Multimodal Word Discovery and Retrieval with Phone Sequence and Image Concepts

Liming Wang[1], *Mark Hasegawa-Johnson*[1,2]

[1]Department of Electrical and Computer Engineering, University of Illinois, Urbana Champaign
[2]Beckman Institute, University of Illinois, Urbana Champaign

lwang114@illinois.edu, jhasegaw@illinois.edu

Abstract

This paper demonstrates three different systems capable of performing the multimodal word discovery task. A multimodal word discovery system accepts, as input, a database of spoken descriptions of images (or a set of corresponding phone transcripts), and learns a lexicon which is a mapping from phone strings to their associated image concepts. Three systems are demonstrated: one based on a statistical machine translation (SMT) model, two based on neural machine translation (NMT). On Flickr8k, the SMT-based model performs much better than the NMT-based one, achieving a 49.6% F1 score. Finally, we apply our word discovery system to the task of image retrieval and achieve 29.1% recall@10 on the standard 1000-image Flickr8k tests set.

Index Terms: unsupervised spoken word segmentation, multimodal learning, neural machine translation, statistical machine translation

1. Introduction

The task of word discovery is to segment and cluster speech or phone sequences into sequence of word units. It is useful for speech technology for unwritten languages and for languages in which obtaining word segmentation and lexicon manually will be prohibitively expensive. Automatic word discovery systems exist (e.g., [1, 2, 3]), but the task is quite challenging, therefore we employ an alternative source of information: images. If each utterance is known to be a spoken description of an image, then the set of concepts visible in the image can be seen as a bag of noisy word labels for the speech.

Several works have used raw audio to discover word units. Methods that imitate child language acquisition often begin by finding recurring patterns in audio [1, 4]. Non-parametric Bayesian hidden Markov models (HMMs) have been widely used in word-unit discovery and various other clustering problem with audio, e.g., a latent Dirichlet process with HMM acoustic models can be used to jointly segment and cluster raw audio into sub-word units [5, 6], or the HMM can be regularized using an L-p norm as sparsity constraint to encourage purer clusters[2]. Using word embeddings as features, it is possible to perform automatic word discovery by modeling each word as a Gaussian mixture model with a Dirichlet prior on its parameters; the model can be trained using expectation maximization (EM), or using a weighted K-means algorithm [3]. The Dirichlet-prior Gaussian mixture model out-performed all other systems by around 10 % (30 % in F-score) during the 2017 zero-resource challenge [7]. Other works have focused on discovering word units from phone sequences or character sequences, such as models based on Pitman-Yor process [8].

A related task to the unsupervised spoken word discovery is query-by-example keyword search in audio, which aims to only search for a collection of keywords and leaves the rest of the speech as background. The most recent widely published benchmark evaluation of this task was the NIST OpenKWS evaluation set on the language Georgian. The Kaldi OpenKWS system [9] trained a DNN-HMM hybrid system and decode the OOV queries by fusing the decoding scores on word-level (with proxy word), phonetic-level and morpheme-level lattices to maximize the ATWV score. The BBN system [10] combined several acoustic models based on DNN, LSTM and CNN on subword units to perform joint decoding and handled OOV queries on the sub-word unit. The STC keyword search system [11] combined 9 different acoustic models based on DNN and GMM with a phone-posterior based OOV decoder [12].

A multilingual approach for spoken word discovery has been proposed by [13], who developed a variant of the IBM model 3 SMT to discover word units of an under-resourced language by aligning parallel texts in a high-resourced language. The same task has been attempted [14] using NMT with attention [15] to align speech or phone sequences to the word labels of the high-resourced language; modifications of the attention mechanism to ensure coverage and richer context. If the true phone sequence in the under-resourced language is unknown, pseudo-phone labels generated by an unsupervised non-parametric Bayesian model [6] can be used as input to the NMT [16].

