IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

mintzai-ST: Corpus and Baselines for Basque-Spanish Speech Translation

Thierry Etchegoyhen, Haritz Arzelus, Harritxu Gete Ugarte,
Aitor Alvarez, Ander González-Docasal, Edson Benites Fernandez

Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
Mikeletegi Pasalekua, 57, Donostia/San Sebastián
{tetchegoyhen,harzelus,hgete,aalvarez,agonzalezd,eebenites}@vicomtech.org

Abstract

The lack of resources to train end-to-end Speech Translation models hinders research and development in the field. Although recent efforts have been made to prepare additional corpora suitable for the task, few resources are currently available and for a limited number of language pairs. In this work, we describe mintzai-ST, a parallel speech-text corpus for Basque-Spanish in both translation directions, prepared from the sessions of the Basque Parliament and shared for research purposes. This language pair features challenging phenomena for automated speech translation, such as marked differences in morphology and word order, and the mintzai-ST corpus may thus serve as a valuable resource to measure progress in the field. We also describe and evaluate several ST model variants, including cascaded neural components, for speech recognition, machine translation, and end-to-end speech-to-text translation. The evaluation results demonstrate the usefulness of the shared corpus as an additional ST resource and contribute to determining the respective benefits and limitations of current alternative approaches to Speech Translation.
Index Terms: Speech Translation, Basque, Spanish, Corpus

1. Introduction

Deep Learning approaches to natural language processing have achieved significant results in the past years, notably in the fields of machine translation [1, 2] and speech recognition [3, 4]. The possibility to train end-to-end models, in particular, has proved fruitful, building on the ability of artificial neural networks to jointly learn different aspects of a task under a data-driven approach.

Speech translation (ST) has been traditionally modelled under a cascaded approach, chaining automated speech recognition (ASR) and machine translation (MT) systems, with task-specific to optimise information communication between components [5, 6, 7]. Although the dominant approach to the task, systems relying on cascaded components are prone to error propagation and end-to-end neural approaches to ST have been explored in recent years as an alternative [8, 9, 10]. While preliminary results with end-to-end systems have shown promises, cascaded systems still obtain better results overall on standard evaluation datasets [11]. One of the main factors for this state of affairs is the scarcity of parallel speech-text and speech-speech corpora, in contrast with the comparatively larger amounts of data available to train separate ASR and MT models, for some language pairs at least.

The paucity of resources to train end-to-end ST systems has led to recent efforts in developing additional parallel corpora suitable for the task, notably the multilingual MuST-C [12] and Europarl-ST [13] corpora. For some language pairs, however, no parallel resources are readily available, as is the case for Basque-Spanish, for instance.[1] This is particularly limiting considering that languages such as Basque exhibit a number of marked linguistic properties, notably rich morphology and relatively free word order, which can represent a challenge for natural language processing tasks in general, and automated translation in particular [15, 14].

In this work, we describe mintzai-ST, a Basque-Spanish parallel speech-text corpus based on the plenary sessions of the Basque Parliament and shared with the scientific community. We evaluate its usefulness for speech translation, by training and evaluating state-of-the art models under both cascaded and end-to-end approaches. To our knowledge, these are the first comprehensive results in terms of speech translation for this language pair, made possible by the production of specific datasets for the task. With approximately 480 hours of audio and 3.3M target words for Spanish to Basque, 190 hours and 1.5M words for Basque to Spanish, the mintzai-ST corpus enables end-to-end training and comparisons with cascaded models, thus providing a basis for further research on alternative approaches to speech translation in a relatively difficult language pair.

The remainder of this paper is organised as follows: Section 2 describes the data acquisition process and statistics of the corpus; in Section 3 we describe the different models that were built for Basque-Spanish speech translation, including cascaded and end-to-end models; Section 4 discusses the results of the comparative model evaluation on the mintzai-ST datasets; finally, Section 5 draws the main conclusions from this work.

2. Corpus description

The corpus was created from the proceedings of the Basque Parliament from 2011 to 2018, a period of time where audio, transcriptions and translations were publicly available. Speakers expressed themselves in either Basque or Spanish, with a majority of interventions in the latter overall; professional transcriptions and translations were then produced into Spanish and Basque, respectively. We describe below the main processes related to data acquisition and preparation of the final corpus.

