was discarded (see Table 1 for detailed statistics). CER figures computed on the selected speeches after Step 2 are also provided in Table 1. These figures are an approximation to a quality assurance measure to ensure that only speeches with little or no noise are included into the corpus. At the end of this process, around 60-80% of the original data was filtered out.

2.2. Source-to-target text alignment

Each selected speech, both transcription and translation, is divided into sentences, using the *sentence-split.pl* script from the Moses toolkit [15], that are aligned using Gargantua [16]. Sentences longer than 20 seconds were split into shorter ones in order to accommodate the data for training purposes. Shorter sentences were generated by computing word-alignments using Fast-align [17] and pairing them to guarantee intra-sentence alignments. The statistics of the remaining data after text-aligning and excluding speeches with no translation into the respective target language are shown in Table 2. As observed in Table 2, this corpus is provided with segmentations, both at the speech and sentence level. The sentence-level segmentation is expected to be devoted to training purposes, while evaluations at the speech level are reported in Section 3.

A speaker-independent train/dev/test partition was defined, devoting approximately 3 hours of audio to each of the dev and test sets, and the rest was left as training data. The dev/test speakers are the same for language directions with the same source language. However, the number of speeches may differ because for some speeches there are translations missing. The training data might be used to fine-tune and adapt out-of-domain models to this specific domain, or even to train basic in-domain ASR, MT and SLT models from scratch.

3. EXPERIMENTS AND RESULTS

This section introduces the setup used for the experiments performed with the Europarl-ST corpus. In addition to ASR and MT experiments, SLT experiments following a cascade approach, in which the output of an ASR system is used as input for an MT system, are reported. First, the performance of models trained on general domain data when applied to the Europarl-ST corpus are evaluated, and second, the usefulness of the Europarl-ST training data for adapting models to the EP specific domain is also assessed. More precisely, results of ASR, MT and SLT experiments are reported using the 4 selected languages (English, German, Spanish and French), for a total of 12 translation directions in the case of translation experiments. Results are reported in terms of WER for ASR experiments, and BLEU [18] for MT and SLT experiments.

In order to properly compute BLEU, both the system hypothesis and the reference translation must have the same number of lines. However, in a SLT experiment, the number of lines will depend on the segmentation applied to the output of the ASR system in the cascade case, and the SLT system in the end-to-end case. Therefore, it is standard to re-segment the system hypothesis in order to get the same number of lines as in the reference. This re-segmentation is performed with the *mwerSegmenter* [19], and then evaluated by computing case-sensitive BLEU (including punctuation signs) with SacreBLEU [20]. All evaluations are carried out at the speech level, so re-segmentation is applied to both, MT and SLT experiments, in order to evaluate them under the same conditions.

3.1. ASR

General-purpose ASR systems for German (De), English (En), Spanish (Es) and French (Fr) were used to generate automatic transcripts for audio speeches in the development and test sets of each language pair. These automatic transcripts are the input text for subsequent MT systems within the SLT cascade approach.

These ASR systems are based on the hybrid deep neural network hidden Markov model (DNN-HMM) approach. Acoustic models, are generated using the TLK toolkit [14] to train feed-forward (FF) DNN-HMM models of three left-to-right tied triphone states, using 48 (De, Es, Fr) or 80-dimensional (En) Mel frequency cepstral coefficients (MFCCs) as input features. State tying was done by applying language-dependent classification and regression trees (CART), which resulted in 10K (Es, Fr) or 18K (De, En) tied triphone states. With the exception of the French ASR system which only features FF-DNNs, these models were used to bootstrap bidirectional long-short term memory (BLSTM) DNN models, the latter model trained using Tensorflow [21]. For German, Spanish and French, we also trained fCMLLR AMs, so that these systems follow a two-step recognition process.

Table 3. Statistics of AM and LM training data.

	Hours (K)	Sentences (M)	Words (G)
De	0.9	71	0.8
En	5.6	532	300
Es	0.8	24	0.7
Fr	0.7	110	1.8

Table 4. ASR results in terms of WER on the test sets.

	De	En	Es	Fr
De	–	19.8	19.8	19.9
En	17.2	–	17.2	17.1
Es	14.6	15.0	–	14.6
Fr	27.3	24.3	27.2	–

On the other hand, regarding the language models (LM), we used a linear combination of several n-gram LMs trained with SRILM [22], combined with a recurrent NN (RNN) LM trained using the RNNLM toolkit [23] (De, Es, Fr), or an LSTM LM trained with the CUED-RNNLM toolkit [24] (En). The vocabulary of these systems was restricted to 200K words. Table 3 shows overall statistics of the amount of training data that were used to train the acoustic models, in terms of speech hours, and the language models, in terms of sentences and words. The number of English words includes 294G words from Google Books counts.

