Evaluating Heuristics for Audio-Visual Translation

Timo Baumann[1], Ashutosh Saboo[1,2]

[1]Department of Informatics, Universität Hamburg, Germany
[2]BITS Pilani, K.K. Birla Goa Campus, Goa, India

Abstract

Dubbing, i.e., the lip-synchronous translation and revoicing of audio-visual media into a target language from a different source language, is essential for the full-fledged reception of foreign audio-visual media, be it movies, instructional videos or short social media clips. In this paper, we objectify influences on the 'dubbability' of translations, i.e., how well a translation would be synchronously revoiceable to the lips on screen. We explore the value of traditional heuristics used in evaluating the *qualitative aspects*, in particular matching bilabial consonants and the jaw opening while producing vowels, and control for *quantity*, i.e., that translations are similar to the source in length. We perform an ablation study using an adversarial neural classifier which is trained to differentiate "true" dubbing translations from machine translations. While we are able to confirm the value of matching *lip closure* in dubbing, we find that the opening angle of the jaw as determined by the realized vowel may be less relevant than frequently considered in audio-visual translation.

Keywords

audiovisual translation, dubbing, lip synchrony, machine translation, ablation study

1. Introduction

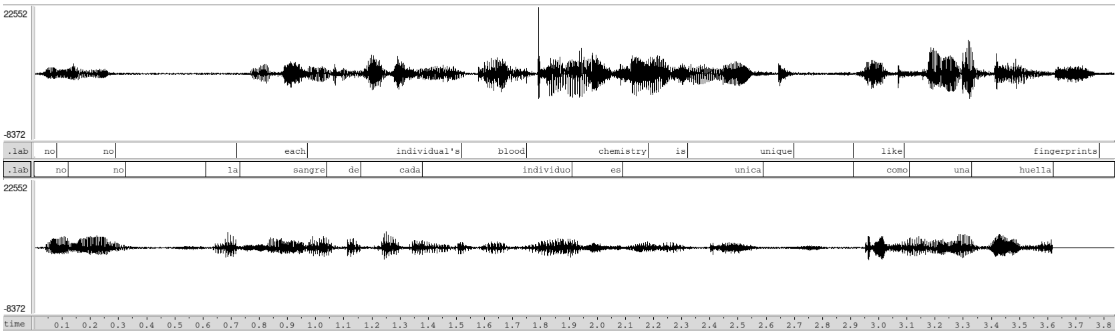
Dubbing is studied in audio-visual translation [1], a branch of translatology, and is at present typically performed manually (although supported by specialized software environments). A major focus is on producing translations that can be spoken in synchrony along with the facial movements (in particular lip and jaw movements) visible on screen. The literature [2, 3] differentiates between quantitative and qualitative aspects of synchrony in dubbing. Both are accepted to be highly relevant but quantity appears as more important than quality. *Quantity* is concerned with the coordination of time of speech and lip movements and is meant to avoid visual or auditory phantom effects. Potentially, the number of syllables or the time estimated to speak in the source and target languages (SL, TL) can be helpful indicators to find translations that enable quantitative synchrony [4].

Quality is important once quantity is established, and is concerned with matching visemic characteristics (i.e., what speech sounds looks like when pronounced) of source and target speech, such as opening angle of the jaw for vowels and lip closure for consonants (e.g., when there is a /b/ in SL, prefer a translation that features one of /m b p/ at that time over one that features /g/, to match lip closure). Quality is often characterized by the heuristic of finding a translation that

CHR 2021: Computational Humanities Research Conference, November 17–19, 2021, Amsterdam, The Netherlands

mail@timobaumann.de (T. Baumann); ashutosh.saboo96@gmail.com (A. Saboo)
https://www.timobaumann.de/work/ (T. Baumann)
0000-0003-2203-1783 (T. Baumann)

CEUR Workshop Proceedings (CEUR-WS.org)



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.`

ideal MT (Google): `No, no.  La química de la sangre de cada individuo es única, como las huellas dactilares.`

Figure 1: Example dubbing from English to Spanish in the show "Heroes" (season 3, episode 1, starting at 29'15", from [5]); MT via Google translate.

