Integration of Dubbing Constraints into Machine Translation

Ashutosh Saboo
BITS Pilani, K.K. Birla Goa Campus*
Goa, India
ashutosh.saboo96@gmail.com

Timo Baumann
Department of Informatics,
Universität Hamburg, Germany
baumann@informatik.uni-hamburg.de

Abstract

Translation systems aim to perform a meaning-preserving conversion of linguistic material (typically text but also speech) from a source to a target language (and, to a lesser degree, the corresponding socio-cultural contexts). Dubbing, i. e., the lip-synchronous translation and revoicing of speech adds to this constraints about the close matching of phonetic and resulting visemic synchrony characteristics of source and target material. There is an inherent conflict between a translation's meaning preservation and its 'dubbability' and the resulting trade-off can be controlled by weighing the synchrony constraints. We introduce our work, which to the best of our knowledge is the first of its kind, on integrating synchrony constraints into the machine translation paradigm. We present first results for the integration of synchrony constraints into encoder decoder-based neural machine translation and show that considerably more 'dubbable' translations can be achieved with only a small impact on BLEU score, and dubbability improves more steeply than BLEU degrades.

1 Introduction

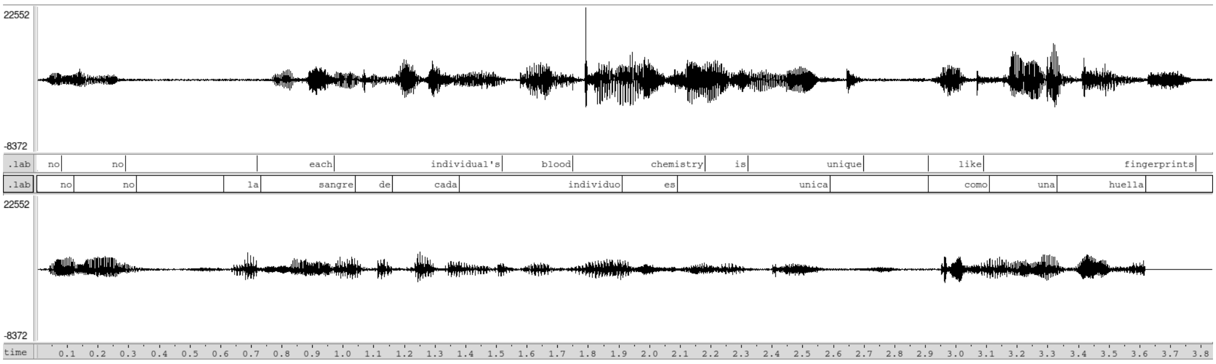
Dubbing, the lip-synchronous translation and revoicing of audio-visual media, is essential for the full-fledged reception of foreign movies, TV shows, instructional videos, advertisements, or short social media clips. Dubbing does not contend for the viewers' visual attention like subtitles (Díaz-Cintas and Remael, 2014) do, and unlike voice-over or asynchronous speech there is no (or only little) mismatch of visual and auditory impression where the resulting cognitive dissonance would otherwise increase the viewers' cognitive load, or even lead to understanding errors (McGurk and Macdonald, 1976). Dubbing is still primarily studied in audio-visual translation (Orero, 2004; Chaume, 2012) and performed manually, unlike textual translation, which is largely being automated or supported by computer-aided translation (Koehn, 2009).

Recent break-throughs in speech-to-speech translation (Jia et al., 2019), do not yield translations that systematically observe dubbing constraints, i. e. do not match phonetically (or rather: visemically) the original source (we call this 'dubbability'). It is our goal to create MT systems where the dubbability of the translation can be controlled so as to optimize the trade-off between translation quality and lip-synchrony of the dubbed speech. We hope that more widely available dubbing across languages will help to stimulate access to foreign media and foster inter-cultural exchange.

We argue that dubbable MT will not simply emerge from training on dubbed audio-visual corpora, i. e. implicitly. By comparison, audio-visual corpora will always remain smaller than pure text-to-text translation corpora. As a result, merely relying on training a conventional MT system on large amounts of dubbing texts is bound to severely limit performance. What's more, the task of dubbing combines the constraints of several areas (meaning-preserving as well as prosodically similar translation) which have different properties. For example, for speech from the off or without the speaker's face visible, there are no limitations on prosodic similarity while it may be critical in close-up scenes; the *translation* system would thus need to consider video as well (but only very selectively so). Thus, we are looking for a flexible weighing of these two aspects which we achieve by introducing *phonetic synchrony constraints* that describe the 'dubbability' of a proposed translation, i. e., how well it is expected to allow for lip-synchronous revoicing in

[*] This work was performed during an internship at Universität Hamburg, Germany.

