We propose a novel extractive summarization method for multi-domain contact center dialogues. We use Class-Speaker Hidden Markov Models (CSHMMs) to simultaneously model domain-specific utterance sequences and common (domain-wide) utterance sequences from multi-domain data.

A CSHMM is basically a concatenation of HMMs trained for each dialogue domain.

For a dialogue in Domain k, extractive summarization is done by selecting utterances that are most likely to have been generated from the HMM for Domain k using Viterbi decoding.

We applied CSHMMs to contact center dialogue transcripts of six different domains. Our method outperformed competitive baselines based on the maximum coverage of important words.

Overview

Three Topologic Variations of CSHMM

(1) Ergodic

(2) Ergodic with common states

(3) Concatenated training

Summarization Procedure

Dialogue Data

Evaluation Results

Our method cannot alter compression rates. We made our method output summaries first and made baselines output summaries with the same lengths.

• Baseline1: Maximum coverage of important words in a dialogue (importance is estimated by TF) by ILP
• Baseline2: Same as Baseline1 but the importance of words is estimated by how much each word is related to the target domain

Evaluation measure: F-measure (accuracy of correctly retrieving content words in the scenarios)