Semantic Frame-Based Document Representation for Comparable Corpora

Hyungsul Kim, Xiang Ren, Yizhou Sun, Chi Wang and Jiawei Han
University of Illinois at Urbana-Champaign
{*hkim21, xren7, sun22, chiwang1, hanj*}*@illinois.edu*

***Abstract*—Document representation is a fundamental problem for text mining. Many efforts have been done to generate concise yet semantic representation, such as bag-of-words, phrase, sentence and topic-level descriptions. Nevertheless, most existing techniques counter difficulties in handling monolingual comparable corpus, which is a collection of monolingual documents conveying the same topic. In this paper, we propose the use of frame, a high-level semantic unit, and construct frame-based representations to semantically describe documents by *bags of frames*, using an information network approach. One major challenge in this representation is that semantically similar frames may be of different forms. For example, "radiation leaked" in one news article can appear as "the level of radiation increased" in another article. To tackle the problem, a text-based information network is constructed among frames and words, and a link-based similarity measure called *SynRank* is proposed to calculate similarity between frames. As a result, different variations of the semantically similar frames are merged into a single descriptive frame using clustering, and a document can then be represented as *a bag of representative frames*. It turns out that frame-based document representation not only is more interpretable, but also can facilitate other text analysis tasks such as event tracking effectively. We conduct both qualitative and quantitative experiments on three comparable news corpora, to study the effectiveness of frame-based document representation and the similarity measure *SynRank*, respectively, and demonstrate that the superior performance of frame-based document representation on different real-world applications.**

***Keywords*-document representation; bag of frames; text information network; link-based clustering**

I. Introduction

Document representation is a fundamental problem for user comprehension and understanding [22], [4], [32], and is also critical to various text processing tasks like text categorization [29] and retrieval [1]. Because of its simplicity and effectiveness, the bag-of-words representation is widely adopted in most of document processing tasks, especially in text categorization [26]. However, there are several areas that other representations outperform the bag-of-words where it is needed to capture complex semantics of text, including phrasal, syntactic and more sophisticated linguistic structures [25], [23], [3].

Analyzing monolingual comparable corpora is one of the areas where the bag-of-words representation has limitations. Monolingual comparable corpora is defined as a collection of documents in the same language (*e.g.*, English) that overlap in the information they convey. In the age of information overload, we can easily collect or access such corpora that cover the same topic such as multiple news reports on the same or similar events from different news agencies, and reviews about the same or similar products or services.

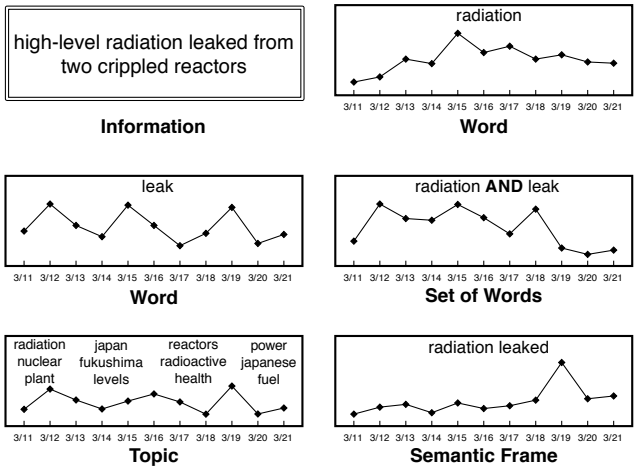


Figure 1. 5 different representations for information and their trend plots

Beyond several studies on monolingual comparable corpora, which study sentence alignments [2] and paraphrasing rules [17], analyzing monolingual comparable corpora has many potential applications. First, the analysis can give a comprehensive summary about one event, fact, or entity because documents in a comparable corpus cover different perspectives of the topic. Second, the analysis can derive a set of consistent information across documents, which helps remove some trivial or misleading information. This application is close related to trustworthiness analysis, where many studies on structured data like movie databases [33] and sensor data [31] have been done, but not in unstructured data like documents. Third, analyzing monolingual comparable corpora can track the trend of information when each document has timestamp.

