End-to-End Speech Translation: a Survey

Bhavani Shankar, Preethi Jyothi, Pushpak Bhattacharyya
Indian Institute of Technology, Bombay
{bhavanishankar, pjyothi, pb}@cse.iitb.ac.in

Abstract

Speech-to-text translation, the process of converting speech in one language to text in another, has numerous applications. SOme of them include hands-free communication, dictation, video lecture transcription, and general translation. Traditional approaches to speech translation rely on a two-step process: Automatic Speech Recognition (ASR) followed by Machine Translation (MT). ASR converts spoken language into written text which is then translated by the MT model. While effective, this cascaded approach suffers from error propagation and high resource requirements. Recognizing these limitations, researchers have focused on developing end-to-end (E2E) speech translation models. These models directly translate speech into another language without relying on intermediate transcription, promising improved accuracy and efficiency.This survey provides a comprehensive overview of the models, metrics, and datasets used in E2E speech translation research.

1 Introduction

Speech-to-text (ST) translation aims to of convert speech in one language to text in another language. It finds widespread application in areas such as automatic subtitling, dictation, video lecture translation, tourism, and telephone conversations. This task can be approached in various ways, including online (simultaneous) or offline translation, depending on the specific application. The ST problem becomes even more complex when dealing with noisy inputs, low-resource or code-mixed languages, and the presence of multiple speakers. Traditional ST translation methods use cascaded approach: automatic speech recognition (ASR) to transcribe the speech followed by machine translation (MT) to translate the resulting text. However, this cascade approach suffers from several drawbacks, including error propagation from ASR to MT, increased cost and training time, and a high resource requirement. To overcome these limitations, researchers have explored end-to-end (E2E) models for ST (Berard et al., 2016; Weiss et al., 2017; Dong et al., 2018; Kano et al., 2017; Berard et al., 2018; Inaguma et al., 2020; Wang et al., 2020c; Zhao et al., 2021). These models use a single neural network trained end-to-end, offering simpler training, lower memory footprint, and reduced cost. This survey offers a comprehensive analysis of existing models, datasets, and approaches for ST translation. The survey is organized into several sections:

- Section 2: Establishes the foundation of the ST task through a formal definition and explores the metrics and loss functions used by researchers.
- Section 3: Provides descriptive statistics of benchmark datasets commonly used in ST research.
- Section 4: Dives deep into the various strategies and approaches employed for ST translation, categorizing them based on frameworks (Sequence-to-sequence and Modality-bridging) and data characteristics (Low-resource, Code-mixed, Streaming, Unsupervised, and Multilingual).

2 Background

This section formally defines the ST task and presents the metrics used to evaluate the performance of ST models.

2.1 Task Definition

The ST task aims to translate an input speech signal, denoted as U, in one language to a translated text, denoted as V, in another language, with the transcription text X. Formally, given a dataset $D = (u_i, x_i, v_i)|i = 1, 2, ..., n$ of pairs of input speech features $u = (u_1, u_2, ..., u_{T_u})$ in one language and output text tokens $v = (v_1, v_2, ..., v_{T_v})$

in another language, the objective of the ST task is to minimize the conditional probability:

$$p(v|u;\theta) = \prod_{t=1}^{T_v} p(v_t|v_{<t}, u; \theta)$$

Here, T_u, T_v, and θ represent the lengths of input features, the number of output tokens, and the model parameters, respectively. The model is optimized for negative log-likelihood over the n parallel sentences in the corpus:

$$l(\theta|D) = -\sum_{i=1}^{n} logP(v_i|u_i;\theta)$$

This optimization is typically solved using an encoder-decoder architecture with an attention mechanism. The encoder maps the speech input to a hidden state representation h, which is then processed by the decoder. The decoder utilizes previously generated text tokens $v_{<t}$, the encoder hidden state h, and an attention vector α (Vaswani et al., 2017). Offline ST translation can process the entire speech signal before producing output text tokens, while online ST begins translating as soon as it receives a few seconds of the speech signal.

2.2 Evaluation Metrics

This section explores various metrics used to evaluate E2E ST models. These metrics are categorized into two primary performance measures: quality and latency.

2.2.1 Quality-based Metrics

Quality metrics measure the closeness of the translation to the target sentence, usually within the range [0, 1]. Most existing literature evaluates these scores on detokenized output.

BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) calculates the weighted average length of the translated sentence that matches the target sentence. The BLEU score is expressed as:

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n logP_n)$$

where BP, w_n, and P_n represent the brevity penalty, positive weights (summing to 1), and n-gram precisions (up to length N), respectively.

TER (Translation Edit Rate) (Snover et al., 2006) measures the amount of modification required to make the translated sentence match the target sentence. The TER is calculated as:

$$TER = \frac{M}{R}$$

where M represents the modifications needed in the translated sentence and R is the average number of reference words.

METEOR (Metric for Evaluation of Translation with Explicit word Ordering) (Lavie and Agarwal, 2007) is an automatic metric that matches unigrams in the translated text to the referenced text. The METEOR score is calculated as:

$$M_Score = FM \cdot (1 - PenaltyM)$$

where $FM = 10\frac{PR}{R+9P}$ and $PenaltyM = 0.5 \cdot \frac{C}{U}$. FM and $PenaltyM$ represent the F-score and the penalty of the METEOR score, respectively. P, R, C, and U denote precision, recall, the number of chunks, and the number of unigrams matched in the translated text, respectively.

BERTScore (Zhang et al., 2020) is another automatic evaluation metric that assesses the similarity between the translated text and the referenced text. It considers Recall (R), Precision (P), and F-score (F), expressed as:

$$R = \frac{1}{|v|}\sum_{v\in v}\sum_{\hat{vj}\in\hat{v}}\frac{1}{|\hat{vj}|},$$

$$P = \frac{1}{|\hat{vi}|}\sum \hat{v} \in \hat{v}\sum_{v_i\in v}\frac{1}{|v|},$$

$$F = \frac{2\cdot R\cdot P}{R+P}$$

where v and $\hat{v}$ represent the tokens of the sentences in the translated text and the ground truth text, respectively.

2.2.2 Latency-based Metrics

For streaming ST tasks, researchers report metrics for measuring latency, which is defined as the delay incurred in starting to produce the translation. Let u, v, and $\hat{v}$ denote the input speech sequence, ground truth text sequence, and system-generated hypothesis sequence, respectively. In streaming ST, the model can only generate output with partial input. If $u_{1:t} = (u_1, ..., u_t)$, $t < T_u$, has been read when generating v_s, the delay in v_s is defined as (Ma et al., 2020):

$$d_s = \sum_{k=1}^{t} T_k$$

where T_k represents the duration of the speech frame u_k. Latency metrics are evaluated using a method that analyzes a sequence of time delays $[d_1, ..., d_{T_v}]$.

Average Proportion (AP) calculates the mean fraction of the source input read during the target prediction generation process:

$$AP = \frac{1}{T_v} \sum_{s=1}^{T_v} \frac{d_s}{T_u}$$

Average Lagging (AL) measures the distance between the speaker and the user based on the number of words used in the conversation (Ma et al., 2018).

$$AL = \frac{1}{\tau(T_u)} \sum_{s=1}^{\tau(T_u)} d_s - \hat{d}_s \qquad (9)$$

where $\tau(T_u) = mins|d_s = \sum_{k=1}^{T_u} T_k$ and $\hat{d}_s$ represents the delays of an ideal policy defined as (Ma et al., 2020):

$$\hat{d}s = (s-1)\frac{\sum k = 1^{T_u} T_k}{T_v} \qquad (10)$$

Differentiable Average Lagging (DAL): One issue with AL is that it is not differentiable due to the min function. To address this, (Cherry and Foster, 2019) introduce a minimum delay of $1/\gamma$ after each operation and define DAL as:

$$DAL = \frac{1}{T_v} \sum_{s=1}^{T_v} d'_s - s - 1 \qquad (11)$$

where:

$$d's = \begin{cases} d_s, & s > 0 \; max(d_s, d's - 1 + 1/\gamma), \\ s > 0 \end{cases}$$

and $\gamma = T_v / \sum_{k=1}^{T_u} T_k$

2.3 Loss Functions

This section explores various loss functions employed for E2E ST models. Let $D = (u, x, v)$ represent a tuple where u, x, and v denote the speech, transcription text, and translation text, respectively.

