Cross-Lingual Clustering Using Large Language Models

Nicole R. Schneider
University of Maryland
College Park, MD, USA
nsch@umd.edu

Avik Das
University of Maryland
College Park, MD, USA
adas1236@terpmail.umd.edu

Kent O'Sullivan
University of Maryland
College Park, MD, USA
osullik@umd.edu

Hanan Samet
University of Maryland
College Park, MD, USA
hjs@umd.edu

Abstract

Text clustering methods traditionally rely on a shared vocabulary and script, which poses a challenge for cross-lingual text clustering problems that arise in a variety of domains including social media, news, finance, and more. Recent approaches to cross-lingual clustering have found success by leveraging latent embedding space representations of neural models and more recently by directly using Large Language Models (LLMs) to do text clustering in zero-shot or few-shot settings. However, much of the recent work focuses on short text, like social media posts. In this paper, we use cross-lingual clustering in the news domain as a case study to test whether LLMs can effectively cluster long documents by extracting and maintaining keyphrases associated with each cluster of documents. We compare the clustering several LLMs produce in a zero-shot setting to a more traditional online clustering method that uses TF-IDF to cluster documents based on their content and time of publication. We find that LLMs tend to cluster the articles based on the text, in particular based on the language of the text more than the content, and ignore the time and location of publication, indicating further work is needed before LLMs can reliably be used in clustering news articles across multiple languages.

CCS Concepts

• **Computing methodologies** → **Natural language generation**; Online learning settings; • **Information systems** → **Clustering**; *Geographic information systems.*

Keywords

Cross-lingual clustering, large language model, news domain, event clustering, spatio-temporal clustering

ACM Reference Format:
Nicole R. Schneider, Avik Das, Kent O'Sullivan, and Hanan Samet. 2024. Cross-Lingual Clustering Using Large Language Models. In *7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GeoAI'24), October 29-November 1, 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3687123.3698280

GeoAI'24, October 29-November 1, 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-1176-3/24/10
https://doi.org/10.1145/3687123.3698280

1 Introduction

Text clustering is a problem that is relevant across a number of domains, including news, social media, health, finance, and more. While there exist a number of traditional clustering methods that work well for textual data, many rely on a shared vocabulary (i.e. TF-IDF [22]). Clustering in a cross-lingual setting is particularly challenging, since there may not be a shared vocabulary or script across the languages, and so this setting has garnered recent attention [30].

Recent approaches to cross-lingual text clustering have found some success by leveraging latent embedding space representations from neural models trained to perform machine translation [1, 6, 18, 31]. More recently Large Language Models (LLMs) themselves have been used to do text clustering in zero-shot or few-shot prompting approaches [38, 40]. The benefit of using LLMs to perform or aid in cross-lingual text clustering is that they contain powerful word models for the languages they were exposed to during training over massive amounts of textual data. Further, LLMs are flexible and capable of generating text on demand across a variety of languages, including ones with different vocabularies and scripts, making them potentially useful for the task of cross-lingual text clustering.

Despite these apparent advantages, much of the recent work in clustering text documents using LLMs has focused on clustering short documents, such as social media posts [4, 7, 28, 38], in a single language, such as English. The challenge remains to automatically cluster longer documents, such as publications or news articles, written in different languages, which is relevant in a number of domains.

In this paper, we explore the viability of applying LLMs to the problem of cross-lingual text clustering by performing a series of experiments using data from the news domain in an online clustering setting. We build on previous work in cross-lingual clustering in the news domain [30] by testing whether LLMs can provide value beyond traditional clustering by extracting key phrases and using them to determine which cluster a new article belongs in. Further, we introduce key metadata fields, including the time of publication and location associated with each article, to aid in clustering and measure how well LLMs are able to account for these pieces of information when deciding which cluster an article belongs in. We compare the cluster assignments several LLMs produce in a zero-shot setting against a more traditional online clustering method that uses machine translation followed by TF-IDF to cluster documents based on their content and time of publication.

We find that in the single-language setting, LLMs produced cluster assignments that had lower Normalized Mutual Information (NMI) than the traditional baseline method. Further, providing the LLMs with article title, time of publication, and location led to worse cluster assignments that using the text alone. In the cross-lingual setting, we found that both the baseline method and the LLMs tended to assign articles to clusters based on the language of the document, rather than its content. Our results indicate that further work is needed before LLMs can be used to reliably cluster large documents, especially ones spanning multiple languages. In addition to our findings, we provide a hand-labeled dataset of English articles and an unlabeled dataset of articles spanning multiple languages to enable future testing of newer LLMs and other prompting strategies and clustering methods on the task of cross-lingual clustering of long documents.

The rest of this paper is organized as follows. In section 2 we describe the news datasets that we create to evaluate mono-lingual and cross-lingual clustering ability. Then, we describe our experimental setup in section 3, and our results in section 4. Next, we discuss the significance of our findings in section 5 and summarize related work in LLM and neural clustering in section 6. Finally, we discuss future work and conclude in section 7.

2 Data Sources

To study how well LLMs can cluster English and cross-lingual news articles, we construct datasets using NewsStand [25, 26, 36], a system designed to allow users to read the news using a map interface. The system ingests articles from thousands of RSS feeds within minutes of publication and presents them to users on a map, with each article's location inferred from its geographic references. The NewsStand interface is dynamic, so articles are collected, processed, and clustered in real time as they are published. Using NewsStand, we construct two datasets: *NewsStandEN* and *NewsStandMULTI*, containing English and non-English news articles, respectively.

2.1 Data Pre-Processing

Both datasets are processed in the following manner. We geo-tag each article to identify geographic terms in the article text. Geotagging consists of three steps: Entity Feature Extraction, Gazetteer Record Assignment, and Geographic Name Disambiguation [36]. The first stage, Entity Feature Extraction, involves identifying important entities in the text and collecting them in an entity feature vector (EFV). This is accomplished using a combination of Part-Of-Speech (POS) tagging and statistical Named-Entity Recognition (NER) tagging [41]. For more details on this process, see [5, 9, 13, 27]. Once extracted, the EFV contains words belonging to proper noun classes, including location. Location entities in the EFV are then assigned a set of matching locations during the Gazetteer Record Assignment stage and the toponyms are resolved during the Geographic Name Disambiguation phase [9–12, 24, 29]. The resolved toponyms are then used to determine the geo-coordinates associated with each news article, using the method outlined in Teitler et al. [36] and Lieberman et al. [13]. Geo-coordinates associated with the strongest geographic reference are assigned to the article, based on a weighted score that emphasizes frequency of the geo-reference within the text, and nearness of the geo-reference to the beginning of the text [19, 20]. Along with the geo-coordinates, the timestamp representing the time of publication of the article is retained to aid with clustering. Additional metadata, including the title of the article, is retained along with the text, timestamp, and geo-coordinates.

We use *Lingua,* [1] a language detection model that uses statistical and rule-based information, to identify the language of the text in each article and assign an appropriate language tag. The language tags are then used to split the articles into two datasets: *NewsStandEN* and *NewsStandMULTI*, containing English and non-English news articles, respectively.

2.2 NewsStandMULTI

The *NewsStandMULTI* [2] dataset consists of 68 unlabeled news articles in 7 different languages spanning a variety of topics, including entertainment, sports, finance, and more. The articles had an average of 965.53 words each, with the lower quartile having 751 or less and the upper quartile having 1169 or more. The lengths of the articles ranged 256 words to 1974 words. Table 1 summarizes the language distribution of the *NewsStandMULTI* dataset.

Language Distribution of *NewsStandMULTI* Dataset

Original Language	ISO ALPHA 2 Code	Number of Articles
German	DEU	47
Spanish	SPA	6
Italian	ITA	6
French	FRA	4
Portuguese	POR	2
Dutch	NLD	2
Czech	CES	1

Table 1: Language distribution of articles in the *NewsStandMULTI* Dataset. The articles sampled were published between June 2019 and April 2020.

2.3 NewsStandEN

The NewsStand dataset containing English-only articles with ground truth cluster labels (termed *NewsStandEN* [3]) contains a sample of 100 news articles and their metadata. The articles had an average of 778.75 words each, with the lower quartile having 482 or less and the upper quartile having 883.5 or more, and range between 30 words and 5545 words, making NewsStandEN articles slightly shorter than NewsStandMULTI articles overall. The pre-processing steps for this dataset include an additional hand labeling phase. Cluster labels for this dataset were hand-annotated, following the method in Zhang et al. [39]. First, one or more keywords describing the news event were assigned to each article by one of three annotators. Then, clusters were initially constructed using the keywords, geo-locations, and timestamps for context. The clusters were then incrementally improved by reviewing the article texts that were

[1] https://github.com/pemistahl/lingua-py
[2] https://github.com/osullik/multilingual_llm_clustering/blob/main/data/raw/spatial_nlp_non_en_detected_language.csv
[3] https://github.com/osullik/multilingual_llm_clustering/blob/main/data/raw/spatial_nlp_en_detected_language.csv

grouped in each cluster to determine if the cluster should be split (articles describe different but similar news events) or combined (multiple clusters contain articles about the same news event). After several rounds of incremental cluster adjustments, the cluster assignments were finalized upon annotator agreement.

3 Method

In this section we describe our experimental setup, including the baselines for comparison and the prompts we use to elicit cluster assignments from the LLMs. All clustering is done in an online fashion to simulate a live environment where news articles are arriving as they are published, in real time. For the single language setting, we use *NewsStandEN*, which contains 100 English articles with hand labeled ground truth cluster assignments.

3.1 Baseline Clustering: NewsStand

In the news domain, clustering is used to group together *story clusters* containing all news articles that describe the same news event. In addition to the requirement that articles in the same cluster share many of the same keywords, they also must be published around the same timeframe. The temporal requirement stems from the emphasis on recency when presenting breaking stories to users. This premise lends itself well to online clustering, which requires less computation than one-shot approaches that involve re-clustering the entire corpus with every new article ingested [36].

To accomplish the clustering, NewsStand employs the vector space model [23], a common approach in text mining and information retrieval. If the articles are not originally written in English, they are first translated into English using machine translation. Then, the articles are converted to term feature vectors in d-dimensional space, where d is the number of distinct terms in every document in a corpus. The term feature vector is extracted using TF-IDF [22]. Elements of the term feature vector represent the frequency of their corresponding term in the document being ingested, where terms that are common in a document but uncommon in the corpus are emphasized. Emphasis is given to location entities and to terms that appear earlier in the article. Since NewsStand is an online system with a dynamic corpus, the term feature vector is computed once for each article at the time it is ingested into the system.

Clustering is also done in an online fashion using a variant of leader-follower clustering [3]. Articles are clustered across two dimensions: the term vector space and the temporal dimension. A term centroid and a time centroid are maintained for each cluster, representing the mean term feature vector and mean publication time of the articles in the cluster, respectively. For each new article ingested, clustering proceeds by checking if there exists a cluster with centroids less than a fixed cutoff distance from the article's term and time values. If so, the article is added to the nearest cluster and its centroids are updated, and if not, a new cluster is created containing only the new article. Term distances are computed using the standard cosine similarity [34], and a Gaussian attenuator is applied to the temporal dimension to favor clusters with time centroids near the article's publication time.

3.2 LLM Clustering

We compare the baseline clustering method to an LLM-based clustering method, where LLMs are directly prompted to cluster articles. On the English-only dataset, for which we have ground-truth cluster labels, we test several prompt variants to determine whether including article titles, timestamps, and geo-coordinates enable LLMs to achieve better clustering results compared to the baseline clustering method. The outcome of those experiments determine the prompts used to conduct clustering on the multi-lingual dataset, for which we do not have ground truth cluster labels.