As the first step toward the analysis of monolingual comparable corpora, we propose the use of *frame*, a high-level semantic feature derived by semantic role labeling (SRL) [14], as the basic unit for document represent in comparable corpora. In Figure 1, we demonstrate the power of semantic frame. Specifically, a collection of news articles about Japan's 2011 Tsunami (which caused radiation leaked from two crippled nuclear reactors in March 19th) is used as a comparable corpora. We use 5 different kinds of

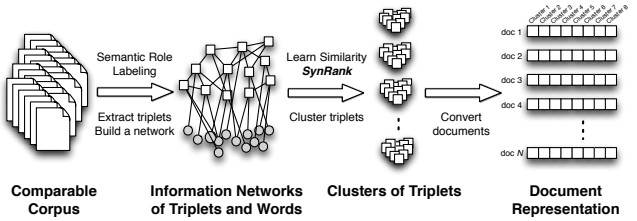


Figure 2. Overview of Our Proposed Framework

representations for this particular information, and measure the popularity using the occurrences of the representations within the corpus, and draw the trends in Figure 1. As shown in the figure, the semantic frame is the only one that isolates the information and detect the peak in March 19th. We will further discuss on this aspect in Section V.

Semantic frame has proved its superiority in various applications including information extraction [6] and question answering [30]. Each frame is a verb-argument structure from a sentence, and is arranged as a subject-verb-object *triplet* where each part is associated with a set of words. By extracting triplets we can find the most important semantic information from a set of documents, and can serve as a better representation for other tasks like event tracking.

However, a higher level document representation usually results in a higher complexity feature space, which leads to sparser document model due to the variational forms. For example, "radiation leaked" in one news article can appear as "the level of radiation increased" in another article. In this paper, we try to resolve the sparsity challenge when dealing with frame-based document representation, by grouping semantically similar frames together.

An information network-based approach is developed to define similarity between frames, by which similar frames can be better grouped together due to the propagation of similarity along different types of network links. We first construct a syntactic structure between each frame-derived triplet and its words. Then, a bi-typed *information network* is built for a corpus by extracting all the nodes and links from different documents, where nodes represent words and triplets, and links exist between them if they are connected in their original syntactic structure. We further propose a link-based similarity measure, called *SynRank*, to calculate the similarity between triplets in an iterative way, where we design different iterative formulas for different types of objects by considering their semantic meanings. Then we can cluster similar frames together according to the obtained similarity. One representative triplet will be selected for one cluster, and documents are represented by the corresponding frames (see Figure 2).

Finally, we validate the effectiveness of our similarity measure comparing with other baselines on several real-world datasets. The results show that the frame-based document representation is more interpretable and comprehensive than baseline methods.

We summarize our contributions of this work as follows:

1) We propose a novel frame-based document representation method which can capture the document semantics and represent comparable corpora in a comprehensive and concise way.
2) We propose to construct an information network from the corpus, and develop a link-based similarity measure called SynRank to capture the similarity between frames and similarity between words jointly, in an iterative and global way.
3) Experiments on real-world datasets show the power of the new document representation method, compared with several baseline approaches.

II. Preliminaries and Method Overview

In this section, we introduce preliminary knowledge about semantic frame and provide an overview of our proposed frame-based document representation method.

Different from bag-of-words representation, which misses the semantic relationships among words, semantic frames aim at capturing the most important elements such as entities and their relationships from a sentence, defined as follows.

Definition 1: A ***semantic frame*** $f \in F$ *is a verb-argument structure in a sentence that describes a type of event, relation, or entity and the participants in it [13].*

This definition is based on the semantic role formalism of PropBank [27]. As seen from Figure 3, extracted frames contain richer information than word and less information (usually single fact, statement, or proposition) than sentences. Notice that, there could be several frames derived from one sentence, and the number of semantic frames in a sentence equals to the number of verbs in the sentence. In this work, we use SRL tool SENNA parser [7] for raw frame extraction, which is reported to have about 74% F_1 measure on CoNLL 2005 benchmark dataset.

