CLUSTERLLM: Large Language Models as a Guide for Text Clustering

Yuwei Zhang Zihan Wang Jingbo Shang*
University of California, San Diego
{yuz163, ziw224, jshang}@ucsd.edu

Abstract

We introduce CLUSTERLLM, a novel text clustering framework that leverages feedback from an instruction-tuned large language model, such as ChatGPT. Compared with traditional unsupervised methods that builds upon "small" embedders, CLUSTERLLM exhibits two intriguing advantages: (1) it enjoys the emergent capability of LLM even if its embeddings are inaccessible; and (2) it understands the user's preference on clustering through textual instruction and/or a few annotated data. First, we prompt ChatGPT for insights on clustering perspective by constructing hard triplet questions <*does A better correspond to B than C*>, where *A*, *B* and *C* are similar data points that belong to different clusters according to small embedder. We empirically show that this strategy is both effective for fine-tuning small embedder and cost-efficient to query ChatGPT. Second, we prompt ChatGPT for helps on clustering granularity by carefully designed pairwise questions <*do A and B belong to the same category*>, and tune the granularity from cluster hierarchies that is the most consistent with the ChatGPT answers. Extensive experiments on 14 datasets show that CLUSTERLLM consistently improves clustering quality, at an average cost of ~$0.6[1] per dataset.

1 Introduction

Text clustering has been studied for years and it has recently gained new significance in identifying public perception from social media (Park et al., 2022), analysing cause of accidents (Xu et al., 2022) or detecting emerging research topics (Martínez et al., 2022). A common practice in text clustering is to apply clustering algorithms (MacQueen, 1967; Zhang et al., 2021a) on top of pre-trained embedders (Muennighoff et al., 2022; Wang et al., 2022; Su et al., 2022) which could achieve higher performance with better pre-training quality. However,

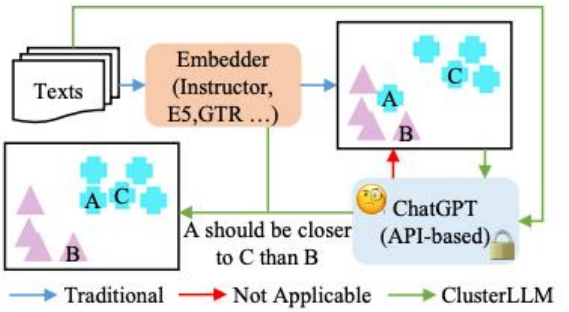

Figure 1: Traditional text clustering often employs clustering algorithms on top of the pre-trained embedders which could produce sub-optimal clustering. LLMs like ChatGPT are not applicable for text clustering directly because of the inaccessible embeddings. CLUSTERLLM resolves the dilemma by leveraging LLM as a guide on text clustering.

recent instruction-tuned LLMs such as ChatGPT, that demonstrated extraordinary language capabilities for various natural language applications by following textual instructions, can only be utilized through the APIs without accessible embedding vectors for clustering. Hence, LLMs can not be directly applied on text clustering tasks.

In this paper, we wish to provide insights on the following question:

- *Can we leverage API-based LLMs to guide text clustering efficiently?*

To approach such a challenge, we first draw inspiration from an observation that *humans represent an instance through comparing with others* (Nosofsky, 2011). For instance, people often classify a new piece of music into a specific genre by relating to familiar ones. In fact, pairwise relationships have been utilized in spectral clustering (Donath and Hoffman, 1972; Cheeger, 1970) before. Nonetheless, naively traversing all the pairs within dataset is obviously intractable for querying ChatGPT.

In this paper, we propose CLUSTERLLM, a generic framework that utilizes LLM to guide a small embedder for finding text clusters with a low

* Corresponding author.

[1]The cost is calculated with `gpt-3.5-turbo`.