ST Loss is defined as the negative log-likelihood of the translation text given the source speech:

$$L_{ST} = - \sum_{(u,v) \in D} logp(v|u)$$

MT Loss is defined as the negative log-likelihood of the translation text given the source transcript:

$$L_{MT} = - \sum_{(x,v) \in D} logp(v|x)$$

ASR Loss is defined as the negative log-likelihood of the transcription text given the source speech:

$$L_{ASR} = - \sum_{(u,x) \in D} logp(x|u)$$

CTC Loss calculates the most likely alignment of the output text sequence given the input speech sequence by summing over all possible output sequence paths:

$$L_{CTC} = - \sum_{(u,x) \in D} \sum_{z \in \phi(x)} logp(z|u)$$

Cross-Modal Adaptation Loss is defined as the sum of all the Mean Squared Errors (MSE) of the speech and transcription texts at both the sequence and word levels:

$$L_{AD} = \sum_{(u,x) \in D} \Big\{ MSE(\bar{h}_u, \bar{h}_x) \; MSE(h_u, h_x)$$

where h_u and h_x represent the speech and word embeddings, and $\bar{h}_u$ and $\bar{h}_x$ are their average embeddings, respectively. MSE represents the difference between the two embeddings.

Cross-Entropy Loss is the negative likelihood of the data combined over all subtasks (ASR, MT, ST) and also from external MT:

$$L_\theta = - \sum_{x,v \in D' \cup D_{MT-ext}} logp(x|v;\theta)$$

where $D' = D_{ASR} \cup D_{MT} \cup D_{ST}$ is the superset of all parallel subsets data.

Contrastive Loss is computed between the speech and transcription text, bringing related pairs closer and pushing unrelated pairs farther apart:

$$L_{CON} = - \sum_{(u,x) \in D} \frac{exp(cos(\bar{h}u, \bar{h}x)/\kappa)}{\sum_{u',x' \in D} exp(cos(\bar{h}u', \bar{h}_{x'})/\kappa)}$$

where cos denotes the cosine similarity and κ represents the temperature hyperparameter.

3 Datasets

The development of robust Speech-to-Text (ST) translation models relies on comprehensive and diverse datasets. This section outlines some of the prominent ST datasets and their characteristics. The creation of these datasets often involves using tools such as Gentle for audio-transcription alignment and BertAlign (github.com/bfsujason/bertalign) for transcription-translation alignment.

3.1 Prominent ST Datasets

- How2 (Sanabria et al., 2018) An ST corpus of English instructional videos with Portuguese translations.
- Augmented Librispeech (Kocabiyikoglu et al., 2018) Derived from the LibriSpeech corpus (Panayotov et al., 2015), a speech recognition repository based on audiobooks from the Gutenberg Project (www.gutenberg.org). This dataset focuses on translating English speech into French text.
- MuST-C (Di Gangi et al., 2019): A large multilingual ST translation corpus compiled from TED Talks. It includes translations from English into fourteen other languages. mTEDx (Salesky et al., 2021)) is a related multilingual dataset from TED talks.
- CoVoST and CoVoST 2 (Wang et al., 2020b): These datasets are based on the Common Voice project (commonvoice.mozilla.org). CoVoST is a many-to-one dataset covering 11 languages, while CoVoST 2 offers one-to-many and many-to-one translations for 15 languages.
- Europarl-ST (Iranzo-Sánchez et al., 2020): A collection of speech and text data from European Parliament proceedings between 2008 and 2012 in four languages. It includes multiple source and target languages for both speech and text.
- VoxPopuli (Wang et al., 2021): An expansion of Europarl-ST that includes data from European parliament sessions from 2009 to 2020.
- Kosp2e (Cho et al., 2021): A Korean (ko) to English (en) ST translation corpus with Korean speech paired with parallel English texts. The corpus comprises data from four domains: news/newspaper (Zeroth), textbooks, AI applications , and COVID-19 diaries.
- GigaST (Ye et al., 2022b): A corpus of speech translations from English to German and Chinese. It is created using the English ASR GigaSpeech (Chen et al., 2021), which features 10,000 hours of transcribed speech from various sources, including audiobooks, podcasts, and YouTube.
- Prabhupadavani (Sandhan et al., 2022): An ST dataset with multilingual and code-mixed speech. English is the primary language, with words and phrases from Sanskrit and Bengali interjected. The text portion contains sentences in 25 languages.

3.2 Additional ST Datasets

In addition to these widely used datasets, several smaller ST datasets exist, including Fisher (Cieri et al., 2004), Call-Home, Gordard Corpus , Glosse Audio Corpus, BTEC, WSJ, IWSLT, CHiME-4 Corpus (Christensen et al., 2010), Miami Corpus, and MSLT Corpus (Federmann and Lewis, 2016).