'best matches phonetically' the source language as it is visible on screen, as estimated by the human audio-visual translator. Although the idea of 'best matching phonetically' is intuitively plausible, there is a research gap on objective and computational measures for the dubbing quality of a given translation, which we aim to fill with this paper. Our long-term goal is to automatically generate a translated script which can be revoiced easily to yield dubbed film that transparently appears as if it had been recorded in the target language all along.

There is some limited recent work [4] on establishing quantitative similarity for dubbing in machine translation (MT). Here, we specifically explore the *qualitative* factors of speech sounds that may be important beyond matching syllable counts while controlling for quantity.

The need for objective measures of dubbing optimality of a given translation arises from the fact that most MT systems are trained on textual material that does not regard dubbing optimality, of which corpora are available that exceed the size of dubbed material by several orders of magnitude. Even subtitles do not fully cover dubbing characteristics. As a result, high-performance MT does not have an implicit notion of dubbing optimality and yields results that are not directly suitable for dubbing, although optimal for textual translations. It is our goal to estimate the importance of qualitative matching between SL and TL and to later add these as constraints to the translation process.

A way of enriching MT with external constraints is described in the following section and builds on heuristics that can be evaluated on partial or full translations of utterances. We use this method to balance MT for quantitative similarity as a basis for our analysis of factors that influence qualitative similarity using an ablation study that employs an adversarial classifier. Our empirical analysis confirms the importance of qualitative similarity and matching lip closures in dubbing. We find that the opening angle of the jaw is comparatively less relevant for dubbing.

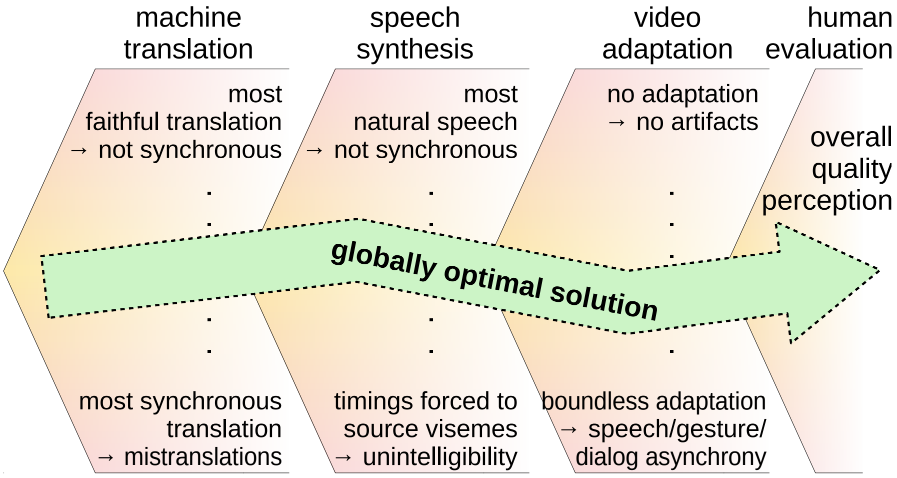

machine translation
speech synthesis
video adaptation
human evaluation
most faithful translation → not synchronous
most natural speech → not synchronous
no adaptation → no artifacts
overall quality perception
globally optimal solution
most synchronous translation → mistranslations
timings forced to source visemes → unintelligibility
boundless adaptation → speech/gesture/ dialog asynchrony

Figure 2: Machine Translation as one part of a full system for dubbing that includes speech synthesis, video adaptation, and considers the perceived quality loss of misaligned speech. Given the complexity of the task, a modular system has clear advantages over an end-to-end monolithic system but requires a notion of dubbing optimality.

2. Dubbing and Translation

Translation from one language to another aims to be a meaning-preserving conversion (typically of text but also of speech) from a source to a target language (and, to a lesser extent, from one socio-cultural context to another). Audio-visual translation adds the constraint that the target language material shall closely match the visemic characteristics of the source to give the impression that a video of the source speaker actually shows the speaker speaking TL when revoiced by a dubbing artist.