To achieve the online clustering effect through zero-shot LLM prompting, we follow the LLM prompting strategy in O'Sullivan et al. [16] and provide an initial system prompt stating the task, and then a series of prompts containing individual articles that should be assigned to a cluster. To retain the context of the previous articles that have already been assigned clusters, we prompt the LLMs to output both article ids and keywords associated with each cluster at each step, and these are fed into the subsequent prompt to accumulate the full clustering of all the articles over the course of each experiment. For experiments that include geo-coordinates and timestamps, we maintain the average geo-coordinate and/or timestamp of the articles in each cluster and also provide that to the LLM to enable location-based and time-based clustering.

Developer	Model	$ per M/Tok
OpenAI	gpt-3.5-turbo	00.50/01.50
	gpt-4o	05.00/15.00
	gpt-4o mini	0.150/0.600
Google	gemini-1.5-flash	00.35/01.05

Table 2: Summary of models evaluated.

3.2.1 Models. We select 4 models from two of the leading LLM developers. Table 2 summarizes the selected models and indicative cost of use. For experimentation, temperature is set to 0 on each model and where available, constant seed values are set to reduce the impact of randomness in generation.

3.2.2 Prompts.

E1. Do LLMs perform better clustering with domain information in the initial system prompt?
To test this we compare the clustering results produced when the LLMs are given a simple prompt stating the goal is to cluster articles vs. a more detailed prompt stating the goal is to form clusters that represent news stories.

```
You are tasked with clustering articles. Cluster
 these articles based on their text content. You will
  be fed articles one by one in the form:

id: {id}

text: {text}

```

```
Respond with a tuple of predicted cluster and a list of keywords associated with the article. If the cluster already exists, update the list of keywords so that it summarizes both the original cluster and the new article. You can use no more than 5 keywords per cluster. If the article does not belong in an existing cluster, create a new one, incrementing the cluster number by 1. The existing clusters are described in the dictionary provided to you after the article.
```

Prompt 1: Generic Initial System Prompt

```
You are tasked with clustering articles for a news system based on whether they express ideas about the same news story or world event. Cluster these articles based on their text content, with the aim to consistently group together articles about the same news event. You will be fed articles one by one in the form:

id: {id}

text: {text}

Respond with a tuple of predicted cluster and a list of keywords associated with the article. If the cluster already exists, update the list of keywords so that it summarizes both the original cluster and the new article. You can use no more than 5 keywords per cluster. If the article does not belong in an existing cluster, create a new one, incrementing the cluster number by 1. The existing clusters are described in the dictionary provided to you after the article.
```

Prompt 2: Detailed Initial System Prompt

For *E1*, we first give each LLM Prompt 1 and then feed it articles from *NewsStandEN* one at a time. Along with each article the output of the previous interaction is provided as the context history, to maintain the cluster assignments made previously. A similar process is repeated using Prompt 2 as the initial system prompt, and the cluster assignments are recorded for each model under each of the two conditions.

E2. Do LLMs perform better at clustering when the article title is provided for additional context?
To test this we compare the clustering results produced when the LLMs are given article text vs. article text and article title.

```
id: {id number}

Text: {text here}
```

Prompt 3: Text Alone Prompt

```
id: {id number}

Title: {title here}

Text: {text here}
```

Prompt 4: Title and Text Prompt

For *E2*, we first give each LLM an initial system prompt following whichever of Prompts 1 or 2 were more successful in *E1*. Then, we feed the LLM articles from *NewsStandEN* one at a time, in one of the two formats specified in Prompts 3 and 4, adjusting lines 3-5 of the initial system prompt to follow Prompt 3 or Prompt 4. With Prompt 3, only the article text is provided to aid with clustering. With Prompt 4, the article title is provided in addition to the article text. Along with each article the output of the previous interaction is provided as the context history, to maintain the cluster assignments made previously. The cluster assignments are recorded for each model under each of the two conditions.

E3. Do LLMs perform better at clustering when provided with the geo-coordinates associated with news articles?
To test this we compare the clustering results produced when the LLMs are given article text vs. article text and article geo-coordinates.

```
id: {id number}

text: {text here}
```

Prompt 5: No Geo-coords Prompt

```
id: {id number}

coordinates: {coords here}

text: {text here}

```

Prompt 6: Geo-coords Prompt

For *E3*, we first give each LLM whichever of Prompts 1 and 2 were more successful in *E1*, and adjust lines 3-5 of the initial system prompt to follow Prompt 5 or Prompt 6. In the case Prompt 6 is used, we also append an additional sentence indicating that the geo-coordinates associated with each article would be provided (and the format it would appear in). Then, we feed it articles from *NewsStandEN* one at a time, in one of the two formats specified in Prompts 5 and 6. With Prompt 5, only the article text is provided to aid with clustering. With Prompt 6, the article's associated geo-coordinates are also provided in addition to the article text. Along with each article, the cluster dictionary produced by the previous interaction is provided as context history, which includes the cluster IDs as keys, the article IDs associated with each cluster, and the keywords associated with each cluster. For experiments using prompt 6, we additionally compute the average geo-coordinate of each cluster after each response of the model, and include it along with the articles assigned to that cluster. The final cluster assignments are recorded for each model under each of the two conditions.