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 94–101
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics



```
source (en): No, no.  Each individual's blood chemistry is unique, like fingerprints.
dubbed (es): No, no.  La sangre de cada individuo es única, como una huella.
   faithful: No, no.  La química de la sangre de cada individuo es única,
                                                como las huellas dactilares.
```

Figure 1: Example dubbing in the show "Heroes" (season 3, episode 1, starting at 29'15", from Öktem et al. (2018)); 'faithful' meaning-preserving translation based on Google Translate.

the target language.[1]

An example of the weighing of lip synchrony and faithful translation in dubbing is shown in Figure 1 which shows an example utterance in the HEROes corpus[2] (Öktem et al., 2018) in its English original and Spanish dubbed revoicing, as well as a meaning-preserving translation. The latter results in about 70 % too many syllables (32 vs. 19 in the source), and would be next to impossible to revoice in a lip-synchronous manner. The human translator (and dubbing expert) resolved the issue by sacrificing some detail in the translation: two terms, "blood chemistry" and "fingerprints" can easily be translated slightly differently (leaving out the "chemistry" and "finger" aspects, as well as singularizing "prints") which reduces the syllable difference down to 20 % without sacrificing the overall meaning conveyed by the utterance.

We describe how synchrony constraints can be included in the MT process, in particular in the search/decoding process of neural MT, in the following section and then describe our implemented system in Section 3 and present results of our experimentation in Section 5. We conclude in Section 6 where we also present our plans for future work.

2 Integration of Dubbing Constraints

Given a source language sentence S, both statistical MT and neural MT perform a search among many different possible candidate utterances C in the target language, wrt. constraints that represent the faithfulness of the translation, $\mathrm{score}_t(C, S)$, with the best scoring candidate picked as the result.

Given the source sentence and a candidate translation, we can compute a phonetic (or visemic) synchrony $\mathrm{score}_p(C, S)$. Then, for dubbing-optimized machine translation, we simply compute a dubbing-optimal score_d that combines both sub-scores using a weight α that indicates the relative importance of phonetic synchrony vs. translation faithfulness:

$$\mathrm{score}_d^\alpha(C, S) = (1 - \alpha) * \mathrm{score}_t(C, S) + \alpha * \mathrm{score}_p(C, S).$$

In application, α can be varied, e. g. according to whether the speaker's face is visible on screen.

MT systems gradually construct and prune the search space as their scoring functions work well locally, i. e., already do well for partial translations.[3] In contrast, synchrony scoring requires a global perspective, in particular for a constraint such as the relative deviation in syllable number between a candidate and the source, i. e. for $\mathrm{score}_p(C, S) = abs(syll(C) - syll(S))/syll(S)$. It is not easy to compute this for only a prefix of C as it is typically unclear which words in the source have already been accounted for and as syllables can be shifted between words (only the total matters).

To integrate phonetic constraints into the search

[1] In this paper, we use the relative difference of syllable count estimates between source and target material as the similarity constraint. We expect that more elaborate constraints, e. g. based on accentuation, stress marks, expected speech durations, articulatory and prosodic features, visemes, etc. will be needed to match human dubbing performance.

[2] http://hdl.handle.net/10230/35572

[3] However, He et al. (2017) use a similar technique as outlined below for BLEU-optimal decoding for NMT.