A perfect dubbing is not always possible given that the same meaning in two languages is expressed with different syntactic structures, different words, and resulting in different speech sounds (and accentuation patterns) that yield different articulatory characteristics (visemes such as the opening or closing of the lips and jaw). Thus, a *tradeoff* must be found between meaning preservation and dubbability. Figure 1 shows, as an example, one original and dubbed utterance in a TV show, as well as the machine translation of the source to the target language via Google translate. We find that MT performs quite well and yields a meaning-preserving translation, which however is substantially longer. In contrast, the dubbed version changes the syntactic structure and uses synonymy to leave out material (ignoring the 'chemistry' aspect of blood and the 'finger' aspect of the print), yielding a more dubbable text.

The translation for dubbing is clearly geared towards more easily 'dubbable' text and it is then the dubbing artist's task to speak the material in such a way that it appears as natural as possible given the video of the original speaker.

A full dubbing system that were to cover both translation *and* speech synthesis as well as potential video adaptation should yield a solution that is *globally optimized* towards the user perception as sketched in Figure 2: it can be wise to choose a sub-optimal translation to yield a better overall synchrony of the system.

Neural machine translation (NMT) has become a popular approach for MT, originally proposed by [6],[7],[8]. NMT trains a single, end-to-end neural network over parallel corpora of SL and TL pairs. Most NMT architectures belong to the encoder-decoder family [7, 9]: after encoding a SL sentence, the decoder generates the corresponding TL sentence word-by-word [7] (and possibly using attention [10] as guidance), thus in a series of locally optimal decisions. Beam search helps to approximate global optimality [11] and is a convenient lever for adding external information into the search process to steer decoding.

In previous work [4], we enriched a translation system with an external dubbing optimality scorer to yield controllable and dubbing-optimal translations, however, only for the *quantity* of TL material produced. We here explore the influence and relative importance of *qualitative* aspects in human dubbing.

3. Measures of Dubbing Optimality

Dubbing optimality is majorly governed by lip synchrony and opening angle of the jaw and, of course, quantity of speech (which is often taken as granted in the research literature on dubbing). We first describe previous results in literature on quantitative measures [4] and then how we use these to establish and analyze qualitative measures based on an adversarial approach in which we aim to train a classifier that attempts to differentiate human gold-standard dubbing from quantitatively re-balanced MT. If this classifier performs poorly, then MT is more difficult to distinguish from gold-standard translation. We then validate various qualitative factors, such as importance of the opening angle of the jaws, closure of the lips, prosody and word boundaries by performing ablation studies with this classifier.

3.1. Enforcing quantitative similarity of phonetic material

In order to allow for an even approximate lip synchrony, the duration of the revoicing should match that of the original speech, in order to avoid audio-visual phantom effects. These can be seemingly 'stray' movements of the mouth in the dubbed version if there is too little to speak, or audible speech while the articulators are not moving if there is too much material to speak. As in [4], we use the number of syllables as the primary indicator of visemic similarity via the standard hyphenation library Pyphen[1] for counting the number of syllables in SL as well as candidates in TL and take the relative difference of the two as the similarity metric.[2] We then rescore the NMT's output by the similarity metric using some weight α. For the experiments below, we report results across the range for all $0 \leq \alpha < 1$; [4] found an $\alpha \sim 0.3$ to yield the best balance between BLEU score (a measure for translation quality, [12]) and quantitative similarity. We will therefore highlight the results in the $0.2 \leq \alpha \leq 0.5$-range.

[1] Pyphen: https://pyphen.org.

[2] This works well for English-Spanish translation; other language pairs may require other quantitative measures, e. g. mora-driven languages.

3.2. Qualitative similarity of phonetic material

Qualitative similarity, i.e., the dubbing artist's voice closely matching the articulatory movements visible on screen, is also highly desirable, beyond quantitative matching. Phonetic aspects of consonants and vowels, such as lip closure and opening angle of the jaw have been reported as being relevant for translations that can be lip-synchronously dubbed, as well as supra-segmental aspects such as prosodic phrasing [13].

