# Cross-lingual text clustering in a large system

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

Jagan Sankaranarayanan
Google
jsan@google.com

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

## ABSTRACT

The multilingual world needs systems that can cluster text written in multiple languages into the same thread or topic. Clustering multilingual text can be accomplished by translating and then clustering text in a canonical language, using multilingual embeddings to cluster articles in a shared embedding space, and via other language-independent methods. The performance and pitfalls of these various methods have not been well studied in the context of real-time clustering across documents written in many languages. We address this problem by generating a large dataset of news articles using a reference architecture that continuously indexed and clustered articles spanning 17 languages over the last 15 years. Through the analysis of these documents and their clusters, the clustering quality is shown to be dependent on the normalization of proper nouns, the types of georeferences, and the overall geographic focus of the document.

## CCS CONCEPTS

• **Information systems → Geographic information systems**.

**ACM Reference Format:**
Nicole R. Schneider, Jagan Sankaranarayanan, and Hanan Samet. 2023. Cross-lingual text clustering in a large system. In *Proceedings of 7th International Conference on Natural Language Processing and Information Retrieval (NLPIR '23)*. ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

The world is multilingual, and systems need to be thus to reach users in their language of preference or competence. This is why one can tweet in any language, and the news is available in many spoken languages. The challenge for online systems is to assemble the news or any text content in various languages into the same "thread" or "topic" to *promote connections and discourse* between people with a wide variety of opinions and perspectives. This makes text clustering across languages an essential step in Cross-Lingual Information Retrieval (CLIR) and related downstream problems [12, 35], and a worthwhile problem to explore in a real ”in the wild” setting.

Clustering algorithms assimilate text into topics or threads based on content and references central to the text, like locations and names of people. There have been many methods for clustering input text data [15, 31, 34, 68] including those designed for online use [2, 4, 5, 13, 17, 18, 23, 24, 28–30, 45, 72, 73]. For domains like news, online clustering is critical since news articles are constantly being generated (and aged, although this is outside our scope) and has to be clustered in real time to provide users with relevant stories as they break. In particular, news documents must be grouped into fine-grained clusters, representing closely-related news articles reporting on a single event or story. However, most traditional clustering techniques are agnostic to the source language of the input text. Furthermore, cross-lingual clustering is a substantially harder problem since different languages do not always share a common vocabulary or script. State-of-the-art cross-lingual text clustering is achieved through large multilingual models like M-BERT [19, 46, 51, 67] and other neural models [8, 47, 62, 63]. An older and simpler approach is to first translate the documents into a common canonical language such as English and then cluster the translated documents post-translation. In this case, translation is usually done using Machine Translation (MT) [22, 44], a bilingual dictionary [7], or a probabilistic model trained on parallel corpora [64], the output of which is fed to the clustering algorithm.

Although cross-lingual clustering has seen recent advances, there are very few studies of the quality of the clusters produced by any cross-lingual clustering method, especially regarding the factors or aspects of the text that influence clustering behavior. We use NewsStand [69] as an example of a large, mature system that clusters millions of news documents written in different languages via the simple translate-then-cluster approach. The NewsStand pipeline is set up to ingest, translate, pre-process, geotag, and cluster articles in real time after pulling them directly from RSS news feeds. The translation is done using a publicly-available cloud-based translation service, and the clustering is accomplished using a simple online clustering algorithm to assign incoming news into text clusters. Our dataset comes from *over 15 years of articles spanning 17 languages* that have been indexed, translated, and clustered by NewsStand.

**Through in-depth analysis of these documents and their clusters, we characterize the issues and phenomena associated with cross-lingual clustering on a massive dataset of millions of articles.** We analyze clusters by measuring their size, in number of documents, and their inter-relatedness, in overlapping terms between the articles. We do so across several document attributes, including original language, proper noun usage, and geographic focus. We find these three attributes to be the dominant factors that influence how translated text clusters post-translation.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*NLPIR '23, December 15–17, 2023, Seoul, South Korea*
© 2023 Association for Computing Machinery.
ACM ISBN ISBN: 979-8-4007-0922-7. . . $15.00
https://doi.org/XXXXXXX.XXXXXXX

We uncover several noteworthy characteristics and phenomena that shed light on the need for more sustained research into cross-lingual text clustering. These points are enumerated below and appear in boldface text throughout the paper:

(1) Single-article clusters and extremely large clusters of loosely related articles are a common phenomenon. The patterns we observe in cluster sizes indicate that we cannot just adjust the hyper-parameters of the clustering algorithm, such as threshold, to improve clustering behavior and eliminate these issues (Section 4.2).
(2) Proper nouns play a critical role in clustering. Inconsistencies in proper noun usage in the text being clustered cause poor entity tagging, which makes Information Retrieval (IR) difficult and adds noise to the cluster formation. Proper nouns with more than one common spelling in a given language pose a particular challenge (Section 6.1).
(3) Location and person proper nouns are typically more critical for clustering than generic proper nouns (Section 6.2).
(4) Articles with a strong local geographic focus tend to cluster well, which may also explain why some languages in NewsStand cluster better than others (Section 7).

Our paper's main contribution is the characterization of previously unexplored problems in cross-lingual text clustering. We identify the pitfalls we see in the clustering behavior of articles in a large system and describe how these remain unresolved in the recent advances in large multilingual models, highlighting the need for continued work in CLIR. We release a subset of documents as a repository[1] to encourage further refinement and improvements to cross-lingual clustering, especially targeting online text clustering which is a critical requirement for news, social media, and other dynamic use-cases. Our paper also sets the stage for follow-up work using clustering behavior as a way to quantitatively evaluate the performance of different translators beyond their fluency and adequacy or to evaluate large multilingual model embeddings, which are commonly used in a variety of applications besides clustering [8].

The rest of this paper is organized as follows. We start by surveying related work in cross-lingual and online text clustering (Section 2). Next, we describe the NewsStand system [69] (Section 3), which performs pre-processing and clusters the articles used in our dataset. Section 4 reviews the clustering landscape in NewsStand and outlines several pitfalls and phenomena we observe. This is followed by an outline of several major factors that we find to be intrinsically associated with clustering outcomes: source language of the article (Section 5), proper noun usage (Section 6), and geographic focus (Section 7). Finally, we discuss the implications of our work and provide directions for future research (Section 8).

## 2 RELATED WORK

There is a substantial body of literature on clustering algorithms, including algorithms designed for clustering time series data [66], or text data. Extensions include clustering cross-lingual text, clustering text in an online and unsupervised fashion, and clustering news and social media text, which we summarize in further detail.

### 2.1 Text Clustering

Clustering, or finding groups of similar objects, is a common data mining problem with many applications [3, 65]. Text data, which is sparse and high-dimensional, poses a particular challenge for clustering, so text data requires specific clustering algorithms beyond the general purpose ones designed for numeric or nominal data [9]. A prominent approach to dealing with the sparsity and high-dimensionality of text data is to preprocess it using the vector-space model Term Frequency-Inverse Document Frequency (TF-IDF) [54]. The TF-IDF representation normalizes data to account for common words that dominate and drown out more discriminative, rarer words, making it a standard tool for representing text data. There are a variety of clustering algorithms that work for text data [15, 68], but many of them are not well suited to clustering a dynamic corpus of articles written in different languages.