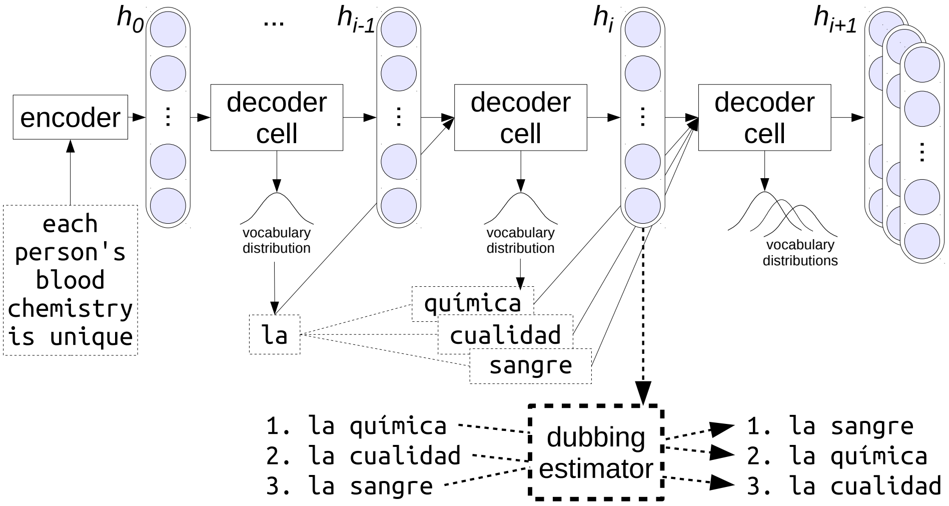


Figure 2: Integration of dubbing constraints into the MT decoder: the beam is re-scored by a combined score of the phonetic similarity of the decoded prefix as well as a heuristic estimate for what remains in the search state.

process, we propose a heuristic dubbing estimator that breaks down the task of phonetic similarity scoring into (a) the known phonetic score for the prefix that has already been generated, and (b) a heuristic $\widehat{\text{score}}_p$ based on the internal state of the decoder for how well the yet untranslated part of the utterance will score. Different prefixes correspond to different decoder states and states are known to capture the remaining length of the translation (Shi et al., 2016). Our method extends over that of Chatterjee et al. (2017), which scores constraints only once all necessary information is available in the decoded prefix. The resulting beam search then performs similarly to A* (Hart et al., 1968).

Figure 2 depicts our method, without loss of generality, for NMT. In the example, the decoding of an utterance at decoding stage i is shown. At i, the decoder may consider to add a word to faithfully translate the phrase "blood chemistry", and as an alternate hypothesis consider translating just "blood" as a shorter form of conveying the same message. All alternatives are placed in the MT system's beam which is then re-scored by the *dubbing estimator* which takes each word sequence in the beam to compute the phonetic score of the prefix, as well as the decoder's hidden state h_i to estimate the score for what will still have to be translated. In this case, we can imagine that "sangre" will re-score to a higher position as its brevity is preferred (whereas the alternatives would still need to add "sangre" in a later decoding stage, thus their states will be estimated as containing more material to come yielding an overall higher estimate and a lower score).

The integration of synchrony constraints into the decoder enables a dubbing-optimal search with very little decoding overhead, however with some implementation effort. In addition, the heuristics $\widehat{\text{score}}_p$ could turn out to be be problematic given little training material or domain mismatches (see below). A similar result at low code complexity but potentially longer run time can be achieved by post-hoc rescoring based on a relatively large beam size from a standard NMT decoder. This approach is implemented in our first prototype which will be described in the next section.

3 Implemented System

We first describe our NMT model and training setup in detail, which yields an MT system that is competitive with the state of the art. Overall, our goal is not to create a heavily optimized system that gives us the highest possible performance in our domain but merely to yield a plausible baseline. We then describe our amendments for dubbing-optimal decoding.

We implement a convolutional encoder-decoder NMT model (Gehring et al., 2017). Given the relatively lesser training data (see below), we use a smaller model than Gehring et al. (2017), inspired by Edunov et al. (2018) and hence adapt certain hyperparameter values as described in Table 1.

We pre-process textual data as follows: we perform tokenization using the scripts from the open-source package Moses[4] (Koehn et al., 2007)

[4] https://github.com/moses-smt/mosesdecoder

followed by a byte-pair encoding compression algorithm to reduce the vocabulary size (Sennrich et al., 2016) using the open-source package `subword-nmt`[5]. We denote words not included in the vocabulary as <UNK>. We do not apply any lowercasing or stemming.

We train our model with `fairseq`[6] (Ott et al., 2019) for the default 34 epochs with training objectives and search settings as found to be optimal by Edunov et al. (2018) for a similar MT task.

Our standard decoder uses a beam-size of 50 (which is larger than typically used, but see next section for results).

For **dubbing-optimal decoding**, we rescore the N-best list from standard decoding $\mathbf{B}_t$ by the method outlined in Section 2: We estimate the number of syllables in each candidate and the source sentence and take the difference ($\text{sylldiff}(C, S) = abs(syll(C) - syll(S))$) and convert this to a $\text{score}_p(C, S) = 1/(1+\text{sylldiff}(C, S))$ that is highest for identical syllable counts. We then reweigh the sub-scores for translation and synchrony with a weight α, yielding a rescored beam $\mathbf{B}_d$ of which we take the best-ranked translation as being the dubbing-optimal translation. The full algorithm for rescoring is described in Algorithm 1. We use `Pyphen`[7] for estimating the syllable count for both English (source language) and Spanish (target language).

