INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Unsupervised phonetic and word level discovery for speech to speech translation for unwritten languages

Steven Hillis, Anushree Prasanna Kumar, Alan W Black

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
{shillis, apkumar, awb}@cs.cmu.edu

Abstract

We experiment with unsupervised methods for deriving and clustering symbolic representations of speech, working towards speech-to-speech translation for languages without regular (or any) written representations. We consider five low-resource African languages, and we produce three different segmental representations of text data for comparisons against four different segmental representations derived solely from acoustic data for each language. The text and speech data for each language comes from the CMU Wilderness dataset introduced in [1], where speakers read a version of the New Testament in their language. Our goal is to evaluate the translation performance not only of acoustically derived units but also of discovered sequences or "words" made from these units, with the intuition that such representations will encode more meaning than phones alone. We train statistical machine translation models for each representation and evaluate their outputs on the basis of BLEU-1 scores to determine their efficacy. Our experiments produce encouraging results: as we cluster our atomic phonetic representations into more word-like units, the amount information retained generally approaches that of the actual words themselves.

Index Terms: speech-to-speech, machine translation, segmentation, unit discovery, low-resource, unwritten languages, Wilderness

1. Introduction

Our work is motivated by the idea of speech-to-speech translation for unwritten languages. Many speech-to-speech translation pipelines today consist of machine translation systems trained on text corpora preceded by a speech recognition component and succeeded by a speech synthesis component. However, this pipeline breaks down when either language has no adequate written form. This is true when a language lacks a written form altogether, but it is also true whenever a language has an inconsistent or inadequate written form. Indigenous languages without a literary tradition, regional dialects without widespread standardization, and conversational dialects without a need for written communication can all be characterized as such. Furthermore, we can imagine a low resource language documentation scenario where audio data is recorded but where a writing system is unknown or unavailable to the field worker. In each of these cases, the standard speech-to-speech pipeline will fail; nonetheless, the technology would be as or more socially and economically valuable to these languages as to any other.

We can reconcile these scenarios by inducing the necessary intermediate textual representations from the unwritten audio itself, in a form of primitive writing system discovery. Once a textual representation has been induced, translation can proceed as normal, with the representation fulfilling the same function as a transcription of written "words". [2] first demonstrated the viability of this approach. When the textual representation consists only of phones as atoms of meaning, translation quality predictably suffers: the semantic density of words and atomic phones differs on orders of magnitude. While we cannot simply construct words from of a phonetic transcription, constructing some higher-order, more word-like representation from the simple transcription is an intuitive approximate.

In pursuit of this idea, we investigate methods to approach the semantic density of orthographic transcriptions by clustering raw phonetic representations of real speech data into more word-like units. We use statistical machine translation to translate the experimental representations of the source language speech to English, our consistent target language. Translation quality is measured for our purposes by BLEU-1 scores, and we assert that a representation producing a better translation has preserved more information from the source speech data.

2. Related Work

Historically, the inventory of a language's phonemes was handcrafted and refined by expert linguists. Today, there exist many methods to automatically discover an phone inventory. This is not the same as a linguistically-motivated phonemic inventory, but findings in [3] suggest that it makes a functional approximate. Many approaches to identifying this inventory involve cross lingual transfer from high-resource languages to lower-resource ones. [4] uses a model trained on a high resource language to predict Articulatory Features for a low resource language, which can then be clustered into phones. This is taken a step further in [5], which derives a phoneme-like segmental representation of the Articulatory Features. This approach is again advanced in [6], where BiLSTMs are used to first identify segment boundaries before extracting and clustering the Articulatory Features.

With similar cross-lingual motivations, [7] produces phonetic transcriptions of low-resource speech by first using English language and acoustic models for ASR, whereby the raw transcription is iteratively refined to more closely align with the speech data. Extending their approach to produce higher-level units, in [8] they cluster the best phonetic transcription using Festival heuristics for syllables and a CRF model for words. A similar approach with an intermediate step is proposed in [9], which extrapolates and adds the low-resource acoustic units that are absent from the high-resource acoustic model before the initial transcription.

Taking a more probabilistic approach towards Acoustic Unit Discovery, [10] innovates on the inference process used in nonparametric phone-loop Bayesian models first proposed by [11] for unit discovery in speech, exchanging Gibbs Sampling for Variational Bayes. They follow this work by refining the process with a Hierarchical Pitman-Yor based bigram

http://dx.doi.org/10.21437/Interspeech.2019-3026

langauge model [12]. [13] combines the process of segmenting and clustering speech data with a efficient approximation of a Bayesian model, representing word segments as acoustic word embeddings and clustering them through K-means. They present a supplemental approach for fully unsupervised acoustic word embeddings in [14].