2.1. Data acquisition

Raw data were first obtained by crawling the web sites where the official plenary sessions are made available: transcriptions and translations were available at http://www.legebiltzarra.eus; videos of the sessions at: https://www.irekia.euskadi.eus.

Texts from the sessions were available as bilingual PDF files, with content in each language provided in a dedicated col-

[1] The only potential speech translation resource we are aware of for this language pair is the EuskoParl corpus [14], which is based on similar sources as mintzai-ST with the main differences that it is not a parallel speech-text corpus and, to our knowledge, is not publicly available.

10.21437/IberSPEECH.2021-41

umn: one for the transcription of the session and the other for its translation. The content was extracted from the PDF files with PDFtoText[2] and boilerplate removal was performed with in-house content-specific scripts. Since the translations were made at the paragraph level for the most part [14], paragraph-level information was maintained.

Videos were provided in different formats over the years (.flv, .webm and .mp4), and audio extraction was performed with FFmpeg[3]. The mapping between videos and reports was performed via inferences from the respective files metadata whenever possible. In most cases, the available information was not sufficient to map video and PDF files with absolute confidence, with multiple and sometimes duplicate videos in many cases[4]; manual revision and mapping were therefore performed throughout this task.

The statistics for the collected raw data are shown in Table 1.

Table 1: *mintzai-ST: raw data statistics*

YEAR	VIDEOS	PDF	HOURS	WORDS
2011	43	21	86.51	132,595
2012	38	21	117.94	173,199
2013	67	38	215.00	306,621
2014	60	30	176.83	252,887
2015	41	27	134.10	195,112
2016	38	21	113.85	170,608
2017	49	33	173.57	250,862
2018	34	26	128.38	207,910
TOTAL	370	217	1,146.18	18,625,252

2.2. Alignment and filtering

As a first step, metadata were filtered from the PDF-extracted text, and source and target files were extracted from the text in the original columns, preserving paragraph-level alignments. Speaker information was usually located at the beginning of a paragraph and was extracted when available.[5]

As a second step, language identification was performed on each paragraph. Since any error at this stage would propagate to subsequent processes, special attention was paid to ensuring correct language identification by employing two separate tools on the content: TextCat and the language identifier of the OpenNER project [16].[6] Paragraphs were discarded if either tool produced different results as their topmost identified language, or if neither tool identified either one of the expected languages; in all other cases, we retained the identified language, by either one or both of the tools.

The third step involved forced alignment, where each word in the source transcription was aligned to a section of the corresponding audio file via source and time indications. This step was performed with the Kaldi toolkit [17], using a bilingual model to reduce the impact of remaining language identification uncertainties. Alignment was performed with different beam sizes (10, 100, 1000 and 10000) and all content was aligned.

At this stage, the source and target files were split on the basis of the previous alignment information, with one paragraph per file. Forced alignment was then applied again, this time with a monolingual model and a small beam size of 1, with a retry beam of 2, to discard alignment errors and non-literal transcriptions.

Since translation models require specific sentence-based training bitexts, the previously aligned paragraphs were further prepared with sentence splitting, tokenisation and truecasing. All operations were performed with the scripts from the Moses toolkit [18]. Sentence-level alignments were then computed with the Hunalign toolkit [19], with an alignment probability of 0.4.

The filtered and aligned data were then randomly split into train, dev and test subsets of triplets consisting of audio, transcription and translation. Triplets were removed from the test sets if the transcription-translation pair appeared in the training set as well. This measure was adopted to account for the fact that even minor acoustic differences might make a triple differ from another, even though the transcription-translation pair would be a duplicate for the machine translation component. Stricter removal along the previous lines allowed for a fair comparison between cascaded and end-to-end models, and made for a more difficult test set as it mainly discarded acoustic variants of greetings and salutations.

The statistics for the mintzai-ST corpus, for Basque to Spanish and Spanish to Basque, are shown in Table 2. The corpus is shared under the Creative Commons CC BY-NC-ND 4.0 license and is available at the following address: https://github.com/vicomtech.[7]

3. Speech translation models

In this section, we describe the different models and components built to measure the usefulness of the prepared corpus as a basis for speech translation, on the one hand, and to identify the relative benefits and limitations of the different modelling alternatives on the other hand. Two main approaches are described in the next section: cascaded models, based on state-of-the art components for speech recognition and machine translation, and end-to-end neural speech translation models.

