INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Lexical Modeling of ASR Errors for Robust Speech Translation

Giuseppe Martucci[1], *Mauro Cettolo*[2], *Matteo Negri*[2], *Marco Turchi*[2]

[1]University of Trento, Italy
[2]Fondazione Bruno Kessler, Trento, Italy

giuseppe.martucci@studenti.unitn.it, {cettolo,negri,turchi}@fbk.eu

Abstract

Error propagation from automatic speech recognition (ASR) to machine translation (MT) is a critical issue for the (still) dominant *cascade* approach to speech translation. To robustify MT to ill-formed inputs, we propose a technique to artificially corrupt clean transcripts so as to emulate noisy automatic transcripts. Our *Lexical Noise* model relies on estimating from ASR data: i) the probability distribution of the possible edit operations applicable to each word, and ii) the probability distribution of possible lexical substitutes for that word. Corrupted data generated from these probabilities are paired with their original clean counterpart for MT adaptation via fine-tuning. Contrastive experiments on three language pairs led to three main findings. First, on noisy transcripts, the adapted models outperform MT systems fine-tuned on synthetic data corrupted with previous noising techniques, approaching the upper bound performance obtained by fine-tuning on real ASR data. Second, the increased robustness does not come at the cost of performance drops on clean test data. Third, and crucial from the application standpoint, our approach is domain/ASR-independent: noising patterns learned from a given ASR system in a certain domain can be successfully applied to robustify MT to errors made by other ASR systems in a different domain.

Index Terms: cascade speech translation, MT robustness.

1. Introduction

Speech translation (ST) is the task of automatically converting utterances in one language into text in another language. This can be done either with a traditional *cascade* architecture [1, 2], or by adopting the more recent *direct* paradigm [3, 4]. The cascade approach relies on the combination of separate, independently trained components – mainly a speech recognition (ASR) and a machine translation (MT) system. Instead, direct systems feature ST without recurring to intermediate representations, by means of a single neural network directly trained on (*audio*, *textual translation*) pairs. Although direct ST is rapidly evolving,[1] its advantage in terms of architectural simplicity is undermined by the scarcity of training corpora to feed data-hungry neural models. Indeed, when it comes to industrial deployment, the wealth of ASR and MT training/adaptation data still makes cascade systems the preferable option by far.

Though still dominant, also the cascade approach has limitations. A well-known issue is error propagation [6]: the adverse effect of ASR errors that cannot be recovered by MT components unable to handle ill-formed inputs [7, 8, 9, 10]. Early solutions to this problem moved from loosely-coupled cascade architectures (with separate, independent components) toward a tight ASR-MT integration based on MT systems capable to decode ASR n-best lists [11], confusion networks [12] or lattices [13]. Later, the attention shifted toward making the MT models robust to ill-formed ASR output by training them on data, containing either *real* or *synthetic* noise.

In [14], real ASR errors are exploited in two ways: *i)* directly, by training the MT on automatically transcribed speech as source, or *ii)* indirectly, by means of an intermediate component trained to “translate” them into error-free transcripts. The first solution is data demanding, as it implies running an ASR system to transcribe a large amount of utterances paired with their translations, which is typically unavailable. The second solution involves the additional cost of training an intermediate correction component. Moreover, both methods are ASR-specific, since models’ training is informed by system-specific transcription errors. More recently, [15] proposed an adversarial learning approach that, different from [14], bypasses the need of data with explicit speech-to-transcription-to-translation alignments and does not involve additional components. However, as acknowledged by the authors, also this solution is ASR-specific and highly dependent on ASR quality.

To overcome this limitation, synthetic errors’ injection has been explored along different directions. Focusing on homophone noise, ASR outputs have been emulated by means of phonologically-motivated algorithms [16], pronunciation dictionaries modeling acoustic confusions [17], or random homophones’ replacement [18, 19]. Homophone noise, however, is only one type of speech recognition error [15] and, according to [20], brings marginal advantages to neural MT. Closer to our work, [20] adopts a more general approach for artificial noise injection on the source side of a parallel training corpus. Inspired by Levenshtein distance, their generative noise model operates on the basis of a manually-set hyper-parameter establishing the amount of noise to be induced by random word substitutions, insertions and deletions. Substitutions and insertions are drawn either uniformly from the vocabulary (“Vanilla” model) or from a unigram distribution (“Unigram” model).[2]