E4. Do LLMs perform better at clustering when provided with time of publication of news articles?
To test this we compare the clustering results produced when the LLMs are given article text vs. article text and article publication timestamp.

```
id: {id number}

text: {text here}
```

Prompt 7: No Timestamp Prompt

```
id: {id number}

timestamp: {timestamp here}

```

```
5    text: {text here}
```

Prompt 8: Timestamp Prompt

For *E4*, we first give each LLM whichever of Prompts 1 and 2 were more successful in *E1*, and adjust lines 3-5 of the initial system prompt to follow Prompt 7 or Prompt 8. In the case Prompt 8 is used, we also append an additional sentence indicating that the timestamp associated with each article would be provided (and the format it would appear in). Then, we feed it articles from *NewsStandEN* one at a time, in one of the two formats specified in Prompts 7 and 8. With Prompt 7, only the article text is provided to aid with clustering. With Prompt 8, the article's associated publication timestamp is also provided in addition to the article text. Along with each article, the cluster dictionary produced by the previous interaction is provided as context history, which includes the cluster IDs as keys, the article IDs associated with each cluster, and the keywords associated with each cluster. For experiments using Prompt 8, we additionally compute the average timestamp after each response of the model, and include it along with the articles assigned to that cluster. The final cluster assignments are recorded for each model under each of the two conditions.

3.2.3 Measurements. We begin with the single-language setting, using the *NewsStandEN* dataset. We record the baseline cluster label assignments from the NewsStand system and record the LLM performances for Prompts 1-8. For these experiments we record the cluster labels assigned to each article by each method or model and using those assignments, we compute the NMI between the recorded cluster assignment, $U = \{U_1, ..., U_p\}$, and the ground truth cluster assignment, $V = \{V_1, ..., V_q\}$, of the N data points as follows [37]. The normalized mutual information, $NMI(U, V)$, is defined as

$$NMI(U, V) = \frac{MI(U, V)}{H(U) + H(V)} \tag{1}$$

where mutual information, $MI(U, V)$, is

$$MI(U, V) = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{n_{ij}}{N} \log \left(\frac{n_{ij}/N}{|U_i||V_j|/N^2} \right). \tag{2}$$

for $n_{ij} = |U_i \cap V_j|$ for $1 \leq i \leq p$ and $1 \leq j \leq q$ and where the entropy H of a cluster assignment $Z = \{Z_1, ..., Z_k\}$ is

$$H(Z) = -\sum_{i=1}^{k} \frac{|Z_i|}{N} \log \frac{|Z_i|}{N}. \tag{3}$$

NMI is a measure of how much information is shared between two distributions [35]. In our context, an NMI of 1.0 means that the recorded clustering matches exactly the clustering in the reference dataset and an NMI of 0.0 means the two clustering assignments are independent. We elect to use NMI as the measure of clustering success as opposed to alternative measurements like B-Cubed [2] since NMI does not reward grouping outliers into one cluster. In the news event clustering setting, outliers should each be given their own new cluster, or should be grouped into the nearest related cluster, but should not be all grouped together into an 'outlier' cluster, which would not correspond to any coherent news event.

For the cross-lingual setting, we record the baseline cluster label assignments of the *NewsStandMULTI* dataset using the NewsStand system. Based on the LLM results from the single-language setting, we select the best performing initial system prompt and metadata settings (with or without title, geo-coordinates, and timestamp) to use in the subsequent cross-lingual LLM experiments. Since there are no ground truth cluster labels for *NewsStandMULTI*, we compare the NMI between the LLMs and the baseline method and qualitatively analyze the cluster results of both methods.

4 Results

In this section we describe the results of the experiments outlined in section 3.

4.1 Single Language Clustering Setting

Table 3 shows the NMI performance results for the baseline method and the LLM models across our five LLM conditions in the single-language setting. All results in this table are compared to hand-labeled ground truth cluster assignments. Comparing the baseline NewsStand clustering method to the LLMs, we find that the baseline outperforms the LLMs in nearly every prompt setup.

The detailed system prompt that includes domain information, stating that the articles being presented are news articles that should be clustered based on whether they describe the same news event, garners better NMI that the generic system prompt that simply asks the LLM to cluster the articles. However, when using the detailed system prompt and including additional information besides article text, like article title, timestamp, or geo-coordinates, the LLMs produced degenerate clusterings, where every article was grouped into one cluster. As such, we report the performance based on cluster assignments made by the LLMs under the generic system prompt, with the additional metadata (title, timestamp, geo-coordinates). The results show the LLMs performed quite similarly with and without each of those pieces of metadata, even declining slightly in performance in some cases by the addition of this information. In other words, adding article title, time of publication, or geo-coordinates associated with each article did not improve the cluster assignments made by the LLMs.

We further observed that some LLMs, like gemini-1.5-flash, occasionally did not assign any keywords to the clusters, making subsequent assignment of articles to that cluster effectively random (i.e. not based on any similarity between the new article and the key phrases describing the cluster). On the flipside, sometimes the LLMs included an excessive quantity of keywords for some clusters, exceeding the prompt's direction to maintain no more than five keywords per cluster. Every LLM tested had instances of assigning more than five keywords, ranging from a few extra to several times the allotted amount. In the most extreme cases, gemini-1.5-flash included over 150 keywords for some clusters. Furthermore, we observed some nonsensical keywords, including article IDs, blank keywords, names of other clusters, single digits, and repeated keywords being assigned for some clusters.