The database used in this paper was first published as an image captioning corpus, for which the baseline system [17] used IBM model I and II [18] combined with Kernel Canonical Component Analysis (KCCA) for mapping both image and text to a joint space. [19, 20] developed a two-branch neural network system to learn the joint representation of image and text. The speech files were first used to train an end-to-end image retrieval system [21, 22], and were then further analyzed to discover word-like units [23]. The task of multimodal word discovery was, we believe, first proposed in [23], where it was performed as a generalization of the image retrieval problem: every possible subsegment of the audio file was tested as a query, and every possible sub-region of the image was tested as a possible retrieval result.

2. Problem formulation

2.1. Word discovery as neural machine translation

Suppose we have a sequence of phone indices $x_1, ..., x_{T_x}$ and a sequence of image concepts $y_1, ..., y_{T_y}$, where $x \in \mathcal{X}, y \in \mathcal{Y} \cup \{NULL\}$. Given $(\mathbf{x}, \mathbf{y})$, the goal of our algorithm is a sequence learning problem to align each phone label $\mathbf{x} = [x_1, x_2, \dots, x_{T_x}]$ to an image concept $\mathbf{y} = [y_1, y_2, \dots, y_{T_y}]$. We assume that $T_y < T_x$ and one phone label is only associated with one or none of the image concepts. In other words, we define an alignment matrix $\mathbf{A} \in \{0,1\}^{T_y \times T_x}, \mathbf{A} = [\mathbf{a}_1 \ldots \mathbf{a}_{T_x}] = [\tilde{\mathbf{a}}_1^\top \ldots \tilde{\mathbf{a}}_{T_y}^\top]^\top$, and we have:

$$\sum_{i=1}^{T_y} a_{it} = 1, \forall t \in \{1, \ldots, T_x\}. \quad (1)$$

Then our model tries to learn to maximize:

$$p(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{A} \in \{0,1\}^{T_y \times T_x}} p(\mathbf{A}|\mathbf{x}) p(\mathbf{y}|\mathbf{x}, \mathbf{A}), \quad (2)$$

for each input phone-concept sequence pair.

Suppose there is a "dominant alignment" $\mathbf{A}^*$such that $p(\mathbf{A}^*|\mathbf{x}) \approx 1$, Eq. (2) is then simplified to one term:

$$p(\mathbf{y}|\mathbf{x}) \approx p(\mathbf{y}|\mathbf{x}, \mathbf{A}^*) \approx \prod_{i=1}^{T_y} p(y_i | y_{1:(i-1)}, \mathbf{x}, \mathbf{A}^*). \quad (3)$$

where $y_{1:(i-1)} = [y_1, \ldots, y_{i-1}]$ is the set of output labels preceding y_i. Another simplification we can make is to compress the one-hot representations of the phone sequence $\mathbf{x}_i, i = 1, ..., T_x$ into a lower-dimensional embedding vector $\mathbf{h}_i, i = 1, ..., T_x$ and learn the dominant alignment with a soft alignment:

$$a_{it}^* := \alpha_{it} = \frac{\exp(e_i(\mathbf{h}(\mathbf{x}_t), \mathbf{s}_{i-1})/T)}{\sum_{j=1}^{T_y} \exp(e_j(\mathbf{h}(\mathbf{x}_t), \mathbf{s}_{j-1})/T)} \quad (4)$$

$$\mathbf{c}_i = \sum_{t=1}^{T_x} a_{it}^* \mathbf{h}_t, \quad (5)$$

where $\mathbf{e}(\cdot)$ can be learned by a neural network, $\mathbf{s}_i$ is the decoder state vector, and $\mathbf{c}_i$ is called the *context vector* in a typical encoder-decoder architecture [15]. T is a temperature term to smooth the softmax [14]. A main difference of our network from the architecture in [15] is that instead of normalizing the energy over the time steps of the input phone sequence, our network normalizes across the attention weights corresponding to the output image concepts for each phone. This helps ensure the assumption in Eq. (1) is satisfied and our alignment for each phone is sparse across image concepts. Let us further assume that y_i depends on $\mathbf{x}$ only by way of its dependence on $\mathbf{c}_i$, and depends on $y_{1:(i-1)}$ only by depending on y_{i-1} and $\mathbf{s}_{i-1}$, therefore

$$\prod_{i=1}^{T_y} p(y_i | y_{1:i-1}, \mathbf{x}, \mathbf{A}^*) \approx \prod_{i=1}^{T_y} p(y_i | y_{i-1}, \mathbf{s}_{i-1}, \mathbf{c}_i). \quad (6)$$