Language independent bottleneck features have also been used to represent speech units [15], [16]. In [17], both LDA and bottleneck features are used to augment the model in [10]. Bottleneck features are learned pairwise over frames in [18] to model word-like segments. Approaches to finetuning bottleneck features are explored in [19], which also comparatively demonstrates their effectiveness.

For a more detailed survey of many of these methods, please reference [20].

Speech translation systems today are becoming more robust to the absence of a stable writing system as they move towards end-to-end models. [21] and [22] do not require an explicit intermediate representation of speech at all—it is latent to the neural model—and a subsequent work uses explicit transcription only during training [23]. The lessened dependency on an explicit intermediate representation of a neural model comes in turn with a demand for data beyond the reality for most unwritten languages, although recent work ([24] and [25]) casts neural models for low resource languages as increasingly attainable.

Particularly relevant to our work is that of [26], which makes identically motivated investigations into optimal methods for clustering phonetic transcriptions of speech with respect to translation results. Their experiments were constrained to Europarl textual data, from which they were forced to simulate acoustic representations. Their acoustic data were generated using speech synthesizers, which means that they did not have the variation of natural speech. Furthermore, the languages in their experiments (English and Spanish) have both a closer relationship and far more resources than would normally be available for such work. We now have the opportunity to work with genuine and deeply different acoustic data, thanks to the Wilderness dataset.

3. Data

The speech and text data for these experiments comes from the CMU Wilderness dataset. The five languages we work with are:

- Avatime/Sidame (AVN)
- Oromo (ORM)
- Hausa (HAU)
- Masaba/Gisu (MYX)
- Lunyole/Nyule (NUJ)

ORM and HAU are languages in the Afroasiatic family, and the remainder are in the Niger-Congo language family. All five languages have over 20 hours of recorded speech. The Wilderness data ranks the quality of the alignment based on how well a grapheme-based speech synthesizer trained on the data can reconstruct a held out test set. We selected languages with a good score because we want to be able to compare the translation results of our acoustically-informed representations with those of our Written Words representation.

MCD stands for Mel Cepstral Distortion [27] is a scaled Euclidean distortion metric used to compare a synthesized utterance with a held out original. Lower is better. In the CMU Wilderness data MCD numbers less that 7 are typically good (it is easy to understand the synthesized output), and when under 6 are very good (the synthesis quality is very easy to understand). Also for ease of comparison all chosen languages use a (mostly) latin-based alphabet. Nothing in our work depends on that, but for it provides initial simplicity when reading the text.

Table 1: *Duration and MCD for each language*

	Duration (h:m:s)	**MCD**
AVN	23:52:43	6.33
ORM	23:48:29	6.48
HAU	20:40:13	5.74
MYX	22:59:17	6.14
NUJ	24:12:56	5.93

4. Methods

We represent the utterances in six different ways, in addition to the oracle orthographic representation of words. We examine two approaches to generating symbolic representations and three approaches to clustering the individual units of these representations into higher-order meaning representations. All methods are unsupervised and applicable for low-resource languages.

Table 2: *Examples of different representations for MYX*

Written Words	Aryo Saulo
Text-Based Phones	pau eh1 r ih1 ow0 s ao1 l ow0
BPE T-BP	paueh1rih1ow0@@ sao1low0@@
Audio-Based Phones	HH AH T EH S AH UW N
BPE A-BP	HHAH@@ TEH@@ SAH@@ UWN@@
Ngram A-BP	HH_AH T_EH S_AH UW_N
Goldwater A-BP	HH AHT EH SAH UWN

4.1. Written Words

The words read by the speakers serve as an upper bound to contextualize the performance of our methods. It is plausible that there exists a higher-order representation of speech that could exceed the semantic density of words for a language's writing system, but we leave that to future work. Written Words is the approach to intermediate representation of speech data for machine translation generally taken for languages with adequate written forms.

4.2. Text-Based Phones

The Festvox software package provides a universal pronunciation model, based on [28] with substantial additions to cover the 700 languages in the Wilderness dataset. These predicted phones, using the X-SAMPA phoneset, remain somewhat simplistic; all the same, they are based on the actual text, while the following sets are derived from the acoustics without any constraints imposed by text. There is no acoustic consideration to creating this Text-Based Phones representation, but it illustrates a lower-level representation of our oracle Written Words approach. This representation still produces a result close to the oracle words.