4 End-to-End Speech-to-Text Translation Models

End-to-end (E2E) models for speech-to-text (ST) translation are gaining popularity compared to traditional cascaded models. This section provides an overview of E2E ST models and categorizes them based on two key aspects: the framework employed and the nature of the data.

4.1 ST Models Based on Frameworks

ST models in the literature utilize the sequence-to-sequence (Seq2Seq) framework either alone or in combination with modality-bridging components.

4.1.1 Seq2Seq Frameworks

A Seq2Seq model generates a sequence of outputs from a sequence of inputs. It consists of an encoder for speech input, a decoder for text output, and an optional shared/semantic decoder connecting the encoder and the decoder. The model is typically optimized for the ST loss.

The core of Seq2Seq models often involves two key components: attention and transformers. This section focuses on models that do not leverage external data for training. Models that utilize external

data for pre-training, such as ASR and/or MT, are discussed in subsequent sections.

Attention Mechanism: Attention mechanisms concentrate on specific sections of the input data during output generation (Vaswani et al., 2017). They have proven successful in achieving state-of-the-art (SOTA) results in natural language processing (NLP) and other areas. Table 1 shows the notable works with attention mechanism.

Transfomers: The transformer architecture, based on multi-headed self-attention (Vaswani et al., 2017), produces contextualized representations of the input. Transformers have surpassed recurrent neural networks (RNNs) in several NLP tasks due to their parallel processing capabilities and contextual representation. Table 2 shows the notable works with attention mechanism.

Findings from Seq2Seq Frameworks:

- Stacked/pyramidal RNNs and alignment smoothing can achieve structural bias.
- Regularizers like transitivity and invertibility improve Character Error Rate (CER).
- Higher-level representations (HLRs) can benefit both transcription and translation.
- Modifying the encoder's self-attention with a logarithmic distance penalty can enhance translation performance.

4.1.2 Seq2Seq with Modality Bridging

Modality bridging refers to learning a combined representation of text and speech. In ST, both speech and text convey the same semantic meaning. Therefore, an effective ST model should learn a representation where embeddings of both modalities for similar speech-text pairs are close together.

Common Approaches for Modality Bridging:

1. Adapters: Small modules integrated with pre-trained networks for specific tasks (Houlsby et al., 2019). They perform at par with fine-tuning approaches while requiring fewer trainable parameters.
2. Contrastive Learning: Approximates "semantic" distance in the input space using a simple distance in the target space after mapping input patterns onto the target space (Chopra et al., 2005). It aims to bring positive instances closer while pushing negative ones apart.
3. Knowledge Distillation: A technique for transferring knowledge from a trained, large "teacher" model to a smaller, more efficient "student" model (Hinton et al., 2015).
4. Optimal Transport: A mechanism for comparing two probability distributions. In ST, speech and text representations can be viewed as probability distributions (Peyré and Cuturi, 2019).
5. Mix-up Strategy: A strategy that mixes speech embeddings and text embeddings into the encoder-decoder of a translation model to bridge the modality gap within a self-supervised learning framework (Fang et al., 2022).

Table 3 shows the notable works with modality bridging.

Findings from Modality Bridging:

- Adapters can shrink speech length and reduce the modality distance between text and speech representations while requiring fewer trainable parameters.
- Contrastive loss often outperforms CTC and L2 loss for modality bridging.
- Combining boundary-based speech length shrinking with contrastive loss can further improve ST task performance.

4.2 ST Models Based on the Nature of Available Data

The previous section examined ST models based on the frameworks used. This section presents an alternative perspective, categorizing E2E ST models based on the nature of the data, such as low-resource, streaming, multilingual, etc.

4.2.1 ST in Low-Resource Settings

Low-resource languages pose unique challenges for ST due to limited speech and/or text data. Pre-training Seq2Seq models on such small datasets can lead to overfitting and poor generalization. Table 4 shows the notable works for Low-resource ST.

Findings for Low-Resource ST:

- Pre-training on ASR data followed by fine-tuning on ST data can effectively address low-resource challenges.
- Generating pseudo-labels using unsupervised cascade models can provide additional training data.