3.1. Cascaded models

The speech-to-text cascaded models are based on separate components for speech recognition and machine translation, each trained on their own datasets, either on the in-domain mintzai-ST corpus only or on a combination of the corpus with additional data. We describe each component in turn below.

For the additional dataset, we favoured publicly available corpora close to the mintzai-ST domain which would allow for a straightforward reproduction of our results. For this language pair, only text-based datasets were available with these characteristics and we selected the OpenDataEuskadi corpus [15], prepared from public translation memories by the translation services of the Basque public administration.[8] This corpus is close enough to the mintzai-ST domain to be meaningfully combined and large enough to contribute significantly to differ-

[2]This specific extraction tool was selected as it preserved column-based alignments.

[3]https://www.ffmpeg.org/

[4]For each session, there were between 1 and 7 videos, and 1 and 2 PDF files.

[5]182 speakers were identified overall.

[6]Available at the following addresses, respectively: https://github.com/Trey314159/TextCat and https://github.com/opener-project/language-identifier.

[7]The providers of the original content have granted permission for its use without additional restrictions.

[8]The corpus is available at the following address: http://hltshare.fbk.eu/IWSLT2018/OpendataBasqueSpanish.tgz

Table 2: *mintzai-ST: final corpus statistics*

SRC	TGT	PARTITION	HOURS	SENTENCES	SRC WORDS	TGT WORDS
ES	EU	TRAIN	468.16	175,826	4,512,294	3,328,172
EU	ES	TRAIN	180.96	85,409	1,149,803	1,536,695
ES	EU	DEV	2.60	1,000	25,359	18,566
EU	ES	DEV	2.23	1,000	13,831	18,673
ES	EU	TEST	7.89	2,788	74,758	55,283
EU	ES	TEST	6.35	2,300	37,706	51,003

ent components of the cascaded models. The corpus amounts to 963,391 parallel sentences, with 23,413,116 words in Spanish and 17,802,212 in Basque.

To connect the components, the best hypothesis of the ASR model was fed to the MT model, after generating punctuation as described in Section 3.1.1. Although considering alternative hypotheses in the n-best ASR output might provide additional robustness and accuracy to the overall system, we left an evaluation along these lines for future research.

3.1.1. Speech recognition

Two speech recognition architectures, based on end-to-end models and Kaldi based systems, were trained and evaluated to test their performance on the new compiled corpus.

The end-to-end speech recognition systems were based on the Deep Speech 2 architecture [20] for both languages, and were set up with a sequence of 2 layers of 2D convolutional neural networks followed by 5 layers of bidirectional GRU layers and a fully-connected final layer. The output corresponds to a *softmax* function which computes a probability distribution over characters. Stochastic Gradient Descent (SGD) was used as optimiser and the input consisted of Mel-scale based spectrograms, which were dynamically augmented through the SpecAugment technique [21]. The models were estimated using only audios lasting less than 40 seconds, due to training memory constraints, with a learning rate of 0.0001 annealed by a constant factor of 1.08 for a total of 60 training epochs.

The Kaldi recognition systems were built with the *nnet3* DNN setup, and using the so-called *chain* acoustic model based on a factorised time-delay neural network (TDNN-F) [22], which reduces the number of parameters of the network by factorising the weight matrix of each TDNN layer into the product of two low-rank matrices. Our TDNN-F models consisted of 16 TDNN-F layers with an internal cell-dimension of 1536, a bottleneck-dimension of 160 and a dropout schedule of '0,0@0.2,0.5@0.5,0'. The number of training epochs was set to 4, with a learning rate of 0.00015 and a minibatch size of 64. The input vector corresponded to a concatenation of 40 dimensional high-resolution MFCC coefficients, augmented through speed (using factors of 0.9, 1.0, and 1.1) [23] and volume (with a random factor between 0.125 and 2) [24] perturbation techniques, and the appended 100 dimensional iVectors.

Language models were 5-gram models with modified Kneser-Ney smoothing, estimated with the KenLM toolkit [25], and were used as either a component of Kaldi-based systems or to rescore the end-to-end models' hypotheses.

