Visualising Conversations

Student: Miguel Angelo Rosales, Supervisor: Professor Richard Harvey
School of Computing Sciences, September 2022-2023

University of East Anglia

Aim and Objectives

Investigate the effectiveness of word embedding tools for off-topic detection in conversations applying the main objectives:

- Collecting a corpus conversations from publicly available datasets
- Application of word embedding tools to word2vec and GloVe to analyse changes in dialogue
- Compare their performances

Data

Gathered 5 conversational corpora containing a different number of topics (5, 7, 9, 11 and 13) from the DailyDialog dataset. Each set contains a unique conversational topic to be trained on using GloVe, word2vec and LDA.

Table 1. 5 Conversational datasets with unique topics

Set	Number of Topics	Topic Types
1	5	Cooking, Flying, Studying, Entertainment, Kids
2	7	Summer, Awards, Birthday, Fruits, Smoking, Renting, Furniture
3	9	Computers, Flowers, Coffee, Church, Dance, Pollution, Disney, House, Weather
4	11	Boating, Skiing, Dentist, Language, Writing, Phones, Fighting, Temperatures, Feminism, Shopping, Laundry
5	13	Morning, Video Games, Human Relations, Newspaper, Skincare, Funeral, Marriage, School, Shopping, Tennis, Coffee, Suits, Sports

Methods

- Similarity measurement approach: Calculated topic relevancies for each conversation in the corpora using cosine similarity values determined from word embeddings, and document-topic & topic-word probabilities gathered from LDA.
- Compared topic relevancy values against each other to determine when the conversation shifted in topic

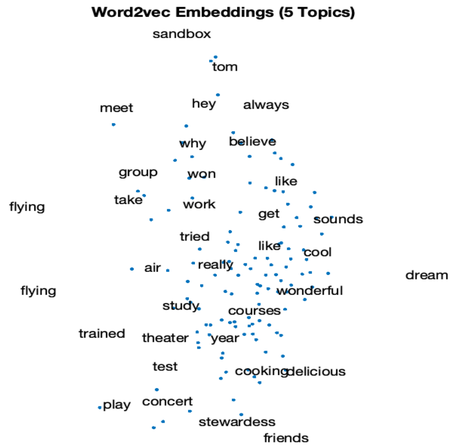

Figure 1. Word2vec embeddings for a conversational corpus

Experiment Design

- Extracted cosine values between LDA-derived feature words (words assigned to a topic) and every word item in the corpus.
- Trained corpus with LDA on MatLab to gather probabilistic values for subsequent topic relevancy calculations

Experiment Results

- Topic Relevancy Comparisons

  High values indicate that a topic is highly related to a conversation and is more dominant than the other topics. Dominance is compared against the whole corpus to quantitatively measure off-topic detection. Some conversations have varied topic relevancy distributions showing that conversations are a mix of topics

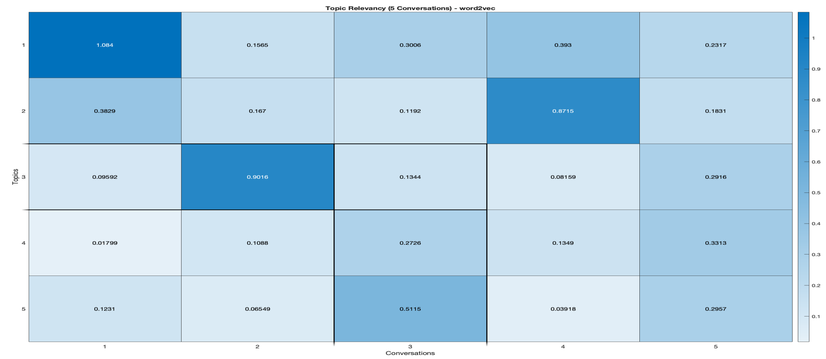

Figure 2. The comparisons of topic relevancies for each conversation in the set containing 5 unique topic dialogues

Conclusions

- Can separate conversations by topic by calculating topic relevancies.
- Differences in values between GloVe and word2vec are based on their cosine values (semantic information) given by their respective word embeddings
- Topic relevancy distributions are influenced by document-topic and topic-word probabilities determined by LDA

MSc Dissertation 2012-13 MSc Course: Computing Science, or ACS, IS, GD, KDD or CmpBio