ASR

Enrich semantics through extending words using wordnet

The transcription from ASR between different videos of one event might include words that are semantically similar but lexically different. Thus, mining and using the similarity properly would make the ASR feature more robust and effective for event detection. Based on the assumption, we explored feature representation to include such information.

In order to min the semantic similarity, we use Word-net to get all the noun synonyms for each word in the vocabulary except for the stop word. Then, we go through the vocabulary list sequentially to get semantic clusters. For example, we take the i th word, compare its synonym list with each synonym list from i+1 th word to the end of the vocabulary. If the two compared synonym lists have common words, the later word will be merged into the cluster of i th word and it won't be considered later.

Through the coarse clustering, we got around 1900 clusters on the vocabulary from med12 dataset, which includes about 20,000 words. This indicates that many words are potentially similar in semantics. But the result also shows that many clusters are not pure enough. They include some unrelated words. The noise mainly caused by the multiple senses and metaphor of common words.

To reveal the semantic similarity information in feature representation, we treat each cluster as new word in vocabulary and including them in calculating the TF-IDF feature. For simplicity, the weight of the cluster is the same with the word occurred in the video, which belongs to the cluster. After this feature expansion, the average Pmiss is 0.7297 on 11746INDEP dataset using transcription directly. The previous best result using transcription directly is 0.7272. The fusion of these achieves 0.7221, which brings slightly improvement.

There are several reasons that affect the effectiveness of feature expansion:
(1) One word usually has several senses and only one sense is applied in the context. We simply include synonyms from all senses and this brings into noise.

There are several solutions for this problem. Though deep analysis of semantic is impractical for the transcription with high word error rate, we can try to apply the part-of-speech information in distinguishing the meaning of the word. Therefore, the word will be correlated with different clusters based on their part-of-speech in the context. Also, if the utterances are too short to even predict part-of-speech, we can just use the meaning with highest frequency given by the Word-net.

(2) Expanding all the words with clusters might weaken the effectiveness of other dimensions since the sum of all dimensions have to be normalized. Therefore, we'll try to expand only words with high weight or those with high posterior probability in ASR.

Future work would also include adding hyponymy for word clustering discovery. Splitting the cluster based on different distribution in different events. How to assign the weight of expanded dimension feature reasonably also deserves further investigation.

Explore Confusion network to improve performance

The soundtrack from video clips is noisy and results to high word error rate for the state-of-art speech recognizer. Considering the large variance of acoustic feature, we explored to use the confusion network to retrieve more candidate guesses from the acoustics to enhance the performance. We use the confidence value provided from the lattice to calculate the bag-of-words feature and evaluate the result on med11 dataset.

As is shown in the table below, using confusion network and removing word appeared only once from the dictionary in the development data achieves improvement than only using the best-1 transcription. The fusion of system using best-1 and confusion network get further improvement. We also investigated to use the speech segment based on the large scale feature based system in ASR. But we didn't get further improvement. The reason might be because there are too many fragments of speech from the system designed for detecting semantic concepts.

ID	name	MAP	Pmiss
1	ASR 1-best	0.1068	0.7277
2	ASR confusion net (cut1)	0.1278	0.7168
3	Fusion of 1 and 2	**0.1293**	**0.7019**
4	ASR confusion net with new segment	0.1106	0.7375

ASR combined with semantic concepts

There is only around 20% of the soundtrack including speech. However, the non-speech part carries rich semantic information. The audio semantic concepts we explored before is an effective way to capture the hidden semantics in the audio. In order to offer an easy-to-use feature that represents the richest semantic information from audio, we investigated to combine the bag-of-words feature from ASR transcription and our semantic concept detection system by simply concatenating the feature vectors.

From the table below, we can observe that the system got significant improvement in terms of both MAP and Pmiss. Therefore, we'll use this combined feature for capturing audio semantics in MED task this year.

ID		MAP	Pmiss
4	4.ASR with semantic concepts (X2)	0.1976	0.5817
5	5.ASR confusion net with semantic concepts (X2)	0.2049	0.5844
6	6.fusion of 4 and 5	**0.2280**	**0.5563**