4.2 Cross-Lingual Clustering Setting

Table 4 shows the NMI between the LLM-based clusterings and the baseline NewsStand clusterings on the *NewsStandMULTI* dataset. The results show that the correspondence between the LLM-based methods and the baseline method are low. Since we lack ground

Method	Model	System Prompt	Title	Timestamp	Geo-coords	NMI
NewsStand (baseline)						0.902
LLM	gpt-3.5-turbo	Generic				0.778
	gpt-4o	Generic				0.839
	gpt-4o mini	Generic				0.850
	gemini-1.5-flash	Generic				0.827
	gpt-3.5-turbo	Detailed				0.806
	gpt-4o	Detailed				0.909
	gpt-4o mini	Detailed				0.784
	gemini-1.5-flash	Detailed				0.809
	gpt-3.5-turbo	Generic	X			0.752
	gpt-4o	Generic	X			0.832
	gpt-4o mini	Generic	X			0.803
	gemini-1.5-flash	Generic	X			0.829
	gpt-3.5-turbo	Generic		X		0.769
	gpt-4o	Generic		X		0.595
	gpt-4o mini	Generic		X		0.836
	gemini-1.5-flash	Generic		X		0.849
	gpt-3.5-turbo	Generic			X	0.770
	gpt-4o	Generic			X	0.788
	gpt-4o mini	Generic			X	0.815
	gemini-1.5-flash	Generic			X	0.815

Table 3: Summary of single-language clustering results for baseline and LLM methods compared to ground-truth cluster assignment.

truth labels for the multi-lingual dataset, to determine which method(s) produced better clusterings, we explore the clusters qualitatively.

4.2.1 Cluster Visualization. We use OpenAI's text-3-embedding-large[4] to construct a t-SNE plot of the articles (Figure 1). Coloring each language differently, it is clear that location in the embedding space is highly correlated with language- with few exceptions, the documents form clear groupings based on their language. There is also a correlation between location in the embedding space and the cluster label assigned by both the baseline method and by GPT-4o.

Plotting the articles by time of publication in Figure 2, we compare the cluster assignments made by the baseline method and by GPT-4o. In the baseline cluster assignment, clusters 3, 6, 7, and 8 are correlated with time, since the articles grouped in those clusters were all published within a few days of each other. On the other hand, there were no clusters by GPT-4o that showed this same property. Instead, the majority of the articles were grouped into one cluster, cluster 3, regardless of time of publication. Most of those articles were written in German, indicating the LLM clustered based on language more than time of publication, which is contrary to the goal of cross-lingual clustering of news articles.

4.2.2 Qualitative Analysis. Qualitatively analyzing the clustering results produced by the LLMs, we further find that clusters typically correspond to the language of the text, rather than the text content. We identified many clusters corresponding to entirely Italian, Spanish, French, or German articles. Likewise, the keywords generated by the LLMs to maintain the cluster topics were in that same language. On the other hand, when analyzing the clusters generated by the baseline method, articles in most clusters were mixed across different languages, but were tied by a common topic, often corresponding directly to a single news event. We also observed a similar phenomenon to the single language setting, where the LLMs sometimes associated nonsensical keywords with some of the clusters, which often corresponded to instances where clusters contained unrelated documents.

[4]https://platform.openai.com/docs/guides/embeddings/faq

5 Discussion

In this section we discuss the significance of the results presented in section 4.

5.1 Single Language Clustering Setting

The baseline NewsStand clustering method involves using machine translation and then constructing a TF-IDF vector and maintaining a time centroid for each cluster. We found that this method generally led to better cluster assignments than using the LLMs we tested to generate cluster assignments. In one case, using GPT-4o with the detailed system prompt, the NMI was higher than the baseline. Surprisingly, adding article title, geo-coordinates, or timestamp information made the cluster assignments worse, even though that

Method	Model	System Prompt	Title	Timestamp	Geo-coords	NMI
NewsStand (baseline)						1
LLM	gpt-3.5-turbo	Generic				0.634
	gpt-4o	Generic				0.425
	gpt-4o mini	Generic				0.575
	gemini-1.5-flash	Generic				0.488
	gpt-3.5-turbo	Detailed				0.632
	gpt-4o	Detailed				0.522
	gpt-4o mini	Detailed				0.229
	gemini-1.5-flash	Detailed				0.561
	gpt-3.5-turbo	Generic	X			0.636
	gpt-4o	Generic	X			0.518
	gpt-4o mini	Generic	X			0.547
	gemini-1.5-flash	Generic	X			0.674
	gpt-3.5-turbo	Generic		X		0.638
	gpt-4o	Generic		X		0.453
	gpt-4o mini	Generic		X		0.555
	gemini-1.5-flash	Generic		X		0.565
	gpt-3.5-turbo	Generic			X	0.614
	gpt-4o	Generic			X	0.407
	gpt-4o mini	Generic			X	0.553
	gemini-1.5-flash	Generic			X	0.671

Table 4: Summary of multi-lingual clustering results for baseline and LLM methods compared to NewsStand cluster assignment

information is intuitively useful for determining which cluster an article belongs to. This finding suggests that it is difficult (under our zero-shot setting) to get current LLMs to leverage multiple pieces of disparate intimation (timestamp, article text, coordinates) to decide an appropriate cluster for the articles. Given the clear room for improvement, we suggest further work on using LLMs to cluster text with an emphasis on incorporating metadata like time and location associated with the text. Using few shot prompting and trying different methods of encoding geo-coordinate and time information are the logical next steps to better leverage LLMs in this capacity.

5.2 Cross-Lingual Clustering Setting

Comparing the cross-lingual cluster assignments by time of publication, we found that the baseline method better accounted for time, with half of the clusters containing only articles that were published very close together in time (as opposed to none with GPT-4o). Instead, GPT-4o grouped many articles together despite being published months apart, with the cluster assignment more correlated to language than time of publication. Based on these results, it seems the baseline method that explicitly accounts for time of publication may be preferred for cross-lingual clustering where the time component is relevant, such as or news publications.