We further formulate each semantic frame into a **triplet** of subjective, verb and objective (see Figure 3), to preserve semantic roles and content in an effective and concise way.

Definition 2: We denote the ***semantic triplet*** *as* $t = (s, v, o)$*, where* s *is* subjective word set *consisting of words with* A_0 *SRL tags in frame* f*,* o *is* objective word set *consisting of words with* A_1 *SRL tags in frame* f*, and* v *is* verb word set *containing verb and all the other arguments such as* A_2*, AM-TMP and AM-LOC, where* A_0 *represents the subjective,* A_1 *represents the objective,* A_2 *represents indirect objective, AM-TMP represents temporal modifier, and AM-LOC represents the location modifier.*

By re-structuring frames into triplets, we have a much clearer structure of each frame. However, these raw triplets cannot be directly used as features to represent documents because there still exists many semantically similar variations (e.g., "earthquake hit Japan" and "quake struck

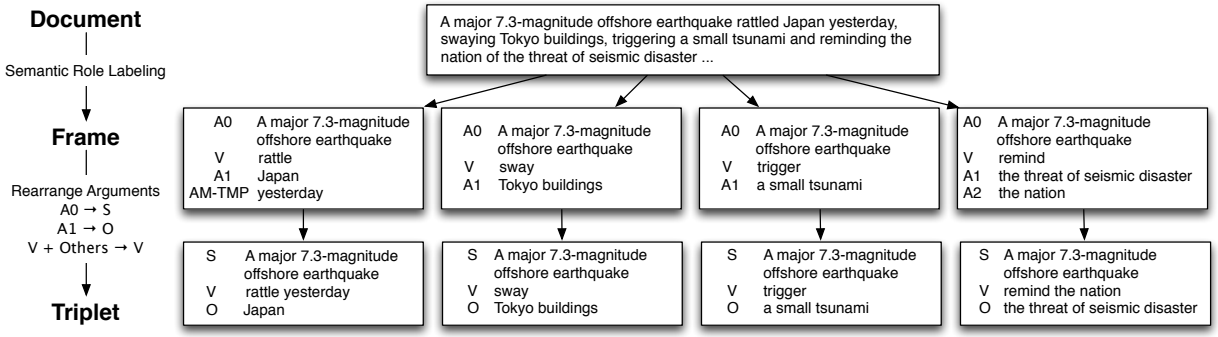


Figure 3. An Illustrative Example for Process of Extracting Triplets from Document.

Japan"), leading to a high-complexity feature space and thus sparse document representation. To resolve this, we first construct a semantic text information network among words and triplets, and then propose a link-based similarity function to measure their similarity. Similar triplets are grouped into clusters based on the similarity and the frame corresponding to representative triplet in each cluster will be selected as the final representation feature for documents.

The overall framework of the process can be summarized into the following three steps (see also Figure 2).

1) **Raw semantic frame extraction.** In this step, raw semantic frames and corresponding semantic triplets are first extracted from sentences in documents through semantic role labeling tool (see Figure 3).
2) **Semantic text information network construction.** We construct a semantic information network for words and triplets extracted from corpus (see Figure 4 and 5), which provides a novel view that different text objects are connected by semantic links.
3) **Link-based triplet clustering for document representation.** Finally, we propose a link-based similarity measure, and cluster triplets into different groups based on it. We select the most representative one in each cluster for final representation of the documents.

The first step is easily done by semantic role labeling tool, we now introduce Step 2 and 3 in following sections.