Finally, the capitalisation models were trained with the recasing tools provided by the Moses open-source toolkit [18], while the punctuation module consisted of a bidirectional RNN model built with out-of-domain data from the broadcast news domain and using the Punctuator2 toolkit [26].

3.1.2. Machine translation

All machine translation models in the experiments reported below were based on the Transformer architecture [2], built with the MarianNMT toolkit [27].

The models consisted of 6-layer encoders and decoders and 8 attention heads. The Adam optimiser was used with parameters $\alpha = 0.0003$, $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. The learning rate was set to increase linearly for the first 16,000 training steps and decrease afterwards proportionally to the inverse square root of the corresponding step. The working memory was set to 8000MB and the largest mini-batch was automatically chosen for a given sentence length that fit the specified memory. The validation data were evaluated every 5,000 steps for models trained on larger out-of-domain datasets and every epoch otherwise; training ended if there was no improvement in perplexity after 5 consecutive checkpoints. Embeddings were of dimension 512, tied between source and target, and all datasets were segmented with BPE [28], using 30,000 operations.

3.2. End-to-end models

End-to-end ST models were trained on the in-domain speech-text corpus, using the Fairseq-ST toolkit[9], which supports different types of sequence-to-sequence neural models [29]. The variant selected for the experiments is the Transformer model enhanced for ST described in [30], more specifically the variant the authors refer to as the S-Transformer.

The model follows the standard Transformer architecture with 6 layers self-attentional encoder and decoder, but adds layers prior to the Transformer encoder, to model 2D dependencies. The audio input is provided to the model in the form of sequences of MEL filters, encoded first by two CNNs to model 2D-invariant features, followed by two 2D self-attention layers to model long-range context. The output of the stacked 2D self-attention layers undergoes a linear transformation, followed by a ReLU non-linearity, and is summed with the positional encoding, prior to feeding the Transformer encoder.

We diverged from the implementation described in [30] on one important aspect. Character-based decoding was replaced with subword decoding, using the previously described BPE models, as the former faced consistent issues, resulting in subpar performance; an identical setup with subwords produced significantly better results overall. Further exploration of these differences between translation pairs is left for future research.

4. Comparative evaluation

We first performed an evaluation centred on cascaded models, where a number of variants could be prepared based on different ASR approaches or different types and volumes of training data.

[9] https://github.com/mattiadg/FBK-Fairseq-S

The variants include: ASR models trained with either an end-to-end neural model (E2E) or the Kaldi toolkit (KAL); ASR and MT models trained on either in-domain data only (IND) or on a combination of in-domain and out-of-domain data (ALL), by integrating the OpenDataEuskadi dataset to train the language and casing models for speech recognition, and the translation models for the MT component; MT models obtained by fine-tuning (FT) models trained on the out-of-domain dataset with the in-domain data.

The results for the cascaded variants on the mintzai-ST test sets, in terms of word error rate (WER) and BLEU [31], are shown in Table 3. All results in the table were computed with ASR output that includes punctuation, generated with the previously mentioned punctuation models. To measure the impact of punctuation on the overall process, differences between BLEU scores obtained with and without punctuation, in that order, are also shown in the table (ΔPUNC).