Though more effective compared to previous solutions, the approach proposed in [20] still suffers from limitations that our work aims to overcome. First, it requires a careful calibration of the type and amount of noise to be induced. Instead, **the solution here proposed avoids any manual calibration step and the adverse effect of wrong decisions.** Second, unigram sampling (in principle more linguistically motivated) underperforms compared to the randomized noise introduced by the Vanilla model, which produces corrupted data far from being representative of plausible ASR-like noise. In contrast, **the improved generative model here proposed (§2)**

[1] The latest International Workshop on Spoken Language Translation (IWSLT 2020) showed that, in specific evaluation conditions (TED-derived data, English as source language), direct systems have almost closed the initially huge performance gap with the cascade ones [5].

[2] Substitutions based on acoustic conditioning are also tested, but without observing noticeable improvements. This suggests that homophone noise yields marginal contributions, advocating for efforts along the lines of [16, 17], which the authors left for future work.

http://dx.doi.org/10.21437/Interspeech.2021-265

results in lexical perturbations that are more compatible with ASR errors, outperforming previous methods with results that approach the scores obtained by fine tuning the MT on real (system-specific) ASR data. Also, from the evaluation standpoint, [20] disregards a key requirement for applying noising techniques in real working conditions: the portability of the approach across ASR models and domains. **Our evaluation (§3)**, instead, **covers a wider range of testing conditions**: not only *in vitro*, by measuring improvements when an adapted MT model is fed either with clean or noisy transcripts, but also *in vivo*, by checking the effect of robustifying a state-of-the-art generic MT system with error patterns learned in a training setting (ASR, domain) different from the actual evaluation setting. Last but not least, **our improvements are obtained with a very limited amount of training data**: even with a few hours of transcribed speech (the equivalent of a typical ASR test set), our lexical model outperforms its competitors.

2. Lexical Noise Model

When comparing automatic and manual speech transcripts, ASR errors occur in the form of spurious insertions, deletions and substitutions. Our goal is to automatically corrupt a clean text so that it resembles ASR output. In this section, we: *i)* propose a lexical model of ASR errors (§2.1), *ii)* outline an algorithm using the model to corrupt a clean text (§2.2), and finally *iii)* discuss the estimation of all the model's statistics (§2.3).

2.1. Model Definition

ASR errors can be modeled by the noisy channel, the central component in the mathematical formulation of a communication system:[3] the corrupted text is the channel output (what is actually seen), while the correct transcript is its input (what should be discovered). In a probabilistic framework, given the ASR output word sequence $\mathbf{a} = a_1 \cdots a_m$, the correct transcript $\mathbf{c} = c_1 \cdots c_n$ can be searched through the decoding process that targets the maximization: $\arg\max_{\mathbf{c}} \Pr(\mathbf{c}|\mathbf{a})$. By applying Bayes' Rule and dropping the constant denominator, we get the un-normalized posterior: $\arg\max_{\mathbf{c}} \Pr(\mathbf{a}|\mathbf{c}) \times \Pr(\mathbf{c})$. Let us focus on the first factor, $\Pr(\mathbf{a}|\mathbf{c})$, of the channel model. Applying the Chain Rule we get:

$$\Pr(\mathbf{a} \mid \mathbf{c}) = \prod_j \Pr(a_j \mid a_1 \cdots a_{j-1}, \mathbf{c})$$

Assuming that words in $\mathbf{a}$ are independent from each other and that each a_j originates from a single correct word, the channel model reduces to:

$$\Pr(a_j \mid a_1 \cdots a_{j-1}, \mathbf{c}) = \Pr(a_j \mid \mathbf{c}) = \Pr(a_j \mid c_i)$$

In this simplified form, our *Lexical Noise* model is defined by a set of conditional distributions, one for each possible correct word c_i, providing the probability of any possible substitute a_j of c_i. Note that insertions can be modeled as well, as long as the distribution $\Pr(a_j|\phi)$ for the empty word ϕ is given.

2.2. Insertion, Deletion and Substitution Algorithm

An algorithm that changes a correct input sequence $\mathbf{c}$ into $\mathbf{a}$ by performing Insertion, Deletion and Substitution (IDS) operations can be represented by a state-based channel [21]. Figure 1 shows how the algorithm works. For each correct word c_i, the algorithm can decide to: *a)* append (i.e. **I**nsert) one or more spurious words to the so-far generated $\mathbf{a}$ according to probability p_I, *b)* **D**elete c_i with probability p_D, or *c)* "**T**ransmit" c_i with probability $p_T = 1 - p_I - p_D$. In this case, c_i can be either appended as is to $\mathbf{a}$ or **S**ubstitued by a_j, one of its possible alternatives selected according to some criterion.

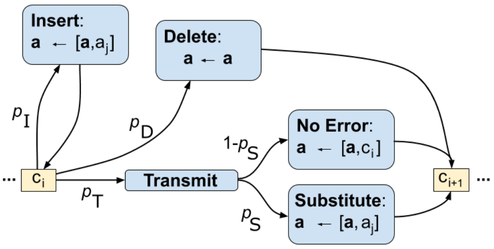

Figure 1: *State-based IDS channel.*

In the Lexical Noise model, the selection criterion used for the *Insert* and *Substitute* operations is based on $\Pr(a_j|\phi)$ and $\Pr(a_j|c_i)$, respectively the lexical conditional distributions of the empty word (ϕ) and of all possible correct words (c_i). If c_i is unknown (out of vocabulary), all operations remain the same except for the substitution, for which a_j is uniformly sampled from the vocabulary as in [20].

With different selection criteria, other instantiations of the Noise model proposed in literature can be defined in terms of our general state-based IDS channel model by sampling:

- uniformly from the vocabulary: Vanilla Noise [20]
- according to the unigram distribution of words in the vocabulary: Unigram Noise [20]
- based on acoustic_similarity(a_j, c_i) measured, for instance, by the distance of the phonetic transcriptions [17] or by the character edit distance [20]: Acoustic Noise.

2.3. Model Estimation

Our Lexical Noise model can be estimated on real data. Given an ASR engine and a benchmark, automatic transcripts' quality is commonly evaluated with WER, which is derived from the Levenshtein distance computed at the word level. First, the recognized word sequence is aligned to the reference transcript; then, looking at that alignment, substitutions, deletions, insertions and correct matches are counted. Such a procedure allows us to build a confusion matrix providing, for each word: *i)* the list of possible substitutes with the corresponding counters, *ii)* the number of times the word has been deleted, and *iii)* the number of times it has been inserted. The Lexical Noise model is estimated by converting those counts into probabilities.

3. Experiments

Most of the experiments discussed below involve fine-tuning an existing neural MT model trained on a large amount of data in one domain using a small amount of data from another domain [22]. "In-domain" data are either noisy (actual ASR outputs or automatically corrupted text) or clean (i.e. manual) transcripts. Though effective on the new in-domain data supplied for model adaptation, fine-tuning typically suffers from drastic performance drops on the general domain ones, unless proper regularization techniques are adopted [23]. Since the goal of our fine-tuning stage is to adapt the MT model to ASR errors preserving its ability to properly translate clean texts, it is crucial to avoid overfitting. We pursue this objective by fine tuning our MT models with dropout [24] and for only one epoch [23].

[3] http://en.wikipedia.org/wiki/Channel_capacity

3.1. Data

ST experiments were carried out on the en-it, de-en and fr-en[4] sections of Europarl-ST [25] (EP-ST henceforth) and on an in-house en-it NEWS ST test set. Adaptation in the NEWS experiments was performed on textual parallel data randomly selected from the OPUS repository.[5] Table 1 shows corpora statistics.

Table 1: *Stats of the EP-ST and NEWS (tokenized) corpora.*

	task	set	time	#segs	#src w	#trg w
EP-ST	en-it	trn	79h 29m	29.6k	792.4k	837.5k
		tst	2h 55m	1,130	29.1k	30.2k
	de-en	trn	30h 08m	12.9k	277.3k	321.7k
		tst	6h 12m	2,631	59.6k	66.6k
	fr-en	trn	31h 48m	12.4k	370.0k	342.3k
		tst	4h 48m	1,804	53.7k	50.0k
NEWS	en-it	opus	na	4.2M	59.6M	57.9M
		tst	58m 2s	196	11.7k	11.3k

3.2. ASR and MT engines

The audio data of each benchmark were automatically transcribed. For EP-ST, we used Hybrid DNN-HMM [26] ASR systems (`asrH` henceforth), whose acoustic models are speaker-independent and wide-band, while the language models are generic 4-grams adapted to the institutional domain. To evaluate our approach in different conditions, the English audio of the NEWS test set was transcribed by two different ASR engines: the same `asrH` used for the English EP-ST audio (apart from the use of a generic language model instead of an adapted one), and an end-to-end (Direct) neural ASR system (`asrD`) based on S-Transformer [27]. While the output of the `asrH` engines is punctuated and cased, `asrD` does not provide punctuation nor casing. ASR performance (%WER) is shown in Table 2.