III. Semantic Text Information Network Construction

In order to merge similar triplets, we need a way to measure similarities between them, which is a problem related to the paraphrase detection task. One of the paraphrase detection methods is leveraging synonyms from a knowledge base such as WordNet [12] to improve the detection performance [24]. However, this kind of approaches are limited for the synonyms in the general usages. For example, the word "threat" is frequently used to refer the word "radiation" in the Japan's tsunami corpus[1], but their similarity in Wordnet is 1.743[2], which is lower than the similarity score of 1.897 between "buildings" and "cars." Thus, it is important to derive a *corpus-based similarity measure* for words in order to measure the similarities of triplet.

To meet this need, we propose to cluster similar triplets using an information network approach, where various text objects and their connections are captured by a *semantic text information network*. As we will show in Section V, it is much more effective to compute the word similarity and frame similarity jointly and globally in a unified framework instead of computing them separately by utilizing links in this text information network.

Definition 3: A ***semantic text information network*** *is a bi-typed undirected graph containing two types of object sets T (triplets) and W (words). For each triplet $t \in T$, it has links to a set of words in W, as well as links to its neighbor triplets as its* context. *The link types are defined by their relations: links from triplet to its contextual triplets belong to* triplet-triplet (TT) *relation; links between triplets and its words belong to* triplet-word TW *relation.*

The network schema of the information network is shown in Figure 4. Notice that words are distinguished by different semantic roles (S,V,O) such as "S: earthquake" and "O: tsunami" in Figure 4.

For a triplet node t in the network, the neighbors of t are denoted by $N_R(t)$, where $R \in \{TT, TW\}$ represents the link type. We denote the context of triplet t as $N_{TT_\sigma}(t)$, where σ is the size of the context window, i.e., the number of nearby triplets that are considered as its context in a document. For simplicity, we denote it by $N_{TT}(t)$.

Based on the semantic text information network, we derive a semantic similarity measure for triplets by analyzing *triplet-triplet* and *triplet-word* links. The intuition behind this measure is that similar text objects share *similar context* around them and *similar content* within them. The details will be introduced in Section IV.

IV. Link-based Triplet Clustering

In this section, we explain in details how link information in the semantic text information network can be leveraged to cluster triplets, where different types of relations, i.e., *triplet-triplet* relation and *triplet-word* relation, are considered simultaneously. We first introduce a novel similarity measure, called SynRank, then show how to compute SynRank, and

[1]There was a nuclear accident and radiation leaks following the tsunami.
[2]This similarity is computed using Leacock & Chodorow [20].

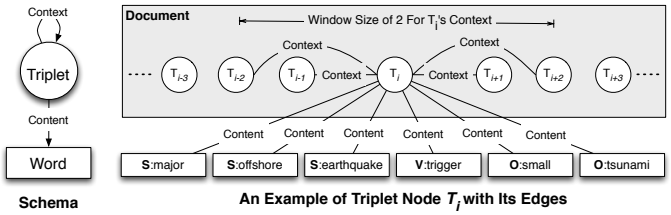


Figure 4. Meta-Schema of Semantic Text Information Network

finally the clustering algorithm for triplets based on this similarity measure.

A. SynRank: A Link-Based Semantic Similarity Measure

Similar to SimRank [18], which measures the similarity between objects in a network based on the assumption that "two objects are similar if they share similar neighbors," we propose our link-based similarity measure, following the intuition that "similar triplets share *similar context* around them and *similar content* within them." In particular, a triplet is most similar to itself, with maximum score 1.

SynRank deals with different types of relations (i.e., triplet-triplet context relation and triplet-word content relation) simultaneously with different updating mechanisms, which distinguishes itself from other link-based measures such as SimRank [18] and P-Rank [34]. Iteratively computing SynRank function can propagate similarities between object pairs in a global manner, *i.e.*, word similarity and triplet similarity are mutually adjusted according to the whole corpus (see Figure 9).