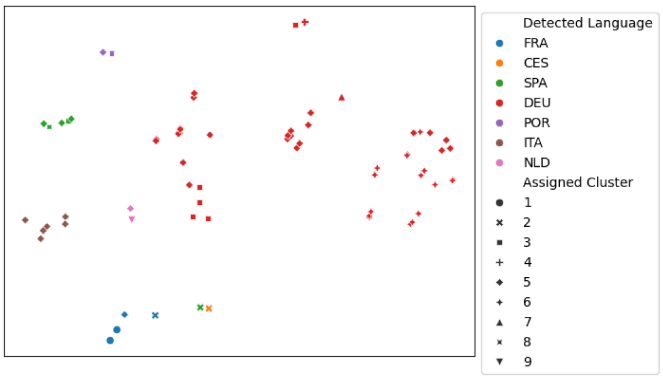
Based on our qualitative analysis of the cluster assignments produced by the LLMs, we found that the LLMs tended to group articles based on the language of the text, rather than the content or news story being described by the article. This behavior was also consistent across our tests that added geo-coordinates or timestamps associated with the articles, for additional language independent context. Overall, we found that besides occasionally producing degenerate cluster assignments, LLMs tended not to leverage the article metadata, which is an area for future improvement that could lead to better clustering results.

6 Related Work

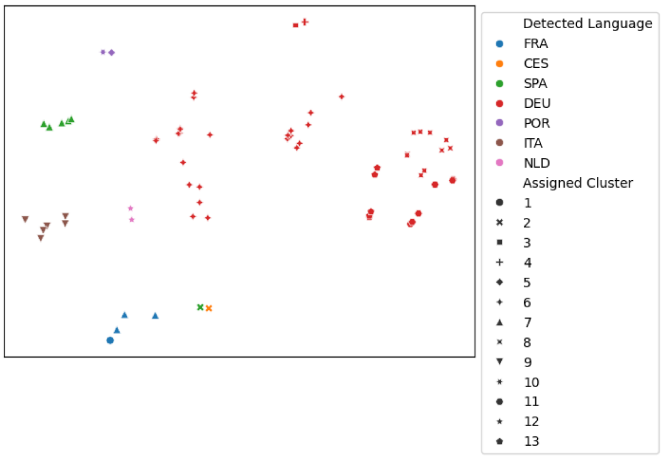
In this section we discuss the related work in text clustering using LLMs and neural methods.

6.1 Neural Embedding Space Clustering

Several language-independent clustering methods that rely on neural embedding spaces have been developed recently. Approaches include using a 3-layer multilingual Bidirectional Long Short Term Memory (BLSTM) encoder to identify nearest neighbor sentences based on similarity in the embedding space, independent of language [31]. Despite being trained on parallel news sentences, named entities like city names and "comma groups" [12] were removed after initial experiments showed that their multilingual distance was not sufficient to reliably distinguish between them. This points to a major issue with using the neural embedding space similarity as a strategy to cluster documents across languages. Previous work on Japanese-Vietnamese news story clustering [6] and cross-lingual

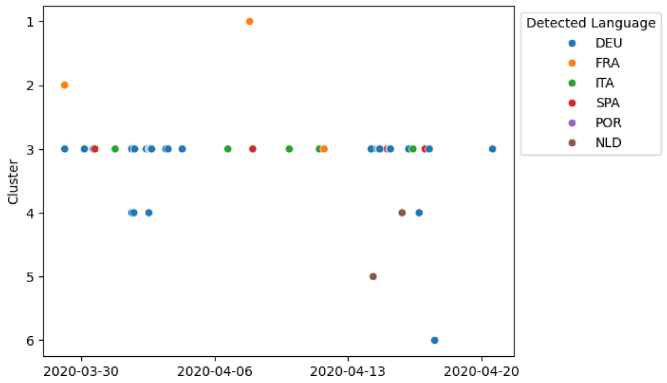


(a) t-SNE plot of the GPT-4o cluster assignments.

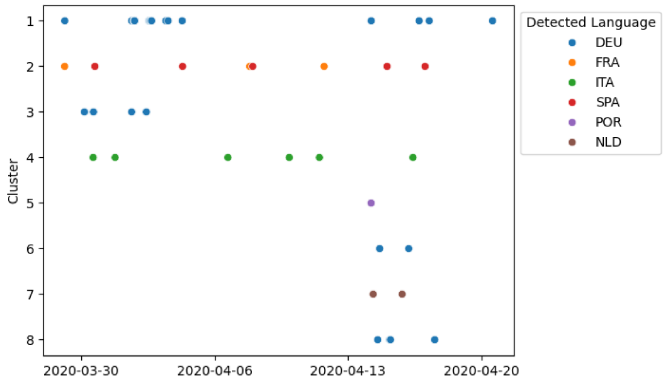


(b) t-SNE plot of the baseline NewsStand cluster assignments.

Figure 1: t-SNE plots of NewsStandMULTI articles by language and cluster label for baseline cluster assignment and GPT-4o cluster assignment. Best viewed in color.



(a) Timeline plot of the GPT-4o cluster assignments. 6 data points omitted to better see detail in different times.



(b) Timeline plot of the NewsStand cluster assignments. 6 data points omitted to better see detail in different times.

Figure 2: Timeline plot of GPT-4o clusters vs baseline NewsStand clusters. Best viewed in color.

news clustering covering 17 languages [30] both show that reasonable cluster formation is highly dependent on the proper nouns in documents, especially location entities.

Other works show that multilingual embeddings [1] and the intermediate state of Neural Machine Translation (NMT) models are promising tools for cross-lingual clustering, particularly in cases with resource-rich language pairs like Japanese-English [32] and when downstream tasks like document classification are the ultimate goal. Similarly, Pires et al. show that transformer-based models like Multilingual-BERT (M-BERT) can map different languages to a shared cross-lingual embedding space, but even using it to cluster articles in a single language does not work well [18, 33]. Generally, neural embedding space methods for cross-lingual clustering show some promise, but lack consistent performance across a variety of languages. Jiang et al. [8] have also attempted to use neural embedding based clustering that accepts time data to try to improve cluster assignments via temporal data, finding success over previous clustering methods.

6.2 LLM Clustering through Prompting

Previous work has attempted to use LLMs to directly cluster articles written in a single language. Zhang et al. [40] showed that LLMs were effective in determining cluster granularity using a pairwise task, as well as showing their effectiveness in topic mining and intent discovery. Viswanathan et. al. [38] further this work by showing that LLMs perform well in entity canonicalization and text clustering, while also showing that LLMs were effective with keyphrase based clustering. However, in both cases, the textual data used for clustering consists of short documents, which limits the applicability to domains like news, where documents are typically longer. Petukhova et al. [17] also find that LLM embeddings are useful in clustering contexts, showing that clustering techniques such as spectral methods and k-means generally perform best in OpenAI embeddings. Some recent work has also focused directly on clustering news articles and newsworthy events. Nakshatri et al. [15] used LLMs to perform event discovery on news articles,

after which they were clustered based on the events discovered. They found LLM methods to be more effective than traditional event detection methods in both entity purity and entity coherence, but focus only on a single language. LLMs have also been shown to be effective in classifying morality, as shown in Roy et al. [21] which used LLMs to identify the moral foundation expressed in political tweets, as well as the moral roles of entities in tweets in few-shot clustering scenarios. In that setting, they find LLMs to be more effective than RoBERTa-based frameworks, but the work is also on shorter text snippets in a single language.

6.3 Using LLMs to Improve Cluster Assignment

In addition to prompting LLMs to perform cluster assignment directly, Viswanathan et. al. [38] used LLMs to correct clustering assignments, and found that LLM corrected clusters generally improved rather than degraded in quality (102 improved vs. 52 degraded). This line of work highlights the flexibility of LLMs to dynamically cluster and re-cluster articles, which is a benefit that many of the more traditional clustering approaches do not provide. Further work could apply similar concepts of cluster correction to the cross-lingual setting to determine if some of the clustering issues we observe in the present paper can be corrected through iterative re-clustering.

7 Conclusion and Future Work

This paper presents experiments measuring LLM performance at the task of cross-lingual clustering of long news articles spanning seven languages. We find that LLMs tend to cluster the articles based on the language of the article text more than the content, and often entirely ignore the time and location of publication of the article, even when explicitly provided that information. Based on these findings, future work, such as developing appropriate few-shot prompting methods, is needed before LLMs can reliably be used in clustering news articles across multiple languages. In addition, further work may improve LLM clustering results by encoding geospatial information in English (e.g. by using the name of a place, such as College Park) rather than in latitude and longitude pairs, which LLMs have been shown to misinterpret [14]. As LLMs improve and methods to use them for cross-lingual clustering are developed, the task presented in this paper can be expanded to include additional steps, such as having the LLM also perform the geo-coordinate extraction, rather than providing it with the article. Work in this direction is promising given the latent world knowledge embedded in LLMs that can be leveraged when clustering articles across languages and locations.

Acknowledgements

This work was sponsored in part by the NSF under Grants IIS-18-16889, IIS-20-41415, and IIS-21-14451, and the Australian-American Fulbright Commission. Its contents do not necessarily represent the official views of the Fulbright Program.

References

[1] W. Ammar, G. Mulcaire, G. Lample, C. Dyer, and N. A. Smith. 2018. C L] 2 1 M ay 2 01 6 Massively Multilingual Word Embeddings.

[2] A. Bagga and B. Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *COLING 1998 Volume 1: The 17th international conference on computational linguistics*.

[3] R. O. Duda and P. E. Hart. 1973. *Pattern Classification and Scene Analysis*. Wiley Interscience, New York.

[4] N. Gramsky and H. Samet. 2013. Seeder finder: Identifying additional needles in the Twitter haystack. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. 44–53.

[5] S. Ho, M. Lieberman, P. Wang, and H. Samet. 2012. Mining future spatiotemporal events and their sentiment from online news articles for location-aware recommendation system. In *Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems*. 25–32.

[6] X. Hong, Z. Yu, M. Tang, and Y. Xian. 2017. Cross-lingual event-centered news clustering based on elements semantic correlations of different news. *Multimedia Tools and Applications* 76 (2017), 25129–25143.

[7] Shadi Jaradat, Richi Nayak, Alexander Paz, Huthaifa I Ashqar, and Mohammad Elhenawy. 2024. Multitask Learning for Crash Analysis: A Fine-Tuned LLM Framework Using Twitter Data. *Smart Cities* 7, 5 (2024), 2422–2465.

[8] H. Jiang, D. Beeferman, W. Mao, and D. Roy. 2024. Topic Detection and Tracking with Time-Aware Document Embeddings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue (Eds.). ELRA and ICCL, Torino, Italia, 16293–16303. https://aclanthology.org/2024.lrec-main.1416

[9] J. L. Leidner and M. D. Lieberman. 2011. Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special* 3, 2 (2011), 5–11.

[10] M.D. Lieberman and H. Samet. 2011. Multifaceted toponym recognition for streaming news. In *Proceedings of SIGIR'11*. Beijing, China, 843–852.

[11] M. D. Lieberman and H. Samet. 2012. Adaptive context features for toponym resolution in streaming news. In *Proceedings of SIGIR'12*. Portland, OR, 731–740.

[12] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. 2010. Geotagging: Using proximity, sibling, and prominence clues to understand comma groups. In *Proceedings of 6th Workshop on Geographic Information Retrieval*. Zurich, Switzerland.

[13] M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. 2007. STEWARD: architecture of a spatio-textual search engine. In *Proceedings of the 15th ACM International Symposium on Advances in Geographic Information Systems*, H. Samet, M. Schneider, and C. Shahabi (Eds.). Seattle, WA, 186–193.