### 2.2 Cross-lingual Text Clustering

Cross-lingual information discovery can be accomplished in various ways, including traditional translate-then-cluster approaches, recent neural embedding space methods, and other language-independent methods.

***Translate-then-cluster methods.*** Traditional methods typically require that documents be translated into a single canonical language, which can be done using MT, a bilingual dictionary, or a probabilistic model trained on parallel corpora [44, 64, 75]. Many studies that measure cluster assignment aim to group documents into a few large clusters, which is insufficient for the problems faced by most real-world IR systems that group documents into threads or event clusters, like NewsStand.

For instance, multi-view clustering [31] uses parallel text views across 5 languages to successfully group 110k Reuters documents into 6 large, coarse-grained clusters based on general topic. Similarly, Wu et al. [71] shows that a bilingual dictionary model yields clustering at least comparable to the translate-then-cluster approach on a similar task of assigning documents to broad category clusters, which aligns with our finding that full translation may not be necessary or sufficient to achieve good clustering performance.

***Neural Embedding Space Methods.*** The recent explosion in large language models has driven the development of several language-independent clustering methods. 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 [62]. Despite being trained on parallel news sentences, named entities like city names and "comma groups" [40] 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 [26] and our present analysis on clustering patterns across 17 different languages both show that reasonable cluster formation is highly dependent on the proper nouns in documents, especially *location* entities.

Other works show that multilingual embeddings [8] and the intermediate state of Neural Machine Translation (NMT) models

[1] https://github.com/nicoleschneider/TranslationClustering

are promising tools for cross-lingual clustering, particularly in cases with resource-rich language pairs like Japanese-English [63] and when downstream tasks like document classification are the ultimate goal. Similarly, Pires et al. [46] show that transformer-based models like Multilingual-BERT (M-BERT) can map different languages to a shared cross-lingual embedding space, but they find that M-BERT does not handle typologically divergent languages well. Even using M-BERT to cluster articles in a single language is a challenge. Stankevičius et al. [67] use M-BERT to perform coarse-grained clustering of Lithuanian documents into 12 topic clusters. They achieved a Matthews Correlation Coefficient (MCC) score of about 0.25 even after fine-tuning, meaning cluster formation was closer to random clustering (score of 0) than perfect clustering (score of 1).

***Other Language-independent methods.*** A handful of other approaches include cross-lingual cluster linking [47], using human-annotated Wikipedia data [32, 33, 53], and using Self-Organizing Maps (SOMs) to automatically cluster cross-lingual terms and documents with similar subjects or concepts [34]. Rupnik et al. [53], building on their previous "Event Registry" system [32, 33], link documents across languages according to a similarity function trained on a human-annotated Wikipedia dataset of English, German, and Spanish linked clusters. However, all of these results are based on a small number of languages, cover limited settings, and do not scale sufficiently to address the problem fully.

## 2.3 Online and Unsupervised Clustering

A dynamic corpus of news articles meant to be browsed "on-the-go" requires clustering to be done in an online fashion [57, 58]. A similar need exists for social media data since new posts are added in real-time, necessitating rapid geoparsing and clustering [21, 38, 43, 60]. Numerous methods for clustering text have been designed to work in an online scenario [2–5, 13, 18, 23, 24, 28–30, 45, 72–74].

In addition to being online, clustering for the news domain must also be completely unsupervised since we must detect new topics, and no training set can accurately predict future events. As such, we also lack knowledge of the number of clusters ahead of time, making many methods infeasible, including popular variants of online spherical k-means (OSKM) [73]. A simple alternative is the basic leader-follower clustering algorithm [17], which assigns a new data point to the nearest existing cluster, or creates a new cluster if none are close enough. This algorithm can be parallelized across multiple GPUs and modified to allow clustering in both content and time, which make it a good choice for NewsStand [70].

## 2.4 Clustering News and Social Media Text

Many works have attempted to cluster short text from social media and news domains, typically with the goal of reducing information overload for end users. Some of the recent success in clustering social media text involves methods that augment the text with outside information. Some works leverage Wikipedia to enrich the text [11, 27], which improves clustering, and others leverage it to label clusters already grouped by human annotators [14].

Other works have attempted to cluster social media text, including [52], which clusters a million tweets covering 30 hashtags into coarse or fine-grained clusters, using hashtags as the gold standard labels. Work has also been done to generate lighter-weight representations of newsworthy social media text using data aggregations that obtain clustering comparable with the full representation [49]. News article recommendation systems have also been developed based on common characteristics like named entities, time of publication, and user preferences and feedback when available [6, 37]. To our knowledge, none of the work in clustering news and social media text has comprehensively studied the clustering behavior of a large set of texts originally written in many different languages.

# 3 NEWSSTAND DATASET

To study the clustering behavior of cross-lingual text documents, we leverage NewsStand [69], 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 as new articles are published, markers are dynamically added to the map in real-time. After assigning a location to each article, NewsStand aggregates the articles into clusters based on the textual content and locations referenced within the document. Critically, this enables articles to be ranked by story significance and displayed to users based on the map position and zoom level selected through the interface.

## 3.1 Preprocessing

The NewsStand dataset we use in this study is preprocessed as follows. Articles are first translated into English using MT (if they are not already in English) and sent through a series of steps to identify geographic terms in the article text or translation output. This occurs during the geotagging process, which consists of four stages: Entity Feature Extraction, Gazetteer Record Assignment, Geographic Name Disambiguation, and Geographic Focus Determination [69]. 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 [76]. The NER tagger is from the LingPipe toolkit [10], which was trained on the brown corpus [20] and additional news data. For more details on this process, see [25, 42, 59]. Once extracted, the EFV contains words belonging to proper noun classes like location, organization, and person. In Section 6, we discuss the relevance of these entity classes to the clustering behavior of translated text. Since location entities are particularly relevant for geotagging, those are marked as geographic features in the EFV and then assigned a set of matching locations during the Gazetteer Record Assignment stage, the toponyms are resolved during the Geographic Name Disambiguation phase [36, 38–40, 56, 61], and a geographic focus is determined for each article.

## 3.2 Clustering the Documents

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

| Language Rankings for Clusterability | | | | |
|---|---|---|---|---|
| Original Language | ISO ALPHA 2 Code | Rank | % Singletons | Num Documents |
| Haitian | HT | 1 | 54.28% | 5944 |
| Japanese | JA | 2 | 59.04% | 4527 |
| Arabic | AR | 3 | 78.93% | 16196 |
| German | DE | 4 | 79.49% | 53673 |
| Hindi | HI | 5 | 82.50% | 21326 |
| Chinese | ZH | 6 | 82.53% | 19965 |
| Persian | FA | 8 | 88.4% | 607 |
| Spanish; Castilian | ES | 9 | 90.0% | 2660 |
| Portuguese | PT | 11 | 90.2% | 3814 |
| French | FR | 12 | 92.4% | 4114 |
| Hebrew | HE | 13 | 93.0% | 3494 |
| Italian | IT | 14 | 93.0% | 16555 |
| Dutch; Flemish | NL | 15 | 93.6% | 3669 |
| Greek, Modern | EL | 16 | 94.7% | 757 |
| Czech | CS | 17 | 95.2% | 759 |
| Russian | RU | 18 | 95.6% | 5381 |

**Table 1: Language rankings by percent of documents clustering as singletons after translation into English, where lower rankings indicate a lower percentage of documents as singletons (better clustering). The rankings are depicted graphically in Figure 2. Languages with fewer than 500 articles in NewsStand are ignored, leaving 17 languages in the dataset.**

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 [69].