Table 1: Custom hyperparameters of our convolutional encoder-decoder model; all other hyperparameters are set as by Gehring et al. (2017).

Hyperparameter	**Value**
Encoder embedding dimension	256
Encoder hidden units in each layer	256
Kernel size for each encoder layer	3
Encoder layers	4
Dropout rate	0.2
Decoder embedding dimension	256
Decoder hidden units in each layer	256
Kernel size for each decoder layer	3
Decoder layers	3

Algorithm 1 N-Best Rescoring with Dubbing Constraints

```
Input: Translation model $P(y|x)$, Test Batch Input $T$, Rescoring Factor $\alpha$
$\mathbf{B}_t \leftarrow \forall_{e \in T} StandardBeamSearch(e)$
for all candidate $C$ in $\mathbf{B}_t$ do
    $\text{score}_t(C) \leftarrow C.score$
    $\text{score}_p(C) \leftarrow 1/(1+ \text{sylldiff}(C, S))$
    $\text{score}_d(C) \leftarrow (1-\alpha) * \text{score}_t(C) + \alpha * \text{score}_p(C)$
Output: Rescored Beam Output $\mathbf{B}_d$
Select: Best-ranked candidate from $\mathbf{B}_d$
```

4 Setup and Evaluation Method

Ideally, a dubbing-optimal translation system should be evaluated on dubbed material. We use the HEROes dubbing corpus (Öktem et al., 2018) a corpus of the TV show with the same name with the source (English) and dubbing into Spanish. The corpus contains a total of 7000 manually aligned utterance pairs in 9.5 hours of speech and based on forced alignment of video subtitles to the audio tracks. The audio material (in both English and Spanish) is not yet used in the experiments reported below.

We find that the HEROes corpus contains 85,767 (resp. 83,561) syllables for English (resp. dubbed Spanish), as computed with `Pyphen`. The average number of syllables per utterance is 12.25 for English and 11.94 for Spanish. We conclude that, on average, both languages use almost the same number of syllables and hence our phonetic similarity measure based on syllables should be useful. (It would be possible, for other language pairs where the notion of syllable differs, e. g. when considering the mora-driven Japanese, to compute some sort of correction factor between the languages. In our case, we simply ignore the relative difference in syllables of $< 3\,\%$ between the languages.)

Although large for a dubbing corpus, the 7,000 utterances are far too little to train an NMT model on. We hence use the English → Spanish parallel data in the Europarl corpus (Koehn, 2005) for training and will evaluate on both the dubbing corpus and a test set based on the Europarl corpus. The genre of science fiction TV shows may differ radically from parliament proceedings. However, this merely results in lower BLEU performance on the out-of-domain data. We believe that model adaptation (e.g. Chu and Wang, 2018) or relatively

[5] https://github.com/rsennrich/subword-nmt

[6] https://github.com/pytorch/fairseq

[7] https://pyphen.org/

more in-domain training material (e.g. Lison and Tiedemann, 2016) would work orthogonal to the dubbing-specific improvements in our paper. Text pre-processing is identical for both corpora.

We measure the translation performance in terms of BLEU (Papineni et al., 2002) as computed with the SacreBLEU software[8] (Post, 2018). Dubbing-optimality of translations in the test-set T is determined by micro-averaging the dubbing-scores as follows: by *synchrony-score* for test-set T defined as:

$$synchrony\text{-}score(T) = \frac{\sum_{e \in T} abs(syll(\mathrm{NMT}(e)) - syll(e))}{\sum_{e \in T} syll(e))}$$

where $\mathrm{NMT}(e)$ is the target translation given by the NMT model $P(y|x)$ (with or without dubbing constraints applied) for English source text e.

As is evident, the lower the synchrony score the better is the dubbing optimality. We run our experiment to analyze the variation of BLEU vs. synchrony score for different rescoring factors α.

We use the trained NMT model as described in the above section. Our decoding algorithm is as described in Algorithm 1, which we use to compute the relation between translation performance and dubbing-optimality of translations.