4.3. Audio-Based Phones

In this approach, we use the method and software introduced by [7] to discover a phonetic representation of speech data. This approach requires only speech data; it is a representation informed purely by acoustics.

We would like to be clear that we do not make any assertions about the quality, integrity, or efficacy of this particular method for phone discovery: the application was available to us, and it meets our need of an atomic, phonetic representation of speech, but there are certainly other methods we could have used. Ultimately, we are not concerned with using the best method for atomic phone discovery; we are instead interested in the best way to improve whichever phonetic-like representation is available by generating units comparable in function to written words from its atomic symbols.

This approach discovers phones by iteratively rebuilding and decoding acoustic models and transcripts from parallel speech-transcription data. In order to do so, it first produces an initial transcript of the speech data by doing ASR using a cross-lingual acoustic model. In our case, the seed acoustic model was an English Wall Street Journal model from Sphinx with the CMU-DICT phone set. That initial transcript is very rawespecially for very different languagesbut it makes the iterative process which follows possible. Again, there may be better ways to derive such phonetic-like units (indeed, we welcome alternative methods for deriving initial phone-level segmentations), but we know this technique is at least reasonable.

The new speech-transcription corpus is used in parallel to train an acoustic model for the target language, which is used in turn to re-decode the audio data. This continues for 10 iterations. After every iteration, a new phonetic transcript and a new target acoustic model have been produced. Each time, these two are used to build a synthetic CLUSTERGEN [29] voice using Festvox. The statistical parametric synthesis is used with consideration to the noise inevitably present in the data. The newly built voice is evaluated over a heldout test set on the basis of MCD scores, where the best automatically and iteratively generated transcription is determined to be the one producing a synthetic voice with the lowest distortion score. This phonetic transcription becomes our atomic phonetic representation for the target language.

4.4. Clustering By Subword Units (BPE)

We use the unsupervised segmentation technique [1] proposed by [30] to merge atomic phones into more word-like units. Clustering by subword units using byte pair encoding [31] works by identifying the most common pairs of consecutive characters, which are then merged together. Subword units are delimited by '@@'. We did not change any default hyperparameters, applying all learned merges with no vocabulary threshold.

4.5. Clustering By Ngrams

In a similar spirit as clustering by subword units using BPE, we cluster atomic phones on the basis of ngram frequency using the method proposed in [26]. We make p passes through our corpus of discovered phones, and for each pass we tabulate the most frequent ngrams. The top k most frequent ngrams (which were almost exclusively bigrams for our atomic phones) are merged into a single unit, separated by an underscore. We took the hyperparameters presented with the method in [26] as-is, with k = 10 and p = 25. Had we tuned the hyperparameters for our data, we may have seen better evaluation results; however, being that we are entirely more interested in relative performance of all methods than in absolute performance of each, we took them as stock.

4.6. Clustering By the Goldwater Approach

The Goldwater approach for word segmentation uses a Dirichlet process/Gibbs sampler algorithm to build a whole word acoustic model to cluster unlabeled speech, as proposed in [32]. We use this approach to cluster our discovered phones, taking their optimal hyperparamters as-is. We made a single modification to their algorithm to account for the length of the utterances in our Discovered Phone representation. They established a threshold of 500 for utterance length, which we had to extend to accommodate our data.

5. Experimental Setup

5.1. Text Matching

While we are pleased to be working with real speech data, doing so becomes much messier than working with synthetic data. In this case, we have the speech and text data from the Wilderness dataset, which comes from Bible.is. Separately, we have a parallel corpus of foreign-English bible data scraped from Bible.com (the same corpus used in [33]), which is unfortunately not assigned verse numbers. This forced us to calculate best matches between the Wilderness text data and the foreign text of the parallel translations in order to create new parallel corpora with our symbolic representations for evaluation.

We tried several different methods for doing this text matching. Simple edit distance would have been intractably slow: we had around 10,000 utterances for each language, and the foreign text in the parallel file averaged approximately 8000 utterances, leaving us with almost 80 million n-cross-n pairs to evaluate.

So, we tried a TF-IDF approach over ngrams [2] where we compared the strings on the basis of cosine distance, which was sufficiently fast but was only able to produce around 2500 matches per language.