The set of probabilities $\{p(y_i|y_{i-1}, \mathbf{s}_{i-1}, \mathbf{c}_i)\}_{i=1}^{T_y}$ can be learned using a recurrent neural net f with state vectors $\mathbf{s}_i$. Now the task reduces to learning the functions h, e, f such that the log-likelihood of the concepts given the phone sequence is maximized.

For the neural-based translator, we used XNMT [24] to implement two networks: One used a standard encoder-decoder structure with attention normalized over the phone sequence, which we will refer to later as the normalized-over-time model; the other has the same bi-directional LSTM encoder with a single 512-dimensional hidden layer but with attention normalized over the concepts. The decoder for the normalized-over-concepts model does not have recurrent state operation (because the state vector $\mathbf{s}_i$ depends on the context vector $\mathbf{c}_i$, which depends on $\mathbf{s}_j$ for times including $j > i$, as shown in Eq. 4), instead, the concept label is fed into the attention network. The attention networks for both models have a single 512-dimensional hidden layer.

2.2. Word discovery as statistical machine translation

Alternatively, we can learn the probability of a phone sequence given a set of image concepts:

$$p(\mathbf{x}|\mathbf{y}) = \sum_{a_1=0}^{T_y} \sum_{a_2=0}^{T_y} \cdots \sum_{a_{T_x}=0}^{T_y} p(\mathbf{A}|\mathbf{y}) p(\mathbf{x}|\mathbf{y}, \mathbf{A}). \quad (7)$$

Following [25], we make use of the following assumptions: 1) $\mathbf{A}$ is integer-valued, specifically $a_{it} = 1$ for $i = i(t)$, else $a_{it} = 0$, 2) all alignments are equally likely given only $\mathbf{y}$: $p(\mathbf{A}|\mathbf{y}) = \frac{\epsilon}{(T_y+1)^{T_x}}$, where ϵ is some normalization constant; 3) given the alignment, each phone depends only on its aligned image concept, thus $p(x_t | x_{1:(t-1)}, \mathbf{A}, \mathbf{y}) = p(x_t | y_{i(t)})$. Eq. (7) is then simplified to:

$$\frac{\epsilon}{(T_y+1)^{T_x}} \prod_{t=1}^{T_x} \sum_{i(t)=1}^{T_y} p(x_t | y_{i(t)}). \quad (8)$$

Optimization with EM results in an iterative formula in terms of the expected counts of a given phone-concept pattern.

The optimal alignment between the phones and image concepts can be then obtained by finding the highest-scored translation pair of a given sentence:

$$i^*(t) = \arg\max_i p(x_t | y_i). \quad (9)$$

To perform image retrieval using the SMT model, for each image example we compute the translation probability of the phone sequence given the image concepts: $P(\mathbf{x}|\mathbf{y})$. For phone-translation pairs unseen before, we simply filled in for the pair a small fixed probability 10^{-12}.