We explore the relative importance of these aspects using an ablation experiment on a classifier that is trained to differentiate human dubbing translations from NMT translations (rescored to yield quantitative similarity). For MT that is ideal for dubbing, this classifier performs poorly, whereas it performs better, the more easily gold-standard dubbing and MT can be differentiated. In essence, if the features that the classifier is deprived of in an ablation setting are not relevant, the performance of the classifier should not reduce (or even improve); if however the classifier is deprived of relevant features, we expect a performance degradation.

We here explore the importance of phonetic/visemic characteristics via different simplifications of the textual material that we feed to the classifier. For example, when we leave out all whitespace and punctuation, the classifier is deprived of morphological and prosodic structure features. If its performance drops (relative to the full input) this reflects the influence on dubbability. Note that Spanish, TL in our experiments, has highly regular grapheme-phoneme correspondences which allows to base our experiment directly on ablations of the graphemic representations.

3.3. Text simplifications for ablation study

We use the following simplifications in addition to passing the full text to the classifier (**full**):

no punctuation tests the influence of phrasing as far as expressed by punctuation in text,

no whitespace (in addition to no punctuation): tests the importance of word boundaries; we hypothesize that word boundaries are of little relevance when dubbing as they are not clearly observable in continuous speech.

In addition to whitepace and punctuation removal:

vowels vs. C we replace all consonants by "*C*" but not the vowels, to test how *opening angle of the jaw* alone (which, to a large extent depends on the vowel produced) helps the model,

consonants vs. V we replace all vowels by "*V*"; as a result, the opening angle of the jaw is *not* observable to the model,

bilabials vs. C vs. V we replace all vowels ("*V*") and consonants ("*C*") except for bilabials ("b, p, m") which are not replaced; thus, *lip closure* is the only consonant characteristic observable to the model,

C vs. V to test if syllable structure alone is valuable for dubbing optimality.

3.4. Model and Training Procedure

For our method, we use an encoder-*en*coder architecture with siamese parameters for the two TL candidates to be compared based on the SL sentence as depicted in Figure 3. We first encode the SL sentence bidirectionally character-by-character using an RNN based on GRUs [14]. Each TL sentence (gold-standard dubbing and NMT) is then also encoded via its characters and GRU

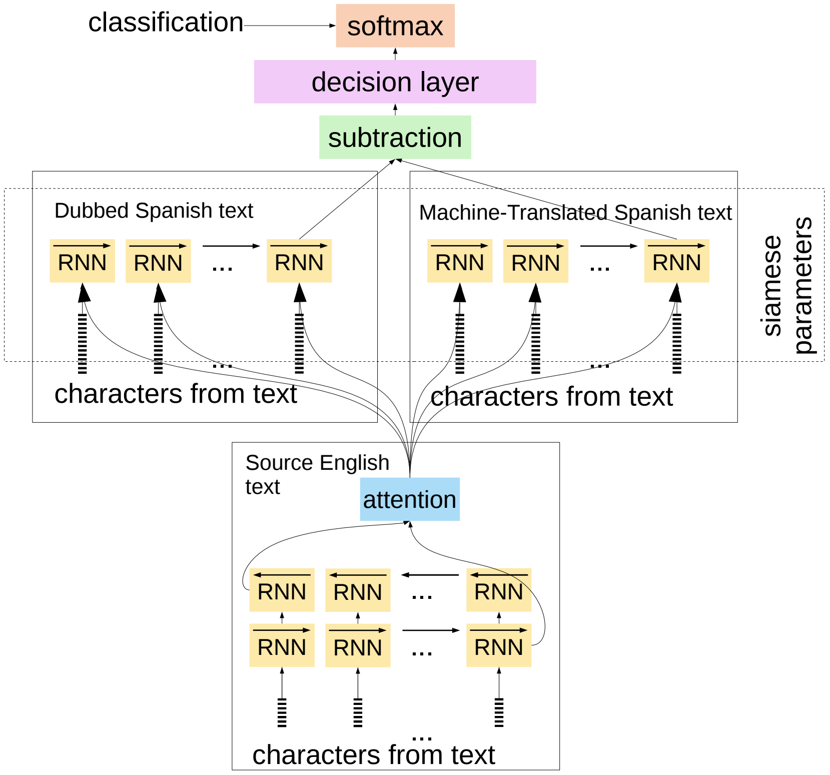


Figure 3: Siamese encoder-encoder classifier for comparing the 'dubbability' of two TL candidates given a SL sentence.