To accomplish the clustering, NewsStand employs the vector space model [55], a common approach in text mining and information retrieval. 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 [54]. 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. 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 [17]. 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 [68], and a Gaussian attenuator is applied to the temporal dimension to favor clusters with time centroids near the article's publication time.

# 4 NEWSSTAND CLUSTER CHARACTERISTICS

We first summarize the cluster landscape in the NewsStand dataset, focusing on two key indicators of cluster behavior: cluster size and inter-relatedness. We use document counts to measure size and use common key terms as a proxy for document inter-relatedness.

## 4.1 Singleton Clusters

The clusters in NewsStand contain approximately 17 million documents in total. Of those, about 7 million comprise what we refer to as *singleton clusters*, or clusters containing only a single document. In some cases, such as for highly particular stories for which there are no other similar articles, a singleton cluster is the appropriate clustering result. However, as we show in Section 5, a very high proportion of articles originating in many of the 17 languages in NewsStand reside in singleton clusters, indicating poor cluster formation. Throughout the subsequent sections of this paper, we present the factors that we find influence the clustering behavior, including the substantial formation of singleton clusters.

## 4.2 Zombie Clusters

The clusters that are not singletons tend to be small in size, containing only a few documents on average. However, some very large clusters contain thousands of documents per cluster. Sometimes these clusters can correspond to major world events that garner the attention of hundreds of publications in a short period. However, in most cases, these very large clusters are what we term *zombie clusters*, or clusters containing a large number of documents with a very small number of different important terms relating those documents to each other. In essence, zombie clusters are very large clusters that grow by picking up articles that are only tangentially related to the existing articles in the cluster.

We can identify such clusters by examining the cluster sizes compared with the cluster inter-relatedness, measured via the number of unique important terms that appear amongst articles in the clusters. Important terms are those which appear frequently in a given document but infrequently in the overall corpus, resulting in a high TF-IDF score. For our purposes, we consider scores above 0.3 to be high, given that the average score for a term is 0.21 in the NewsStand data. Figure 1 shows the cluster sizes measured in the number of documents on the y-axis against the number of unique important terms on the x-axis. Zombie clusters are those with few, if any, important terms (left side of the plot) tying thousands of articles together (upper portion of the plot).

For instance, we observed a zombie cluster of 3525 articles that were clustered together based on 2 important terms: "beer" and "wurst". This cluster contained unrelated articles that referred to these terms but did not describe any common news event or story that should justify them being clustered together. Zombie clusters like this one represent a poor clustering outcome since the goal is to obtain clusters that describe the same news event, and zombie clusters, by definition, do not accomplish that goal.

The phenomenon of a large proportion of singleton clusters existing alongside zombie clusters indicates that simply adjusting the distance cutoff in the clustering algorithm would not improve the clustering outcomes overall. Increasing the distance cutoff would lead to more zombie clusters forming since it would be even easier for unrelated articles to cluster together by chance. On the other hand, decreasing the threshold would lead to stricter clustering and an even greater proportion of singleton clusters. **This tension motivates our research into what aspects of the articles influence how they cluster, particularly across languages, when translation may further complicate clustering behavior.**

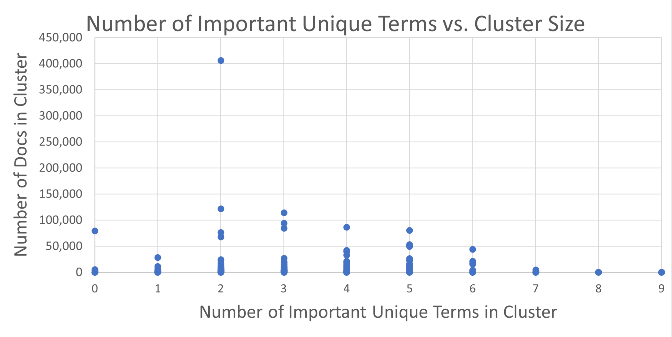

**Figure 1: Plot of the number of unique important terms (greater than 0.3 TF-IDF score) appearing in clusters vs. the cluster size. Large clusters with few important terms tying the articles together are considered Zombie clusters and tend to appear in the upper left region of the plot.**

## 5 FACTOR 1: SOURCE LANGUAGE

The first factor we observe that influences clustering is the source language of the original documents. The NewsStand articles are distributed across 17 different languages. Table 1 describes the 17 languages, their abbreviations that will be used in graphs throughout this paper, and the number of articles per language in the NewsStand data. The non-English articles are translated upon ingest into the NewsStand system, but metadata indicating the original language is retained along with the original and translated versions of the text, allowing us to analyze how original language plays a role in how the articles cluster.

Figure 2 shows the distribution of cluster sizes for articles originating in each language in Table 1. Since NewsStand contains more documents for some languages than others, the plot is normalized over the total number of articles for each language. This gives an overall snapshot of how well articles are clustering based on their original language. The languages in this plot are ranked from left to right (top to bottom in the legend) based on the proportion of articles from that language that reside in singleton clusters. These rankings are also enumerated in Table 1.

For example, of the 17 languages analyzed, Russian (RU) has the highest proportion of singletons, with greater than 95% of Russian-original articles residing in singleton clusters. On the other hand, Japanese (JA) and Haitian (HT) articles show completely different clustering behavior from the other languages, with 59% and 54% of articles in singleton clusters, respectively. This indicates a much better clustering outcome, with closer to half of all articles actually being grouped in some fashion. Throughout the rest of the paper, we explore some of the other factors that influence these differences in clustering behavior.

## 6 FACTOR 2: PROPER NOUN TRANSLATION

As described in Section 3.1, NewsStand's geotagging process involves identifying proper noun entities within each document it ingests. The entities are labeled with their class, indicating the type of noun identified. However, these classes are not disjoint. For example, the class proper noun (NNPP) is generic and encapsulates other more specific classes like location (LOC) and person (PER). Figure 3 shows the distribution of entity tag classes in the NewsStand data. Although there are 35 classes of entities tagged in the NewsStand data, NNPP, LOC, PER, and organization (ORG) are by far the most commonly tagged categories, representing 84.6% of the tags generated by NewsStand.

Proper nouns often convey important information in news stories, including the subject matter, people or organizations involved, and key locations where the story took place. As such, it is natural to suspect that the proper nouns in an article also play a key role in determining how that article clusters. To explore this hypothesis, we consider proper noun tag density and proper noun tag class.

### 6.1 FACTOR 2A: Entity Tag Density

We define entity tag density as the proportion of proper nouns in an article compared to the overall length of the article. This metric can be measured by counting the number of entities tagged in the article at the time of ingest and dividing by the number of words or characters in the article text. Intuitively, a higher (lower) entity tag density indicates that an article references many (few) places, people, etc., compared to other articles of similar length. However, since the metric is calculated by counting the number of *tagged* entities, the ability of the tagger to identify proper nouns

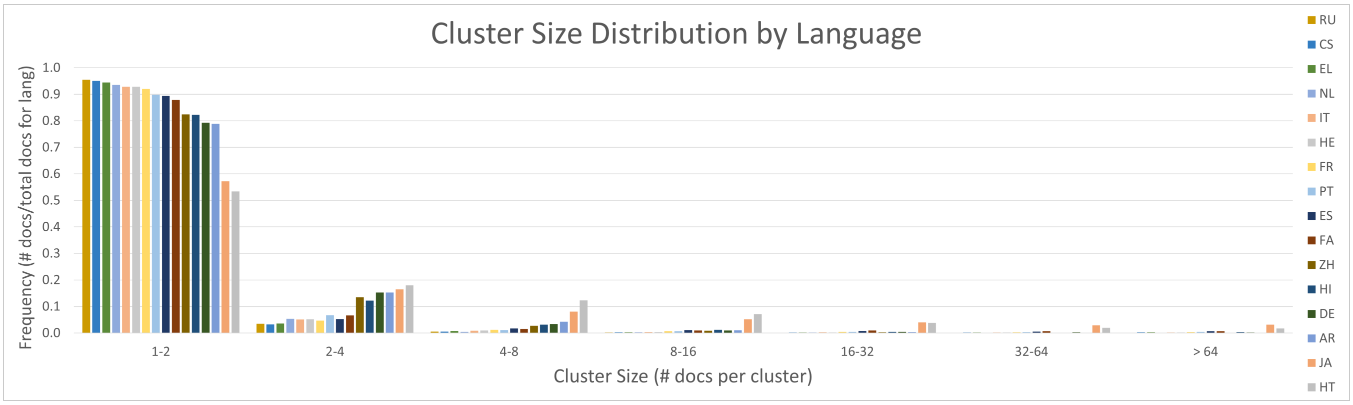

**Figure 2: Distribution of binned cluster sizes across languages besides English. Languages are sorted such that languages with higher percentages of singletons appear towards the left for each bin on the x-axis, and languages with a lower percentage of singletons appear towards the right of each bin.**

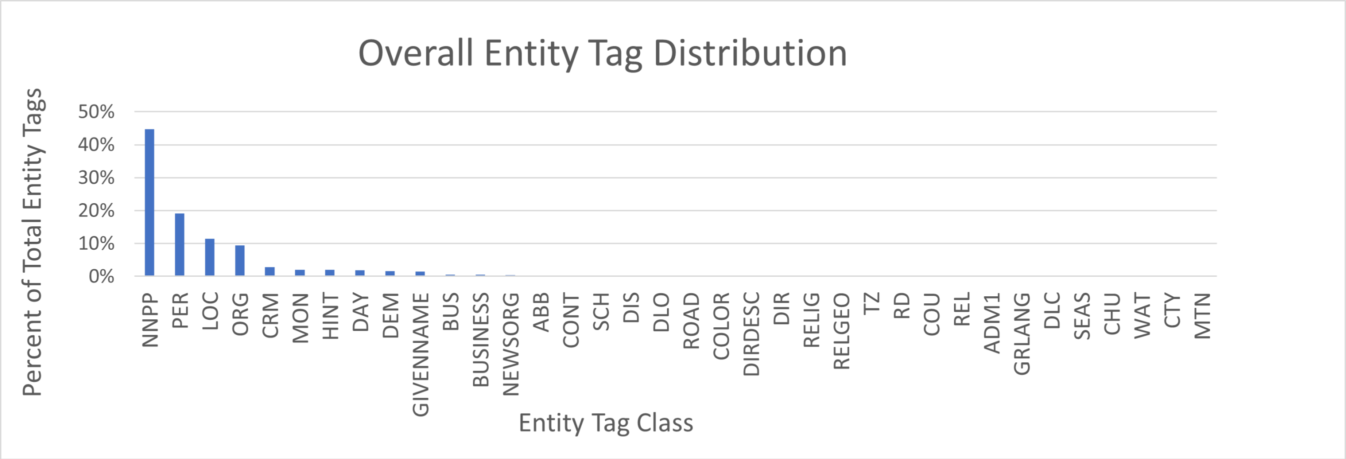

**Figure 3: Distribution of entity tags identified in the text of the articles ingested by the NewsStand pipeline. Proper noun (NNPP) is the most frequently identified entity tag class, followed by person (PER), location (LOC), and organization (ORG). The other 31 classes are rarely identified in comparison to NNPP, PER, LOC, and ORG.**

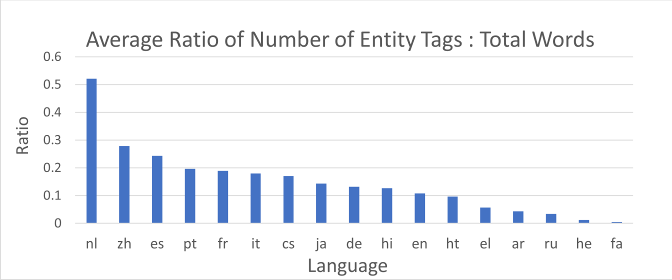

**Figure 4: Average entity tag density per language after translation into English. Entity tag density is the ratio of the number of entity tags to the number of words in the translation output.**

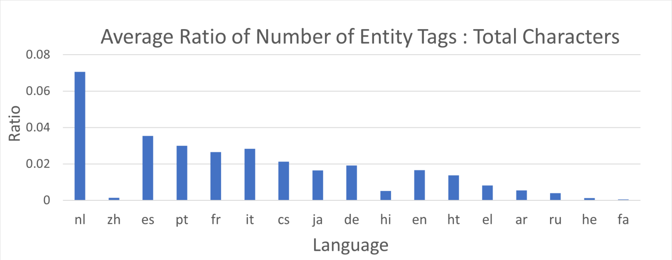

**Figure 5: Average entity tag density is the ratio of the number of entity tags to the number of characters in the translation output.**

when they appear in the text is a critical factor. Articles containing entities that are misspelled or otherwise not recognized will have a disproportionately low entity tag density.