3. Experimental setup

3.1. Datasets

Our dataset consists of 7996 images that are present in both the Flickr8k [17] and Flickr30k corpora. We used the Flickr30kEntities dataset to extract the image concepts and phone sequences primarily because it contains bounding boxes with phrases from the caption that describes the particular region, like "a girl in a white shirt". The phrase-level segmentation is used as our groundtruth alignments. Only the images that also appeared in Flickr8k are used, so that we can compare our results to other speech-to-image [21] and text-to-image [19] retrieval systems. In order to make sure that our image retrieval results use the same dataset as [23] and [19], we divided the data into a training set (6996 images) and a test set (the same 1000 images that are used as the test set in [23, 19]). For word discovery, we restricted our attention to a list of concepts that appeared at least 10 times in our training dataset. By using the WordNet [26] synsets to merge similar concepts, we were able to find a list of 1547 distinct image concepts, each of which occurs at least 10 times. We merged repeated concepts in a sentence in order to maintain the many-to-one mapping between the phone sequence to concept. The captions are converted into phone sequences via the CMU pronunciation dictionary [27], which contains 39 distinct phonetic symbols and 69 symbols in total with the stress symbol. Compound words such as "fingerpaint" or "skateboarder" do not have an entry in the dictionary, and we simply replace them with an `UNK` symbol. We used NLTK [28] as the interface to both Wordnet and CMU dictionary. Notice that we did not use an image classifier to compute the probabilities of the image concepts and instead used the hard label from Wordnet because we believe the task of detecting the image label is largely separate from our main task; future work will seek to integrate such an image classifier into our system.

3.2. Evaluation metrics

For the word discovery task, we evaluated our results by comparing the predicted alignment with the groundtruth. Several metrics are used to evaluate the system: The word intersection-over-union (IoU) measures the similarity between the span of a discovered phone sequence and the span of the groundtruth image concept at the same location in the phone string by measuring the ratio of the length of their intersection, divided by the length of their union. Accuracy is measured as the percentage of phones that are assigned to the correct image concept. To visualize the tradeoff between the true positive rate and false positive rate, we also plot the receiver operating characteristic (ROC) curve using the alignment probability for the SMT and the attention weights for the NMT for each class. The ROC is defined by treating the word discovery as a retrieval problem where the query is an image concept and the retrieved documents are the relevant words; using this setup we measure the recall, precision and F-measure of the system. This helps us to fairly evaluate the system because our dataset is unbalanced in amount of instances for each image concept. For the image retrieval task, we followed [21, 19] in using recall@1, 5, 10, which treats the entire caption as the query, and the set of test images as the database. We also follow [21, 19] in assuming a one-to-one mapping between captions and images, despite the large number of image pairs with similar concepts.

4. Results

Table 1: *SMT vs NMT: Word Discovery Results*

	SMT	NMT (norm. over concepts)	NMT (norm. over time)
Word-IoU	6.00	**46.0**	21.0
Accuracy	**43.8**	23.0	41.5
Recall	**52.9**	18.0	29.2
Precision	**46.7**	12.1	33.0
F-Measure	**49.6**	14.5	31.0

The word discovery results of our system are shown in Table 1. The SMT model performs the best in almost all the evaluation metrics. It may be that SMT outperforms NMT because the size of our dataset ($\approx$ 100000 phones and 40000 image concepts) is more suitable for SMT than NMT. However, we notice that the NMT normalized-over-time has an accuracy almost equal to that of SMT, although its recall, precision, and F-measure are low. The disparity between F-measure and accuracy may indicate that the disparity between SMT and NMT has less to do with the amount of training data than with the imbalanced numbers of training tokens for different image concepts. For example, the concept with the highest number of groundtruth aligned phones is the NULL class; it is possible for NMT to achieve high accuracy but low F-score by simply aligning all phones to the NULL class.

The NMT generally performs better at IoU scores than the SMT. A plausible explanation of this finding is illustrated in Fig. (1): about 40 % of the pseudo-words discovered by the SMT are less than 3 phones long, and 90 % of the pseudo-words are within 10 phones long. By comparison, the NMT normalized over concepts tends to generate pseudo-words that are much longer than an actual word, potentially merging various words together. Since the IoU penalizes longer sequences less than it penalizes shorter sequences, NMT has a higher score than SMT. The NMT with attention normalized over time seems to have a length distribution closest to the ground truth.

