

Multimodal Word Discovery with Phone Sequence and Image Concepts

Liming Wang¹, Mark Hasegawa Johnson^{1,2}

¹Department of Electrical and Computer Engineering
University of Illinois, Urbana Champaign

²Beckman Institute
University of Illinois, Urbana Champaign

Interspeech, September 2019

Table of Contents

Motivation

Problem Formulation

Model Description

Experiments and Results

Table of Contents

Motivation

Problem Formulation

Model Description

Experiments and Results

Motivation

1. Lack of official orthographic system for many languages in the world
2. Lack of lexical-level and word-level transcriptions for training ASR systems for majority of existing languages, e.g., Mboshi
3. Link between low-resource speech learning and early language acquisition process: Information sources besides speech (e.g., vision and taste)?

Table of Contents

Motivation

Problem Formulation

Model Description

Experiments and Results

Multimodal Word Discovery

► Inputs:

1. $x_1, \dots, x_{T_x}, x \in \mathcal{X}$: phone sequences that the infant hears
2. $y_1, \dots, y_{T_y}, T_y < T_x, y \in \mathcal{Y} \cup \{NULL\}$: a set of image concepts that the infant sees

► Output:

1. Alignment matrix: word unit = consecutive alignments to the same concept

$$A \in [0, 1]^{T_x \times T_y} = [a_1^\top, \dots, a_{T_x}^\top]^\top = [\tilde{a}_1 \dots \tilde{a}_{T_y}]$$

► Assumptions:

1. One concept per phone: $\sum_{i=0}^{T_y} a_{ti} = 1, t = 1, \dots, T_x$
2. Unaligned phones: $a_{t0} = 1$

Table of Contents

Motivation

Problem Formulation

Model Description

Experiments and Results

SMT vs NMT

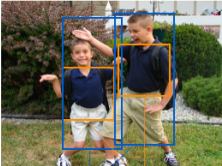
- ▶ Statistical Machine Translation (SMT): first introduced by Brown et. al. 1993 [1]

1. Learning goal: $p(x|y) = \sum_{A \in \{0,1\}^{T_y \times T_x}} p(A|y)p(x|y, A)$
2. Inference: EM algorithm, iteratively computing $p(x_t|y_i)$ in terms of the expected counts:

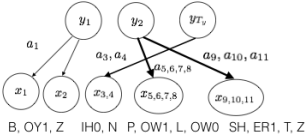
$$\langle c(x_t|y_i; x, y) \rangle := \mathbb{E}_A[\delta_{i(t)i}|x, y]$$

3. Output: hard alignment $i(t) := \arg \max_i p(a_{ti} = 1|x, y)$
- ▶ Neural Machine Translation (NMT) with attention: introduced by Bahdanau et. al. 2014 [2]:
 1. Learning goal: $p(y|x) \approx p(y|x, A^*)$ (dominant path assumption)
 2. Inference: backpropagation + batched gradient descent
 3. Output: soft alignment α_{ti}

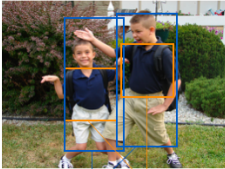
Models: Machine Translation



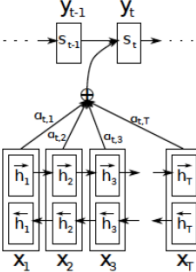
Male-child Polo-shirt NULL



"Boys in blue polo shirts"



Male-child Polo-shirt </s>



B OY1 Z IH1 N B L UW1 P OW1 L OW0 SH ER1 T S </s>

"Boys in blue polo shirts"

NMT Attention Mechanism

1. Normalize-over-time model

(Bahdanau et. al., [2]):

$$a_{it}^* := \alpha_{it} = \frac{\exp(e_i(h(x_t), s_{i-1})/T)}{\sum_{j=1}^{T_x} \exp(e_j(h(x_t), s_{j-1})/T)}$$

2. Normalize-over-concept model:

$$a_{it}^* := \alpha_{it} = \frac{\exp(e(h(x_t), y_i)/T)}{\sum_{j=1}^{T_y} \exp(e(h(x_t), y_j)/T)}$$