Authors (Year)	Technique	Problem Solved	Dataset	Language Pair	BLEU
Berard et al. (2016)	Convolutional Attention	Eliminate Transcript	BTEC	Fr→En	46.7
Duong et al. (2016)	Phone-to-Text Alignment	Word-spotting	CallHome	CallHome	21.2
Anastasopoulos et al. (2016)	FastAlign + DTW	Low-resource	Fisher	Griko→It	30.8
Weiss et al. (2017)	Multitask Learning	Improved Performance	CallHome	Es→En	53.8
Kim et al. (2017)	Joint CTC-Attention Model	Long Inputs	Fisher	Es→En	48.7
Berard et al. (2018)	Encoder + Decoder	Cascading	WSJ0	Es→En	17.4

Table 1: Notable works on E2E Speech translation using Attention mechanism.

Authors (Year)	Technique	Problem Solved	Dataset	Language Pair
Gangi et al. (2019)	Transformer-based Seq2Seq with Attention	High Training Time	IWSLT	En→Fr
Alastruey et al. (2022)	Drop Weights that Attention Discards	Attention for Long Sequences	MuST-C	En→De

Table 2: Notable works on E2E Speech translation using Transformers.

Authors (Year)	Technique	Problem Solved	Dataset
Liu et al. (2019))	Knowledge Distillation	Improve Text Translation Performance	Librispeech
Baevski et al. (2020)	M-Adapter + Wave2Vec 2.0 + mBart	Training Gap Between Pre-training	
Han et al. (2021)	Chimera	Projecting Audio	MuST-C
Ye et al. (2021)	ConST (XSTNet + Contrastive Loss)	Closes Modality Gap	MuST-C
Gállego et al. (2021)	Wave2Vec 2.0 + mBart + Adapter	Slow Convergence Speed	
Ouyang et al. (2023)	WACO	Limited Parallel Data (1-hour)	MuST-C
Zeng et al. (2023)	AdaTrans	Closing Gap Between Speech and Text	Librispeech
Fang et al. (2022)	STEMM	Speech and Text Mixup	MuST-C
Le et al. (2023)	CTC Loss + Optimal Transport (Siamese-PT)	Limited Knowledge Transfer Ability	IWSLT

Table 3: Notable works on E2E Speech translation using Modality Bridging.

Authors (Year)	Technique	Problem Solved	Dataset	Language Pair	BLEU
Bansal et al. (2019)	Pre-training on High-resource ASR Data	Low-resource	Fisher	Es→En	20.2
Wang et al. (2023)	Unsupervised ST with Pseudo-Labels	Pseudo-labels	Godard Corpus	Mboshi→Fr	7.1

Table 4: Notable works for Low-Resource E2E ST.

4.2.2 Code-Mixed ST

Code-mixing refers to speech where a primary language is used, but words or phrases from other (embedded) languages are also included. Table 5 shows the notable works for code-mixed ST.

Findings for Code-Mix ST:

- Standard ST models often achieve good results on code-mixed data with low-resource settings and no fine-tuning, especially when pre-trained encoders and decoders are used, and multilingual models are employed.

4.2.3 ST in Streaming Setting

Streaming ST, also known as simultaneous translation, involves translating the input as it arrives without waiting for the entire input (Fügen et al., 2007; II et al., 2014; Ive et al., 2021).

Common Approaches for Streaming ST:

- Wait-K Policy: The model waits for *k* input speech segments before starting translation (Ma et al., 2019).
- Segmentation: The encoded speech is segmented to identify word, sub-word, or phone boundaries.
- Integrate-and-Fire (IF) Neuron: A neural mechanism that fires above a threshold when sufficient context is developed.
- Transducers/Unidirectional Transformers: Models capable of producing text sequences given input speech sequences in an online/streaming fashion (Graves, 2012).

Table 6 shows the notable works for streaming ST.

Findings from Streaming ST:

- RNN-Ts can reduce memory footprint but at the expense of translation delay.

Authors (Year)	Technique	Problem Solved	Dataset	Language Pair	BLEU
Weller et al. (2022)	Joint Transcription	Code-Mix Speech	Fisher	Es/En→En	26.2

Table 5: Notable works for Code-mixed E2E ST.

Authors (Year)	Technique	Problem Solved	BLEU
Ren et al. (2020)	SimulSpeech + Attention-level KD	Online Streaming Setting	22.49
Dong et al. (2022)	MoSST	Finding Boundaries for Acoustic Units	24.9
Liu et al. (2021)	CAAT	Policy	35.3
Xue et al. (2022)	Transformer Transducer (TT)	High Inference Latency	22.3
Wang et al. (2022)	LAMASSU	Joint ASR	30.7

Table 6: Notable works for Streaming ST.