Table 2: *%WER of ASRs either considering all original tokens (ori) or in case-insensitive no-punctuation condition (ci np).*

EP-ST		asrH			
task	set	ori	ci np		
en-it	en trn	28.36	19.40		
	en tst	28.80	20.44		
de-en	de trn	30.42	22.04		
	de tst	31.21	22.86		
fr-en	fr trn	28.62	19.08		
	fr tst	28.77	19.18		
NEWS		asrH		asrD	
task	set	ori	ci np	ori	ci np
en-it	en tst	29.02	18.41	39.00	25.81

Our MT engines are built using ModernMT,[6] which is based on the state-of-the-art Transformer [28] architecture. The baselines are Big Transformer models (as defined in [28], $\approx 2.1 \times 10^8$ params) trained on generic domain data, again from the OPUS repository, unless otherwise stated. MT quality is measured in terms of BLEU scores [29] computed by means of SacreBLEU with default signatures.[7]

3.3. Results

Effectiveness of the Lexical Noise model (*in vitro*). The first set of experiments aims at verifying the effectiveness of our Lexical Noise model under controlled conditions. For that sake, on a standard benchmark, we compare the performance of various instances of the same MT model, each adapted to a different type of transcripts. In Table 3, the first two rows show the BLEU scores of our generic MT system (FBK) and of Google Translate (GT) on the EP-ST test set. The other rows refer to the FBK model fine-tuned on the EP-ST training data, where the source side contains: manual transcripts (`man`), manual transcripts concatenated to either ASR transcripts (`man+asr`) or transcripts synthetically noised by means of the Lexical (`man+lex`) or Vanilla [20] (`man+van`) models. Results are grouped according to two variants of the test source sentences to be translated: manual (`man`) or automatic transcripts generated by `asrH`. The scores of `FBK_man+lex` and `FBK_man+van` are averages of three noising runs. The main outcomes are:

Table 3: *BLEU scores on EP-ST test sets, using either manual or asrH transcripts as input to MT models fine tuned on differently noised data.* $^\blacktriangledown$ *and* $^\triangledown$ *indicate a statistically significant drop at p<0.05 and p<0.10 below the* ***value*** *in bold in the same column; no symbol means no statistically significant difference. Statistical significance is computed as in [30].*

MT model	MT input					
	man			asrH		
	en-it	de-en	fr-en	en-it	de-en	fr-en
GT	32.22$^\blacktriangledown$	32.91$^\blacktriangledown$	40.92$^\blacktriangledown$	22.93$^\blacktriangledown$	24.98$^\blacktriangledown$	30.78$^\blacktriangledown$
FBK	34.10$^\blacktriangledown$	34.47$^\blacktriangledown$	40.52$^\blacktriangledown$	24.63$^\blacktriangledown$	25.26$^\blacktriangledown$	30.92$^\blacktriangledown$
FBK_man	**34.78**	**37.61**	**44.00**	25.21$^\blacktriangledown$	27.87$^\blacktriangledown$	33.78$^\blacktriangledown$
FBK_man+asr	34.38$^\blacktriangledown$	37.37	43.68$^\triangledown$	**26.09**	**29.32**	**35.18**
FBK_man+lex	34.56	37.21$^\triangledown$	43.86	26.05	28.88$^\triangledown$	34.91
FBK_man+van	34.48$^\triangledown$	37.19$^\triangledown$	43.69	25.44$^\blacktriangledown$	28.41$^\blacktriangledown$	34.35$^\blacktriangledown$

(1) When translating ASR transcripts (`asrH` columns), fine-tuning on data corrupted with our Lexical Noise model (`FBK_man+lex`) allows to consistently outperform the adaptation on manual transcripts (`FBK_man`) in all the language settings. The primary goal of any noising technique to robustify the MT model to ill-formed ASR inputs is thus achieved.
(2) On the same input type, `FBK_man+lex` achieves results that do not statistically differ from those obtained by MT adaptation to actual ASR transcripts (`FBK_man+asr`) in two out of three tasks. In other words, our Lexical Noise model approaches the upper bound performance obtained by fine tuning on real ASR data. The Vanilla Noise model (`FBK_man+van`), in contrast, always underperforms with a significance level of 0.05.
(3) When translating error-free manual transcripts (`man` columns), our Lexical model statistically guarantees the same quality level of the best model (`FBK_man`) in two out of three tasks. Overall, improved robustness does not come at the cost of undesired performance drops on clean data.