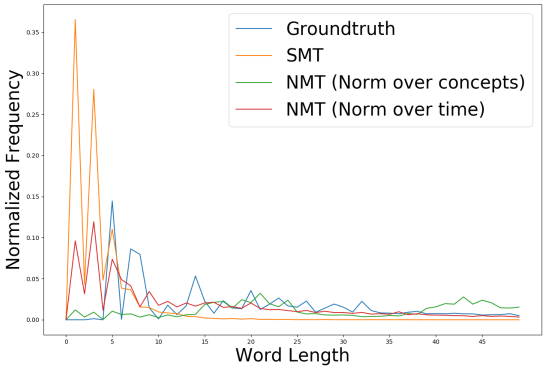

Figure 1: *Normalized frequency of the lengths of words from the groundtruth and discovered by the models*

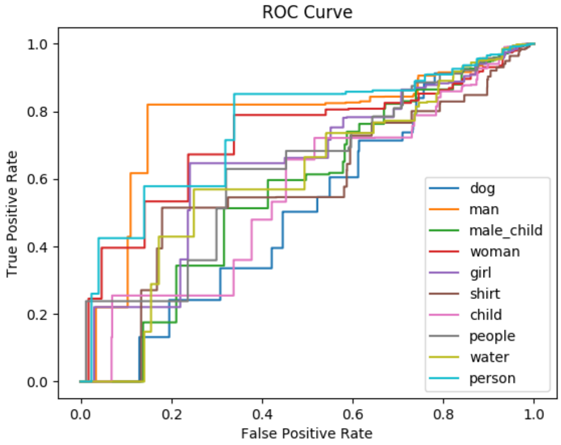

Figure 2: *ROC curve of the statistical word discoverer for the top-10 concepts (excluding NULL) in the flickr30k dataset*

The ROC of the SMT and NMT (normalized over time) for the top-10 image concepts is shown respectively in Fig. (2) and Fig.(3). These figures show that the SMT is capable of detecting some frequent image concepts, like “girl” or “man”, accurately with a small false positive rate. The model, however, fails to control the false positives beyond chance level for the extremely common concept “dog”. NMT also has trouble rejecting false positive of the “dog” concept, but seems to be much better at detecting other common concepts such as “woman” and “shirt”.

Between the two NMT models, the normalized-over-time model performs better than the normalized-over-concept model, even though the normalized-over-time model fails to satisfy Eq. (1). One explanation may be that the normalized-over-

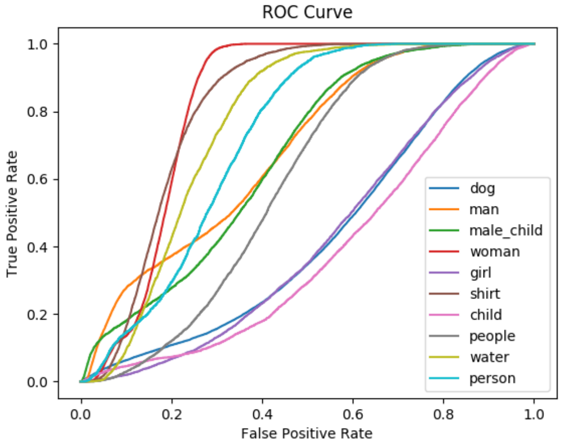

Figure 3: *ROC curve of the attention-based word discoverer for the top-10 concepts (excluding </s>) in the flickr30k dataset*

concept model ignores spatial proximity relationships, such as the tendency for "hand" to be visible near "human", since its decoder does not use the state vector of the decoder network to calculate attention. The difference in performance between the normalized-over-time and normalized-over-concepts models might therefore be interpreted to mean that spatial proximity relationships between image concepts can be useful for multimodal word discovery.

Table 2: *Image retrieval rates using SMT multimodal word discovery, compared to speech-based and text-based published results for the same corpus.*

	Recall@1	Recall@5	Recall@10
SMT	9.42%	21.1%	29.1%
Harwath&Glass [21]	-	-	17.9%
Karpathy [19]	10.3%	31.4%	42.5%

Image retrieval scores for these systems are shown in Table 2. Our SMT system achieves 29% recall@10 score in a database of 1000 images from Flickr8k. This is better than the results reported in [21], though the comparison is not entirely fair: We assume the image and phone labels are extracted perfectly, while [21] does not have such an assumption. Our results are worse than [19], partly because their system is trained on words with known boundaries and on a much larger flickr30k dataset, though they do not assume perfect image labels. Another reason is that our system used only entity labels of the image and no information about the action and pose of the objects in the image is used. The results suggest that our model needs to be extended to incorporate richer context to get higher performance on the retrieval task.

