JMLR: Workshop and Conference Proceedings 53:1–16, 2016 BIGMINE 2016

Contextual Embedding for Distributed Representations of Entities in a Text Corpus

Md Abdul Kader MKADER@MINERS.UTEP.EDU
The University of Texas at El Paso, El Paso, TX 79968

Arnold P. Boedihardjo ARNOLD.P.BOEDIHARDJO@USACE.ARMY.MIL
U. S. Army Corps of Engineers, Alexandria, VA 22315

Sheikh Motahar Naim SNAIM@MINERS.UTEP.EDU
The University of Texas at El Paso, El Paso, TX 79968

M. Shahriar Hossain MHOSSAIN@UTEP.EDU
The University of Texas at El Paso, El Paso, TX 79968

Abstract

Distributed representations of textual elements in low dimensional vector space to capture context has gained great attention recently. Current state-of-the-art word embedding techniques compute distributed representations using co-occurrences of words within a contextual window discounting the flexibility to incorporate other contextual phenomena like temporal, geographical, and topical contexts. In this paper, we present a flexible framework that has the ability to leverage temporal, geographical, and topical information of documents along with the textual content to produce more effective vector representations of entities or words within a document collection. The framework first captures contextual relationships between entities collected from different relevant documents and then leverages these relationships to produce inputs of a graph, or to train a neural network to produce vectors for the entities. Through a set of rigorous experiments we test the performance of our approach and results show that our proposed solution can produce more meaningful vectors than the state-of-the-art methods.

Keywords: Contextual Embedding; Distributed Representation of Entity

Introduction

Modern text mining tasks extensively rely on lower dimensional representations of documents. Many systems consider words as the unit of text, as well as many frameworks leverage language ontology (Chen and Manning, 2014), sentence structures (Finkel et al., 2005; Angeli et al., 2015), annotations (Toutanova et al., 2003), and natural language processing techniques to conceptualize text for better reflection of the context, thus making the tools heavily language-dependent. The task of generating contextual representations of text requires a generalized approach that can capture latent relationships between information pieces without any exhaustive usage of dictionary or linguistic tools. Language independent mechanisms are gradually becoming essential with the increasing appearance of domain-specific terminologies and derivative acronyms in modern text data. With complex textual information, meta data, and latent themes, it has become more challenging to compute relationships between entities because co-occurrence is no more the sole indicator of *relevance* between entities. From the perspective of document similarity, the use of overlap of terms to compute the similarity is not sufficient to capture contextual relevance. This paper aims at generating distributed representations of elements of text, especially entities,

to capture latent but contextual relevance even when entities do not appear in the same document.

The general aim of a distributed representation is to capture syntactic and semantic relationships. Current distributed representation generation techniques for text datasets, e.g., *word2vec* (Mikolov et al., 2013b,c), *doc2vec* (Le and Mikolov, 2014), and *topic2vec* (Shamanta et al., 2015), rely on a sliding window over the contents of the documents to create a context. This context window is used to create the input and output samples for a neural network. The most prominent feature of these frameworks is the ability to generate word vectors that preserve syntactic context of the words. The use of a sliding window as the context still limits the potential of the techniques because of the assumption that contextual words lie solely within a window or within a document. While training the model for generating the distributed vectors, *word2vec*-family of algorithms look into one document (one line, to be precise) at a time — thus ignoring the order and interdependence of the documents. In reality, and also based on our observation, an event or a topic is historically covered by a group of articles, and inclusion of that group information in training could improve the quality of the word vectors that are contextually relevant to a particular event. In addition, time plays an important role in contextual drift of the vocabulary. Most text datasets (such as, news articles and scientific publications) are nowadays time-stamped. As an example of how context of a word may change over time — the context of the word *cloud* before the year 2000 was relevant to weather, while today it might be more relevant to cloud computing and cloud storage. Moreover, the context is tightly coupled with the topics of the documents where the word *cloud* was seen.

In addition to time, geographical locations related to a document may have a great influence on the context of the entities involved. For example, Figure 1 shows entities surrounding President Barack Obama in three different articles published in 2010, 2012, and 2014. Notice that the entities surrounding *Obama* in the three entity relationship graphs of Figure 1 create geographical contexts — Afghanistan, Russia, and Middle East. This is an indication that the geographical scope and context related to an entity may vary over time.

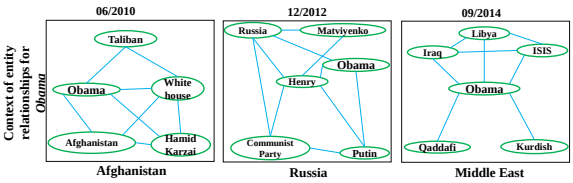


Figure 1: Contextual pieces of information around entity *Obama*. Three entity relationship graphs show three different geographical contexts (Afghanistan, Russia, and Middle East) of three different years (2010, 2012, and 2014).

In this paper, we describe a new mechanism to compute contextually relevant entities of each document of a corpus. The contextual information is bound by temporal, geographical, thematic information retrieved from each document. Our proposed framework generates distributed vectors taking the contextual information into account. As a result of the contextual relevance of the vectors, our method is able to discover causal and evidential relevance between entities. For example, *Cholera* and *Flood* or *Storm* are relevant. Contextually, entities like *Burma* and *Myanmar* are the same, which is better captured by our framework over state-of-the-art methods.

In summary, the contributions of this paper are as follows.

- We propose an optimization framework that can, for each document of a corpus, flexibly generate contextual information constrained by time, geographical location and latent topics. The retrieved contextual information pieces do not solely rely on

co-occurrence of the entities in other documents, rather they depend in relationships of entities seen in other relevant pairs of documents.
- We demonstrate two techniques to effectively generate low dimensional vector representations of entities by leveraging the discovered contextual relationships.
- We conduct a set of experiments to evaluate the generated distributed entity vectors. We also demonstrate how to leverage the generated vectors for traditional clustering and classification problems. The quality of the vectors are evaluated using a benchmark word-analogy dataset as well.

Problem Description

From our empirical analysis (Section 5.1), we observe that the context of a document is influenced by the topics published recently. As such context detected around an entity may change over time as the relevant topics surrounding that entity change, we exemplify this phenomena in Figure 2 which displays four news articles related to President Barack Obama and a relevant entity relationship-graph for each of these four documents generated by our proposed system. Our system uses each document as a seed to retrieve relationships between entities from recently published relevant documents (the mechanism is described later in Section 4). As a result, the entities in the relationship-graphs of Figure 2 may not appear in the seed documents shown in the figure. The figure shows that a document published in November 2008 describes the relationships between contemporary Senator Barack Obama and Senator Hillary Clinton. The relationship graph reveals contextual relationship between President Bush, White House and Illinois, Barack Obama, and Hillary Clinton. Another document published in June 2010, which contains President Obama, shows a relationship graph that is different than the one published in 2008. This is because President Obama's context relevant to the document published in 2010 is surrounded by different entities than in 2008. In September 2012, the surrounding context of the President Barack Obama switches to the election where Mitt Romney was the opponent leader from the Republican Party. The relationship graph in Figure 2 includes journalist Eric Fehrnstrom who was related to Romney as a top donor for the election campaign. The document placed in 2012 in Figure 2 does not contain the entity *Eric Fehrnstrom* but our system includes him because of his relevance with the topic of the document. President Obama's surrounding context through

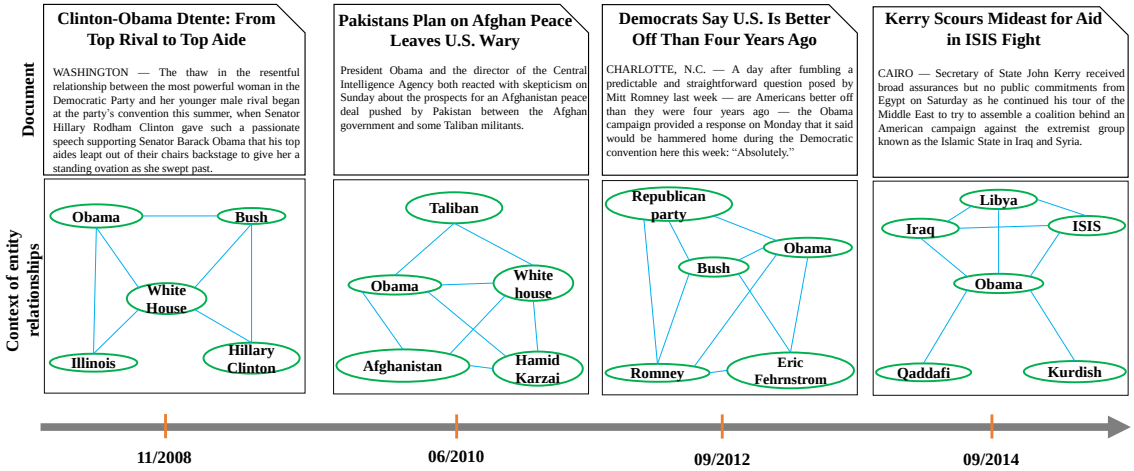


Figure 2: Variation of context of the entity *Obama* from November 2008 to September 2014.

a seed document published in 2014 shows that the concentration shifted toward entities like *Libya*, *Iraq*, *ISIS*, *Qaddafi*, and *Kurdish*.

The example of Figure 2 demonstrates that the surrounding context around a entity may change over time. Along with many parameters, the context is influenced by the topic of the documents where an entity is observed. Our framework retrieves the relevant entities through evidence seen in recently published articles, establishes relationships between pairs of entities seen in different (but relevant) documents, and finally builds a holistic contextual representation for each entity leveraging the relationships. We formally describe the associated problem in the following subsection.

Problem Formulation

Let $\mathcal{D} = \{d_1, d_2, \ldots, d_{|\mathcal{D}|}\}$ be the set of documents and $\mathcal{E} = \{e_1, e_2, \ldots, e_{|\mathcal{E}|}\}$ be the set of entities in the corpus. Each document $d \in \mathcal{D}$ has a set of entities $\mathcal{E}_d \subset \mathcal{E}$, which we refer to as the textual content of d. In addition to its textual content, every document also includes a set of extra information. Let the geographical location related to each document $d \in \mathcal{D}$ be G_d and the publication date of d be T_d. Also, inspired by topic modeling (Blei et al., 2003) techniques, we assume that every document is a mixture of l latent topics. Let $\mathcal{T}_d$ be the topic distribution in document d.

Our primary focus in this paper is on news articles. The higher level task from an analytic point of view is to generate vectors for all entities relevant to a subset of documents $\mathcal{D}_s \subset \mathcal{D}$ that represents a particular event of interest, e.g., *Ebola outbreak* or *Cholera*. That is, $\mathcal{D}_s$ is the user input for the generation of relevant entities. Ideally, $\mathcal{D}_s$ can be the set of returned documents from a search query, or a set of documents prepared by an expert. Notice that $\mathcal{D}_s$ is the set of seed documents but the scope of relevant entities of the seed documents may span the entire corpus $\mathcal{D}$. Combining the seed information with textual, geographical and temporal relevance of each document, we define a series of tasks to generate distributed vectors for each entity.

1. For each given seed document $d \in \mathcal{D}_s$, identify the set of nearest neighbors $\mathcal{N}_d \subset \mathcal{D}$.
2. For a given seed document d, find a set of documents $\vartheta_d \subset \mathcal{D}$ that are published before d. Each document in ϑ_d satisfies some topical, geographical, and temporal constraints. For each nearest neighbor of d, $d' \in \mathcal{N}_d$, list documents $\vartheta_{d'} \subset \mathcal{D}$ that are published before d' and that satisfy the same topical, geographical, and temporal constraints.
3. Identify a set of entity relationships $R = \{\rho_1, \rho_2, \ldots, \rho_{|R|}\}$ where each relationship $\rho_i \in R$ is a pair of entities (e_1, e_2) such that e_1 is observed in d and d' where d is the seed document and d' is a nearest neighbor of d. Additionally, e_2 is observed in a document of ϑ_d and another document of $\vartheta_{d'}$. The more evident ρ_i is among d' and documents of $\vartheta_{d'}$ the stronger ρ_i is.
4. Transform the entity relationship set R generated for every seed document $d \in \mathcal{D}_s$ to generate distributed representations of every entity traced by the steps above.

In the next section we describe our proposed framework that carries out these tasks.

Related Research

Due to the superiority of distributed representation in capturing generalized view of information over local representations, it has been successfully used in diverse fields of scientific research (Chalmers, 1992; Hinton, 1986; Elman, 1991; Hummel and Holyoak, 1997; Pollack, 1990). A pioneering work of Rumelhart et al. (1988) on distributed representation in language modeling targets learning of representations by back-propagating errors using

a neural network. Later, a more sophisticated neural probabilistic model (Bengio et al., 2003) was proposed by Bengio et al., which uses a sliding window based context of a word to generate compact representations. Recently Mikolov et al. (2013a) introduced continuous bag-of-words (CBOW) and skip-gram models to compute continuous vector representations of words efficiently from very large data sets. The skip-gram model was significantly improved in (Mikolov et al., 2013b), both in terms of speed and quality of the generated vectors, by replacing hierarchical softmax with a more efficient negative sampling technique and including phrase vectors along with words. Le and Mikolov (2014) then extended the CBOW model to learn distributed representation of higher level texts like paragraphs and documents. Unlike the word embedding methods discussed above that produce a singular representation of a word or a phrase, Huang et al. (2012) propose a language model that incorporates both local and global document context and learn multiple embeddings per word to account for homonymy and polysemy.

Although these word embedding frameworks use different approaches and address multiple aspects of a language to generate better context of the words, they completely rely on the textual content of the documents. In our framework, we look beyond text merely appearing withing a document by incorporating temporal, geographical and topical information. We argue that these additional information pieces are useful to understand the context of a unit (word or entity) better, and thus can be used to generate word embeddings capturing subtle difference in the context.

Methodology

The proposed framework consists of a number of components. First, we develop a document model where each document is represented as a probability distribution over the set of entities in the corpus. Second, for each seed document, we find a set of nearest neighbor documents. Third, for a seed document and each of its nearest neighbors, we generate another set of documents constrained by some criteria. Fourth, we formulate an optimization problem to extract the relationships between the entities in the sets of documents retrieved using the previous steps. Finally, we compute distributed vectors for the entities encountered in all the documents selected for all seeds. We leverage two methods to generate the vectors, one is focused on a graph based approach and the other one is driven by the machinery commonly seen in neural network based distributed vector generation (Mikolov et al., 2013b; Shamanta et al., 2015). In the following subsections we describe each of these steps in more detail. For the convenience of the readers, we list the symbols used in this paper in Table 1.

Table 1: List of symbols

Symbol	Description
$\mathcal{D}$	Set of documents
$\mathcal{E}$	Set of entities
$\mathcal{D}_s$	Set of seed documents
G_d	Geo location of document d
T_d	Publication date of document d
l	Number of latent topics in the corpus
$\mathcal{T}_d$	Topic distribution of document d
$\mathcal{N}_d$	Nearest neighbors of d
ϑ_d	Documents published before d bound by topical, geographical, and temporal constraints
α	Geographical context threshold
β	Temporal context threshold

Document Modeling

Our approach focuses on entities detected from the text instead of considering words as the primary feature unit. The motivation behind the use of entities comes from the analytic necessity of proximity measures among pairs of entities like people, organization, and location. The process described in this paper is generic in nature and can be adapted for unigrams or words without no modification. We use standard Named Entity Recognizers (Alias-i.; Stanford NLP Group) to extract entities from each news articles of the corpus. The probability distribution $P^d = \{p_1^d, p_2^d, \ldots, p_{|\mathcal{E}|}^d\}$ of each document $d \in \mathcal{D}$ over the set of entities $\mathcal{E}$ can be computed as:

$$P_i^d = \frac{W(e_i, d)}{\sum_{e' \in \mathcal{E}} W(e', d)} \tag{1}$$

where e_i is the i^{th} entity of the entity set $\mathcal{E}$ and $W(e, d)$ is the weight reflecting the associativity between document d and entity e. We compute the association between each document $d \in \mathcal{D}$ and each entity $e \in \mathcal{E}$ using a normalized form of *TF-IDF* (Hossain et al., 2012).

$$W(e, d) = \frac{(1 + log(tf_{e,d}))(log\frac{|\mathcal{D}|}{df_e})}{\sqrt{\sum\limits_{e' \in \mathcal{E}_d} \left((1 + log(tf_{e',d}))(log\frac{|\mathcal{D}|}{df_{e'}})\right)^2}} \tag{2}$$

where $tf_{e,d}$ is the frequency of entity e in document d, df_e is the number of documents containing entity e, and $\mathcal{E}_d$ is the set of entities detected in document d.

Expansion from a Seed Document

As described earlier, the basic idea of a seed document comes from the fact that context of any entity appears from a document. The context of the same entity seeding from two different documents may vary. Later in Section 4.3, we describe how our framework discovers relevant entities (or, entity relationships, to be more precise) given a seed document. Figure 3 outlines the process of expanding a seed document. For each seed document $d \in D_s$ where $D_s \subset D$, we select k nearest neighbors $\mathcal{N}_d = \{d_1, d_2, \ldots, d_k\}$ from D. In Figure 3, the seed document d is denoted by d_0 for consistency in pictorial representation. The k-nearest neighbors are selected based on KL-divergence (Kullback and Leibler, 1951) between the probability distribution of d and the distribution of each of the documents in D. The set of $k+1$ documents, $\mathcal{N}_{d_0} = \{d_0, d_1, d_2, \ldots, d_k\}$, ideally represents a coherent set of textually similar documents. For example, the seed document d_0 in Figure 3 (illustrated in the half right side of the figure) is about *Cholera outbreak in Haiti*, and the other three documents are the nearest neighbors containing similar events.

Once we have documents $\mathcal{N}_{d_0}$ similar to the seed document, our approach seeks a prior event or entity relationships that most likely led to the event described in the seed document. As described later in Section 5.1, the theme of a particular news article is more prominent in its recent past. We use a popular topic modeling algorithm, Latent Dirichlet Allocation (LDA) (Blei et al., 2003), to estimate the topic distribution $\mathcal{T}_d$ in each document $d \in D$. For every document $d_i \in \mathcal{N}_{d_0}$ we create a set of candidate documents $\vartheta_{d_i} = \{d_1^i, d_2^i, \ldots, d_{|\vartheta_{d_i}|}^i\}$ where each candidate document $d_j^i \in \vartheta_{d_i}$ satisfies the following three constraints.

- **Topical divergence:** Relevant events are expected to have some commonality in their topics. Therefore, d_i should have a certain level of topical similarity to $d_j^i \in \vartheta_{d_i}$.

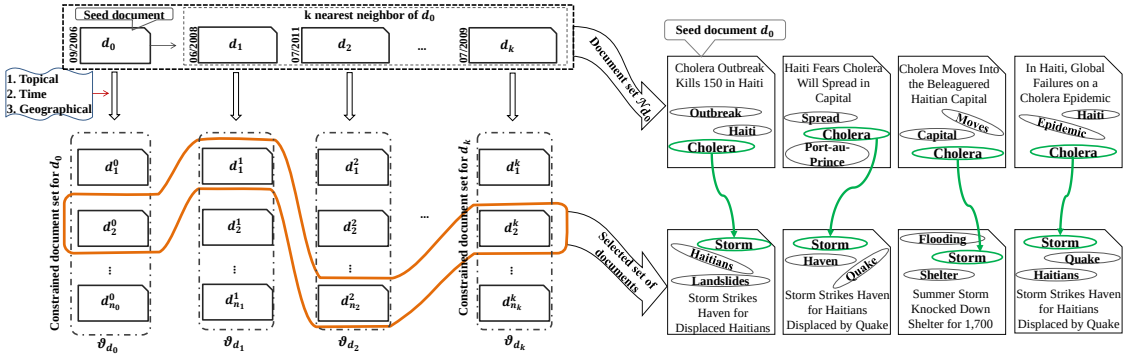


Figure 3: *(left)* Candidate generation process from a seed document d_0. *(right)* Cholera outbreak and its preceding related events.