5 Experiment and Results

It has previously been pointed out that NMT performance suffers from a beam search size beyond 5 or 10 (Koehn and Knowles, 2017; Tu et al., 2017) and numerous methods have been proposed to circumvent this (Huang et al., 2017; Ott et al., 2018; Yang et al., 2018). However, for our present way of dubbing-optimization based on N-best rescoring, high beam sizes are essential for the dubbing-rescoring described in Algorithm 1 to have some material to work with. With only few candidates to be rescored, it might not necessarily give us the most ‘dubbable’ result.

We experimented with various beam sizes and found no BLEU degradation for a beam size of 50. Larger beams may eventually lead to a degradation and run time would become overly long as it linearly increases with the beam size. Owing to the best of both worlds, we resort to a beam size of 50 for the experiments reported below.

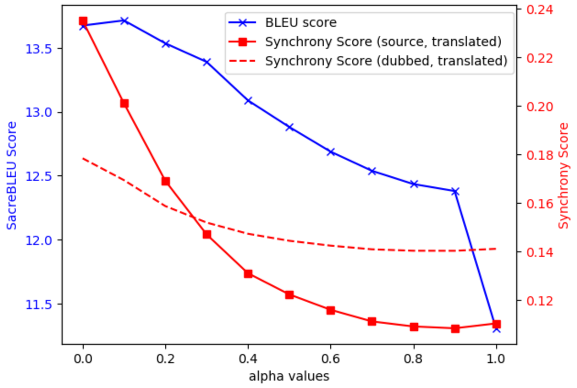


Figure 3: Evaluation results for the HEROes corpus.

5.1 Evaluation on Dubbing Material

Figure 3 shows BLEU scores (left scale, higher is better) and synchrony score (right scale, lower is better) of our proposed system for a range of α between 0 and 1. Notice that $\alpha = 0$ corresponds to no rescoring, i. e. the baseline system.

The relatively low BLEU score of 13.67 for the baseline system reflects the domain-mismatch between HEROes and Europarl.[9] We find that BLEU score is impacted only moderately for relatively low values of α, with a relative decrease of 2 % for $\alpha = .3$. At the same time, we find the synchrony score to improve drastically already with small values of α: while the difference in syllables between source and target is almost one quarter in the baseline system, this is almost halved, down to 14 % for $\alpha = .3$.

Figure 3 also contains the synchrony score of the proposed translations vs. the actual gold-standard dubbed texts (dotted line in the figure). As can be seen, the similarity increases up to about $\alpha = .3$ and then flattens out. This is in line with our observation that, while source and target number of syllables correlate highly, there is no perfect match, indicating that our synchrony constraint has only limited value. However, it also points to the fact that a human dubbing expert needs to find the middle ground between faithful translation and perfect synchrony. Given that two differing linguistic systems are involved, a perfect synchrony is simply impossible if the meaning is to remain approximately correct.

[8] https://github.com/mjpost/sacreBLEU

[9] To the great relief of the authors, the European parliament does *not* speak like supernatural figures in a mystery TV show that was scrapped after only 4 seasons due to the harsh criticism on its ludicrous nature.

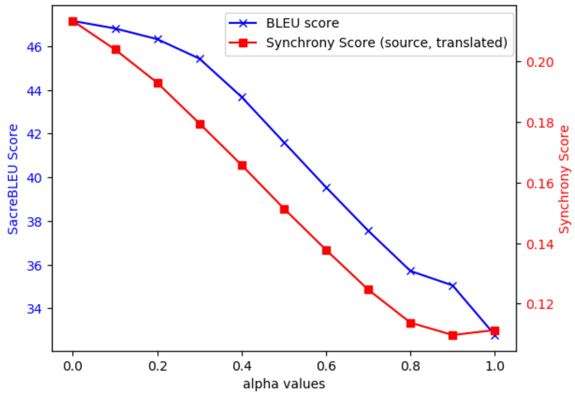


Figure 4: In-domain evaluation results for Europarl.

5.2 In-Domain Evaluation

We also evaluate our method *in-domain*, on test data sampled from Europarl (excluded from training). In particular, we use those source sentences for which multiple reference translations are contained in the corpus (about 18k instances). Europarl translations, of course, are not transcripts of lip-synchronously dubbed speech. Thus, our expectations for synchrony constraints are somewhat lower. However, testing in-domain still helps greatly to validate our out-of-domain results above.