To summarize: in this first set of experiments the Lexical Noise model has proved capable of satisfying all the main requirements of noise models, performing better (most of the cases) or on par with the closest alternative solution.

Data requirements for model estimation. The second set of experiments aims at analysing the relation between the amount of data required to estimate the Lexical Noise model and downstream translation performance. Ideally, the less the data needed to train an effective perturbation model, the better. To let any behaviour differences emerge, in these experiments the MT baseline is a Small Transformer model ($\approx 38 \times 10^6$ params) built

[4] ISO 639-1 language codes
[5] http://opus.nlpl.eu
[6] http://github.com/modernmt/modernmt
[7] case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.14

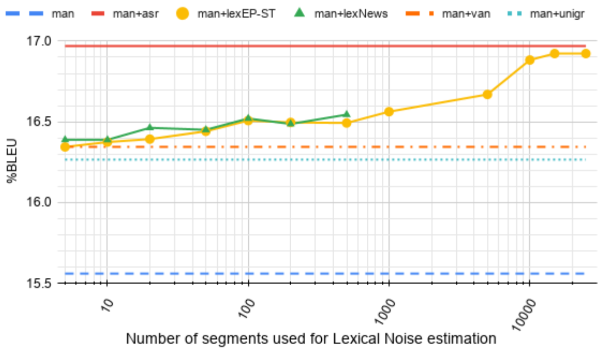

Figure 2: *BLEU scores of translations of en-it EP-ST asrH test set from various MT models trained on en-it EP-ST data only.*

on the EP-ST original training data only.[8] Contrastive models were trained on the concatenation of two instances of the EP-ST training set: one with the manual transcripts on the source side, the other with transcripts noised either with actual ASR errors or with synthetic errors generated by different noise models – Vanilla, Unigram (§2.2) and our Lexical model. The latter was estimated on increasing amounts of either EP-ST or NEWS[9] data, in order to also quantify the impact of in-/out-of-domain noise model estimation with respect to the translation domain. Figure 2 plots all the curves. Each point of the two "lex" curves is the average of three noising runs. The main outcomes are:

(4) With very few data for its estimation, the Lexical model behaves similarly to Vanilla and Unigram models, as expected. By enriching the training set, however, it performs increasingly better: 10k segments allow for estimating the model at its best, approaching the results obtained by using real ASR errors.
(5) The Lexical model estimation is domain independent given that, with the same number of segments used for training, the noise models built on in-domain (`man+lexEP-ST`) and out-of-domain data (`man+lexNews`) perform the same.
(6) Vanilla and Unigram models achieve similar MT performance, which is slightly better for the former as shown in [20]; this supports our choice to compare the Lexical model mostly to the Vanilla model, instead of the Unigram one.

In brief, these experiments have shown that the Lexical model starts to be effective even if trained on out-of-domain data, of size equivalent to that of a typical ASR test set (here, two hours), and that it quickly reaches its best performance.

Effectiveness of the Lexical Noise model (*in vivo*). The third set of experiments targets the challenging application scenario where in-domain audio data (and, in turn, their automatic transcripts) are not available to directly train an emulator of ASR errors. In this scenario, we aim at assessing whether an MT system can be adapted with artificial noise introduced in in-domain parallel data by a Lexical Noise model trained on out-of-domain audio data. Table 4 shows the BLEU scores (again, averaged over three noising runs) obtained on the en-it NEWS test set by fine-tuning the FBK model on (in-domain) NEWS training data noised either by the Lexical model trained on (out-of-domain) EP-ST audio data or by the Vanilla model. Different from the first round of experiments (Table 3), here the fine-tuning is done only for making the MT models robust to ASR errors, not for domain shifting too, given that the domain of the original MT model and of the test set is the same (NEWS): this is why fine-tuning is performed only on noised data, ignoring the original correct text. The main outcomes are:

Table 4: *BLEU scores on en-it NEWS test set, using either manual or ASR transcripts as input, with our FBK MT model fine tuned on different amounts of and differently noised data.*

	man	asrH	asrD
GT	40.07	26.09	19.83
FBK	38.99	24.71	18.67

fine-tun.	FBK_lexical			FBK_vanilla		
#segs	man	asrH	asrD	man	asrH	asrD
5k	38.99	24.99	19.24	38.37	23.42	16.57
50k	39.05	25.72	20.62	39.26	24.88	19.16
100k	39.10	25.94	21.03	39.24	25.07	19.71
500k	38.77	26.66	21.55	39.44	25.42	20.51
2M	38.92	27.04	21.60	39.53	25.90	20.48
4M	39.10	26.98	21.86	39.81	25.53	20.57

(7) Both Lexical and Vanilla models allow to make MT models more robust to ASR errors; the robustness is greater the more noised data for fine-tuning are used, although from a certain point on performance seems to reach a plateau.
(8) On both ASR transcripts, the Lexical model consistently outperforms the Vanilla model; their best performance on `asrH` and `asrD` differs by 1.14 and 1.29 absolute BLEU points.
(9) Despite the original MT model (FBK) performs worse than GT by 1.38 and 1.16 BLEU points when translating `asrH` and `asrD` transcripts, the Lexical model does not only allow to fill the gap but even to outperform GT by about 1 (27.04 vs. 26.09) and 2 (21.86 vs. 19.83) points on the two automatic transcripts.
(10) The adaptation to data generated by the Lexical model does not degrade the quality of the original MT model in the translation of clean manual transcripts.

To recap: this final set of experiments shows that a Lexical model trained on errors made in another domain (EP-ST) by a different ASR has been capable to robustify a state-of-the-art generic MT better than the Vanilla approach, while preserving the original translation quality on error-free transcripts.

4. Conclusions

Current solutions to the error propagation problem in cascade ST focus on robustifying the MT component via adaptation to real ASR errors (costly and not always available) or synthetic material emulating them. Synthetic noising techniques should satisfy three key requirements. The first one is to produce useful corrupted text with agile and scalable solutions that are not too demanding in terms of training data. Second, the generated noise should not affect MT quality in presence of error-free transcripts. Third, to avoid the proliferation of system-specific noising components in the long run, the underlying approach should be general enough to be domain- and ASR independent. With an eye at these three requirements, the *Lexical Noise* model proposed in this paper achieves coherent results on three language pairs, outperforming previous solutions.

5. Acknowledgements

The authors would like to thank PerVoice S.p.A. for the technical support and its staff for the helpful discussions.

[8] The motivation for this simplified scenario is that the MT models described in §3.2 achieve highly competitive results (see row `FBK_man` in Table 3) that reduce the room for improving via adaptation. Tiny models trained with less data, instead, are more suitable to make the actual contribution of different adaptation techniques visible.

[9] To make this experiment more informative, we added to the NEWS test set of Table 1 another hour of English audio in the NEWS domain, which was manually transcribed.

6. References

[1] F. Stentiford and M. Steer, "Machine translation of speech," *British Telecom Technology Journal*, vol. 6, no. 2, pp. 116–122, 1988.

[2] A. Waibel, A. N. Jain, A. E. McNair, H. Saito, A. G. Hauptmann, and J. Tebelskis, "Janus: a speech-to-speech translation system using connectionist and symbolic processing strategies," in *Proc. of ICASSP*, Toronto, Canada, 1991, pp. 793–796.

[3] 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 *Proc. of the NIPS Workshop on end-to-end learning for speech and audio processing*, Barcelona, Spain, 2016.

[4] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-Sequence Models Can Directly Translate Foreign Speech," in *Proc. of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2625–2629.

[5] E. Ansari, A. Axelrod, N. Bach, O. Bojar, R. Cattoni *et al.*, "Findings of the IWSLT 2020 Evaluation Campaign," in *Proc. of IWSLT*, Virtual Event, 2020, pp. 1–34.

[6] M. Sperber and M. Paulik, "Speech translation and the end-to-end promise: Taking stock of where we are," in *Proc. of ACL*, Virtual Event, 2020, pp. 7409–7421.

[7] E. Cho, J. Niehues, and A. Waibel, "NMT-Based Segmentation and Punctuation Insertion for Real-Time Spoken Language Translation," in *Proc. of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2645–2649.