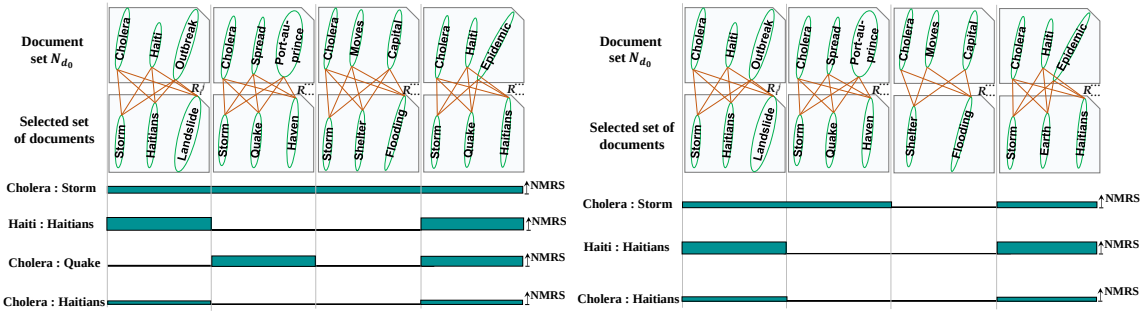


Figure 4: *(left)* The relationship *Cholera:Storm* is found in both documents of each of the k columns indicating a strong contextual relevance. *(right)* None of the relationships is present in all k pairs of documents indicating that the relationships are not very evidential. The objective function will favor the left set of selected documents because it reveals coherent relationships.

d_j^i is included in ϑ_{d_i} only if $KLDiv(\mathcal{T}_{d_i}, \mathcal{T}_{d_j^i}) \leq \alpha$, where α is the topical context threshold.

- **Geographical context:** Relevant events are likely to happen around similar geographical locations. G_d represents the set of all location entities in document $d \in \mathcal{D}$. d_j^i is included in ϑ_{d_i} only if $G_{d_i} \cap G_{d_j^i} \geq \beta$, where β is the geographical context threshold.
- **Temporal Order:** Based on our observation regarding dominance of a topic of a document in the recent past, our time constraint for the selection of d_j^i is $T_\theta < T_{d_j^i} < T_{d_i}$, where $T_\theta = T_{d_i} - \theta$ is the date θ days prior to T_{d_i}.

The left half part of Figure 3 represents the process described in this subsection and the other half provides sample documents.

Construction of Entity Relationships

The intuition behind generating entity relationships using separate documents is that, if a relationship $\rho = (e_1, e_2)$, where $e_1 \in d_i$ and $e_2 \in d_j^i$, is repeatedly observed between document pairs (d_i, d_j^i) such that $d_i \in \mathcal{N}_{d_0}$ and $d_j^i \in \vartheta_{d_i}$, then the relationship ρ is an important one because multiple pairs of documents support ρ. In the top row of the right half of Figure 3, we present four documents of $\mathcal{N}_{d_0} = \{d_0, d_1, d_2, d_3\}$. From the corresponding

sets of ϑ_{d_0}, ϑ_{d_1}, ϑ_{d_2}, and ϑ_{d_3} we select one document d_j^i from each set ϑ_{d_i}. The four best representative documents, d_j^i, are presented in the bottom row of the right half of Figure 3.

Figure 4 shows two scenarios with two different sets of selected documents. In the left side, each document pair (d_i, d_j^i) contains the entity relationship $\{$*Cholera:Storm*$\}$ whereas in the right side none of the entity relationships exists in all pairs of (d_i, d_j^i) documents. This indicates that the set of documents selected in the left side of Figure 4 provides a more coherent evidence of entity relationships than the one in the right side. Now, a crucial question that can be asked is why we advocate selection of at most one document from each ϑ_{d_i} to prepare the selected set of documents. Based on empirical studies during the development of the objective function described in this subsection, rarely seen entity relationships between d_i and documents of ϑ_{d_i} that are evident in all i's are more important than frequently seen entity relationships even if they are observed in all i's of d_i and ϑ_{d_i} document-pairs. For example, the relationship $\{$*Cholera:Basket Ball*$\}$ might be very frequent for most i's of (d_i, d_j^i) document pairs but the abundance of such relationship results from the fact that regardless of time and event under analysis there will be always sports, fashion, technology sections in most news papers. Our objective here is not to find a context using the most frequent relationships observed many times, rather to discover more accurate entity relationships within ϑ_{d_i} that are observed many times for documents similar to d_i, i.e., $\mathcal{N}_{d_0}$, through the selection of rare entities. Therefore, during the selection process, our objective function should seek for a document $d_j^i \in \vartheta_{d_i}$ that creates a rare set of entity pairs with d_i that are evident at similar level of scarcity in other $(d_{i'}, d_j^{i'})$ document pairs where $d_j^{i'} \in \vartheta_{d_{i'}}$ and $i \neq i'$.

This subsection outlines how we select the best representative document d_j^i from ϑ_{d_i} to best capture the entity relationships. Notice that each (d_i, d_j^i) pair, in the example of Figure 3, repetitively contains the relationship *(Cholera, Storm)* indicating that *Cholera* appeared after *Storm* because each document $d_i \in \mathcal{N}_{d_0}$ is published after any document $d_j^i \in \vartheta_{d_i}$ was published.

Given a hypothetical probability distribution over all entities X that can be considered as a synthetic document, we can construct a membership probability distribution $\mathbf{v}_i^X = \{\mathbf{v}_{i_1}^X, \mathbf{v}_{i_2}^X, \ldots, \mathbf{v}_{i_{n_i}}^X\}$ for the documents in ϑ_{d_i}. Each $\mathbf{v}_{i_j}^X$ will represent how probable it is that X and $P^{d_j^i}$ are the same compared to all the documents in ϑ_{d_i}.

$$\mathbf{v}_{i_j}^X = \frac{exp(-\|X - P^{d_j^i}\|)}{\sum_{j'=1}^{n_i} exp(-\|X - P^{d_{j'}^i}\|)} \tag{3}$$

Since our aim is to select one document from each set ϑ_{d_i} our objective function should reward for a non-uniform distribution of $\mathbf{v}_i^X$. We measure the non-uniform nature of a distribution using the following formula:

$$C(\mathbf{v}_i^X) = \frac{\|U(\frac{1}{n_i}) - \mathbf{v}_i^X\|_1}{2 - \frac{2}{|\mathbf{v}_i^X|}} \tag{4}$$

$C(\mathbf{v}_i^X)$ will generate a scalar in the range from 0 to 1 where larger scores indicate high probabilities associated with only a few documents of ϑ_{d_i}.