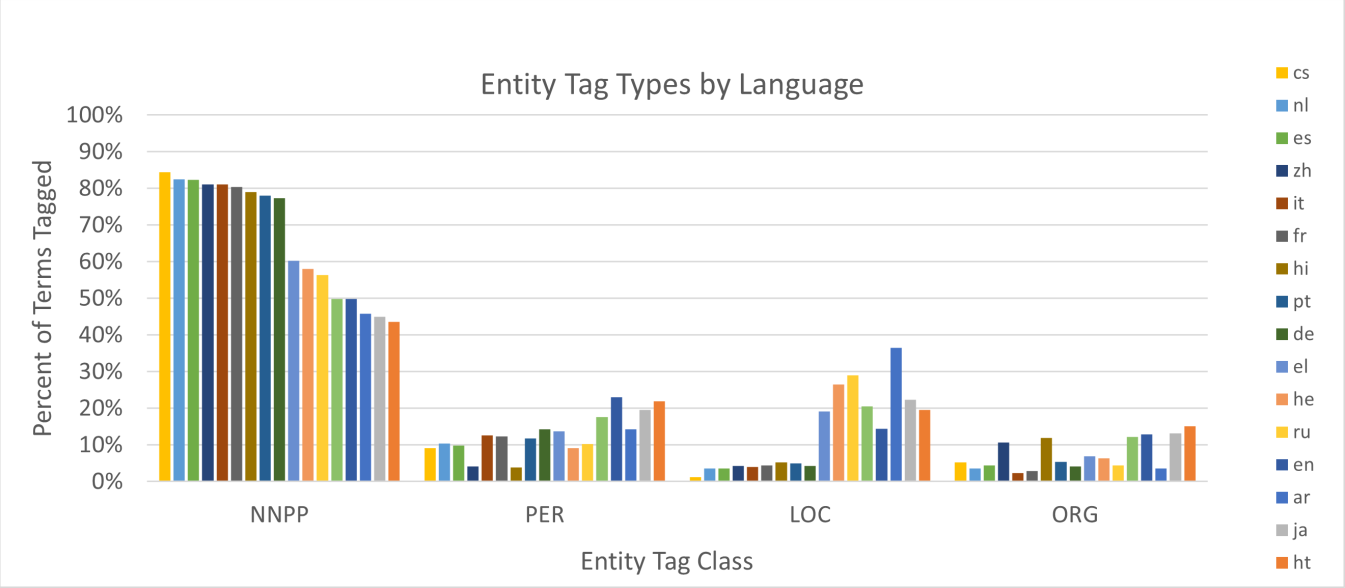

**Figure 6: Entity tag class by language after translation into English. This represents the baseline distribution of entity tag classes per language in NewsStand, not considering the role any of these terms play in clustering. Classes are filtered to the 4 most common (Proper Noun, Person, Location, Organization), representing 84.6% of the entity tags in NewsStand.**

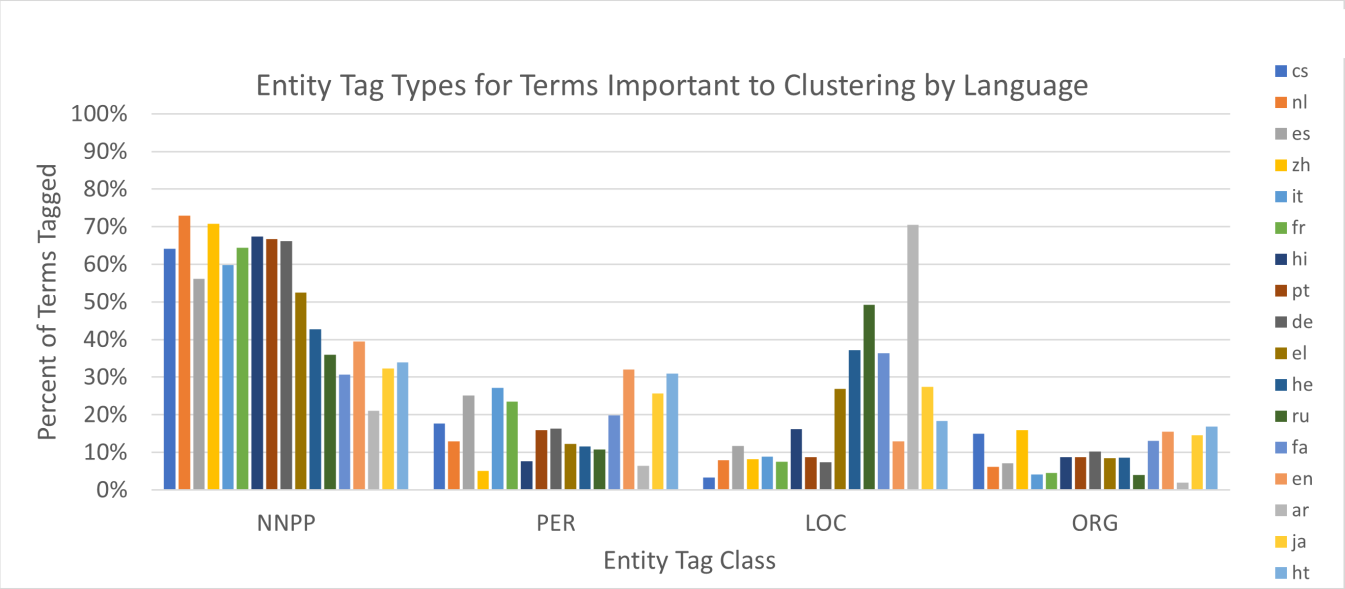

**Figure 7: Entity tag class by language for very important terms (TF-IDF score greater than 0.4). These terms are important for clustering since their TF-IDF scores are well above the average for NewsStand. Classes are filtered to the 4 most common (Proper Noun, Person, Location, Organization), representing 84.6% of the entity tags in NewsStand.**

Figures 4 and 5 show the average word and character entity tag density, respectively, for each language in NewsStand. The metric is calculated on the translated (into English) text, but in some cases, poor translation leads to many source words being carried directly into the translation output. For languages with *logographic* writing systems, this can make word count an unreliable measurement of article length. For example, we observed that poorly translated Chinese (ZH) articles with a mix of logograms and English words in the output have an unusually low word count and, therefore, a high entity tag : total word ratio, despite relatively few entity tags being recognized. Lee et al. [34] point out that Chinese words may contain several characters, but words are not separated by spaces, meaning

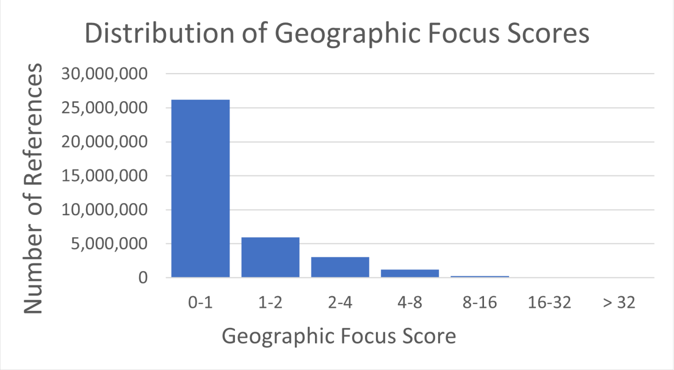

**Figure 8: Distribution of geographic focus scores for georeferences. Higher scores indicate local focus; lower scores indicate a lack of local focus.**

determining word count requires more complicated techniques for Chinese text than for English text. To account for this phenomenon, we instead take character count as the denominator of the ratio, which yields a metric more in line with the intuitive notion of the density of entities tagged even for cases when the source language is retained in the output.