We formulate above intuition into a link-based similarity measure function, called **SynRank**, which takes the recursive form as follows. For two triplets nodes t_i and t_j, at the k-th iteration of SynRank, if $t_i = t_j$, then $s_T^{(k)}(t_i, t_j)$ is set to be 1; otherwise,

$$s_T^{(k)}(t_i, t_j) = C \cdot \Big[(1-\lambda) \cdot s_{TW}^{(k)}(t_i, t_j) + \lambda \cdot s_{TT}^{(k)}(t_i, t_j)\Big], \tag{1}$$

where $s_{TW}^{(k)}(t_i, t_j)$ and $s_{TT}^{(k)}(t_i, t_j)$ denote content similarity based on *triplet-word* (TW) relation and contextual similarity based on *triplet-triplet* (TT) relation at k-th iteration, respectively. λ is a trade-off parameter, and constant $C \in [0, 1]$ is a damping factor similar as the one in SimRank [18].

Note that for a semantic text information network with $|T|$ triplets, a set of $|T|^2$ SynRank equations needs to be computed. We use $\mathbf{S}_T \in \mathbb{R}^{|T|\times|T|}$ to denote the triplet similarity matrix, where $\mathbf{S}_T(i, j) = s_T(t_i, t_j)$.

Other essential updating formula, including content-based triplet similarity $s_{TW}(t_i, t_j)$, context-based triplet similarity $s_{TT}(t_i, t_j)$, and word similarity S_W, are further introduced as follows.

*1) **Content-based Triplet Similarity**:* Given a pair of triplets t_i and t_j, their content-based similarity, $s_{TW}(t_i, t_j)$, is defined according to the similarity between their content neighbors $N_{TW}(t_i)$ and $N_{TW}(t_j)$.

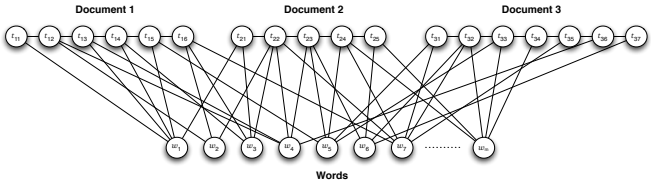


Figure 5. An Example of Semantic Text Information Network on Three Documents and with Context Window Size 1 ($\sigma = 1$).

Example 1 (Similar triplets with similar content):

$t_1 = (\text{S:}\{\text{An earthquake}\}, \text{V:}\{\text{unleashed}\}, \text{O:}\{\text{7.3m waves}\});$
$t_2 = (\text{S:}\{\text{A 8.9 quake}\}, \text{V:}\{\text{unleashed}\}, \text{O:}\{\text{a tsunami wave}\})$

Just like above example, triplets are thought to be similar if they have same/synonymous terms in subjectives, verbs, and objectives, respectively.

Assumption 1: In semantic text information network, two triplet nodes t_i and t_j are said to be content-based similar *if many of their linked words $a \in N_{TW}(t_i)$ and $b \in N_{TW}(t_j)$ are similar.*

Following the assumption, a recursive equation for updating $s_{TW}(t_i, t_j)$ can be derived. If $t_i = t_j$, then $s_{TW}^{(k)}(t_i, t_j) = 1$; otherwise,

$$s_{TW}^{(k)}(t_i, t_j) = \sum_{a \in N_{TW}(t_i)} \sum_{b \in N_{TW}(t_j)} \frac{f_{t_i,a} \cdot f_{t_j,b} \cdot s_W^{(k-1)}(a, b)}{F_{TW}(t_i) F_{TW}(t_j)}, \tag{2}$$

where $f_{t_i,a}$ denotes the occurrence frequency of word a in triplet t_i, and $F_{TW}(t_i)$ denotes total word occurrence in t_i, *i.e.*, $F_{TW}(t_i) = \sum_{a \in N_{TW}(t_i)} f_{t_i,a}$. Here, $s_W(\cdot, \cdot)$ is the similarity between words, which will be introduced in Section IV-A3. We rewrite Equation (2) into matrix form

$$\mathbf{S}_{TW}^{(k)} = \mathbf{D} \cdot \mathbf{S}_W^{(k-1)} \cdot \mathbf{D}^T, \tag{3}$$