5. Error analysis

The attention/alignment matrices of a typical caption is shown in Fig. (4). As shown in the plots, many of the errors in our systems are caused by the alignment of phones that should belong to the NULL/< /s> class. This may be due to an underestimate of the prior probabilities of the NULL symbol. Conversely, the NMT tends to overestimate the probability of the end-of-sentence symbol </s>, and therefore it mis-aligns phones that belong to a true concepts to the end symbol. This example also suggests that the accuracy score of the normalized-over-time model may be inflated by true negative examples, i.e., by the allocation of phones whose true label should be NULL (the first column of each matrix in the figure) to other classes.

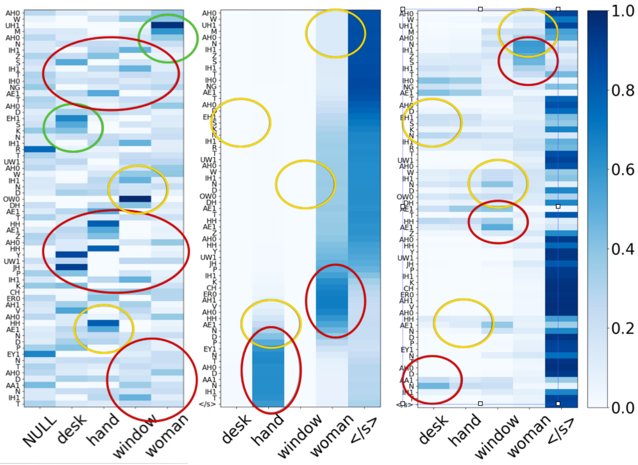

Figure 4: *Comparison of the attention/alignment probabilities for the SMT and NMT models for the caption "A **woman** is sitting at a **desk** near to a **window** that has a huge picture of a **hand** painted on it". The correctly discovered words are circled in green, the false negative or partially correct words in yellow and the false positive words in red. 1) Left: the SMT model correctly discovers the words "woman" and "desk" and partially discovered "hand" and "window", while generating many false positives; 2) Middle: the normalized-over-concept NMT model detects "woman" and "hand" but is overpowered by the < /s> symbol; 3) Right: the normalized-over-time NMT model almost correctly discovers the concept "woman" and "window" but not "desk" and "hand".*

The main source of error of our image retrieval system is the confusion between images with very similar sets of objects. This is partly because, unlike [19], our network does not use a max-margin loss objective, so the classification boundaries between similar image concepts are not optimally generalizable. Also richer contexts may be useful to augment the difference between the representation for different images.

6. Conclusions

This paper describes three systems for the task of discovering word units from phone labels and image concepts. With the amount of data we have, the SMT-based model performs much better than the NMT-based ones, according to our evaluation metrics. Finally, we applied our word discovery system to the task of image retrieval and show that it can achieve performance that is about halfway between the published scores for speech-based and text-based image retrieval.

7. Acknowledgements

The idea of this work originated from the 2017 JSALT workshop at Carnegie Mellon University, Pittsburgh. We would like to thank [17] for releasing the Flickr8k dataset.

8. References

[1] O. J. Rašaňen, G. Doyle, and M. C. Frank, "Unsupervised word discovery from speech using automatic segmentation into syllable-like units," in *Interspeech*, 2015.

[2] S. Bharadwaj, M. Hasegawa-Johnson, J. Ajmera, O. Deshmukh, and A. Verma, "Sparse hidden markov models for purer clusters," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[3] H. Kamper, K. Livescu, and S. Goldwater. (2017) An embedded segmental k-means model for unsupervised segmentation and clustering of speech. [Online]. Available: https://arxiv.org/pdf/1703.08135.pdf

[4] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[5] 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*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 40–49.