Another consequence of source words being carried over or translated poorly is that they are not likely to be recognized by the entity tagger, even if they represent proper nouns. If this happens systematically across articles of a given language, it will manifest in a low average entity tag density for the language. Recalling Figure 2, the poor clustering observed for Russian (RU) and Greek (EL) articles (the highest and third highest proportion of singleton clusters, respectively, out of 17 languages) may be explained by the systematically low entity tag density measured for those articles, shown in Figure 5. **In other words, a low number of entity tags being recognized by the tagger means an article has a poor chance of clustering well with other articles in NewsStand.**

To make this phenomenon more concrete, we take the following example from a Russian article in NewsStand. The sentence originally written as "Не так давно известный политолог Досым Сатпаев поделился мнением об экологических рисках Казахстана." [1] was translated by NewsStand as "Not long ago, a political scientist opinion about environmental risks Казахстана...". The word "Казахстана" was carried over into the translation output, causing it to be missed by the entity tagger. The correct translation is *Kazakhstan*, which is a location that would have ideally been tagged and used to help cluster the article. Later in the article, another reference is made to Kazakhstan, but this time it is written as "Казахстан" in the original text, and the output is correctly translated, allowing the entity tagger to recognize the location. Taken together, these two instances illustrate an important issue for clustering translated text. **A proper noun that has multiple common spellings in a non-English source language can be problematic for entity tagging and clustering if the spellings are not translated consistently in the target language.**

## 6.2 FACTOR 2B: Entity Tag Class

Knowing that proper nouns are critical for clustering, we observe how the relative rarity (TF-IDF score) of different types of proper nouns contribute to an article's clusterability. Figure 6 shows the baseline distribution of entity tag classes per language in NewsStand, not accounting for the role those terms play in clustering. Proper noun is the dominant class, with many languages having, on average, between 70% and 80% of their entity tags in this category. Comparatively, far fewer instances of the more specific person, location, and organization tags are recognized. All entity classes besides the four most common classes, representing 84.6% of the entities tagged, are filtered out for clarity.

Looking at the term feature vectors, we can determine which of the terms in Figure 6 are important for clustering by considering their TF-IDF scores. Figure 7 shows the distribution of entity tags with high TF-IDF scores (greater than 0.4), broken out by class and language. These terms we consider very important for clustering since their TF-IDF scores are well above the average for NewsStand.

Interestingly, the Haitian and Japanese language articles have the lowest proportion of generic proper noun tags of all the languages indexed by NewsStand. Other languages, like Arabic, stand out for having an unusually high proportion of one entity tag class, in this case, location. This may be due to Arabic naming practices, which often include place names that might be incidentally tagged as location entities rather than person entities by the geotagger.

Across most languages, we observe that the person and location tag classes appear with higher frequency in Figure 7 than in Figure 6. Similarly, the more generic proper noun class appears with much lower frequency in Figure 7 than in Figure 6. **This indicates that location and person proper nouns are typically more important for clustering than generic proper nouns.** This finding coincides with an interesting problem with the recent multilingual embedding-based cross-lingual clustering approaches, namely that they are limited in their ability to differentiate between specific named entities like cities [62]. While mapping articles to a multilingual embedding space to cluster them seems like a reasonable method, our analysis shows that the proper noun entities, not the majority of the common language in the article, are what matters for clustering.

# 7 FACTOR 3: GEOGRAPHIC FOCUS

The nature of the geographic references in an article also contributes to how the article clusters.

## 7.1 Local and Global Georeferences

Following Quercini et al. [48], we define a reader's spatial lexicon as the limited set of locations that the reader can identify and place on a map [41, 48]. This is further broken down into the *local lexicon* and the *global lexicon*. The local lexicon refers to the set of small, highly local places familiar to an audience based on proximity. These places are commonly referenced in local newspapers, which have a localized and specific *geographic focus*. On the other hand, the global lexicon includes geographically distant but highly prominent places, such as major international cities, that are known by nearly everyone. For example, an article referring to "Paris" could be referring to the prominent "Paris, France", which is part of the global lexicon of places known by almost everyone, or it could be referring to one of the many smaller cities like Paris, Texas, which is part of the local lexicon for people in the surrounding areas.

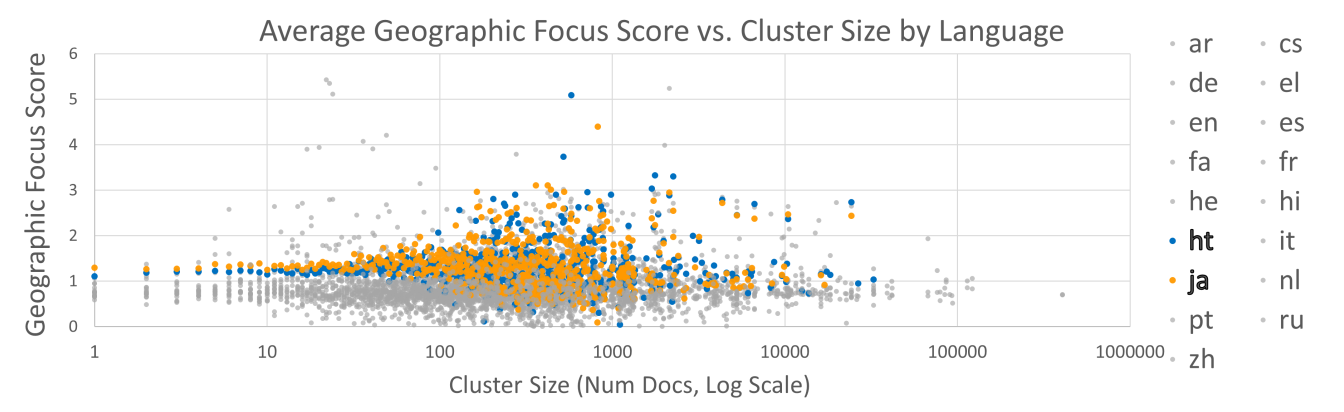

**Figure 9: Average geographic focus score for articles belonging to clusters of different sizes. Languages that cluster well, Haitian (HT) and Japanese (JA) are highlighted in the legend and colored blue and orange, respectively, in the plot. Best Viewed in color.**

NewsStand quantifies the overall locality versus globality of the georeferences in each article it ingests using a geographic focus score [16]. Figure 8 shows the distribution of geographic focus scores for individual georeferences in the NewsStand data. Higher scores indicate local focus and lower scores indicate a lack of local focus. The skew indicates most georeferences in NewsStand articles refer to places in the global lexicon, and comparatively fewer georeferences refer to places in the local lexicon.

This observation matches the intuition behind the geotagging framework MetaCarta [50], which assumes that toponyms correspond to the most prominent interpretation, such as Paris, France, about 95% of the time, and thus reasonably good geotagging can be achieved by always choosing the prominent location barring strong evidence to the contrary. However, Lieberman et al. [41] show that establishing local lexicons, which is what NewsStand's preprocessing pipeline does, leads to more accurate spatial indexes.

## 7.2 Article Focus

After identifying and disambiguating the geographic entities in the news articles, NewsStand's geotagger determines which georeferences are relevant to the article's overall geographic focus and which are mentioned in passing. The relevance of each georeference to its article's focus is computed using a linearly decreasing weighted frequency ranking, which is motivated by the fact that important georeferences are often made early in an article's text. With this weighting, an occurrence of a georeference $g$ that appears closer to the beginning of an article's text gives more weight to $g$'s ranking than a similar reference that appears near the end [59, 69].