units, which take as additional input the attended-to output from the SL encoder. This attention layer conditions on the TL recurrent state and we expect that it will be able to learn the relation of source words to target words, or even of textually observable phonetic sub-word features (like bilabial consonants), thereby computing the matching of TL and corresponding SL material in one encoding. We train the TL encoders for each candidate TL sentence in a *siamese* setup [15], where parameters are shared, and then subtract the resulting representations in order to yield the difference between the two candidates. The multi-dimensional difference is then passed to a final decision layer. We train this setup and report results for each kind of experimental text simplification in order to find out the value of different kinds of information (expressed as relative performance penalty of leaving out the corresponding feature).

4. Data and Experiments

We use the HEROes corpus [5], 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 utterance pairs in 9.5 hours of speech that are based on forced alignment of video subtitles to the audio tracks. The results have been manually checked and re-aligned to each other.

We trained an NMT system on the OpenSubtitles corpus [16] with fairseq [17] with settings as described in [4]. The NMT yields a BLEU score of 26.31 on our data which degrades to 25.43 after rescoring with an α value of 0.3. We produce rescored translation results for all α weighings.

Our classifier is trained on the triples of SL, TL dubbing, and TL NMT candidate for all text simplifications and all values of α, using 10-fold cross validation on the corpus. We report the overall accuracy for each classification setting as well as the *standard deviation* across folds. The classifier is implemented in DyNet [18].[3] We use 20-dimensional character encodings, 20-dimensional RNN states, and 20-dimensional attention. We train with the Adam method [19] for 10 iterations and using a dropout of 0.2. This is not the result of an extensive hyper-parameter search but a mixture of best guesses and experience.

5. Results and Discussion

The results of the experiments are presented in Figure 4, where the x-axis denotes the control over quantitative similarity (via the rescoring factor α) and the y-axis denotes the classifier accuracy percentage (an accuracy around 50 % indicates that the classifier is unable to differentiate true dubbing). The standard deviation across folds for each of the values is in the range of 1-4 percentage points. Thus, while we have not performed significance tests across folds, we feel confident that differences reported below are likely 'real'.

The figure shows that the classifier performs best with **full** input. Translations in the relevant α-range that yields good quantity (but still reasonable translations) are more difficult, indicating

[3] The code and full experimental data are available at https://github.com/timobaumann/duboptimal.

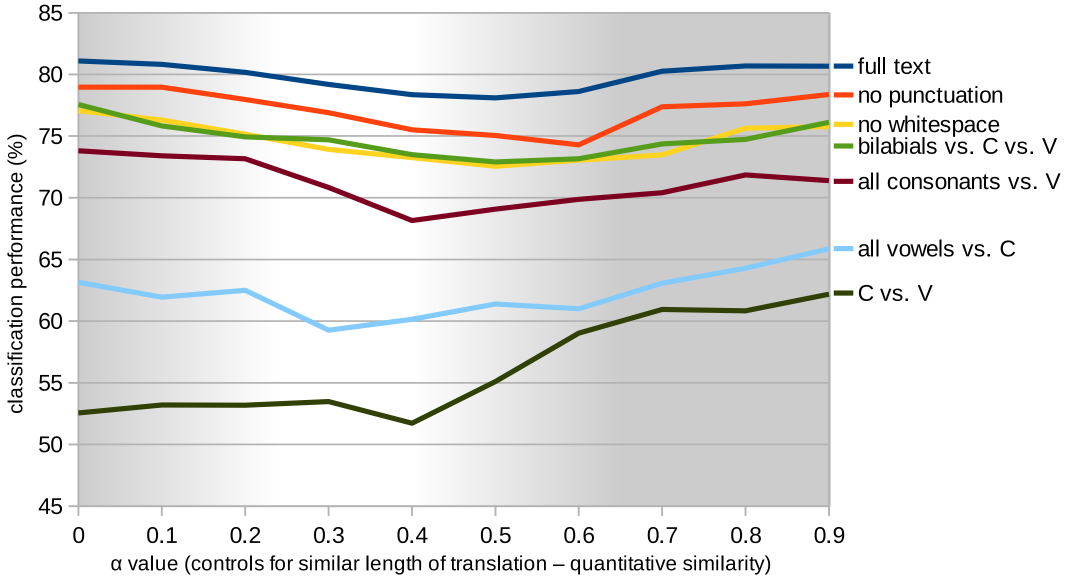


Figure 4: Plots of classifier performance (in %) for the ablation settings across all values for the control of quantitative similarity (α). The less relevant ranges for α are shaded in gray.

that the adversarial task is particularly difficult under these circumstances. Only retaining the *syllabic structure* (**C vs. V**) yields the worst performance (only marginally above chance) and can be considered as hardly helpful – this debunks the common misunderstanding that all that needs to be kept in dubbing is the right number of syllables.

Leaving out **punctuation** and **whitespace** has some but not radical effects (probably within the margin of error) indicating that both *prosodic phrasing and lexicomorphology* do not need to be strictly retained while translating for dubbing; instead, these allow for some degree of freedom to better match other aspects.

Regarding lip closure and jaw movements, we find that (a) removing vowel information (**all consonants vs. V**) only hurts a little, whereas retaining only vowel information (**all vowels vs. C**) leads to a considerable performance drop. From this we conclude that matching the *opening angle of the jaw* is at least not done through vowel choice, and may be less critical than described in the literature.

(b) In contrast, removing vowel information and even reducing the consonant information to whether it's bilabials or not (**bilabials vs. C vs. V**) yields surprisingly high performance (even better than retaining all consonants, possibly because the model learns more easily with fewer input symbols) which indicates that *lip closure* is indeed observed tightly in the dubbing corpus.

6. Summary and Conclusion

We have studied the importance of *aspects of qualitative similarity* in dubbing, in particular when quantitative similarity is controlled for. The literature in translatology for dubbing posits that jaw movement and lip closure are critical aspects to be observed in dubbing. However we found no study prior to ours to investigate the relative importance of these aspects, measure the importance in an objective way, or investigate the importance of further potential influences such as lexicomorphology and prosodic phrasing. We have presented an ablation study to try to find those features that are particularly relevant to discern qualitatively ignorant NMT from true dubbing using a neural siamese classifier.

We find that we can confirm the importance of matching lip closures in dubbing. We therefore conclude that good dubbing requires a good matching of lip closures. By comparison, the opening angle of the jaw (which intrinsically varies between different vowel types) appears to be far less important. Our quantification of dubbing constraints leads the way towards a further optimization of machine translation for dubbing as it enables the training or adaptation procedure to take into account these constraints. Additionally, our classifier could directly be included into NMT via an adversarial learning procedure.

Our experiments yield objective evidence about the importance of qualitative aspects for dubbing. However, we acknowledge that further research is needed. In particular, our study is restricted to the textual form and does not include the speech signal in the corpus, which would allow for a better temporal alignment analysis. Furthermore, our analysis uses the full corpus rather than only those parts where the face is visible on-screen (and hence qualitative aspects matter).[4] Finally, the ultimate evaluation gold standard for dubbing would be a user study that compares different dubbing alternatives. This could be used to directly optimize towards

[4] A tool for on- vs. off-screen detection has become available only very recently [20].

human judgements of dubbing alternatives (which might even differ with user preferences), or information retention for educational material to estimate distraction of less-than-ideal dubbing.

More broadly, we believe that ablation studies are a suitable tool in computational humanities research as they can help to objectively analyse and quantify the various aspects of existing humanistic theories for complex phenomena such as in this study.

Acknowledgments

The second author's work was performed during an internship at Universität Hamburg which was partially supported by Volkswagen Foundation under the funding codes 91926 and 93255. We thank the anonymous reviewers for their insightful remarks.

References

[1] P. Orero, Topics in audiovisual translation, volume 56, John Benjamins Publishing, 2004.

[2] X. Martínez, Film dubbing: Its process and translation, in: P. Orero (Ed.), Topics in Audiovisual Translation, John Benjamins Publishing, 2004, pp. 18–22.