Table 4 shows, for each SLT test set, WER figures computed from the ASR part only. Rows represent source (ASR) languages, whilst columns represent target (MT) languages. It is important to remind that the set of source speeches, though mostly overlapping, are different because the correspoding target text translation may not exist. Results show that most WER figures are below 20%, except in those pairs having French as input language. This is explained because the French ASR system does not feature BLSTM acoustic models, and it is the language with least acoustic resources.

3.2. MT

A Neural Machine Translation (NMT) system was built for each translation direction mainly using publicly available corpora from OPUS [25] and excluding the Europarl corpus to avoid data overlapping. The training data used in each language pair is shown in Table 5. This includes the list of corpora and the total number of sentences.

The corpora were preprocessed by applying 40K BPE [26] operations, learnt jointly over the source and target data. The

Table 5. Training data used for the MT systems

Pair	Corpora	# sents(M)
De↔En	DGT,eubookshop TildeMODEl, Wikipedia	21.0
De↔Es	DGT, eubookshop, JRC-Acquis, TildeModel	14.3
De↔Fr	eubookshop, JRC-Acquis, TildeModel	14.3
En↔Es	commoncrawl, eubookshop, EU-TT2, UN, Wikipedia	21.1
En↔Fr	commoncrawl, giga, undoc, news-commentary	38.2
Es↔Fr	DGT, eubookshop, JRC-Acquis, UNPC	37.2

Table 6. BLEU scores of out-of-domain MT systems with reference transcriptions as input and fine-tuning BLEU scores between parenthesis.

	De	En	Es	Fr
De	–	32.6 (**36.3**)	26.8 (**29.3**)	23.2 (**27.1**)
En	33.6 (**37.6**)	–	46.3 (**48.2**)	34.7 (**39.2**)
Es	20.9 (**24.8**)	39.2 (**41.8**)	–	29.3 (**33.1**)
Fr	23.3 (**26.3**)	38.7 (**42.3**)	34.8 (**36.3**)	–

models follow the Transformer NMT architecture [27] and are trained using the Transformer BASE configuration using 4GPU machines and an initial learning rate of $5\mathrm{e}{-4}$, decayed using the inverse square root scheme. Once the training converges, a fine-tuning step was carried out using the training data generated in Section 2. To do so, we fix the learning rate to $5\mathrm{e}{-5}$, and we use a standard SGD optimizer instead of Adam. We measure performance on the dev set and stop training once the perplexity stops decreasing. Table 6 shows BLEU scores of the out-of-domain MT systems compared with those obtained by fine-tuning with the Europarl-ST training data shown between parenthesis. These MT systems are evaluated on automatic outputs generated from reference transcriptions as a standalone MT task.

The results vary depending on the amount of resources used for each system as well as the intrinsic difficulty of each translation direction. As observed, the fine-tuned systems trained on the Europarl-ST corpus provide very significant improvements over the out-of-domain systems, ranging from +1.9 up to +4.0 BLEU, which confirms the quality and usefulness of the training data.

3.3. SLT

Table 7. BLEU scores of cascade-based SLT experiments with fine-tuned models assessed on the test sets.

	De	En	Es	Fr
De	–	21.3	17.5	15.7
En	22.4	–	28.0	23.4
Es	15.6	26.5	–	22.0
Fr	15.3	25.4	23.2	–

This section presents the results of the SLT experiments following the cascade approach, in which we use the output of the ASR system as input for the MT system. The output of the ASR system is segmented based on detected silences. For this task, we will combine the ASR and MT models described in Sections 3.1 and 3.2. We use the fine-tuned MT systems as they outperform the out-of-domain systems in all cases. The results of the SLT experiments are shown in Table 7.

Table 7 shows that BLEU scores in the SLT experiments are lower than those in the MT experiments. This is to be expected, as the MT system has to cope not only with error propagation from incorrect transcriptions, but also with a sub-optimal segmentation of the input which might not correspond with whole sentences. This could be improved with a specific segmentation and punctuation module [2]. As expected, although the overall BLEU scores are lower, the ranking of the performance across translation directions is preserved, with MT systems that obtained the highest scores in the MT experiments, also obtaining the highest scores in the SLT experiments, and vice versa. Although SLT results are constrained by the complexity of this task, these results serve as a good starting baseline for future developments.

4. CONCLUSIONS

We have presented a novel SLT corpus built from European Parliament proceedings. The experiments presented have shown how our proposed filtering pipeline is able to extract good quality data that is useful both for evaluating the performance of out-of-domain systems in this task, as well as for system adaptation to the specific domain of parliamentary debates. We believe that the release of this multi-source and multi-target corpus will enable further research into multilingual SLT.