Considering the local and global nature of georeferences and the weighted frequency raking that emphasizes references appearing early in the text, a typical example of an article with a high geographic focus score is one that references a local town or city at the start of the article. On the other hand, an article would garner a low focus score by referring to several prominently known locations, like major cities, throughout the article text.

Figure 9 shows the average overall geographic focus score for all articles of language $l$ residing in a cluster of size $n$, where $n$ is plotted on the x-axis in Log Scale and $l$ is shown by color. Haitian and Japanese languages, which showed better clustering behavior than the other languages in NewsStand in Section 5, also tend to have higher geographic focus scores. This points to one explanation for why Haitian and Japanese articles in NewsStand tend to cluster well- they tend to have a higher geographic focus, meaning they refer to places in the local lexicon early in the article text. In other words, these articles tend to focus on a location, thereby improving their chances of clustering with other articles. This aligns with our earlier finding that accurate translation of proper nouns, *particularly location references*, is critical to successful clustering. **When georeferences are not translated correctly, they are not recognized by the geotagger, and do not contribute to the focus score of the article or to the articles' ability to cluster with other similar articles referencing the same place.** This is one explanation for the high proportion of singleton clusters observed across many languages in NewsStand.

## 8 CONCLUSIONS AND FUTURE WORK

To better map the cross-lingual text clustering landscape, we evaluated the document clusters in a large system that has been performing cross-lingual text clustering on news articles for over a decade. In doing so, we found that the following factors influence the quality of the clustering behavior: the document's original language, proper noun usage and type, and geographic focus. Articles were more likely to form coherent clusters when they were originally written in certain languages, contained specific classes of proper nouns like location and person entities, and had a strong local focus tied to a particular geographic region rather than a global focus. By analyzing the clusters formed using a simple translate-then-cluster method, we highlight the apparent pitfalls associated with cross-lingual information retrieval (CLIR) in the news domain and point out that many of these issues are not solved by recent advances in CLIR, especially the use of multilingual embeddings and other language-independent methods for clustering, which are known to poorly distinguish between proper noun entities. Future work in this domain includes a detailed comparison of more complex cross-lingual clustering schemes, like large language model based methods, and further improvements to multilingual-embeddings to address their inability to distinguish between named entities like city names, which we show to be an important factor for clustering news articles. Finally, this work opens the door to using clustering to evaluate translation or multilingual embedding quality.

## ACKNOWLEDGMENTS

This work was sponsored in part by the NSF under Grants IIS-18-16889, IIS-20-41415, and IIS-21-14451.

## REFERENCES

[1] [n. d.]. Караван. https://www.caravan.kz/news/smertelnyjj-smog-zasukha-i-ugroza-goloda-o-kakikh-ehkologicheskikh-problemakh-kazakhstana-chashhe-vsego-pishut-v-zapadnykh-smi-753179/. Accessed: 2022-12-01.

[2] C. Aggarwal and P. Yu. 2006. A Framework for Clustering Massive Text and Categorical Data Streams, Vol. 2006. https://doi.org/10.1137/1.9781611972764.44

[3] C. Aggarwal and C. Zhai. 2012. A survey of text clustering algorithms. *Mining Text Data* (2012), 77–128. https://doi.org/10.1007/978-1-4614-3223-4_4

[4] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. 2003. A Framework for Clustering Evolving Data Streams. In *Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29* (Berlin, Germany) *(VLDB '03)*. VLDB Endowment, 81–92.

[5] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. 2004. A Framework for Projected Clustering of High Dimensional Data Streams. In *VLDB*.

[6] M. Agrawal, M. Karimzadehgan, and C. Zhai. 2009. An online news recommender system for social networks. In *Proceedings of the Workshop on Search in Social Media (SSM 2009), co-located with ACM SIGIR 2009 Conference on Information Retrieval, Boston*. Citeseer.

[7] M. Aljlayl and O. Frieder. 2001. Effective Arabic-English Cross-Language Information Retrieval via Machine-Readable Dictionaries and Machine Translation. In *Proceedings of the Tenth International Conference on Information and Knowledge Management* (Atlanta, Georgia, USA) *(CIKM '01)*. Association for Computing Machinery, New York, NY, USA, 295–302. https://doi.org/10.1145/502585.502635

[8] 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.

[9] R. Baeza-Yates and B. Ribeiro-Neto. 2011. *Modern Information Retrieval the Concepts and Technology Behind Search.*

[10] B. Baldwin and B. Carpenter. [n. d.]. LingPipe. http://alias-i.com/lingpipe/. Accessed: 2022-07-13.

[11] S. Banerjee, K. Ramanathan, and A. Gupta. 2007. Clustering Short Texts Using Wikipedia *(SIGIR '07)*. Association for Computing Machinery, New York, NY, USA, 787–788. https://doi.org/10.1145/1277741.1277909

[12] M. Bansal, J. DeNero, and D. Lin. 2012. Unsupervised Translation Sense Clustering. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Montréal, Canada, 773–782. https://aclanthology.org/N12-1095

[13] P. S. Bradley, U. M. Fayyad, and C. Reina. 1998. Scaling Clustering Algorithms to Large Databases. In *KDD*.

[14] D. Carmel, H. Roitman, and N. Zwerdling. 2009. Enhancing Cluster Labeling Using Wikipedia. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 139–146. https://doi.org/10.1145/1571941.1571967

[15] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. 2017. Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections. *SIGIR Forum* 51, 2 (aug 2017), 148–159. https://doi.org/10.1145/3130348.3130362

[16] J. Ding, L. Gravano, and N. Shivakumar. 2000. Computing geographical scopes of web resources. (2000).

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

[18] F. Farnstrom, J. Lewis, and C. Elkan. 2000. Scalability for Clustering Algorithms Revisited. *SIGKDD Explor. Newsl.* 2, 1 (jun 2000), 51–57. https://doi.org/10.1145/360402.360419

[19] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. 2022. Language-agnostic BERT Sentence Embedding. arXiv:2007.01852 [cs.CL]

[20] W. Nelson Francis. 1965. A Standard Corpus of Edited Present-Day American English. *College English* 26, 4 (1965), 267–273. http://www.jstor.org/stable/373638

[21] J. Gelernter and S. Balaji. 2013. An algorithm for local geoparsing of microtext. *GeoInformatica* 17, 4 (Oct. 2013), 635–667.

[22] S. Green, N. Andrews, M. R. Gormley, M. Dredze, and C. D. Manning. 2012. Entity Clustering Across Languages. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Montréal, Canada, 60–69. https://aclanthology.org/N12-1007

[23] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. 2003. Clustering data streams: Theory and practice. *IEEE Transactions on Knowledge and Data Engineering* 15, 3 (2003), 515–528. https://doi.org/10.1109/TKDE.2003.1198387

[24] Q. He, K. Chang, E. Lim, and J. Zhang. [n. d.]. *Bursty Feature Representation for Clustering Text Streams.* 491–496. https://doi.org/10.1137/1.9781611972771.50 arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611972771.50

[25] 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.