[14] R. Manvi, S. Khanna, G. Mai, M. Burke, D. Lobell, and S. Ermon. 2024. GeoLLM: Extracting Geospatial Knowledge from Large Language Models. arXiv:2310.06213 [cs.CL] https://arxiv.org/abs/2310.06213

[15] N. Nakshatri, S. Liu, S. Chen, D. Roth, D. Goldwasser, and D. Hopkins. 2023. Using LLM for Improving Key Event Discovery: Temporal-Guided News Stream Clustering with Event Summaries. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 4162–4173. https://doi.org/10.18653/v1/2023.findings-emnlp.274

[16] K. O'Sullivan, N. R. Schneider, and H. Samet. 2024. Metric Reasoning in Large Language Models. In *Proceedings of the 32nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Atlanta, USA). Association for Computing Machinery.

[17] A. Petukhova, J. P. Matos-Carvalho, and N. Fachada. 2024. Text Clustering with LLM Embeddings. arXiv:2403.15112 [cs.CL] https://arxiv.org/abs/2403.15112

[18] T. J. P. Pires, E. Schlinger, and D. Garrette. 2019. How Multilingual is Multilingual BERT?. In *Annual Meeting of the Association for Computational Linguistics*.

[19] G. Quercini and H. Samet. 2014. Uncovering the spatial relatedness in Wikipedia. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. Dallas, TX, 153–162.

[20] G. Quercini, H. Samet, J. Sankaranarayanan, and M. D. Lieberman. 2010. Determining the spatial reader scopes of news sources using local lexicons. *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 43–52.

[21] S. Roy, N. Sridhar Nakshatri, and D." Goldwasser. 2022. Towards Few-Shot Identification of Morality Frames using In-Context Learning. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*. Association for Computational Linguistics, Abu Dhabi, UAE, 183–196. https://doi.org/10.18653/v1/2022.nlpcss-1.20

[22] G. Salton and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24, 5 (1988), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

[23] G. Salton, A. Wong, and C. S. Yang. 1975. A Vector Space Model for Automatic Indexing. *Commun. ACM* 18, 11 (nov 1975), 613–620. https://doi.org/10.1145/361219.361220

[24] H. Samet. 2014. Using minimaps to enable toponym resolution with an effective 100% rate of recall. In *Proceedings of GIR'14*. Dallas, TX.

[25] H. Samet, M. D. Adelfio, B. C. Fruin, M. D. Lieberman, and B. E. Teitler. 2011. Porting a web-based mapping application to a smartphone app. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic*

Information Systems. Chicago, IL, 525–528.

[26] H. Samet, B. E. Teitler, M. D. Adelfio, and M. D. Lieberman. 2011. Adapting a map query interface for a gesturing touch screen interface. In *Proceedings of the Twentieth International Word Wide Web Conference (Companion Volume)*. Hyderabad, India, 257–260.

[27] H. Samet, B. E. Teitler, M. D. Lieberman, J. Sankaranarayanan, D. Panozzo, and J. Sperling. 2009. *Reading News with Maps: The Power of Searching with Spatial Synonyms*. Technical Report. Computer Science Department, University of Maryland, College Park, MD. submitted for publication.

[28] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. 2009. TwitterStand: News in Tweets *(GIS '09)*. Association for Computing Machinery, New York, NY, USA, 42–51. https://doi.org/10.1145/1653771.1653781

[29] N. R. Schneider and H. Samet. 2021. Which Portland is It? A Machine Learning Approach. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Recommendations, Geosocial Networks and Geoadvertising* (Beijing, China) *(LocalRec '21)*. Association for Computing Machinery, New York, NY, USA, Article 8, 10 pages. https://doi.org/10.1145/3486183.3491066

[30] N. R. Schneider, J. Sankaranarayanan, and H. Samet. 2024. Cross-lingual Text Clustering in a Large System. In *Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval* (Seoul, Republic of Korea) *(NLPIR '23)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3639233.3639356

[31] H. Schwenk. 2018. Filtering and Mining Parallel Data in a Joint Multilingual Space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 228–234. https://doi.org/10.18653/v1/P18-2037

[32] K. Seki. 2018. Exploring Neural Translation Models for Cross-Lingual Text Similarity. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) *(CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 1591–1594. https://doi.org/10.1145/3269206.3269262

[33] L. Stankevivcius and M. Lukovsevivcius. 2020. Testing pre-trained Transformer models for Lithuanian news clustering.

[34] M. Steinbach, G. Karypis, and V. Kumar. 2000. A Comparison of Document Clustering Techniques. *Proceedings of the International KDD Workshop on Text Mining* (06 2000).

[35] A. Strehl and J. Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research* 3, Dec (2002), 583–617.

[36] B. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. 2008. NewsStand: A new view on news, W. G. Aref, M. F. Mokbel, H. Samet, M. Schneider, C. Shahabi, and O. Wolfson (Eds.). Irvine, CA, 144–153.

[37] N. X. Vinh, J. Epps, and J. Bailey. 2010. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research* 11, 95 (2010), 2837–2854. http://jmlr.org/papers/v11/vinh10a.html

[38] V. Viswanathan, K. Gashteovski, K. Gashteovski, C. Lawrence, T. Wu, and G. Neubig. 2024. Large Language Models Enable Few-Shot Clustering. *Transactions of the Association for Computational Linguistics* 12 (2024), 321–333. https://doi.org/10.1162/tacl_a_00648

[39] J. Zhang, A.-T. Kuo, N. R. Schneider, J. Peters, and H. Samet. 2023. BroadcastSTAND: Clustering Multimedia Sources of News. In *Proceedings of the 7th ACM SIGSPATIAL Workshop on Location-Based Recommendations, Geosocial Networks and Geoadvertising* (Hamburg, Germany) *(LocalRec '23)*. Association for Computing Machinery, New York, NY, USA, 33–36. https://doi.org/10.1145/3615896.3628347

[40] Y. Zhang, Z. Wang, and J. 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*. 13903–13920.

[41] G. Zhou and J. Su. 2002. Named Entity Recognition Using an HMM-Based Chunk Tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, USA, 473–480. https://doi.org/10.3115/1073083.1073163