If X is the free variable of an optimization routine then the following objective function would result in a high probability document in each set ϑ_{d_i}.

$$f(X) = \sum_{i=0}^{k} C(\mathbf{v}_i^X)$$

$f(X)$ will basically provide the best X for which there is a relevant document (without any confusion) in each set ϑ_{d_i}. If each ϑ_{d_i} has an importance factor that is additionally determined as a free variable $A = \{a_1, a_2, \ldots, a_k\}$ such that $\|A\|_1 = 1$, then the objective function becomes

$$f(X, A) = \sum_{i=0}^{k} a_i \times C(\mathbf{v}_i^X) \tag{5}$$

Equation 5 is a suitable objective function to ensure a common theme between the selected documents of each set ϑ_{d_i}, given that each selected document has the highest $\mathbf{v}_{i_j}^X$ after the optimization routine converges. However, this does not guarantee that the entity relationships R_i^j observed between $d_i \in \mathcal{N}_{d_0}$ and a selected $d_j^i \in \vartheta_{d_i}$ for a particular i are also observed for other i values. At this stage, we will modify Equation 5 to incorporate such relationships.

A set of relationships R_i^j between two documents $d_i \in \mathcal{N}_{d_0}$ and $d_j^i \in \vartheta_{d_i}$ is composed of the set of all possible relationships $\rho = (e_1, e_2)$ such that $e_1 \in d_i$ and $e_2 \in d_j^i$. We compute the shared information between two sets of relationships R_i^j and R_l^k using Normalized Mutual Relationships Score (NMRS):

$$NMRS(R_i^j, R_l^k) = \sum_{\rho \in R_i^j \cup R_l^k} p(\rho|R_i^j, R_l^k) log \frac{p(\rho|R_i^j, R_l^k)}{p(\rho|R_i^j)p(\rho|R_l^k)} \tag{6}$$

where the probability $p(\rho|R_i^j)$ of a relationship $\rho = (e_1, e_2)$ given the set of relationships R_i^j is computed using the following formula

$$p(\rho|R_i^j) = \frac{f_{e_1,d_i} * f_{e_2,d_j^i} + 1}{\sum_{\rho' \in R_i^j}(f_{e'_1,d_i} * f_{e'_2,d_j^i} + 1)} \tag{7}$$

where f_{e_1,d_i} is the frequency of entity e_1 in document d_i.

Similarly, the probability $p(\rho|R_i^j, R_l^m)$ of the relationship ρ given the set of relationships R_i^j and R_l^m is calculated by

$$p(\rho|R_i^j, R_l^m) = \frac{min(f_{e_1,d_i} * f_{e_2,d_j^i}, f_{e_1,d_l} * f_{e_2,d_m^l}) + 1}{\sum_{\rho' \in R_i^j \cup R_l^m}(max(f_{e'_1,d_i} * f_{e'_2,d_j^i}, f_{e'_1,d_l} * f_{e'_2,d_m^l}) + 1)} \tag{8}$$

We modify the objective function in Equation 5 to incorporate the relationships in the following new objective function.

$$\begin{aligned} f(X, A) &= \sum_{i=1}^{K} C(\mathbf{v}_i^X) \sum_{j=1}^{n_i} C(\mathbf{v}_{(i+1)}^X) \times \sum_{m=1}^{n_{i+1}} a_i \mathbf{v}_{i_j}^X a_{i+1} \mathbf{v}_{(i+1)_m}^X NMRS\Big(R_i^j, R_{i+1}^m\Big) \\ &= \sum_{i=1}^{K} C(\mathbf{v}_i^X) C(\mathbf{v}_{(i+1)}^X) a_i a_{i+1} \times \sum_{j=1}^{n_i} \sum_{m=1}^{n_{i+1}} \mathbf{v}_{i_j}^X \mathbf{v}_{(i+1)_m}^X NMRS\Big(R_i^j, R_{i+1}^m\Big) \end{aligned} \tag{9}$$

Similar to the objective function of Equation 5, the objective function of Equation 9 will result in a common theme between the selected documents of each set ϑ_{d_i}, given that each selected document has the highest $\mathbf{v}_{i_j}^X$. In addition, the objective function of Equa-

tion 9 maximizes the entity relationships R_0^j observed between $d_0 \in \mathcal{N}_{d_0}$ and a selected $d_j^0 \in \vartheta_{d_0}$ over all R_i^j sets with subsequent i values. The objective function is smooth and continuous and any local optimization routine will be able to maximize it over the set of variables X and A. We used Python to implement the objective function and leveraged `scipy.optimize.minimize` as our optimization routine.

Vector Generation from Relationships

Using the optimization formula described in Section 4.3, after the selection of the best (d_i, d_j^i) pairs of documents, we obtain a set of entity relationships. The objective function was maximized for this set of relationships. These entity relationships basically form a context and can be represented as edges of a graph for every seed document, as shown in Figures 1 and 2. These transformations are done to extract the latent features contained by the aggregated relationships. Now, the task of vector generation for each entity, given the contextual set of entity relationships for every seed document, can be performed in one of the two ways, (a) compose all the entity relationships in a weighted graph and apply an orthogonal transformation of the weighted graph adjacency matrix to form vectors for the entities, and (b) use the entity relationships discovered for every seed document to train a neural network to generate neural entity embeddings. The first approach uses spectral graph theory (Chung, 1997) and Principal Component Analysis (PCA) to transform the $|\mathcal{E}| \times |\mathcal{E}|$ adjacency matrix to a $|\mathcal{E}| \times C$ matrix of C principal components. The second approach resembles the method used in Word2Vec (Mikolov et al., 2013b,c). At each step of the training of Word2Vec, a set of consecutive words from a document is given to the network where it takes one word from that set as input and attempts to predict the remaining words in the set. We leverage this model to create vectors of entities by feeding each observed entity relationship (a pair of entities) to the network — one entity is used as input to predict the other one. Figure 5 shows that $entity_i$ is given as the input of the two-layer neural network to predict $entity_j$ for a relationship $\rho = (entity_i, entity_j)$.

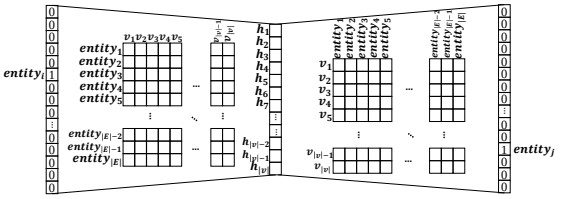


Figure 5: Two layers neural network for entity vectorization.

Experimental Results

In this section, we seek to answer the following questions[1].

1. What is the justification for using the temporal, geographical and topical constraints during the optimization relevant to each seed document? (Section 5.1)
2. How effective are the generated entity relationships? (Section 5.2)
3. How good are the generated vectors in capturing the context of entities? (Section 5.3)
4. Can the entity vectors be used to produce high-quality clusters? (Section 5.4)
5. How useful are the entity vectors in classifying documents? (Section 5.5)

We used approximately 54,000 New York Times articles that are categorized as politics. For supervised evaluations, we used the 20 Newsgroups dataset (Lang, 1995), which contains approximately 20 thousand newsgroup documents.