Then, we substituted sent2vec [3] embeddings in for TF-IFD and again compared on the basis of lowest cosine distance. With this method, we were able to match approximately 3000 verses for each language using the Actual Word representations, which we sanity-checked by keyword matching proper nouns between the foreign and English texts. There was a significant spread across the number of matches we were able to make for each language, and they were all certainly lower than we would have liked, but we determined them to be sufficient for our experiments.

Table 3: *Number of matched utterances for each language*

AVN	ORM	HAU	MYX	NUJ
3232	3301	3318	2853	2966

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

[2] https://bergvca.github.io/2017/10/14/super-fast-string-matching.html

[3] https://github.com/epfml/sent2vec.git

5.2. Translation

We use the Moses statistical translation system introduced in [34] to evaluate our results. We apply their tokenization and truecasing, and tune using their implementation of Minimum Error Rate Training (MERT) on 300 heldout development verses. We produce test results on 300 separately heldout test verses, always translating from one of our foreign languages into English. We acknowledge the instability inherent in evaluating translation results on a corpus of this size, but we are working within the confines of our data.

6. Results

Generally, machine translation results are evaluated using the BLEU metric [35]. However, this metric imposes fluency constraints based on higher-order ngram scores which are not appropriate for our purposes. Instead, we evaluate on the basis of BLEU-1 scores, utilizing unigrams to measure the adequacy of information retained in translation. These results are far more salient to our investigation of the relative semantic densities of symbolic representations of speech.

Table 4: *Translation Results (BLEU-1)*

	AVN	**ORM**	**HAU**	**MYX**	**NUJ**	**Mean**
Written Words	0.18	0.16	0.19	0.18	0.17	0.18
Text-Based Phones	0.17	0.06	0.18	0.15	0.13	0.14
BPE T-BP	0.12	0.09	0.12	0.12	0.11	0.11
Audio-Based Phones	**0.08**	0.02	0.02	0.02	0.10	0.05
BPE A-BP	0.08	**0.13**	0.09	0.13	0.09	0.10
Ngram A-BP	0.04	0.12	**0.17**	**0.17**	**0.16**	**0.13**
Goldwater A-BP	0.04	0.09	0.15	0.14	0.14	0.11

7. Discussion

For four out of our five languages, we were able to improve upon the translation results of our symbolic representations of audio by clustering the atomic phonetic representations into higher-order, word-like transcriptions. For three out of these four, the most effective approach was the ngram clustering technique from [26]; for the remainder, ORM, the most effective approach was the BPE subword segmentation of the discovered phones. AVN contradicts our hypothesis, as the information retained after translation refused to improve from the Discovered Phone representation.

In generating these experiments, we were forced to confine ourselves to several uncertainties. Firstly, we cannot be certain that the Wilderness source data originates from the same version of the New Testament as the target English text (or as the foreign text we had to match the Wilderness data with, for that matter). Similarly, different translations of the same New Testament often separate verses differently. Furthermore, we cannot be certain that our text matching was optimal: over a relatively small corpus and a concentrated domain, there were almost always multiple candidate matches with very similar cosine distances in the embedding space. Finally, we cannot be certain that our translation results are stable given our data. Despite these uncertainties, the trend of the results seems clear: by clustering atomic representations of speech in unwritten languages, we can approach the semantic density of traditionally transcribed words. Our results strongly indicate that building models with automatically clustered sequences of acoustic units improves translations.

In the future, we envision creating a lexicon of word-like speech representations and then leveraging that lexicon to train a language model, which could in turn be used as a prior to constrain whichever phone discovery method we chose to employ. Then, after clustering a new transcription into word-like units, the new and improved lexicon could be used to similarly improve phone discovery, and so-forth. In this way, we hope to move towards the translation performance of written words as an intermediate representation for speech-to-speech translation for languages without that luxury.

8. References

[1] A. W. Black, "CMU wilderness multilingual speech dataset," in *ICASSP*, 2019.

[2] L. Besacier, B. Zhou, and Y. Gao, "Towards speech translation of non written languages," in *2006 IEEE Spoken Language Technology Workshop*. IEEE, 2006, pp. 222–225.

[3] T. Kempton and R. K. Moore, "Discovering the phoneme inventory of an unwritten language: A machine-assisted approach," *Speech Commun.*, vol. 56, pp. 152–166, Jan. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2013.02.006

[4] P. K. Muthukumar and A. W. Black, "Automatic discovery of a phonetic inventory for unwritten languages for statistical speech synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2594–2598.

