Unsupervised Two-Stage Keyword Extraction from Spoken Documents by Topic Coherence and Support Vector Machine

Yun-Nung Chen, Yu Huang, Hung-Yi Lee, and Lin-Shan Lee

Summary

- Proposed Two stage keyword extraction:
  - 1st stage ranks words by topic coherence and term significance to select examples used for SVM training in the 2nd stage.
  - 2nd stage trains an SVM for keyword detection.

1st Stage

Main idea: keywords describe specific topics:
- Included in documents with similar content
- Only relate to a few latent topics

- Topic Coherence measure:
  - For each word \( t_i \), retrieve relevant documents \( R(t_i) \) from document database.
  - Database: (a) target spoken archive, (b) Google, (c) Wikipedia, (d) using (a)+(b)+(c).
  - Keywords retrieve similar documents, and documents retrieved by non-keywords have diverse content.
  - topic coherence: average pair-wise similarity for relevant documents \( R(t_i) \)

- Term significance measure:
  - Entropy of PLSA latent topic distribution for each word
  - Focus on few topics: Low entropy
  - Scatter on many topics: High entropy
  - For each word \( t_i \), retrieve relevant documents \( R(t_i) \) from document database.
  - Keywords: \( W_1 \) topics, \( W_2 \) non-keywords
  - Term significance measure: inverse of the latent topic entropy.

- Rank words by the combination of topic coherence measure and term significance measure.

2nd Stage

- Training example selection for SVM:
  - Positive examples: top N words on the list
  - Negative examples: bottom N words on the list
- Each word is represented by a feature vector for SVM training[1]:
  - Prosodic: pitch, duration, energy
  - Lexical: TF, IDF, TF-IDF, PoS
  - Semantic features based on PLSA model
  - Train an SVM classifier
- All the words on the list are classified by the SVM classifier

Experiments

- Corpus: 45.2 hours course lectures (single instructor) in Mandarin Chinese and English code-mixing
- ASR accuracy: 78.15% (Mandarin char), 53.44% (English word), and 76.26% (overall)

1st Stage Keyword Extraction (ASR)

- Select top K words on the ranking list from 1st stage.
- Topic coherence based on Google \( \rightarrow \) high recall
- Topic coherence based on Wikipedia \( \rightarrow \) high precision
- Topic coherence based on all had best F-measure and precision \( \rightarrow \) used in 2nd stage

Two-Stage Keyword Extraction (ASR)

- 2nd-stage offers improvement on Recall and F-measure
- Two stages keyword extraction outperform K-means Exemplar[2]

Reference: