Improving Hierarchical Text Clustering with LLM-guided Multi-view Cluster Representation for Interaction Drivers in Contact Centers

Anup Pattnaik **Cijo George** **Rishabh Kumar Tripathi**[†]
Sasanka Rani Vutla[†] **Jithendra Vepa**
Observe.AI, India
{anup.pattnaik, cijo.george, rishabh.tripathi}@observe.ai
{sasanka.vutla, jithendra}@observe.ai

Abstract

In this work, we present an approach that introduces different perspectives or views to improve the quality of hierarchical clustering of interaction drivers in a contact center. Specifically, we present a multi-stage approach that introduces LLM-guided multi-view cluster representation that significantly improves the quality of generated clusters. Our approach improves average Silhouette Score by upto 70% and Human Preference Scores by 36.7% for top-level clusters compared to standard agglomerative clustering for the given business use-case. We also present how the proposed approach can be adapted to cater to a standard non-hierarchical clustering use-cases where it achieves state-of-the-art performance on public datasets based on NMI and ACC scores, with minimal number of LLM queries compared to the current state-of-the-art approaches. Moreover, we apply our technique to generate two new labeled datasets for hierarchical clustering. We open-source these labeled datasets, validated and corrected by domain experts, for the benefit of the research community.

1 Introduction

Contact centers record interactions between their agents and customers and store them in the form of text transcripts for multiple downstream use cases like quality assurance, business analytics and insights. The primary reason for an interaction, often referred to as an *interaction driver* is an essential data point for some of these downstream use cases. Identifying these drivers at an interaction level can be automated with multiple techniques ranging from simple classification based on key phrases (Jindal and Liu (2006)) to Large Language Model (LLM) based generation in recent times (Casanueva et al. (2020a)). However, with contact centers handling hundreds of thousands of interactions a day, each with a unique driver, making sense of this data for downstream business use-cases is tedious and time-consuming.

Contact center interaction drivers are often thought of as having a two-level hierarchical structure consisting of a few main categories with several sub-categories under each main category. These are often referred to as level-1 (L1) and level-2 (L2) categories respectively. This makes it conducive to applying hierarchical clustering techniques to organize them into L1 & L2 clusters. However, while contact center business use cases call for the best quality especially at the top-level for L1 clusters, current state-of-the-art clustering techniques fall short of this. Specifically, L1 categories surfaced by existing methodologies often fail to bring out the right abstraction and multifaceted similarities within the L2 categories. Meanwhile, we notice that capturing this abstraction comes naturally to humans, and in recent times the best of Large Language Models (LLMs) (AI@Meta, 2024) have demonstrated this capability as well.

Different businesses have unique perspectives on how they prefer to cluster their L1 and L2 drivers. Table 1 illustrates the results of the generic clustering technique under the first perspective, where all queries related to a tourist destination form an L1 cluster. While this approach is logical from a modeling standpoint, businesses typically require more granular clustering to separate inquiries, booking modifications, and cancellations into distinct clusters for their downstream use cases, as shown under Perspective 2 in Table 1.

In this context, we present a methodology that generates L1 clusters that are better and more aligned with human-preferences from given L2 clusters for contact center interaction drivers. Specifically, we present a multi-stage clustering approach that introduces an LLM guided multi-view representation of L2 clusters to improve quality of L1 clusters. Our method employs standard agglomerative clustering to first derive the L2 clusters, and

[†] Equal contribution as third authors.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 719–727

Documents	Perspective 1		Perspective 2	
	Cluster Name	Cluster Description	Cluster Name	Cluster Description
Customer called in to cancel Niagara falls tour	Niagara Falls Tour	Customers calling in to inquire, book, modify and cancel bookings for Niagra Falls	Tour Cancellation	Drivers related to tour cancellations due to various reason
Customer wanted to add one more family member to the Niagara Falls Boating Experience	Niagara Falls Tour	Customers calling in to inquire, book, modify and cancel bookings for Niagra Falls	Tour Modification	Customer calling in to modify a booking they made with the tours company

Table 1: Different Perspectives on Clustering Interaction Drivers for the Travel Domain

then introduces a weighted multi-view embedding representation of the L2 clusters to explicitly capture its different facets before generating the L1 clusters. The latter step captures and incorporates the semantic abstractions that drive different human perspectives into the clustering process. The proposed approach improves Silhouette Scores on our internal datasets by up to 70%, and Human Preference Scores by up to 36.7%.

We also present how this approach can be adapted for standard non-hierarchical clustering approaches as an alternative to current state-of-the-art approaches (Zhang et al. (2023); Raedt et al. (2023)). Benchmarking experiments on public intent-classification datasets which are used for evaluation of clustering techniques, specifically Banking77 (Casanueva et al. (2020b)) and CLINC150 (Larson et al. (2019)) shows that our approach achieves close to state-of-the-art performance measured by Normalized Mutual Information (NMI) (Strehl and Ghosh (2002), Danon et al. (2005)) and Clustering Accuracy (ACC) (Kuhn (1955)) scores. Our approach is also cost effective with number of LLM queries limited to twice the number of L2 clusters, while LLM queries needed by the above mentioned existing approaches increases linearly with number of documents in the dataset.

Moreover, we apply our approach to generate L1 clusters on top of existing base clusters for Banking77 and CLINC150 datasets. The new datasets consist of 7 and 15 L1 clusters respectively, and we call them Banking7 and CLINC15. The L1 clusters generated are validated and corrected by domain experts, and we open-source these datasets as part of this work for the benefit of the research community.

To summarize, below are our specific contributions in this work.

- We introduce the problem space and motivation for generating L1/L2 clusters for interaction drivers in contact centers and the challenge of lack of perspectives with existing approaches.
- We propose a novel multi-stage clustering approach that introduces multi-view representations for L2 clusters to improve the quality of L1 clusters and better align with human-preferences.
- We demonstrate how the proposed approach can be adapted for standard non-hierarchical clustering use-cases to achieve state-of-the-art performance compared to recent LLM-guided approaches while being cost effective.
- We open-source two new hierarchical clustering datasets derived from existing intent classification datasets for the benefit of the research community.

2 Methodology

The proposed approach consists of four key stages as illustrated in Figure 1 and detailed below.

2.1 Stage 1: Deriving L2 Clusters with Agglomerative Clustering

We first employ standard agglomerative clustering (Murtagh and Legendre (2014)), a bottom-up hierarchical clustering approach that starts by treating each document as an individual cluster and then iteratively merge the closest pairs of clusters until a predefined number of clusters K is reached.

The resulting clusters $\{C_i, C_2, \ldots, C_K\}$ represent the L2 clusters.

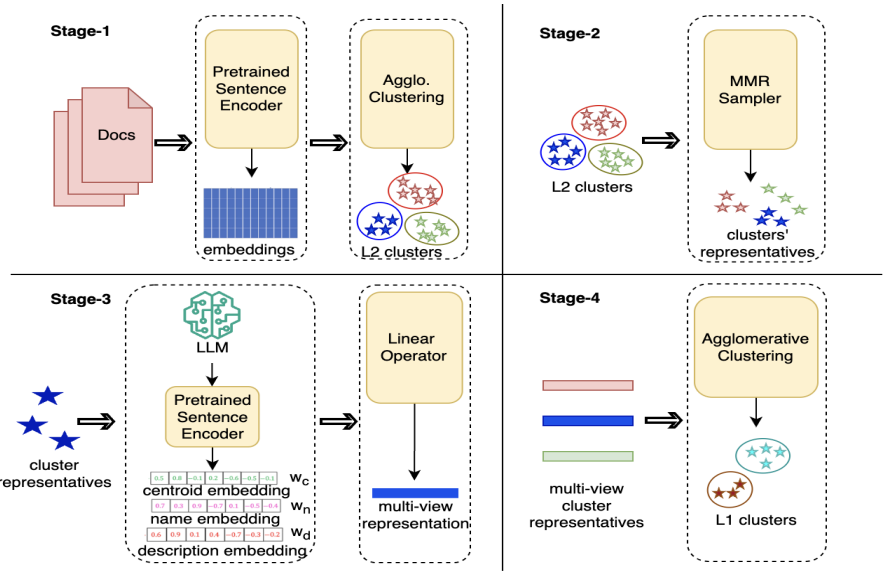

Figure 1: The end-to-end process of Multi-view Hierarchical Clustering. Stage 1 involves encoding documents and applying agglomerative clustering to generate L2 clusters. Stage 2 uses MMR sampling to select representative documents from each cluster. Stage 3 leverages an LLM to refine cluster representations through multi-view embeddings. Finally, Stage 4 applies agglomerative clustering to form L1 clusters from the multi-view representations

2.2 Stage 2: Sampling Representative L2 Cluster Documents

For each L2 cluster C_k, we sample a subset of representative documents $R_k \subset C_k$ from the set of documents x_{c_k} belonging to that cluster. To ensure that the documents are representative of the cluster, we sample based on Maximal Marginal Relevance (MMR) (Carbonell and Goldstein (1998)), which balances relevance and diversity in information retrieval. This technique iteratively select documents based on a trade-off between their relevance to the query and their dissimilarity to the documents already selected.

2.3 Stage 3: Generating LLM-guided Weighted Multi-view Representations

Using the representative documents R_k, we leverage an in-house LLM to generate a concise cluster name CN_k (of 3-5 words), and a cluster description CD_k for each L2 cluster (less than 50 words), using tailored prompts for each task. The in-house LLM is a Llama-3 Instruct 8B model by AI@Meta (2024)), supervised fine-tuned (SFT) on 60K data points generated using the GPT-4-0314 API (OpenAI (2024)), with human-in-the-loop validation. For each L2 cluster C_k, we create 3 different views: The centroid embedding e_{c_k} obtained by taking the average of all documents in a L2 cluster, the cluster name embedding e_{n_k} and the cluster description embedding e_{d_k}. These embeddings are combined into a single multi-view representation e_k using a weighted average as follows:

$$e_k = w_c e_{c_k} + w_n e_{n_k} + w_d e_{d_k},$$

where w_c, w_n and w_d are the weights assigned to the centroid, name and description embeddings, respectively. These weights can be tuned to optimize clustering performance.

Incorporating representations of cluster names and descriptions as different views bring in abstract semantic information about the intermediate clusters (L2) into their embedding representations. This helps align the next higher level clusters (L1) better with human preferences.

2.4 Stage 4: Generating L1 Clusters using Weighted Multi-view Representations

The weighted multi-view representations $\{e_1, e_2, \ldots, e_K\}$ are then input to an agglomerative clustering algorithm to derive the broader L1 categories. The algorithm clusters these multi-view representations into M clusters $\{L_1, L_2, \ldots, L_M\}$, representing the L1 categories. To enhance explainability, we apply the same strategy to generate cluster names and descriptions for the L1 clusters as well.

Internal Datasets	M	K
Quick Commerce	8	52
Education	6	48
Travel	5	40

Table 2: Pre-defined number of L1 clusters (M) and L2 clusters (K) across internal datasets

3 Evaluation on Internal Datasets

3.1 Experimental Setup

We evaluate the effectiveness of the proposed approach on internal datasets* from three distinct domains: Quick Commerce, Education, and Travel. Each dataset is composed of interaction drivers generated by an in-house LLM from 5,000 real contact center interactions within the respective domains. The interactions are sampled from both live chat sessions and call transcripts, providing a diverse representation of customer communications. These interactions encompass a wide range of user queries and issues, providing a robust test bed for our clustering approach. In our internal dataset experiments, we set pre-defined values for M and K, as presented in Table 2. These values are based on specific business requirements and operational workflows, ensuring that the experimental setup aligns with practical use cases and domain-specific needs.

For each of these datasets, we compute both L2 and L1 clusters using our proposed methodology. We set the parameters for agglomerative clustering that gives L2 clusters with the best silhouette score for a given domain. Multi-view representation is generated for each of the L2 clusters as described in Section 2.3, and they are clustered again using agglomerative clustering to arrive at L1 clusters.

We benchmark our approach against standard agglomerative clustering without multi-view representation. We employ two embedding models for evaluation: Sentence Transformer MPNet (all-mpnet-base-v2) (Reimers and Gurevych (2019)) and the Instructor Model (Wang et al. (2020)). Sentence Transformer MPNet is recognized for its superior performance in semantic textual similarity tasks, making it suitable for capturing the nuanced differences in interaction drivers. Instructor Model, on the other hand, is designed to incorporate instructional data, enhancing its ability to understand and categorize complex interactions.

Given the lack of ground truth labels for the L1/L2 clusters in our internal datasets, we use Silhouette Score (Rousseeuw (1987)) as the evaluation metric. Specifically, we use the average score across samples in a cluster, where the score for a sample i is given by:

$$\text{silhouette\_score}(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))},$$

where a_i is the average distance of sample i to all other samples in its cluster, and b_i is the average distance of sample i to all samples in the cluster nearest to i.

For further validation, we also use domain experts to annotate the quality of the clusters that we are generating. We sample 50 driver documents from each L1 cluster and provide the domain experts with the following: Interaction Driver text, L1 cluster name, L1 cluster description and List of top 3 most similar clusters to the tagged L1. Top 3 most similar clusters for each L1 cluster are obtained based on cosine similarity between the L1 cluster centroids.

The domain experts are posed with the following question - *Does the given interaction driver belong to the given cluster?* and they have to annotate on a 5 point Likert scale (Jebb et al., 2021) where 5 is *Strongly Agree* and 1 is *Strongly Disagree*. We average the scores on the Likert scale to come up with the Human Preference Score. Domain experts are provided with comprehensive guidelines to ensure labeling consistency across the datasets. Each data point is independently labeled by five annotators, achieving inter-annotator agreement with a Kappa score of 0.76. While the authors define the annotation guidelines, they do not participate in the actual annotation process.

3.2 Results

Results in Table 3 show that the proposed approach leads to at least 47% and up to 70% better average Silhouette Scores across the domains, compared to standard agglomerative clustering. There was also significant increment of 36.7%, averaged across all datasets, on the human preference score, which is the critical business metric.

We conduct the following ablation studies and evaluation of the impact of variations in configurable parameters.

*The dataset cannot be released/open-sourced due to proprietary reasons.

Approaches	Quick Commerce		Education		Travel	
	Silhouette Score	HPS	Silhouette Score	HPS	Silhouette Score	HPS
Std. Agglomerative w/ MPNet	0.035	3.262	0.038	3.39	0.039	3.411
Std. Agglomerative w/ Instructor	0.044	3.423	0.040	3.445	0.043	3.484
Proposed Approach w/ MPNet	**0.053**	4.412	**0.059**	4.563	**0.064**	4.57
Proposed Approach w/ Instructor	**0.065**	4.682	**0.068**	4.711	**0.071**	4.728

Table 3: Silhouette and Human Preference Scores (HPS) of L1 clusters across different approaches and domains

w_c	w_n	w_d	Q. Comm.	Education	Travel
1.0	0.0	0.0	0.0458	0.049	0.054
0.0	1.0	0.0	0.032	0.04	0.051
0.0	0.0	1.0	0.03	0.038	0.044
0.5	0.5	0.0	0.046	0.042	**0.064**
0.0	0.5	0.5	0.034	0.039	0.046
0.5	0.0	0.5	0.044	0.04	0.048
0.34	0.33	0.33	0.05	**0.059**	0.056
0.5	0.25	0.25	**0.053**	0.053	0.058

Table 4: Impact of different views

# docs sampled	Sampling Strategy	
	Random	MMR w/ 0.4 diversity
5	0.015	0.022
10	0.018	0.026
20	0.015	**0.053**
50	0.013	0.029

Table 5: Silhouette Scores w/ Different Sampling Strategies

3.2.1 Impact of Views

To understand the impact of the three views introduced in multi-view cluster representation, we vary the weights of each of the views. Note that varying the weights do not significantly impact the overall costs, as this process occurs after the LLM has been invoked to generate the L2 cluster names and descriptions. Results from this exercise shown in Table 4 show the following trends.

- The highest silhouette scores across all domains are achieved through multi-view clustering rather than any single view, underscoring the critical importance of integrating multiple perspectives.
- Removing centroid view significantly reduces average silhouette scores across all domains, showing the importance of this view across all domains
- Name view contributes more significantly to the clustering quality than the description view based on these domains.

3.2.2 Impact of Sampling Strategy

The sampling strategy employed to select representative documents of each L2 cluster for name and description generation consists of two factors - the sampling algorithm, and the number of documents sampled. For the former, we study the impact of random sampling compared to MMR. For the latter, we vary the number of documents sampled well.

Results presented in Table 5 show that MMR sampling consistently outperforms random sampling across all evaluation metrics. Increasing the number of sampled documents generally improves performance up to a certain point, with the most significant improvements observed from 10 to 20 documents. Beyond 20, the cluster quality declines. One possible reason for this could be LLM's limitations in handling large contexts effectively.

4 Extending to Non-Hierarchical Clustering

Text clustering research in recent times proposing LLM-guided approaches have reported state-of-the-art performance on labeled public datasets. While our methodology in this paper is primarily targeted towards hierarchical clustering, we posit that this approach can be adapted to improve quality of clusters generated for standard clustering use-cases, and can provide a more efficient alternative to the current state-of-the-art techniques.

To adapt our approach to standard clustering use-cases, we assume that the final output clus-

ters required are L1 clusters, and there exists a layer of hidden L2 clusters. We apply the proposed multi-view cluster representation to the hidden L2 clusters before they are clustered again to generate the L1 clusters. This approach brings in different perspectives through multi-view representation as a light-weight one-time intervention during a bottom-up clustering process to improve quality of the final clusters generated.

4.1 Datasets and Baselines

To benchmark the proposed approach for improving standard clustering, we take the following popular labeled intent classification datasets - Banking77 and CLINC150. Banking77 comprises of 13083 customer service queries from banking domain labeled with 77 intents. CLINC150 comprises of 150 intents and 23700 samples across 10 domains. We consider the labeled intent classes as the L1 clusters and assume a hidden L2 layer with 500 clusters.

We evaluate the alignment between the generated L1 clusters and the labeled intents using NMI and ACC scores. To establish a robust baseline for our approach, we draw comparisons with two recent methodologies in intent discovery and text clustering that report state-of-the-art performance: IDAS Raedt et al. (2023) and ClusterLLM Zhang et al. (2023). IDAS highlights the efficacy of using abstractive summaries for intent discovery, while ClusterLLM demonstrates the advantages of integrating LLM feedback for improving clustering accuracy and granularity.

4.2 Results

Our approach improves NMI scores by 10.3% and 9.2% over standard agglomerative clustering, using MPNet and Instructor embedding models respectively. Corresponding increase in ACC is 11.3% and 11.7%. We achieve state-of-the-art NMI and ACC scores of 94.2 and 86.2 respectively on CLINC150 dataset, and are very close to numbers reported by ClusterLLM on Banking77 dataset. The reported performance is with the number of L2 clusters set to 500. We observe a variation of less than 2% for NMI and ACC scores with number of L2 clusters varying from 500 to 1000. Our primary objective is to demonstrate the feasibility of the proposed approach for non-hierarchical datasets. The consistency of results across different number of L2 clusters reinforces the robustness of our method.

Moreover, the number of LLM queries required for IDAS and ClusterLLM increases linearly with the number of documents being clustered. In contrast, our approach requires LLM queries proportional to twice the number of intermediate L2 clusters, and is independent of the total number of documents. Hence, we argue that our approach is more cost effective while still achieving state-of-the-art clustering results.

5 Open-source Labeled Dataset for Two-level Hierarchical Clustering

We applied our proposed approach to Banking77 and CLINC150 datasets to generate a 7 and 15 L1 clusters respectively. The number of L1 clusters is determined by optimizing for silhouette scores. As this optimization is performed after generating names and descriptions using the LLM, this step do not significantly impact our computational costs. The generated clusters were validated/ corrected through the following annotation process. Annotators were given names and descriptions of the existing intent classes, derived using our proposed approach along with text samples from the intent class and the corresponding L1 cluster generated. They were asked to verify if the tagging of an intent class to an L1 cluster was correct and if not, to reassign the intent class to the correct L1 cluster. We open-source the labeled two-level hierarchical dataset thus created as an additional contribution to the community[†].

6 Related Works

Subjectivity in definition of multi-view: Supported by Chao et al. (2017) which states that multi-view data is useful in solving real-world applications in the big data era. Prior works (Kumar and III, 2011; Kumar et al., 2011) utilized different language representations of the same unit to represent its diversified views. Similarly, in the multimedia domain, Petkos et al. (2014) used various modalities to represent unique perspectives of the same entity. In this work, each view is derived on the basis of cluster attributes, particularly name, description and its centroid.

Evolution of LLM-guided clustering: Prior works like Wang et al. (2023) proposed a Propose-Assign-Select strategy demonstrating the use of

[†]https://github.com/Observeai-Research/hierarchical-clustering-data-corpus

Approach	Banking77			CLINC150		
	NMI	ACC	Silhouette Score	NMI	ACC	Silhouette Score
Std. Agglomerative w/ MPNet	73.2	58.6	0.072	81.2	74.2	0.083
Std. Agglomerative w/ Instructor	76.4	60.1	0.085	84.5	76.1	0.092
IDAS	82.84	67.43	-	93.82	85.48	-
ClusterLLM w/ Instructor	**85.15**	**71.2**	-	94	83.8	-
Proposed Approach w/ MPNet	82.9	67.5	0.108	92.9	82.6	0.12
Proposed Approach w/ Instructor	84.9	69.6	**0.12**	**94.2**	**86.2**	**0.145**

Table 6: Evaluation on Public Intent Classification Datasets

gpt-4 (proposer) and claude-v1.3 (assigner) to indicate whether or not text samples should belong to a particular cluster. Similarly, motivated by the fact that LLM like chatgpt can't be used for clustering due to unavailability of its embeddings, Zhang et al. (2023) proposed using LLM as a guide for sensibly decide merging of two data points at each step of clustering. Furthermore, Viswanathan et al. (2024) extended LLM-guided clustering to semi-supervised setup by targeting the low-confidence points in the clusters and use LLM guidance to assign them to most relevant cluster.

Exploring Hierarchical Datasets: Prior works demonstrated the evolution of data corpora by introducing hierarchy in labels, hence, extending the research opportunities for hierarchical clustering. For instance, Web of Science (Kowsari et al., 2017) was released in varying sizes and number of parent-child categories, covering diverse scientific domains: WOS-11967, WOS-46985, WOS-5736. Similarly, Petukhova and Fachada (2022) released the Multi-labeled News Dataset (MN-DS), a hierarchical dataset for news classification with categories defined in two-levels of hierarchy. However, these data corpora have not been extensively explored by the research community, hence, making it challenging to benchmark experimental results.

7 Limitations and Future Work

Our research showcases the effectiveness of the proposed methodology in generating hierarchical clusters, but there are several key areas for future exploration and limitations to consider.

First, we limited our experiments to agglomerative clustering. However, our methodology is clustering algorithm-independent, suggesting that future work could investigate various algorithms, such as k-means, DBSCAN, or spectral clustering, to enhance L1 and L2 cluster formations across diverse datasets.

Second, our current framework employs a specific external LLM for generating cluster names and descriptions. Future research could benchmark different LLM architectures and sizes to determine which configurations yield the most meaningful cluster representations.

Lastly, determining the optimal number of L2 clusters remains a challenge in unsupervised clustering. Future work could focus on developing efficient methods for this task, potentially employing advanced heuristics or hybrid approaches to improve robustness and applicability.

In summary, while our study provides a strong foundation, there are ample opportunities to extend this research by exploring diverse clustering techniques, evaluating LLM performance, and optimizing the clustering process

8 Conclusion

In this paper, we present a multi-stage approach for the hierarchical clustering of interaction drivers in contact centres that achieves significantly better quality for the top-level clusters. We propose to leverage LLM-guided multi-view intermediate cluster representations as part of the clustering process to obtain more coherent and meaningful top level clusters. Our approach despite using out-of-the-box embedding models and requiring minimal LLM queries (twice #L2 clusters), achieves better Silhouette Scores for our internal datasets, and state-of-the-art NMI and ACC scores on public datasets. We also release two labeled datasets for hierarchical clustering for the benefit of the research community.

9 Ethical Considerations

1. The data used in this work include contact center conversations between agents and customers that often contains sensitive PCI/PII information. We ensure that all such sensitive information is redacted at the source before they are processed through our pipeline. Moreover, all of our computation happens in-house and no data is sent out to any external services.

2. We use Language Models in this work, which can potentially exhibit biases. We take proactive measures to prevent such bias including carefully designing prompts to prevent biases, ensuring that any data used for fine-tuning language models are free from such biases and systematic audit of model outputs.

References

AI@Meta. 2024. Llama 3 Model Card.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020a. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020b. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Guoqing Chao, Shiliang Sun, and Jinbo Bi. 2017. A survey on multi-view clustering. *CoRR*, abs/1712.06246.

Leon Danon, Albert Díaz-Guilera, Jordi Duch, and Alex Arenas. 2005. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008–P09008.

Andrew T Jebb, Vincent Ng, and Louis Tay. 2021. A review of key likert scale development advances: 1995–2019. *Frontiers in psychology*, 12:637547.

Nitin Jindal and Bing Liu. 2006. Mining comparative sentences and relations. volume 2.

Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, , Matthew S Gerber, and Laura E Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*. IEEE.

Harold W. Kuhn. 1955. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97.

Abhishek Kumar and Hal Daumé III. 2011. A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 393–400. Omnipress.

Abhishek Kumar, Piyush Rai, and Hal Daumé III. 2011. Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1413–1421.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Fionn Murtagh and Pierre Legendre. 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of classification*, 31:274–295.

OpenAI. 2024. GPT-4-0314 API.

Georgios Petkos, Symeon Papadopoulos, Emmanouil Schinas, and Yiannis Kompatsiaris. 2014. Graph-based multimodal clustering for social event detection in large collections of images. In *MultiMedia Modeling - 20th Anniversary International Conference, MMM 2014, Dublin, Ireland, January 6-10, 2014, Proceedings, Part I*, volume 8325 of *Lecture Notes in Computer Science*, pages 146–158. Springer.

Alina Petukhova and Nuno Fachada. 2022. Mn-ds: A multilabeled news dataset for news articles hierarchical classification.

Maarten De Raedt, Fréderic Godin, Thomas Demeester, and Chris Develder. 2023. Idas: Intent discovery with abstractive summarization. *Preprint*, arXiv:2305.19783.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Peter Rousseeuw. 1987. Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65. *Journal of Computational and Applied Mathematics*, 20:53–65.

Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.

Vijay Viswanathan, Kiril Gashteovski, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2024. Large language models enable few-shot clustering. *Trans. Assoc. Comput. Linguistics*, 12:321–333.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *CoRR*, abs/2002.10957.

Zihan Wang, Jingbo Shang, and Ruiqi Zhong. 2023. Goal-driven explainable clustering via language descriptions. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. ClusterLLM: Large language models as a guide for text clustering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13903–13920, Singapore. Association for Computational Linguistics.