- Adaptive read-write policies improve ST task performance.

4.2.4 Unsupervised ST

Unsupervised ST leverages unlabeled speech and text data to train ST models, avoiding the cost of manual annotation and parallel corpus creation. Table 7 shows the notable works for unsupervised ST ST.

Findings from Unsupervised ST:

- Pre-trained acoustic and language models combined with pseudo-labels using self-training demonstrate promising results for unsupervised ST translation.
- Domain adaptation is another technique for unsupervised ST

4.2.5 Multilingual ST

Multilingual ST models aim to translate between multiple speech input/output languages. The translation can be one-to-many, many-to-one, or many-to-many.

Common Approaches for Multilingual ST:

- Language ID (LID): Identification labels used to identify the target language and translate speech simultaneously.
- Dual Decoder: Transformers with two decoders, one for each ASR and ST, and a dual-attention mechanism.
- Pre-trained Multilingual Models: Pre-trained encoders and decoders for acoustic modeling and language modeling, respectively.

Table 8 shows the notable works for multilingual ST.

Findings from Multilingual ST:

- Language ID tokens often work well with the encoder.
- Mixed data training with language ID can transfer learning from high-resource languages (HRLs) to low-resource languages (LRLs).
- LID training with ASR data on unrelated languages often performs poorly.
- Adapters can help boost multilingual pre-trained model translation performance.

5 Summary

The survey presented so far shows significant improvements in E2E ST models. These advancements are likely due to leveraging pre-trained ASR/MT models or their respective corpora to train ST encoders/decoders. Weakly labeled/pseudo labels are another approach for generating additional training data for ST models. Modality bridging techniques, including contrastive learning, mix-up strategies, adapters, and optimal transport, have emerged as effective methods for closing the gap between speech and text representations.

1. Unsupervised ST: Pre-trained acoustic and language models combined with pseudo-labels using self-training demonstrate promising results for unsupervised ST translation. Domain adaptation is another technique for unsupervised ST.
2. Streaming ST: The k-wait policy, used in conjunction with segmentation and a dual-attention mechanism in multilingual settings, yields strong results for streaming ST.
3. Modality Bridging: Adapters are used in both modality bridging and multilingual settings

Authors (Year)	Technique	Dataset	BLEU
Kahn et al. (2020)	Teacher-Student Model (Wave2Vec 2.0 + Self-Training + Decoder Without LM)	CoVoST-2	27.2

Table 7: Notable work on Unsupervised ST.

Authors (Year)	Technique	Problem Solved	Dataset	BLEU
Inaguma et al. (2019)	LID + Mixed Data Training	One-to-Many	Fisher	46.3
Le et al. (2020)	Dual-Decoder (Esp-Net)	Joint ASR	CALLHOME	17.3
Liu et al. (2020)	LNA Fine-tuning + Zero-Shot	Unseen Languages	Librispeech	17.6
Le et al. (2021)	Adapters	Multilingual Pre-trained Performance	MuST-C	23.63
Wang et al. (2020a)	One-to-Many	CoVoST-2	28.12	

Table 8: Notable works for Multilingual ST.

with pre-trained models, improving performance.

4. Code-Mix ST: Standard ST models perform well on code-mixed data without fine-tuning. However, research on code-mixed data is still limited.

5.1 Overall Performance Trend of E2E ST Approaches in Common Benchmarks

This section investigates the performance trends of Speech Translation (ST) models using the MuST-C dataset, visualized in Figure 1. The MuST-C dataset was chosen due to its widespread use in the research community since its introduction in 2019. Figure 1 demonstrates a steady improvement in ST model performance over time. A significant advancement is observed in June 2021, attributed to the work of Ye et al. (2022a). This model achieved superior results by incorporating several key elements:

- Replacing fBank with Wav2Vec 2.0 for feature representation.
- Integrating a pre-trained Machine Translation (MT) decoder trained on extensive parallel MT data.
- Implementing multi-task fine-tuning across Automatic Speech Recognition (ASR), MT, and ST data.