Significance of Constraints

In Section 4.2, we explained how a seed document can be expanded by first taking its k-nearest documents and then generating a set of candidate documents for each of those

1. Codes and data are provided here: http://dal.cs.utep.edu/projects/storyboarding/bigmine16/

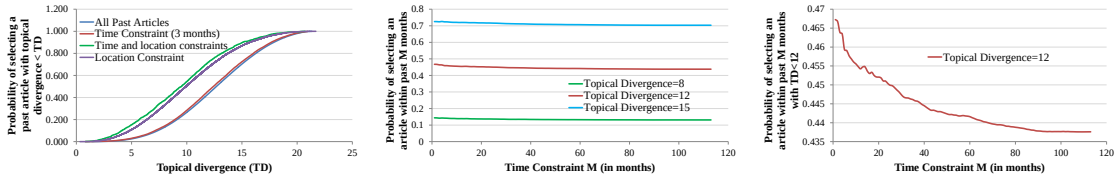


Figure 6: (left) Addition of constraints increases the likelihood of having topically similar documents. (middle and right) The effect of time constraints on topical evolution.

k documents. The candidate documents are selected by enforcing temporal, geographical, and topical constraints. In this section, we provide empirical justification for using such constraints while generating the candidate documents. Figure 6 (left) shows that the probability of selecting a topically similar document published prior to a seed document increases when selection is constrained by both time and location, as evident through the green line of the plot. Figure 6 (middle) demonstrates that longer spans in time as the temporal constraint dilute topics resulting in higher topical divergence with the seed. A similar evidence is found in the experiment with Figure 6 (right). It shows that longer temporal span in the past for the selection of the candidate documents leads to lower probability of finding topically similar documents. The probability is the ratio, number of documents satisfying topical constraint to total number of documents satisfying time constraint. The topical divergence between the seed document and a document published in the past is measured by computing the KL-divergence between the topic distribution of these two documents. These divergences are averaged over the number of pairs observed during each experiment.

All the experiments of Figure 6 illustrate that the selection process of candidate documents from a seed is well founded by natural topical trends observed in news articles.

Contextual Relationships for a Seed Document

After we select the k-nearest documents and the corresponding sets of candidate documents for each seed document d_0, we formulate an optimizer in Section 4.3 that produces highly probable entity relationships and the corresponding set of selected documents that carry a

Table 2: Selected set of documents and corresponding relationships for a seed document that describes cholera outbreak.

$d_i \in \mathcal{N}_{d_0}$	Set of relationships	Selected set of documents	a_i
Cholera Outbreak Kills 150 in Haiti	–	New Flood Warnings Raise Fears in Pakistan	0.12
Haiti Fears Cholera Will Spread in Capital	'the world health organization : the world health organization', 'the world health organization : health', 'the world health organization : world health organization', 'world health organization : the world health organization', 'world health organization : health'	Evacuations Continue in Southern Pakistan	0.12
Vaccinations Begin in a Cholera-Ravaged Haiti	'world health organization : cholera', 'health : cholera', 'health : port-au-prince', 'world health organization : port-au-prince', 'world health organization : health'	In Haiti, Global Failures on a Cholera Epidemic	0.12
Pattern of Safety Lapses Where Group Worked to Battle Ebola Outbreak	'balakrish nair : haitians', 'balakrish nair : haitian', 'balakrish nair : paul farmer', 'balakrish nair : h.i.v', 'balakrish nair : haiti'	Botswana Doctor Is Named to Lead W.H.O. in Africa	0.12
In Haiti, Global Failures on a Cholera Epidemic	'thomas r : partners', 'thomas r : sierra leone', 'thomas r : ebola', 'thomas r : sierra leones', 'thomas r : he'	In a Gang-Ridden City, New Efforts to Fight Crime While Cutting Costs	0.12
Ebola Could Strike 20,000, World Health Agency Says	'montereys : balakrish nair', 'montereys : nepal', 'montereys : blame', 'montereys : the lancet', 'montereys : tropical medicine'	Health Officials Try to Quell Fear of Ebola Spreading by Air Travel	0.08
Cholera Moves Into the Beleaguered Haitian Capital	'the world health organization : health', 'world health organization : health', 'titus naikuni : ebola', 'titus naikuni : liberia', 'titus naikuni : the world health organization'	Amid Cholera Outbreak in Haiti, Misery and Hope	0.09
Medical Need Climbs Alongside Death Toll in Yemen	'health : borders', 'diarrhea emergency : humanitarian', 'diarrhea emergency : the world health organization', 'diarrhea emergency : marie-evelyne louis', 'diarrhea emergency : christine antoine'	Pakistani Lawmakers Urge Diplomacy in Yemen Conflict but Decline Combat Role	0.12
U.N., Fearing a Polio Epidemic in Syria, Moves to Vaccinate Millions of Children	'unicef : yemens', 'unicef : yemenis', 'unicef : the world health organization', 'unicef : abdu rabbu mansour hadi', 'unicef : houthi'	40 Years After War, Israel Weighs Remaining Risks	0.11

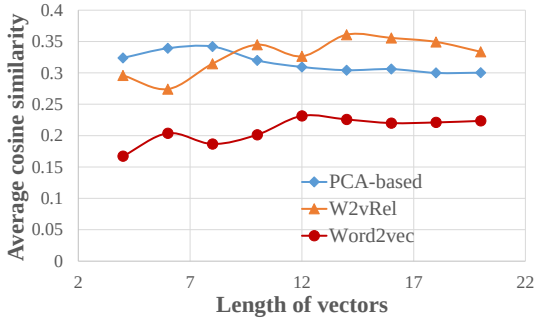
common theme. An example of such a set of entity relationships is shown in the second column of Table 2 for a Cholera related seed document. The relationships in i-th row of the table are characterized by high $NMRS(R^j_{i-1}, R^k_i)$ scores in Equation 6, i.e., they share significant mutual information in the document pairs of i-th row and $(i-1)$-th row. The first document in the first column of Table 2 is the seed document that describes cholera outbreak in Haiti. The other documents in the first column are the k nearest neighbors of the seed document. The third column records a selected document, which yields highly probable relationships, from the candidate pool of each document in the first column. A few notable relationships are *'the world health organization : world health organization'*, *'unicef : the world health organization'*, and *'health : borders'*. The last column shows the final importance factor or weight of each row, as determined by the optimizer (variable A in Equation 9). In this specific case, for the nine pairs of documents in nine rows of the table, the weights varied from 0.08 to 0.12.

Given a seed document, our system is able to discover contextual entity relationships from an automatically crafted set of documents selected from the entire corpus. Table 2 shows the outcome for one seed document. For every seed document, our system generates pairs of contextual entities that might not directly appear in the seed document, or the relationships might not even appear in one single document in the entire corpus.

