February 3, 2020

Dear Prof. Tan Lee and Mr. Jackson,

Thank you for providing the opportunity for a second review of our paper "Multimodal Word Discovery and Retrieval with Spoken Descriptions and Visual Concepts". We appreciate the time and effort you and the reviewers have devoted into providing detailed and valuable feedback to our manuscript. We greatly benefit from the insightful comments of the reviewers while refining our manuscript. We had tried our best to make modifications to reflect as much of the suggestions by the reviewers as possible. Below is a point-to-point response to the reviewers' comments and suggestions.

Comments from Reviewer 1:

- **Comment**: *"... However, the organization of this paper should be improved. Specifically, the paper is consisted of lots of components, but the relationship among them is not clearly addressed"*
  Response: Thank you for pointing this out. We have added a figure to explain the goal of the models proposed in the paper, and divided the section VI (B) into three sections: "the static subword-level segmentation approach", "the static word-level segmentation approach" and "the dynamic segmentation approach" and added more explanation about the relations between the models (such as the last sentence in section VI (B)(2) at page 6); we also visualize the distributions of concept-level F1 scores across different models, which intend to give a more fine-grained comparison between the models.
- **Comment**:*"Please try to enhance the quality of all the figures in this paper. All of them are not clear when printed them out!"*
  Response: Thank you for the suggestion. We have improved the quality of the figure by directly converting the pictures to pngs instead of taking screenshots.
- **Comment**:*"In the middle part of section VIII-A, the reference for wordnet is missing."*
  Response: Thank you for pointing this out. The problem should be

fixed now.

- **Comment**: *"For the evaluation metric, i.e., section VIII-B, the paper claims that the accuracy is defined as the percentage of phones that align., thus the results may overestimate. This is because that the evaluation seems to be a bag-of-words method, so the ordering information is ignored! If I make a mistake on this part, please also add some descriptions for the issue."*

  Response: This is an important and valid concern. However, since the order of the phonetic symbols in a caption is known to the model, we do not need to worry about their order, as long as we can predict the alignment between each phone symbol to the corresponding image concept correctly. Therefore, even though the evaluation metric is agnostic of the order of the phone symbols, it does not overestimate the result.

Comments from Reviewer 2:

- **Comment**: *In the abstract: "A multimodal word discovery system accepts, as input, a database of spoken descriptions of images (or a set of corresponding phone transcriptions) and learns a mapping from phone strings to their associated image concepts." This does not need to be the case, because many previously published systems discover words directly at the waveform level and do not rely on phonetic transcripts.*

  Response: This is an important point. We have modified the statement to "learn a mapping from waveform segments (or phone strings)", which should include the case for word discovery at the waveform level.

- **Comment**: *"... but the take-home message needs to be better defined, and the presentation better focused on that. It is hard for me to determine exactly what the take-home message is in the current paper. I think this is partially due to the fact that the paper doesn't provide an apples-to-apples baseline comparison, and partly due to the fact that the word discovery evaluation metrics are somewhat opaque. Boiling down the word discovery evaluation to corpus-level P/R/F1 numbers and ROCs does not tell me how many unique words/concepts were learned, which ones were learned and which ones were not, and so on. Maybe a histogram of per-concept or per-word F1 scores would be more illuminating here?"*

Response: Good point and great suggestions. We have incorporated your suggestions as follows. First, we improved the structure of the paper by adding a figure to explain the goal of the models proposed in the paper and dividing the long section VI (B) into subsections; second, we visualize the distributions of concept-level F1 scores across different models as suggested in Figure (4)(5)(6), which should give a more fine-grained comparison between the models. As for the reviewer's comment about the evaluation metric, we have added some of the 2017 ZRSC evaluation metrics to the paper as well as more explanation of the definition of the evaluation metrics in section VIII (B). From this additional analysis, the take-home message for the paper is that segmental models seem to work better than frame-wise models in both phone-level and audio-level.