6 Conclusion

This survey paper provides a comprehensive overview of recent advancements in end-to-end (E2E) speech-to-text (ST) translation. We delve into the models, evaluation metrics, and datasets used in training ST systems, reviewing various frameworks and highlighting key research contributions in the field. We categorize ST models based on the data they handle and the models employed. Furthermore, we discuss promising future directions for enhancing ST translation performance. Our analysis reveals that the gap between cascaded and E2E system performance, in both online and offline settings, is steadily closing. However, for certain language pairs, a significant gap persists, indicating the need for further research. This survey aims to provide valuable insights into the field of ST translation, stimulate further research, and ultimately drive advancements in this area.

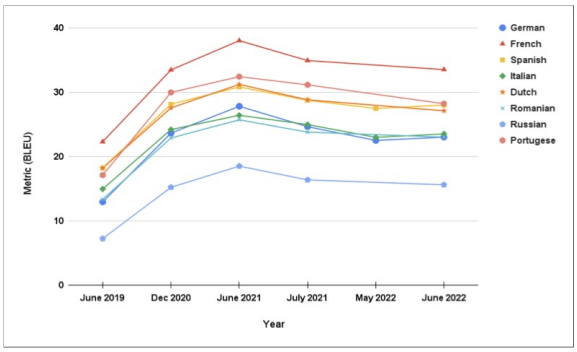

Figure 1: Performance of ST models on MuST-C data over a period of three years.

References

Belen Alastruey, Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. 2022. On the locality of attention in direct speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 402–412. Association for Computational Linguistics.

Antonios Anastasopoulos, David Chiang, and Long Duong. 2016. An unsupervised probability model

for speech-to-translation alignment of low-resource languages. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1255–1263. The Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 58–68. Association for Computational Linguistics.

Alexandre Berard, 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 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 6224–6228. IEEE.

Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *CoRR*, abs/1612.01744.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 3670–3674. ISCA.

Won-Ik Cho, Seok Min Kim, Hyunchang Cho, and Nam Soo Kim. 2021. kosp2e: Korean speech to english translation corpus. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 3705–3709. ISCA.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 539–546. IEEE Computer Society.

Heidi Christensen, Jon Barker, Ning Ma, and Phil Green. 2010. The chime corpus: a resource and a challenge for computational hearing in multisource environments. pages 1918–1921.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. 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.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888.

Qian Dong, Yaoming Zhu, Mingxuan Wang, and Lei Li. 2022. Learning when to translate for streaming speech. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 680–694. Association for Computational Linguistics.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 949–959. The Association for Computational Linguistics.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. STEMM: Self-learning with speech-text manifold mixup for speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7050–7062, Dublin, Ireland. Association for Computational Linguistics.

Christian Federmann and William D. Lewis. 2016. Microsoft speech language translation (MSLT) corpus: The IWSLT 2016 release for English, French and German. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Christian Fügen, Alex Waibel, and Muntsin Kolss. 2007. Simultaneous translation of lectures and speeches. *Mach. Transl.*, 21(4):209–252.

Gerard I. Gállego, Ioannis Tsiamas, Carlos Escolano, José A. R. Fonollosa, and Marta R. Costa-jussà. 2021. End-to-end speech translation with pre-trained models and adapters: UPC at IWSLT 2021. In *Proceedings of the 18th International Conference on Spoken Language Translation, IWSLT 2021, Bangkok, Thailand (online), August 5-6, 2021*, pages 110–119. Association for Computational Linguistics.

Mattia Antonino Di Gangi, Matteo Negri, Viet-Nhat Nguyen, Amirhossein Tebbifakhr, and Marco Turchi. 2019. Data augmentation for end-to-end speech translation: Fbk@iwslt '19. In *International Workshop on Spoken Language Translation*.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *CoRR*, abs/1211.3711.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2214–2225. Association for Computational Linguistics.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Alvin Grissom II, He He, Jordan L. Boyd-Graber, John Morgan, and Hal Daumé III. 2014. Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1342–1352. ACL.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual end-to-end speech translation. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, pages 570–577. IEEE.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. ESPnet-ST: All-in-one speech translation toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311, Online. Association for Computational Linguistics.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchís, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 8229–8233. IEEE.

Julia Ive, Andy Mingren Li, Yishu Miao, Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2021. Exploiting multimodal reinforcement learning for simultaneous machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3222–3233. Association for Computational Linguistics.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, Tatiana Likhomanenko, Gabriel Synnaeve, Armand Joulin, Abdelrahman Mohamed, and Emmanuel Dupoux. 2020. Libri-light: A benchmark for ASR with limited or no supervision. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 7669–7673. IEEE.