Evaluation through Entity Analogy

In this subsection, we compare the generated vectors for entities using two methods as described in Section 4.4, PCA and neural network based approaches, to Google's Word2vec in terms of contextual analogy of entities.

In the first experiment, we evaluate the ability of the distributed vectors obtained from our methods in capturing context. We leverage the database of capital-country, country-currency and city-state pairs provided with the code base of Word2vec (Mikolov et al., 2013b) as ground truth in this experiment. We calculate the average cosine similarities among entity pairs of all capital-country, country-currency and city-state pairs. Figure 7 (a) shows that our methods, referred as *PCA-based* and *W2vRel*, outperform the Word2vec method in terms of average similarity between the pairs. Interestingly, after a certain length of vector, our *W2vRel* method surpasses *PCA-based* method but the Word2vec method still performs worse than both of our methods. Figure 7 (a) provides an evaluation using ground



(a) Evaluation using set of analogous words shows that our methods perform significantly better for making similar vectors for entities that are contextually analogous.

Entity pair	*Analogous?*	*PCA-based*	*W2vRel*	*Word2vec*
burma : myanmars	✓	***0.84***	***0.99***	***0.64***
burma : yangon	✓	0.713	0.86	0.66
myanmars : yangon	✓	0.954	0.86	0.93
myanmars : tripoli	✗	***0.02***	***-0.06***	***0.45***
burma : tripoli	✗	0.056	-0.06	0.17
republican : al gore	✗	-0.124	0.08	0.7
republican : obama	✗	***-0.111***	***0.09***	***0.31***

(b) Sample cosine similarities between pairs of vectors generated by three methods. Our approaches capture similarities/dissimilarities better than Word2vec both for analogous and non-analogous pairs.

Figure 7: Experimental results for analogous pairs of entities.

truth analogous entities. As an additional analysis, we examined pair samples to evaluate vectors generated both for analogous and non-analogous pairs. In Figure 7 (b), we present cosine similarity scores of seven pairs of entity vectors, out of which three pairs are analogous and four pairs are non-analogous. The table shows that the cosine similarity between *Burma* and *Myanmars* is high using both our approaches than Word2vec. Given that *Burma* is the former name of *Myanmars*, our approaches tend to capture this relationship better than Word2vec. Similarly, our approaches capture better analogous relationships than Word2vec for cases, {burma : yangon} and {myanmars : yangon}. For all the non-analogous samples — {myanmars : tripoli}, {burma : tripoli}, {republican : al gore}, and {republican : obama} — cosine similarity scores resulting from pairs of vectors using our approaches are lower than the scores using Word2vec. This indicates that our approaches are able to distinguish non-contextual pairs better than Word2vec.

We also examined entities of interest by computing their 10-nearest neighbors using all three methods. Table 3 compares the top ten contextually similar entities of *Qaddafi* retrieved by these three methods. In Table 3, *PCA-based* and *W2vRel* refer to the two approaches we use to generate vectors for entities. Obviously, *Word2vec* refers to Google's Word2vec approach. It should be explained in this space how entity vectors are generated using Word2vec instead of words only. To make the systems comparable, we made the text input for Word2vec a list of entities as they appear in the text documents instead of using word units. All three methods retrieve correlated entities to some extent as the most similar entities to *Qaddafi*. Cosine similarity between vectors was used to compute proximity. The entities retrieved by two of our methods produced better results than the ones retrieved by the baseline Word2vec approach. For example, *Colonel Qaddafi* appears as the most similar entity to *Qaddafi* using both our approaches but Word2vec lists *Colonel Qaddafi* as the ninth nearest entity to *Qaddafi*. Our observation in this case is that the PCA-based method retrieved most contextual entities for *Qaddafi*. Highly relevant entities to *Qaddafi* are marked in the table in bold.

Similarly, Table 4 shows the top ten entities contextually similar to *Burma*, the former name of the country *Myanmar*. The PCA-based and W2vRel methods retrieved several related entities that are highlighted in bold. Word2vec could not retrieve any entity that is related to *Myanmar*, to the best of our knowledge.

Table 3: Top 10 contextually similar entities for *Qaddafi*.

Qaddafi					
PCA-based		**W2vRel**		**Word2vec**	
colonel qaddafi	0.983	**colonel qaddafi**	0.99	**tripoli**	0.81
the a.p	0.969	**tripoli**	0.973	zimbabwe african national union-patriotic front	0.79
tripoli	0.938	zliten	0.96	ice	0.77
libyan	0.909	alain jupp	0.959	daniel malan	0.770
monica garca prieto	0.824	laurence hart	0.958	keeb	0.768
libyans	0.816	baghdadi al-mahmoudi	0.955	curiosity of ice	0.759
solidarity	0.811	the a.p	0.948	guantnamo	0.756
thirachai phuvanat-naranubala	0.807	jupp	0.947	kabul international	0.754
nature	0.722	mustapha abdul jalil	0.934	**colonel qaddafi**	0.734
jay carney	0.708	bad boy	0.933	james g	0.724

Table 4: Top 10 contextually analogous entities for *Burma*.

Burma					
PCA-based		**W2vRel**		**Word2vec**	
myanmar.he	0.973	student generation	0.999	teams	0.853
association of **south-east asian** nations	0.973	**myanmars**	0.999	clegg	0.837
nobutaka machbimura	0.972	kenji nagai	0.998	stanford hospital	0.827
min zaw	0.972	**burma** media association	0.998	asahi glass foundation	0.813
kenji nagai	0.971	association of **south-east asian** nations	0.998	shaw	0.803
ibrahim gambari	0.971	lee hsien loong	0.998	van	0.802
shwe	0.84	gambari	0.998	mcdonnell young	0.801
myanmars	0.84	**myanmar.he**	0.997	central district of california	0.792
sheik nabil qaouk	0.84	u nyan win	0.997	yavlinsky	0.788
tyre	0.84	min zaw	0.996	kenji nagai	0.786

Evaluation using Clusters

The previous section (Section 5.3) describes that our approaches generate vectors that are easily distinguishable for non-analogous pairs, as well as detectable for analogous pairs. Vectors with such capabilities tend to produce good clustering outcomes. In this section we evaluate the generated vectors in terms of clustering quality. We cluster the entities, given a generated vector for each entity, using k-means clustering. We apply k-means on three different sets of entity vectors generated by three methods (a) our *PCA-based* approach, (b) our neural network based approach referred to as *W2vRel* in the figures, and (c) benchmark *Word2vec* approach. We measure the quality of clustering outcomes using two standard cluster evaluation measures: Silhouette coefficient (Rousseeuw, 1987) and Dunn index (Dunn, 1973). For both the measures, larger values are better. Figure 8 (left) shows that our two proposed methods outperform Word2vec in terms of the average Silhouette coefficient. Negative average Silhouette coefficient for Word2vec indicates lack of structure in the clustering outcome. Both our approaches have positive Silhouette coefficients. Figure 8 (right) shows that our neural network based method, referred as *W2vRel* in the figure, performs better than the Word2vec and our PCA based in terms of Dunn index. Our PCA based method performs marginally better than baseline Word2vec method.