In terms of future work, the presented filtering pipeline can be extended to cover additional languages in the future. Additionally, we will study new filtering techniques to increase the amount of hours available per each language pair.

Finally, we also plan on gauging the performance of end-to-end models for this task, and compare it with cascade systems that use MT models adapted to the translation of ASR output. This adaptation can be carried out by training MT systems on real ASR output as source input [28] or on simulated ASR output by applying noising techniques to the source side [1].

5. REFERENCES

[1] Matthias Sperber, Jan Niehues, and Alex Waibel, "Toward robust neural machine translation for noisy input

sequences," in *IWSLT 2017*.

[2] Eunah Cho, Jan Niehues, and Alex Waibel, "NMT-Based Segmentation and Punctuation Insertion for Real-Time Spoken Language Translation," in *Interspeech 2017*.

[3] E. Matusov, P. Wilken, P. Bahar, J. Schamper, P. Golik, A. Zeyer, J.A. Silvestre-Cerdà, A. Martínez-Villaronga, H. Pesch, and J. Peter, "Neural Speech Translation at AppTek," in *IWSLT 2018*.

[4] Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen, "Sequence-to-sequence models can directly translate foreign speech," in *Interspeech 2017*.

[5] Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel, "Attention-passing models for robust and data-efficient end-to-end speech translation," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 313–325, Mar. 2019.

[6] Elizabeth Salesky, Matthias Sperber, and Alan W Black, "Exploring phoneme-level speech representations for end-to-end speech translation," in *ACL 2019*.

[7] Jan Niehues, Roldano Cattoni, Sebastia Stker, Mauro Cettolo, Marco Turchi, and Marcello Federico, "The IWSLT 2018 Evaluation Campaign," in *IWSLT 2018*.

[8] Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu, "Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model," in *Interspeech 2019*.

[9] Mattia Antonino Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi, "MuST-C: a Multilingual Speech Translation Corpus," in *NAACL-HLT 2019*.

[10] Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur, "Improved speech-to-text translation with the Fisher and Callhome Spanish–English speech translation corpus," in *IWSLT 2013*.

[11] Marcely Zanon Boito, William N. Havard, Mahault Garnerin, Eric Le Ferrand, and Laurent Besacier, "Mass: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the bible," in *LREC 2020 (accepted)*.

[12] Astrid Van Aggelen, Laura Hollink, Max Kemman, Martijn Kleppe, and Henri Beunders, "The debates of the European Parliament as linked open data," *Semantic Web*, vol. 8, no. 2, pp. 271–281, 2017.

[13] Mickael Rouvier, Grégor Dupuy, Paul Gay, Elie el Khoury, Téva Merlin, and Sylvain Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," in *Interspeech 2013*.

[14] Miguel A. del Agua, Adrià Giménez, Nicolás Serrano, Jesús Andrés-Ferrer, Jorge Civera, Alberto Sanchís, and Alfons Juan, "The Translectures-UPV Toolkit," in *IberSpeech 2014*.

[15] Philipp Koehn et al., "Moses: Open source toolkit for statistical machine translation," in *ACL 2007*.

[16] Fabienne Braune and Alexander M. Fraser, "Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora," in *COLING 2010*.

[17] Chris Dyer, Victor Chahuneau, and Noah A. Smith, "A simple, fast, and effective reparameterization of IBM model 2," in *NAACL-HLT 2013*.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *ACL 2002*.

[19] Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney, "Evaluating machine translation output with automatic sentence segmentation," in *IWSLT 2005*.

[20] Matt Post, "A call for clarity in reporting BLEU scores," in *WMT18*.

[21] "Tensorflow," `https://www.tensorflow.org/`.

[22] A. Stolcke, "SRILM – an extensible language modeling toolkit," Denver, CO, USA, Sept. 2002, pp. 901–904.

[23] "The RNNLM Toolkit," `http://www.fit.vutbr.cz/~imikolov/rnnlm/`.

[24] Xi Chen, Xin Liu, Y. Qian, Mark J. F. Gales, and Philip C. Woodland, "CUED-RNNLM An open-source toolkit for efficient training and evaluation of recurrent neural network language models," in *ICASSP 2016*.

[25] Jörg Tiedemann, "Parallel data, tools and interfaces in OPUS," in *LREC 2012*.

[26] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Neural machine translation of rare words with subword units," in *ACL 2016*.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS 2017*.

[28] Stephan Peitz, Simon Wiesler, Markus Nußbaum-Thom, and Hermann Ney, "Spoken language translation using automatically transcribed text in training," in *IWSLT 2012*.