[6] L. Ondel, P. Godard, L. Besacier, E. Larsen, M. Hasegawa-Johnson, O. Scharenborg, E. Dupoux, L. Burget, F. Yvon, and S. Khudanpur, "Bayesian models for unit discovery on a very low resource language," in *Proc. ICASSP*, 2018.

[7] E. Dunbar, X. Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera, and E. Dupoux, "The zero resource speech challenge 2017," *CoRR*, vol. abs/1712.04313, 2017. [Online]. Available: http://arxiv.org/abs/1712.04313

[8] M. Johnson, T. Griffiths, and S. Goldwater, "Adaptor grammars: A framework for specifying compositional nonparametric bayesian models," in *Neural Information Processing Systems*, 2007.

[9] J. Trmal, M. Wiesner, V. Peddinti, X. Zhang, P. Ghahreman, Y. Wang, V. Manohar, H. Xu, D. Povey, and S. Khudanpur, "The kaldi openkws system: Improving low resource keyword search," in *Interspeech*, 2017.

[10] T. Alume, D. Karakos, W. Hartmann, R. Hsiao, L. Zhang, L. Nguyen, S. Tsakalidis, and R. Schwartz, "The 2016 bbn georgian telephone speech keyword spotting system," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[11] I. Medennikov, A. Romanenko, A. Prudnikov, V. Mendelev, Y. Khokhlov, M. Korenevsky, N. Tomashenko, and A. Zatvornitskiy, "Acoustic modeling in the stc keyword search system for openkws 2016 evaluation," in *Interspeech*, 2017.

[12] Y. Khokhlov, N. Tomashenko, I. Medennikov, and A. Romanenko. (2017) Fast and accurate oov decoder on high-level features. [Online]. Available: https://arxiv.org/pdf/1707.06195.pdf

[13] F. Stahlberg, T. Schlippe, S. Vogel, and T. Schultz, "Word segmentation through cross-lingual word-to-phoneme alignment," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2012, pp. 85–90.

[14] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2016, pp. 949–959. [Online]. Available: http://aclweb.org/anthology/N16-1109

[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[16] P. Godard, M. Z. Boito, L. Ondel, A. Berard, A. Villavicencio, and L. Besacier, "Unsupervised word segmentation from speech with attention," in *Interspeech*, 2018.

[17] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: data, models and evaluation metrics," in *Journal of Artificial Intelligence Research*, 2010.

[18] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263 – 311, 1993.

[19] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Neural Information Processing Systems*, 2014.

[20] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the 2015 Conference on Computer Vision and Pattern Recognition, CVPR*, 2015.

[21] D. Harwath and J. Glass, "deep multimodal semantic embeddings for speech and images," *Automatic Speech Recognition and Understanding*, 2015.

[22] D. Harwath, A. Torralba, and J. Glass, "unsupervised learning of spoken language with visual context," *Neural Information Processing Systems*, 2016.

[23] D. Harwath and J. Glass, "Learning word-like units from joint audio-visual analysis," 55^{th} *Annual Meeting of the Association for Computational Linguistics*, 2017.

[24] G. Neubig, M. Sperber, X. Wang, M. Felix, A. Matthews, S. Padmanabhan, Y. Qi, D. S. Sachan, P. Arthur, P. Godard, J. Hewitt, R. Riad, and L. Wang, "XNMT: The extensible neural machine translation toolkit," in *Conference of the Association for Machine Translation in the Americas (AMTA) Open Source Software Showcase*, Boston, March 2018.

[25] P. Brown, V. J. D. Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, 1992.

[26] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995. [Online]. Available: http://doi.acm.org/10.1145/219717.219748

[27] A. Rudnicky. (2014) The carnegie mellon pronouncing dictionary. [Online]. Available: https://github.com/cmusphinx/cmudict

[28] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009.