As can be seen in Figure 4, we see a similar decrease in BLEU scores (and only very gradually for small α values) and more strongly improving synchrony scores. This again points towards a useful trade-off when combining synchrony constraints with the requirement of meaning-preserving translations. There is a range of possible reasons why our method does not work as well for Europarl as for the HEROes corpus. In particular, Europarl is not transcribed speech and hence may be less 'dubbable' by nature; many phrases in Europarl may translate to phrases with a different number of syllables in the target language, yet the model is reluctant to give up this translation in the in-domain condition; the proxy-target of syllables may work less well for longer, more specific words as found in legal texts, where a focus on only accentuated syllables may be more useful.

6 Conclusion and Future Work

We have explored the task of *dubbing-optimal* machine translation, i. e. machine translation that unifies the constraints of faithfulness in translation with the constraint of lip-synchrony for revoicing of audio-visual media. We have, so far, limited our synchrony constraint to counting syllables (which acts as a proxy to jaw openings that would be a major factor in visemic characteristics of speech).

We have outlined how one can integrate *synchrony constraints* into to the search during decoding by estimating the amount of syllables that are still remaining in the hidden state of the encoder-decoder model. We have implemented a simpler prototype system that instead rescores a conventional system's final N-best list.

Using the (as far as we know) largest corpus of dubbed speech available, the HEROes corpus (Öktem et al., 2018), we have shown our method to yield much more 'dubbable' translations than those that result from a standard MT system. In fact, while the manual dubbing for the sentence in Figure 1 abbreviates the phrase "blood chemistry" to plain "sangre", our model instead chooses "la *química* de cada persona es única" which is still a reasonable translation of "blood chemistry" and comes very close in terms of syllable count.

In the future, we intend to implement the fully integrated search as described in Section 2, as well as implement more powerful synchrony metrics that could also ground in the source audio (e. g. to find out what syllables were stressed) or the source video (e. g. to find out how well the face is visible), and could also consider detailed aspects of the target speech (e. g. via speech synthesis cost estimates for forcing the target text on the observed visemes).

One interesting and relevant aspect of teaching humans interpreting is the task of rewording material in the target language (Gile, 2005). A model that can be trained towards an ability of coming up with alternate wordings for the same concept (but with different synchrony-related properties) would potentially yield much better candidates for 'dubbability' assessment.

Acknowledgments

This work is partially supported by Volkswagen Foundation under the funding codes 91926 and 93255. We thank the 3 anonymous reviewers for their insightful remarks.

References

Rajen Chatterjee, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia, and Frédéric Blain. 2017. Guiding neural machine translation decoding with external knowledge. In *Proceedings of the Second Conference on Machine Translation*, pages 157–

168, Copenhagen, Denmark. Association for Computational Linguistics.

Frederic Chaume. 2012. *Audiovisual translation: Dubbing*. St. Jerome Publishing, Manchester.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, USA. Association for Computational Linguistics.

Jorge Díaz-Cintas and Aline Remael. 2014. *Audiovisual Translation, Subtitling*. Routledge, London.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Classical structured prediction losses for sequence to sequence learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 355–364, New Orleans, USA. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pages 1243–1252.

Daniel Gile. 2005. Teaching conference interpreting. In *Training for the New Millennium*, pages 127–151. John Benjamins Publishing Company, Amsterdam.

P. E. Hart, N. J. Nilsson, and B. Raphael. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107.

Di He, Hanqing Lu, Yingce Xia, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2017. Decoding with value networks for neural machine translation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 178–187.

Liang Huang, Kai Zhao, and Mingbo Ma. 2017. When to finish? optimal beam search for neural text generation (modulo beam size). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2134–2139, Copenhagen, Denmark. Association for Computational Linguistics.

Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. *CoRR*, abs/1904.06037.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Philipp Koehn. 2009. A process study of computer-aided translation. *Machine Translation*, 23(4):241–263.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Canada. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Harry McGurk and John Macdonald. 1976. Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

Alp Öktem, Mireia Farrús, and Antonio Bonafonte. 2018. Bilingual Prosodic Dataset Compilation for Spoken Language Translation. In *Proc. IberSPEECH 2018*, pages 20–24.

Pilar Orero, editor. 2004. *Topics in Audiovisual Translation*. John Benjamins Publishing Company.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3956–3965, Stockholmsmässan, Stockholm Sweden. PMLR.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, USA. Association for Computational Linguistics.

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 on Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Xing Shi, Kevin Knight, and Deniz Yuret. 2016. Why neural translations are the right length. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2278–2282, Austin, USA. Association for Computational Linguistics.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction.

Yilin Yang, Liang Huang, and Mingbo Ma. 2018. Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3054–3059, Brussels, Belgium. Association for Computational Linguistics.