where we define matrices $\mathbf{D} \in \mathbb{R}^{|T|\times|W|}$, $\mathbf{S}_{TW} \in \mathbb{R}^{|T|\times|T|}$, and $\mathbf{S}_W \in \mathbb{R}^{|W|\times|W|}$ as $\mathbf{D}(i, j) = f_{t_i,w_j}/F_{TW}(t_i)$, $\mathbf{S}_{TW}(i, j) = s_{TW}(t_i, t_j)$, and $\mathbf{S}_W(i, j) = s_W(w_i, w_j)$, respectively. $|W|$ denotes number of unique words in the corpus. The computational complexity of Equation (3) is $\mathcal{O}(|T|^2 L^2)$, where L is the maximum number of words in a triplet.

*2) **Context-based Triplet Similarity**:* It is not sufficient to fully measure semantic similarity between two triplets by only their contents. In some cases, there could be only a few words inside the two triplets that are same/synonymous. We then propose to evaluate contextual similarity between two triplets, $s_{TT}(t_i, t_j)$, based on their contextual neighbors $N_{TT}(t_i)$ and $N_{TT}(t_j)$.

Example 2 (Similar triplets with similar context):

$t_1 = (\text{S:}\{\text{The first wave}\}, \text{V:}\{\text{hit}\}, \text{O:}\{\text{coasts in Japan}\});$
$t_2 = (\text{S:}\{\text{A wave over 5 feet}\}, \text{V:}\{\text{struck}\}, \text{O:}\{\text{there}\})$

Many articles in our Japan Tsunami news dataset reported not only the tsunamis in Japan, but also the Hawaii's tsunamis. Thus, by merely looking at t_2, we have no idea about where the wave struck. Intuitively, we can seek context of t_1 and t_2 as complementary reference. More specifically, context of a triplet is defined by neighbor triplets within a size σ window in its document (see Figure 5). For example, if the contexts of t_1 and t_2 are both about "Japan coasts," t_1 and t_2 become similar to each other.

Assumption 2: In semantic text information network, two triplet nodes t_i and t_j are said to be context-based similar *if their linked triplets in the context windows $a \in N_{TT}(t_i)$ and $b \in N_{TT}(t_j)$ are similar.*

Remind that for content-based measure of Equation (2), each word in triplet t_i will be compared with each word in t_j. However, in context-based measure, it may be meaningless to compare t_i's neighbor that talks about current fact with t_j's neighbor which can be a quotation. Our method in Equation (4) is to compare each of t_i's neighbor a only with the neighbor of t_j that is *most similar* to a. With above intuition, we derive a recursive equation for $s_{TT}(t_i, t_j)$ by iterating over neighbors of t_i and t_j. At k-th iteration,

$$s_{TT}^{(k)}(t_i,t_j) = \eta(t_i,t_j)\cdot\left(\sum_{a\in N_{TT}(t_i)}\max_{b\in N_{TT}(t_j)} s_T^{(k-1)}(a,b) + \sum_{a\in N_{TT}(t_j)}\max_{b\in N_{TT}(t_i)} s_T^{(k-1)}(a,b)\right), \quad (4)$$

where $\eta(t_i,t_j) = \frac{1}{|N_{TT}(t_i)|+|N_{TT}(t_j)|}$ denotes the number of triplet pairs in summation, which will scale the final similarity score into [0, 1].

The computational complexity for calculating Equation (4) for all triplets is $\mathcal{O}(|T|^2\sigma^2)$. Note that by using some pruning strategy, we actually do no have to compute pairwise similarity for triplets. Due to space limit, we do not discuss the pruning issue in details here.

*3) **Corpus-based Word Similarity**:* Recall that in Equation (2), content-based triplet similarity $\mathbf{S}_{TW}$ is measured based on word similarity $\mathbf{S}_W$. In this section, we will address the problem of how to define a good word similarity $\mathbf{S}_W$.