[5] P. Baljekar, S. Sitaram, P. K. Muthukumar, and A. W. Black, "Using articulatory features and inferred phonological segments in zero resource speech processing," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] M. Müller, J. Franke, A. Waibel, and S. Stüker, "Towards phoneme inventory discovery for documentation of unwritten languages," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5200–5204.

[7] S. Sitaram, S. Palkar, Y.-N. Chen, A. Parlikar, and A. W. Black, "Bootstrapping text-to-speech for speech processing in languages without an orthography," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7992–7996.

[8] S. Sitaram, G. K. Anumanchipalli, J. Chiu, A. Parlikar, and A. W. Black, "Text to speech in new languages without a standardized orthography," in *Eighth ISCA Workshop on Speech Synthesis*, 2013.

[9] O. Scharenborg, F. Ciannella, S. Palaskar, A. Black, F. Metze, L. Ondel, and M. Hasegawa-Johnson, "Building an asr system for a low-resource language through the adaptation of a high-resource language asr system: Preliminary results," *Proceedings of ICNLSSP, Casablanca, Morocco*, 2017.

[10] L. Ondel, L. Burget, and J. ernock, "Variational inference for acoustic unit discovery," *Procedia Computer Science*, vol. 81, no. C, pp. 80–86, 2016.

[11] C.-y. Lee and J. Glass, "A nonparametric bayesian approach to acoustic model discovery," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 40–49.

[12] L. Ondel, L. Burget, J. ernock, and S. Kesiraju, "Bayesian phonotactic language model for acoustic unit discovery," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 5750–5754.

[13] H. Kamper, K. Livescu, and S. Goldwater, "An embedded segmental k-means model for unsupervised segmentation and clustering of speech," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 719–726.

[14] H. Kamper, "Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models," *arXiv preprint arXiv:1811.00403*, 2018.

[15] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Twelfth annual conference of the international speech communication association*, 2011.

[16] K. Veselỳ, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 336–341.

[17] C. Liu, J. Yang, M. Sun, S. Kesiraju, A. Rott, L. Ondel, P. Ghahremani, N. Dehak, L. Burget, and S. Khudanpur, "An empirical evaluation of zero resource acoustic unit discovery," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5305–5309.

[18] Y. Yuan, C.-C. Leung, L. Xie, H. Chen, B. Ma, and H. Li, "Extracting bottleneck features and word-like pairs from untranscribed speech for feature representation," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 734–739.

[19] E. Hermann and S. Goldwater, "Multilingual bottleneck features for subword modeling in zero-resource languages," *arXiv preprint arXiv:1803.08863*, 2018.

[20] O. Scharenborg, L. Besacier, A. Black, M. Hasegawa-Johnson, F. Metze, G. Neubig, S. Stüker, P. Godard, M. Müller, L. Ondel *et al.*, "Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the speaking rosetta jsalt 2017 workshop," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4979–4983.

[21] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *arXiv preprint arXiv:1612.01744*, 2016.

[22] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," *arXiv preprint arXiv:1703.08581*, 2017.

[23] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6224–6228.

[24] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Low-resource speech-to-text translation," *arXiv preprint arXiv:1803.09164*, 2018.

[25] ——, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," *arXiv preprint arXiv:1809.01431*, 2018.

[26] A. Wilkinson, T. Zhao, and A. W. Black, "Deriving phonetic transcriptions and discovering word segmentations for speech-to-speech translation in low-resource settings." in *INTERSPEECH*, 2016, pp. 3086–3090.

[27] T. Toda, A. Black, and K. Tokuda, "Voice converstion based on maximum-likelihood estimation of speech parameter trajectory," *IEEE Transaction of Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2236, 2007.

[28] T. Qian, K. Hollingshead, S.-y. Yoon, K.-y. Kim, and R. Sproat, "A Python toolkit for universal transliteration." in *LREC Malta*, 2010.

[29] A. W. Black, "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling," in *Ninth International Conference on Spoken Language Processing*, 2006.

[30] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *arXiv preprint arXiv:1508.07909*, 2015.

[31] P. Gage, "A new algorithm for data compression," *C Users J.*, vol. 12, pp. 23–28, 1994.

[32] S. Goldwater, T. L. Griffiths, and M. Johnson, "A bayesian framework for word segmentation: Exploring the effects of context," *Cognition*, vol. 112, no. 1, pp. 21–54, 2009.

[33] C. Malaviya, G. Neubig, and P. Littell, "Learning language representations for typology prediction," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2529–2535. [Online]. Available: https://www.aclweb.org/anthology/D17-1268

[34] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, 2007, pp. 177–180.

[35] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.