[26] 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.

[27] J. Hu, L. Fang, Y. Cao, H. Zeng, H. Li, Q. Yang, and Z. Chen. 2008. Enhancing Text Clustering by Leveraging Wikipedia Semantics. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Singapore, Singapore) *(SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 179–186. https://doi.org/10.1145/1390334.1390367

[28] G. Hulten, L. Spencer, and P. Domingos. 2001. Mining Time-Changing Data Streams. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California) *(KDD '01)*. Association for Computing Machinery, New York, NY, USA, 97–106. https://doi.org/10.1145/502512.502529

[29] A. Kalogeratos, P. Zagorisios, and A. Likas. 2016. Improving Text Stream Clustering Using Term Burstiness and Co-Burstiness. In *Proceedings of the 9th Hellenic Conference on Artificial Intelligence* (Thessaloniki, Greece) *(SETN '16)*. Association for Computing Machinery, New York, NY, USA, Article 16, 9 pages. https://doi.org/10.1145/2903220.2903229

[30] D. Kifer, S. Ben-David, and J. Gehrke. 2004. Detecting Change in Data Streams. In *VLDB*.

[31] Y. Kim, M. Amini, C. Goutte, and Patrick Gallinari. 2010. Multi-View Clustering of Multilingual Documents. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Geneva, Switzerland) *(SIGIR '10)*. Association for Computing Machinery, New York, NY, USA, 821–822. https://doi.org/10.1145/1835449.1835633

[32] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik. 2014. Cross-lingual detection of world events from news articles.. In *ISWC (Posters & Demos)*. 21–24.

[33] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik. 2014. Event Registry: Learning about World Events from News. In *Proceedings of the 23rd International Conference on World Wide Web* (Seoul, Korea) *(WWW '14 Companion)*. Association for Computing Machinery, 107–110. https://doi.org/10.1145/2567948.2577024

[34] C. Lee and H. Yang. 2003. A Multilingual Text Mining Approach Based on Self-Organizing Maps. *Applied Intelligence* 18, 3 (2003), 295–310. https://doi.org/10.1023/a:1023250105036

[35] K. Lee, K. Kageura, and K. Choi. 2002. Implicit Ambiguity Resolution Using Incremental Clustering in Korean-to-English Cross-Language Information Retrieval. https://doi.org/10.3115/1072228.1072314

[36] 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.

[37] L. Li, D. Wang, T. Li, D. Knox, and B. Padmanabhan. 2011. SCENE: A Scalable Two-Stage Personalized News Recommendation System. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 125–134. https://doi.org/10.1145/2009916.2009937

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

[39] 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.

[40] 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.

[41] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. 2010. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Proceedings of the 26th IEEE International Conference on Data Engineering*. Long Beach, CA, 201–212.

[42] 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.

[43] F. Liu, M. Vasardani, and T. Baldwin. 2014. Automatic identification of locative expressions from social media text: A comparative analysis. In *Proceedings of LocWeb'14*. Shanghai, China, 9–16.

[44] D. W. Oard. 1998. A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval. In *Machine Translation and the Information Soup*, David Farwell, Laurie Gerber, and Eduard Hovy (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 472–483.

[45] L. O'Callaghan, A. Meyerson, R. Motwani, N. Mishra, and S. Guha. 2002. Streaming-data algorithms for high-quality clustering. *Proceedings 18th International Conference on Data Engineering* (2002), 685–694.

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

[47] B. Pouliquen, R. Steinberger, C. Ignat, E. Käsper, and I. Temnikova. 2004. Multilingual and cross-lingual news topic tracking. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. COLING, Geneva, Switzerland, 959–965. https://aclanthology.org/C04-1138

[48] 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.

[49] M. Quezada and B. Poblete. 2019. A Lightweight Representation of News Events on Social Media *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1049–1052. https://doi.org/10.1145/3331184.3331300

[50] E. Rauch, M. Bukatin, and K. Baker. 2004. A Confidence-Based Framework for Disambiguating Geographic Terms. *Proceedings of the HLT-NAACL, Workshop on Analysis of Geographic References (WS9)* (03 2004). https://doi.org/10.3115/1119394.1119402

[51] N. Reimers and I. Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing*.

[52] K. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking. 2011. Topical clustering of tweets. *Proceedings of the ACM SIGIR: SWSM* 63 (2011).

[53] J. Rupnik, A. Muhič, G. Leban, B. Fortuna, and M. Grobelnik. 2017. News across Languages: Cross-Lingual Document Similarity and Event Tracking. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (Melbourne, Australia) *(IJCAI'17)*. AAAI Press, 5050–5054.

[54] 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

[55] 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

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

[57] 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.

[58] 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.

[59] 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.

[60] J. Sankaranarayanan, H. Samet, B. Teitler, M. D. Lieberman, and J. Sperling. 2009. TwitterStand: News in tweets, D. Agrawal, W. G. Aref, C.-T. Lu, M. F. Mokbel, P. Scheuermann, C. Shahabi, and O. Wolfson (Eds.). Seattle, WA, 42–51.

[61] 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

[62] 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

[63] 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

[64] P. Sheridan and J. Ballerini. 1996. Experiments in multilingual information retrieval using the SPIDER system. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. 58–65.

[65] S. K. Shinde and U. V. Kulkarni. 2011. Hybrid personalized recommender system using fast k-medoids clustering algorithm. *Journal of Advances in Information technology* 2, 3 (2011), 152–158.

[66] P. Singh, M. Wątorek, A. Ceglarek, M. Fąfrowicz, and Paweł Oświęcimka. 2022. Analysis of fMRI time series: neutrosophic-entropy based clustering algorithm. *Journal of Advances in Information Technology* 13, 3 (2022).

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

[68] 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).

[69] 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.

[70] B. E. Teitler, J. Sankaranarayanan, and H. Samet. 2010. *Online document clustering using the GPU*. Technical Report TR–4970. Computer Science Department, University of Maryland, College Park, MD.

[71] K. Wu and B. Lu. 2007. Cross-Lingual Document Clustering. In *Advances in Knowledge Discovery and Data Mining*, Zhi-Hua Zhou, Hang Li, and Qiang Yang (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 956–963.

[72] W. Xu, Y. Li, and J. Qiang. 2021. Dynamic clustering for short text stream based on Dirichlet process. *Applied Intelligence* 52, 4 (2021), 4651–4662. https://doi.org/10.1007/s10489-021-02263-z

[73] S. Zhong. 2005. Efficient online spherical k-means clustering. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, Vol. 5. 3180–3185 vol. 5. https://doi.org/10.1109/IJCNN.2005.1556436

[74] S. Zhong. 2005. Efficient streaming text clustering. *Neural Networks* 18, 5 (2005), 790–798. https://doi.org/10.1016/j.neunet.2005.06.008 IJCNN 2005.

[75] D. Zhou, M. Truran, T. Brailsford, V. Wade, and H. Ashman. 2012. Translation Techniques in Cross-Language Information Retrieval. *ACM Comput. Surv.* 45, 1, Article 1 (dec 2012), 44 pages. https://doi.org/10.1145/2379776.2379777

[76] 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