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. 2017. Structured-based curriculum learning for end-to-end english-japanese speech translation. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 2630–2634. ISCA.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 4835–4839. IEEE.

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, May 7-12, 2018*. European Language Resources Association (ELRA).

Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, page 228–231, USA. Association for Computational Linguistics.

Hang Le, Juan Miguel Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2020. Dual-decoder transformer for joint automatic

speech recognition and multilingual speech translation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3520–3533. International Committee on Computational Linguistics.

Hang Le, Juan Miguel Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 817–824. Association for Computational Linguistics.

Phuong-Hang Le, Hongyu Gong, Changhan Wang, Juan Pino, Benjamin Lecouteux, and Didier Schwab. 2023. Pre-training for speech translation: CTC meets optimal transport. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 18667–18685. PMLR.

Dan Liu, Mengge Du, Xiaoxi Li, Ya Li, and Enhong Chen. 2021. Cross attention augmented transducer networks for simultaneous translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 39–55. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-end speech translation with knowledge distillation. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1128–1132. ISCA.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3025–3036. Association for Computational Linguistics.

Siqi Ouyang, Rong Ye, and Lei Li. 2023. WACO: word-aligned contrastive learning for speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3891–3907. Association for Computational Linguistics.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Gabriel Peyré and Marco Cuturi. 2019. Computational optimal transport. *Found. Trends Mach. Learn.*, 11(5-6):355–607.

Yi Ren, Jinglin Liu, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. Simulspeech: End-to-end simultaneous speech to text translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3787–3796. Association for Computational Linguistics.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. The multilingual tedx corpus for speech recognition and translation. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 3655–3659. ISCA.

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. *CoRR*, abs/1811.00347.

Jivnesh Sandhan, Ayush Daksh, Om Adideva Paranjay, Laxmidhar Behera, and Pawan Goyal. 2022. Prabhupadavani: A code-mixed speech translation data for 25 languages. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 24–29, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Changhan Wang, Hirofumi Inaguma, Peng-Jen Chen, Ilia Kulikov, Yun Tang, Wei-Ning Hsu, Michael Auli, and Juan Pino. 2023. Simple and effective unsupervised speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10771–10784. Association for Computational Linguistics.

Changhan Wang, Juan Miguel 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, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4197–4203. European Language Resources Association.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Changhan Wang, Anne Wu, and Juan Miguel Pino. 2020b. Covost 2: A massively multilingual speech-to-text translation corpus. *CoRR*, abs/2007.10310.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020c. Curriculum pre-training for end-to-end speech translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3728–3738, Online. Association for Computational Linguistics.

Peidong Wang, Eric Sun, Jian Xue, Yu Wu, Long Zhou, Yashesh Gaur, Shujie Liu, and Jinyu Li. 2022. LAMASSU: streaming language-agnostic multilingual speech recognition and translation using neural transducers. *CoRR*, abs/2211.02809.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 2625–2629. ISCA.

Orion Weller, Matthias Sperber, Telmo Pires, Hendra Setiawan, Christian Gollan, Dominic Telaar, and Matthias Paulik. 2022. End-to-end speech translation for code switched speech. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1435–1448, Dublin, Ireland. Association for Computational Linguistics.

Jian Xue, Peidong Wang, Jinyu Li, Matt Post, and Yashesh Gaur. 2022. Large-scale streaming end-to-end speech translation with neural transducers. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 3263–3267. ISCA.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 2267–2271. ISCA.

Rong Ye, Mingxuan Wang, and Lei Li. 2022a. Cross-modal contrastive learning for speech translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5099–5113, Seattle, United States. Association for Computational Linguistics.

Rong Ye, Chengqi Zhao, Tom Ko, Chutong Meng, Tao Wang, Mingxuan Wang, and Jun Cao. 2022b. Gigast: A 10, 000-hour pseudo speech translation corpus. *CoRR*, abs/2204.03939.

Xingshan Zeng, Liangyou Li, and Qun Liu. 2023. Adatrans: Adapting with boundary-based shrinking for end-to-end speech translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2353–2361. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jiawei Zhao, Wei Luo, Boxing Chen, and Andrew Gilman. 2021. Mutual-learning improves end-to-end speech translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3989–3994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.