[3] F. Chaume, Audiovisual translation: Dubbing, St. Jerome Publishing, 2012.

[4] A. Saboo, T. Baumann, Integration of dubbing constraints into machine translation, in: Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), Association for Computational Linguistics, Florence, Italy, 2019, pp. 94–101. URL: https://www.aclweb.org/anthology/W19-5210. doi:10.18653/v1/W19-5210.

[5] A. Öktem, M. Farrús, A. Bonafonte, Bilingual prosodic dataset compilation for spoken language translation, in: Proc. IberSPEECH 2018, ISCA, 2018, pp. 20–24. doi:10.21437/IberSPEECH.2018-5.

[6] N. Kalchbrenner, P. Blunsom, Recurrent continuous translation models, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2013, pp. 1700–1709. URL: https://aclanthology.org/D13-1176.

[7] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, in: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14, MIT Press, Cambridge, MA, USA, 2014, pp. 3104–3112.

[8] K. Cho, B. van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder–decoder approaches (2014) 103–111. URL: https://aclanthology.org/W14-4012. doi:10.3115/v1/W14-4012.

[9] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation (2014) 1724–1734. URL: https://aclanthology.org/D14-1179. doi:10.3115/v1/D14-1179.

[10] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Proceedings of International Conference on Learning Representations (ICLR), volume abs/1409.0473, 2014. URL: http://arxiv.org/abs/1409.0473. arXiv:1409.0473.

[11] X. Hu, W. Li, X. Lan, H. Wu, H. Wang, Improved beam search with constrained softmax for NMT, in: Proceedings of Machine Translation Summit XV: Papers, Miami, USA, 2015, pp. 297–309. URL: https://aclanthology.org/2015.mtsummit-papers.23.

[12] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[13] A. Öktem, M. Farrús, A. Bonafonte, Prosodic phrase alignment for machine dubbing, in: Proc. Interspeech 2019, 2019, pp. 4215–4219. URL: http://dx.doi.org/10.21437/Interspeech.2019-1621. doi:10.21437/Interspeech.2019-1621.

[14] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1724–1734. URL: https://aclanthology.org/D14-1179. doi:10.3115/v1/D14-1179.

[15] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, R. Shah, Signature verification using a “siamese" time delay neural network, in: J. Cowan, G. Tesauro, J. Alspector (Eds.), Advances in Neural Information Processing Systems, volume 6, Morgan-Kaufmann, San Francisco, USA, 1994, pp. 737–744. URL: https://proceedings.neurips.cc/paper/1993/file/288cc0ff022877bd3df94bc9360b9c5d-Paper.pdf.

[16] P. Lison, J. Tiedemann, OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles, in: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), European Language Resources Association (ELRA), Paris, France, 2016. URL: https://aclanthology.org/L16-1147.

[17] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, M. Auli, 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), Association for Computational Linguistics, Minneapolis, USA, 2019, pp. 48–53. URL: https://www.aclweb.org/anthology/N19-4009. doi:10.18653/v1/N19-4009.

[18] G. Neubig, C. Dyer, Y. Goldberg, A. Matthews, W. Ammar, A. Anastasopoulos, M. Ballesteros, D. Chiang, D. Clothiaux, T. Cohn, K. Duh, M. Faruqui, C. Gan, D. Garrette, Y. Ji, L. Kong, A. Kuncoro, G. Kumar, C. Malaviya, P. Michel, Y. Oda, M. Richardson, N. Saphra, S. Swayamdipta, P. Yin, Dynet: The dynamic neural network toolkit., CoRR abs/1701.03980 (2017). URL: http://arxiv.org/abs/1701.03980. arXiv:1701.03980.

[19] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, (ICLR 2015), 2015. URL: http://arxiv.org/abs/1412.6980.

[20] S. Nayak, T. Baumann, S. Bhattacharya, A. Karakanta, M. Negri, M. Turchi, See me speaking? differentiating on whether words are spoken on screen or off to optimize machine dubbing, in: ICMI Companion: 1st Int. Workshop on Deep Video Understanding, ACM, 2020, pp. 130–134. doi:10.1145/3395035.3425640.