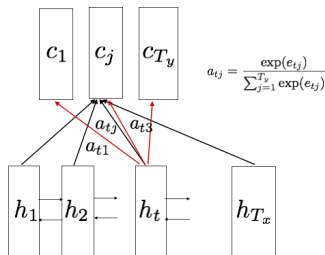
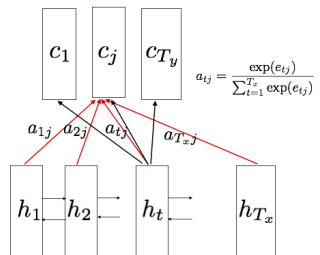


Table of Contents

Motivation

Problem Formulation

Model Description

Experiments and Results

Dataset Extraction

1. Raw data: Flickr8k and Flickr30k with object bounding boxes and phrase-level boundaries; English as “simulated” low-resourced language
2. Image concept extraction: merge similar noun phrases Flickr30kEntities using Wordnet synsets and select concepts with frequency > 10
3. Utterances selection: captions with all image labels having frequencies > 10 , ≈ 8000 captions in total
4. Caption transcription: transcribe text into phone sequence via CMU dictionary
5. Dataset split: same test set as in (Karpathy 2014)

Model parameters

- ▶ SMT: initialized counts with indicator for co-occurences
- ▶ NMT: written with XNMT toolkit [3]; 512-dimensional embedding layer, 512-dimensional one-layer BLSTM encoder and LSTM decoder, 512-dimensional fully-connected attention; 0.5 dropout

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 |

| | Recall@1 | Recall@5 | Recall@10 |
|-------------------|----------|----------|-----------|
| SMT | 9.42% | 21.1% | 29.1% |
| Harwath&Glass [4] | - | - | 17.9% |
| Karpathy [5] | 10.3% | 31.4% | 42.5% |

Analysis - ROC Curve

- ▶ ROC curve: visualize tradeoff between false positive and true positive rate for one-versus-all classification of concepts
- ▶ Rougher transition from SMT; higher variances from NMT

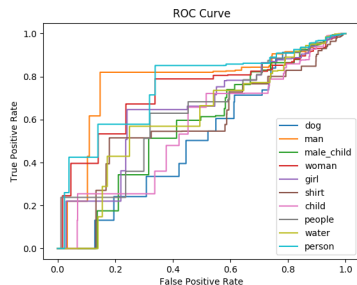


Figure: ROC plot for SMT

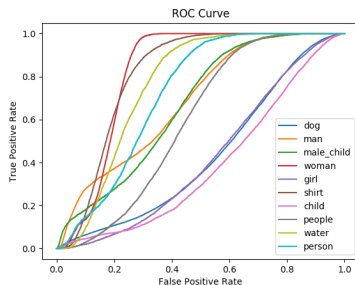


Figure: ROC plot for NMT

Analysis - Soft Alignment Plots

- ▶ Soft alignment matrix: A $T_x \times T_y$ matrix with each entry as $p(a_{ti} = 1|x, y)$ for SMT and attention weights for NMT
- ▶ “A woman is sitting at a desk near to a window that has a huge picture of a hand painted on it”

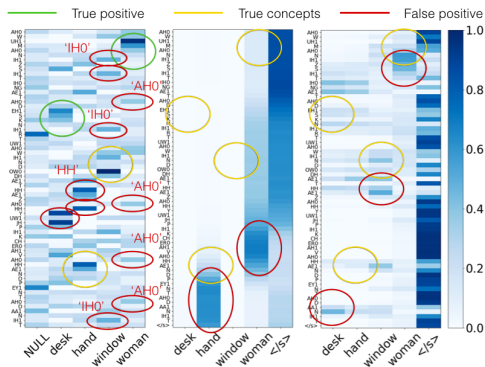






Figure: Left: SMT, Middle: normalized-over-concept, Right: normalized-over-time

Conclusion and Future Works

1. SMT performs superior to NMT on our low-resource multimodal setting
2. SMT learns meaningful units from image concepts
3. Future directions: multimodal word discovery beyond mixture models; word discovery with raw audio and image

Thank you ! The code will be available at
<https://github.com/lwang114/MultimodalWordDiscovery>

-  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.
-  D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *ICLR*, 2015.
-  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.
-  D. Harwath and J. Glass, “deep multimodal semantic embeddings for speech and images,” *Automatic Speech Recognition and Understanding*, 2015.



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