[8] N. Ruiz, M. A. D. Gangi, N. Bertoldi, and M. Federico, "Assessing the Tolerance of Neural Machine Translation Systems Against Speech Recognition Errors," in *Proc. of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2635–2639.

[9] N.-T. Le, B. Lecouteux, and L. Besacier, "Disentangling ASR and MT Errors in Speech Translation," in *Proc. of MT Summit*, Nagoya, Japan, 2017.

[10] Y. Belinkov and Y. Bisk, "Synthetic and natural noise both break neural machine translation," in *Proc. of ICLR*, Vancouver, Canada, 2018.

[11] V. H. Quan, M. Cettolo, and M. Federico, "Integrated n-best re-ranking for spoken language translation," in *Proc. of INTERSPEECH*, Lisbon, Portugal, 2005, pp. 3181–3184.

[12] N. Bertoldi, R. Zens, and M. Federico, "Speech translation by confusion network decoding," in *Proc. of ICASSP*, Honolulu, US-HI, 2007, p. 1297–1300.

[13] C. Dyer, S. Muresan, and P. Resnik, "Generalizing word lattice translation," in *Proc. of ACL*, Columbus, US-OH, 2008, pp. 1012–1020.

[14] S. Peitz, S. Wiesler, M. Nussbaum-Thom, and H. Ney, "Spoken language translation using automatically transcribed text in training," in *Proc. of IWSLT*, Hong Kong, 2012, pp. 276–283.

[15] Q. Cheng, M. Fang, Y. Han, J. Huang, and Y. Duan, "Breaking the data barrier: Towards robust speech translation via adversarial stability training," in *Proc. of IWSLT*, Hong Kong, 2019.

[16] Y. Tsvetkov, F. Metze, and C. Dyer, "Augmenting Translation Models with Simulated Acoustic Confusions for Improved Spoken Language Translation," in *Proc. of EACL*, Gothenburg, Sweden, 2014, pp. 616–625.

[17] N. Ruiz, Q. Gao, W. Lewis, and M. Federico, "Adapting Machine Translation Models toward Misrecognized Speech with Text-to-Speech Pronunciation Rules and Acoustic Confusability," in *Proc. of INTERSPEECH*, Dresden, Germany, 2015, pp. 2247–2251.

[18] X. Li, H. Xue, W. Chen, Y. Liu, Y. Feng, and Q. Liu, "Improving the Robustness of Speech Translation," 2018, arXiv:1811.00728.

[19] H. Liu, M. Ma, L. Huang, H. Xiong, and Z. He, "Robust neural machine translation with joint textual and phonetic embedding," in *Proc. of ACL*, Florence, Italy, 2019, pp. 3044–3049.

[20] M. Sperber, J. Niehues, and A. Waibel, "Toward Robust Neural Machine Translation for Noisy Input Sequences," in *Proc. of IWSLT*, Tokyo, Japan, 2017, pp. 90–96.

[21] M. C. Davey and D. J. MacKay, "Reliable Communication over Channels with Insertions, Deletions, and Substitutions," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 687–698, February 2001.

[22] M.-T. Luong and C. D. Manning, "Stanford Neural Machine Translation Systems for Spoken Language Domains," in *Proc. of IWSLT*, Da Nang, Vietnam, 2015, pp. 76–79.

[23] A. V. Miceli Barone, B. Haddow, U. Germann, and R. Sennrich, "Regularization techniques for fine-tuning in neural machine translation," in *Proc. of EMNLP*, Copenhagen, Denmark, 2017, pp. 1489–1494.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[25] 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 *Proc. of ICASSP*, Barcelona, Spain, 2020, pp. 8229–8233.

[26] G. E. Hinton, L. Deng, D. Yu, G. Dahl *et al.*, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97, 2012.

[27] M. A. Di Gangi, M. Negri, and M. Turchi, "Adapting Transformer to End-to-end Spoken Language Translation," in *Proc. of INTERSPEECH*, Graz, Austria, 2019.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Proc. of NIPS*, Long Beach, US-CA, 2017, pp. 5998–6008.

[29] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proc. of ACL*, Philadelphia, US-PA, 2002, pp. 311–318.

[30] P. Koehn, "Statistical Significance Tests for Machine Translation Evaluation," in *Proc. of EMNLP*, Barcelona, Spain, 2004, pp. 388–395.