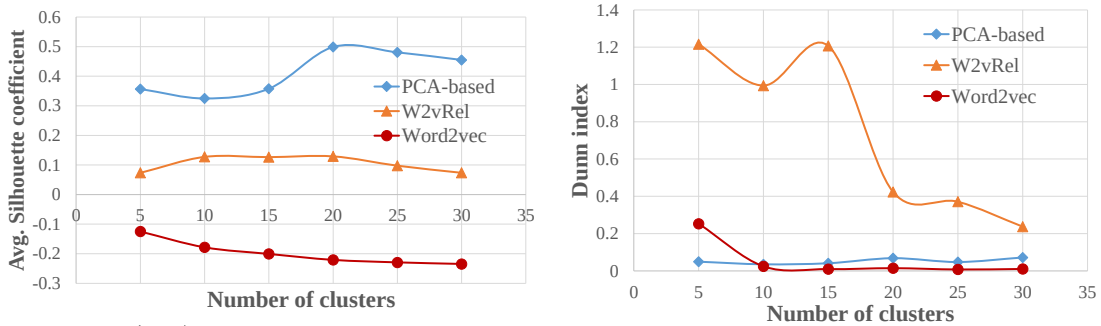


Figure 8: *(left)* Our approaches exhibit positive and higher average Silhouette coefficient than Word2Vec. *(right)* Vectors generated by our neural network based method provides the best Dunn index.

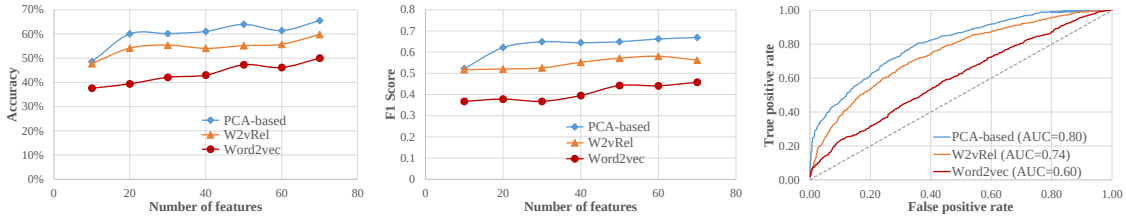


Figure 9: Accuracy, F1 score and ROC curve for classifying documents from 20newsgroups dataset based on the entity vectors produced by the methods.

Evaluation using Classification

In this section, we compare the quality of the vectors by the three methods through a classification task. We use 20newsgroups (Lang, 1995) dataset for this purpose that contains 18,828 news articles divided into 20 exclusive classes related to topical categories. The purpose of our methods and Word2vec is to generate vectors for entities. We construct feature vectors for the documents for classification by first clustering the entity vectors into c groups using k-means clustering algorithm. Then we create a c-dimensional feature vector for each document d_i where the j^{th} element of the feature vector is the number of entities in document d_i that belong to the j^{th} cluster of entities.

We use Support Vector Machine (SVM) to classify the documents. We use 10-fold cross validation for the evaluation. In a previous section we have observed that the entity vectors generated by our methods return better clustering of entities. As a result, the entity vectors contribute towards better document classification as shown in Figure 9. The left and middle plots in Figure 9 show that our methods (marked as PCA-based and W2vRel) outperform the Word2vec method in terms of classification accuracy and F-measure. Figure 9 (right) shows the corresponding ROC curve for each method. To combine multiple class ROC we use macro averaging. Macro averaging is appropriate in this example because the 20 newsgroups dataset contains almost equal number of documents for each group. Both our approaches result in higher Area under the curve (AUC) than that of the Word2vec method.

Conclusion

Our framework leverages contextual information available in a corpus to generate distributed representations for entities observed in each document. Experimental results in this paper depict comparative analyses of different word embedding techniques, studies of effectiveness of the generated distributed vectors in several data mining applications, and qualitative analyses of the contexts generated for entities. Although within the scope of this paper, we considered only geographical, temporal and topical information as bounding context of our objective function, the framework is designed in such a flexible way that other types of information, if available, can be easily integrated. Our study in this paper was limited to news articles. We will expand our analyses on the scholarly literature, study context and paradigm shifts over time, and investigate how distributed representations can be time dependent.

Acknowledgments

This material is based upon work supported by the U.S. Army Engineering Research and Development Center under Contract No. W9132V-15-C-0006.

References

Alias-i. LingPipe 4.1.0. Accessed: May 13, 2016, http://alias-i.com/lingpipe/.

Gabor Angeli, Melvin Johnson Premkumar, and Christopher D Manning. Leveraging linguistic structure for open domain information extraction. In *ACL-IJCNLP*, 2015.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *Machine Learning Research*, 3:1137–1155, 2003.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *JMLR*, 2003.

David J Chalmers. Syntactic transformations on distributed representations. In *Connectionist Natural Language Processing*, pages 46–55. 1992.

Danqi Chen and Christopher D Manning. A fast and accurate dependency parser using neural networks. In *EMNLP*, pages 740–750, 2014.

Fan RK Chung. *Spectral graph theory*, volume 92. 1997.

Joseph C Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. 1973.

Jeffrey L Elman. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7(2-3):195–225, 1991.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, 2005.

Geoffrey E Hinton. Learning distributed representations of concepts. In *CogSci'86*, 1986.

M. Shahriar Hossain, Patrick Butler, Arnold P. Boedihardjo, and Naren Ramakrishnan. Storytelling in entity networks to support intelligence analysts. In *KDD '12*, 2012.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *ACL*, 2012.

John E Hummel and Keith J Holyoak. Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104(3):427, 1997.

S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 1951.

Ken Lang. Newsweeder: Learning to filter netnews. In *ML95*, pages 331–339, 1995.

Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML'14*, pages 1188–1196, 2014.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS'13*, 2013b.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751, 2013c.

Jordan B Pollack. Recursive distributed representations. *Artificial Intelligence*, 46(1), 1990.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Cognitive Modeling*, 5, 1988.

Debakar Shamanta, Sheikh Motahar Naim, Parang Saraf, Naren Ramakrishnan, and M Shahriar Hossain. Concurrent inference of topic models and distributed vector representations. In *ECML PKDD*, pages 441–457. 2015.

Stanford NLP Group. Stanford NER. Accessed: May 13, 2016, http://nlp.stanford.edu/software/CRF-NER.shtml.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL '03*, 2003.