Topic Detection

Marek Sarvaš

November 17, 2024

1 Approach

Based on the assignment that the data are from call center and domain or topic knowledge about it is unknown, the approach was to create dataset from publicly available datasets and apply unsupervised methods.

1.1 Dataset

I found two publicly available datasets: **DeepPavlov topics** and **MultiWOZ**. Both have some advantages and disadvantages. DeepPavlov topics dataset have data from various topics (30) which is more interesting to choose from. However, the data are mainly scraped internet comments and discussion and don't resemble a conversation.

MultiWOZ is a large dataset containing conversations on multiple topics such as: **restaurant, hotel, attraction, taxi, train, hospital and police**. The downside is that the dialogues contain turns from multiple domains, but usually every turn (sentence or couple of sentences) is from one domain.

In the end I chose MultiWOZ which format and content is, in my opinion, more suitable for this task. For my dataset 100 random dialogues were sampled from the MultiWOZ and stored as (turn, topic) pairs.

1.2 Methods

When following the assignment, there should be no knowledge about the domain and I don't have any labels. Therefore, unsupervised methods has to be used to found topics in these conversations. For this I chose a 3 step process: embedding extraction, clustering and keyword extraction. The specific methods chosen were inspired by Stanik et al.[1]

Embedding extraction

Embedding extraction from text data to gain a representation of each turn in conversation. For this I used a pre-trained transformer encoder model. This model was trained with sentence similarity objective and, therefore, should be able to obtain a good representation of each turn.

Clustering

The clustering of the high dimensional embeddings, where the clusters are created based on the conversation content. Embeddings in the same cluster should be semantically similar so they represent the same or similar topics. DBSCAN was chosen as the clustering method, because it is fast and it automatically infers number of clusters. On top of that some data points can be labeled as outliers, which should filter out very rare conversations.

Before clustering, there is need to reduce the dimensions of extracted embeddings, for this UMAP method was used and the embeddings where reduced to 20 dimension.

Keyword extraction

After the clustering, 5 biggest clusters were chosen as the 5 most common topics. All dialogue data from each cluster was concatenated, normalized and keywords and key-phrases were extracted. For the extraction process 3 methods were used resulting in various quality (evaluated subjectively). The methods were: 10 most common words in the cluster, key-phrases extracted using nltk library and key-phrases extracted with pre-trained transformer model.

[1] https://arxiv.org/pdf/2108.08543

2 Results

The results are heavily influenced by the clustering method, specifically **eps** parameter for DBSCAN. This parameter was calculated from distances between neighbors found by the the K-Nearest Neighbors algorithm on embeddings of various dimensionality sizes. For each **eps** a Silhouette coefficient was calculated to evaluate cluster quality as shown in Table1.

Table 1: Silhouette score and corresponding DBSCAN **eps** value for different embedding sizes.

Embedding dim	eps	Silhouette score
2	0.0003631114959716797	0.189
20	0.0006346702575683594	**0.659**
32	0.0006922483444213867	0.598
64	0.0008952021598815918	0.491

With these parameters the discovered topics are as following:

Topic 1

- Labels in cluster: mainly **Train**
- Key-words: train, would, like, need, cambridge, arrive, book, arrives, time, departing

Topic 2

- Labels in cluster: **Booking**, **Restaurant** and some **Train, Taxi**
- Key-words: reference, book, number, need, people, would, like, table, booking, nights

Topic 3

- Labels in cluster: N/A - mainly filler sentences
- Key-words: thank, great, welcome., thanks, goodbye., that's, you., day., need, good

Topic 4

- Labels in cluster: mainly **Attraction**
- Key-words: would, like, looking, centre, town., cambridge, go, phone, museum, area

Topic 5

- Labels in cluster: mainly **Hotel**
- Key-words: free, hotel, like, would, need, price, star, guest, parking., 4