Table 3: *Evaluation results on cascaded variants*

LANG	ASR	MT	WER	BLEU	ΔPUNC
EU-ES	E2E IND	IND	14.43	28.4	+1.0
EU-ES	E2E ALL	IND	14.12	28.4	+0.8
EU-ES	E2E IND	ALL	14.43	33.3	+2.4
EU-ES	E2E ALL	ALL	14.12	33.4	+2.3
EU-ES	E2E IND	FT	14.43	33.3	+2.4
EU-ES	E2E ALL	FT	14.12	33.4	+2.6
EU-ES	KAL IND	IND	12.07	29.2	+0.9
EU-ES	KAL ALL	IND	11.78	29.4	+1.1
EU-ES	KAL IND	ALL	12.07	34.7	+2.6
EU-ES	KAL ALL	ALL	**11.78**	**34.7**	+2.7
EU-ES	KAL IND	FT	12.07	33.7	+2.5
EU-ES	KAL ALL	FT	11.78	33.9	+2.6
ES-EU	E2E IND	IND	8.26	20.6	+1.3
ES-EU	E2E ALL	IND	8.15	20.6	+1.3
ES-EU	E2E IND	ALL	8.26	22.0	+1.0
ES-EU	E2E ALL	ALL	8.15	22.0	+1.1
ES-EU	E2E IND	FT	8.26	21.5	+1.3
ES-EU	E2E ALL	FT	8.15	21.5	+1.2
ES-EU	KAL IND	IND	7.23	20.9	+1.4
ES-EU	KAL ALL	IND	7.21	20.9	+1.3
ES-EU	KAL IND	ALL	7.23	22.5	+1.2
ES-EU	KAL ALL	ALL	**7.21**	**22.7**	+1.5
ES-EU	KAL IND	FT	7.23	21.9	+1.2
ES-EU	KAL ALL	FT	7.21	22.0	+1.4

Overall, cascaded models trained on all data performed significantly better than their in-domain counterparts, with improvements of up to 5 and 1.6 BLEU points for EU-ES and ES-EU, respectively. These results were mostly due to improvements obtained on the MT components, as was expected from adding significantly larger amounts of training data to the small in-domain datasets. For the ASR components, the impact in terms of WER was minor, with around .3 gains in either language, mainly due to the use of the same data for acoustic modelling in all cases.

Punctuation had a significant impact on the results, with systematic improvements of up to 2.6 and 1.5 BLEU points in EU-ES and ES-EU, respectively. This trend is not entirely surprising, since the translation models were trained on data that include punctuation marks; the impact of punctuation was amplified for models trained on larger amounts of data.

Regarding the overall translation quality, as measured in terms of BLEU scores at least, the results are in line or higher than typical results in similar tasks [11]. One explanation for higher marks is the domain specificity of the corpus, with recurrent topics and typical expressions. Nonetheless, the corpus also features challenging characteristics for automated speech translation, such as the use of Basque dialects or the idiosyncratic properties of the two languages at hand.

From the previous evaluation, we selected the best cascaded variants based on either in-domain or all data and compared with the end-to-end speech translation models, in both translation directions. The comparative results on the mintzai-ST test sets are shown in Table 4, where BP indicates the brevity penalty computed within the BLEU metric.

Table 4: *Results on cascaded and end-to-end models*

LANG	MODEL	ASR	MT	WER	BLEU	BP
EU-ES	CAS	IND	IND	12.07	29.2	0.913
EU-ES	CAS	ALL	ALL	**11.78**	**34.7**	0.978
EU-ES	E2E	-	-	-	17.0	1.000
ES-EU	CAS	IND	IND	7.23	20.9	0.954
ES-EU	CAS	ALL	ALL	**7.21**	**22.7**	0.969
ES-EU	E2E	-	-	-	12.9	1.000

The most notable result from this evaluation is the large difference in terms of BLEU between the cascaded and the end-to-end variants under similar conditions, i.e. using only the in-domain data. Under these conditions, the end-to-end variant was outperformed by 12.2 and 8 BLEU points in EU-ES and ES-EU, respectively. Since the conditions were similar, with relatively small amounts of training data, this large gap may be attributed to the relative dependency of the end-to-end model on larger volumes of training data. Given the noticeably better results obtained with cascaded end-to-end components, alternative end-to-end ST approaches that increase in terms of robustness in low-resource scenarios will need to be further explored.

Interestingly, the end-to-end model produces translation which are closer in length to the human references, as shown by results in terms of brevity penalty. Although further analyses of these aspects will be warranted, these results indicate that the end-to-end systems built for these experiments may be modelling aspects of the speech translation process which are not fully captured by their cascaded counterparts.

5. Conclusions

We described mintzai-ST, the first publicly available corpus for speech translation in Basque-Spanish, shared with the community to support research in the field. The corpus enables development and evaluation of end-to-end or cascaded ST models, and we presented the first results along these lines for this challenging language pair.

In future work, we will prepare variants of the corpus with varying alignment constraints, to measure the impact of larger amounts of training data, in particular for end-to-end models. We will also carry additional analyses of the relative strengths and weaknesses of different ST architectures, building upon the preliminary results described in this paper. Finally, we will explore speech-speech variants of this corpus, by means of speech synthesis, to provide further support to research on speech-to-speech translation.

6. Ackowledgments

This work was supported by the Department of Economic Development and Competitiveness of the Basque Government under project MINTZAI (KK-2019/00065). We wish to thank the anonymous IberSpeech reviewers for their insightful feedback.

7. References

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the International Conference on Learning Representations*, 2015.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 6000–6010.

[3] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of the IEEE international conference on acoustics, speech and signal processing*, 2013, pp. 6645–6649.

[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 4960–4964.

[5] H. Ney, "Speech translation: Coupling of recognition and translation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999, pp. 517–520.

[6] E. Matusov, S. Kanthak, and H. Ney, "On the integration of speech recognition and statistical machine translation," in *Proceedings of the Ninth European Conference on Speech Communication and Technology*, 2005, pp. 3176–3179.

[7] G. Kumar, G. Blackwood, J. Trmal, D. Povey, and S. Khudanpur, "A coarse-grained model for optimal coupling of ASR and SMT systems for speech translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1902–1907.

[8] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 949–959.

[9] A. Bérard, O. Pietquin, L. Besacier, and C. Servan, "Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation," in *Proceedings of NIPS Workshop on end-to-end learning for speech and audio processing*, 2016.

[10] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," in *Proceedings of INTERSPEECH*, 2017, pp. 2625–2629.

[11] J. Niehues, R. Cattoni, S. Stüker, M. Negri, M. Turchi, E. Salesky, R. Sanabria, L. Barrault, L. Specia, and M. Federico, "The IWSLT 2019 Evaluation Campaign," in *Proceedings of the 16th International Workshop on Spoken Language Translation*, 2019.

[12] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: a Multilingual Speech Translation Corpus," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 2012–2017.

[13] J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan, "Europarl-st: A multilingual corpus for speech translation of parliamentary debates," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 8229–8233.

[14] A. Pérez, J. M. Alcaide, and M.-I. Torres, "Euskoparl: a speech and text spanish-basque parallel corpus," in *Proceedings of INTERSPEECH*, 2012, pp. 2362–2365.

[15] T. Etchegoyhen, E. Martínez Garcia, A. Azpeitia, G. Labaka, I. Alegria, I. Cortes Etxabe, A. Jauregi Carrera, I. Ellakuria Santos, M. Martin, and E. Calonge, "Neural Machine Translation of Basque," in *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, 2018, pp. 139–148.

[16] R. Agerri, M. Cuadros, S. Gaines, and G. Rigau, "OpeNER: Open polarity enhanced named entity recognition," *Procesamiento del Lenguaje Natural*, no. 51, pp. 215–218, 2013.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *Proceedings of IEEE Workshop on automatic speech recognition and understanding*, 2011.

[18] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 2007, pp. 177–180.

[19] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy, "Parallel corpora for medium density languages," in *Proceedings of Recent Advances in Natural Language Processing*, 2005, pp. 590–596.

[20] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proceedings of the International Conference on Machine Learning*, 2016, pp. 173–182.

[21] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of INTERSPEECH*, 2019, pp. 2613–2617.

[22] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Proceedings of INTERSPEECH*, 2018, pp. 3743–3747.

[23] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proceedings of INTERSPEECH*, 2015, pp. 3586–3589.

[24] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proceedings of INTERSPEECH*, 2015.

[25] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 2011, pp. 187–197.

[26] O. Tilk and T. Alumäe, "Bidirectional recurrent neural network with attention mechanism for punctuation restoration," in *Proceedings of INTERSPEECH*, 2016.

[27] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast neural machine translation in C++," in *Proceedings of ACL 2018, System Demonstrations*, 2018, pp. 116–121.

[28] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 1715–1725.

[29] M. A. Di Gangi, M. Negri, and M. Turchi, "Adapting transformer to end-to-end spoken language translation," *Proceedings of INTERSPEECH*, pp. 1133–1137, 2019.

[30] M. A. Di Gangi, M. Negri, R. Cattoni, R. Dessi, and M. Turchi, "Enhancing transformer for end-to-end speech-to-text translation," in *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, Aug. 2019, pp. 21–31.

[31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.