- **Comment**:" *It might be good to also include the references to and the numbers from . . .*"
  Response: Agree. We have added the related references.
- **Comment**:"*Section VIII first paragraph: Should cite [17] for the Flickr8k corpus and add the citation for the Flickr30k corpus, and Flickr30kEntities*"
  Response: Agree. We have added the citations mentioned above to the bibliography.
- **Comment**:"*Section VIII first paragraph: WordNet citation missing*"
  Response: Thank you for pointing this out. Should be fixed now.
- **Comment**:"*Section VIII first paragraph: 'captions in which every concepts appear at least 10 times': I assume this means retaining only the concepts that appear at least 10 times in the training set?*"
  Response: Actually it is the concepts that appear at least 10 times including both the training, validation and test set of the Flickr8k dataset. That is, we split the dataset after determining the set of concepts to use for our systems.
- **Comment**:"*Figure 4: The caption does not make it possible to distinguish which ROC curve belongs to which system.*"
  Response: Thank you for pointing this out. We have replaced the ROC curve with the F1 histograms for each model, which should help distinguishing clearly the performance of different systems.
- **Comment**:"*Throughout paper: "groundtruth" should be "ground truth"*"
  Response: Fixed.

- **Comment**:"*Throughout paper: Figure and Table caption fonts don't match*"
  Response: Thank you for the suggestion. But we believe that the fonts for the tables and figures are set by the template provided by the IEEE Transaction, so we do not change the font type.

Comments from Reviewer 3:

- **Comment**:"*The work is built on the assumption that the description of an image has a connection with the concepts existing in the image thus new words could be found by finding alignment between the visual concepts and phonetic description or acoustic features. This assumption is not necessarily valid in reality since a text describing an action in an image may have no connection with the objects present in the image. For example, the description of an image showing traffic in a city could only describe the concept of traffic and not necessarily the cars, traffic lights etc which are the main concepts in an image.*"
  Response: That is an important point. The model does only consider concepts that correspond to visual objects without taking into account the actions between them and makes simplifying assumptions when it comes to higher-level concepts such as "traffic" in the reviewer's example. But we believe that the model can still learn the word "traffic" if the phonetic strings of the word "traffic" appear most frequently when concepts such as "traffic light", "car", "highway" are present in the image, by aligning the relevant phone strings to those concepts.
- **Comment**:"*The authors, however, use the Flickr8K dataset which provides multiple descriptions for each image that each describe the concepts in the image and contains the links between the entities in the caption and the bounding boxes which guarantee the existence of the image concepts. Although, this could also be contrary to the claim that the authors make in the introduction about their framework being unsupervised and that the image labels can be seen as a bag of - noisy - word labels for the speech.* "
  Response: You have raised an important question, which is what level of supervision our models use. Our systems actually do not have access to links between entities in the caption and the bounding boxes. Instead, our models only accept as inputs a bag of bounding box labels

and *unsegmented* sequences of phone labels or audio features, without direct information about which bounding box corresponds to which part of the caption. Further, the image labels of the bounding boxes do not necessarily correspond to the actual words appeared in the caption but an abstract language-independent category extracted using WordNet. Therefore, the image labels can indeed be seen as a bag of noisy word labels for the speech.

- **Comment**:"*Subsection B in section VI, could be split into more subsection for legibility reasons.*"
  Response: We agree with the suggestion. Accordingly, we have split the section VI (B) into three subsections: "the static subword-level segmentation approach", "the static word-level segmentation approach" and "the dynamic segmentation approach".
- **Comment**:"*The inputs and outputs in figure 1 are not clear and can be confusing for the reader.*"
  We agree with the suggestions. In the revised paper we have added more descriptions about the input and output of the models in the new Figure (1) as well as more descriptions in the caption of the new Figure (2), which is the Figure (1) from the previous version of the paper.
- **Comment**:"*There is no reference to figure 2 in the paper, and caption is also missing .*"
  Response: The caption has been added and the figure is now mentioned in the last sentence of the introduction of Section VI.

In addition, all grammatical and spelling errors and mistakes in the equations have been corrected.

We look forward to hearing from you regarding our new submission and any further questions and comments you may have.

Sincerely,

Liming Wang
Mark Hasegawa-Johnson