The most straightforward way to calculate S_W is simply using the identity matrix, which only leverages the fact that a word is only similar to itself. A better strategy might be using some predefined thesaurus such as WordNet [12] to capture more sophisticated similarity structure between words. However, these methods are not able to capture the corpus-specific information. For example, "Japan" and "Tsunami" should be treated more similar in a news corpus about Japan Tsunami than in a corpus about the study of Tsunami's nature. Also, words in semantic text information network are distinguished by different semantic roles (S,V,O) denoting different semantic information, which cannot be well distinguished by knowledge-based approaches.

To address this problem, we propose to adaptively and iteratively update word similarity so that $\mathbf{S}_W$ and $\mathbf{S}_T$ can mutually enhance each other. Intuitively, a good word similarity should generate content-based triplet similarity $\mathbf{S}_{TW} = \mathbf{D}\mathbf{S}_W\mathbf{D}^T$ consistent with triplet similarity $\mathbf{S}_T$.

*Assumption 3: In semantic text information network, corpus-specific information (*i.e.*, context of triplet) is well embedded into word similarity $\mathbf{S}_W$ if content-based triplet similarity $\mathbf{S}_{TW}$ is consistent with triplet similarity $\mathbf{S}_T$.*

Suppose at the k-th iteration of SynRank, the triplet similarity $\mathbf{S}_T^{(k)}$ is derived by Equation (1), based on above assumption, we update $\mathbf{S}_W$ by approximately solving the optimization problem as follows:

$$\mathbf{S}_W^{(k)} = \text{argmin}_{\mathbf{S}_W}\mathcal{L}(\mathbf{S}_W) = \|\mathbf{S}_T^{(k)} - \mathbf{D}\mathbf{S}_W\mathbf{D}^T\|_F^2, \quad (5)$$

where $\|\mathbf{X}\|_F = (\sum_{i,j} X_{ij}^2)^{\frac{1}{2}}$ is matrix Frobenius norm for measuring how consistent the two matrices are. Objective function $\mathcal{L}(\mathbf{S}_W)$ in Equation (5) measures the difference between content-based triplet similarity $\mathbf{D}\mathbf{S}_W\mathbf{D}^T$ and current triplet similarity $\mathbf{S}_T^{(k)}$. By minimizing it, we have the optimal solution as follows:

$$\hat{\mathbf{S}}_W^{(k)} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{S}_T^{(k)}\mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}, \quad (6)$$

whose computational complexity is $\mathcal{O}(|W||T|^2)$.

In order to enforce word similarity to fall in the range of $[0, 1]$, post-processing $\hat{S_W}^{(k)}$ is further performed by

$$s_W^{(k)}(t_i,t_j) = \begin{cases} 1, & \text{if } t_i = t_j. \\ \max(\frac{\hat{\mathbf{S}}_W^{(k+1)}(t_i,t_j)}{\|\hat{\mathbf{S}}_W^{(k)}\|_F^2},\ 0), & \text{otherwise;} \end{cases} \quad (7)$$

If $\mathbf{D}^T\mathbf{D}$ is not invertible, $\mathbf{S}_W$ can be updated approximately based on gradient descent method

$$\mathbf{S}_W^{(k)} = \mathbf{S}_W^{(k-1)} - \alpha\cdot\left\{(\mathbf{D}^T\mathbf{D})\mathbf{S}_W^{(k-1)}(\mathbf{D}^T\mathbf{D}) - \mathbf{D}^T\mathbf{S}_T^{(k)}\mathbf{D}\right\} \quad (8)$$

where α is the step size. In our experimental setting, we have $|T| \gg |W|$, and thus $\mathbf{D}^T\mathbf{D}$ is in practice of full rank and invertible.

B. Algorithm for SynRank

Similar to SimRank, solution to the SynRank equations can be derived by iterations leading to a fixed-point. Starting with $\mathbf{S}_W^{(0)} = \mathbf{I}_{|W|}$ and $\mathbf{S}_T^{(0)} = \mathbf{I}_{|T|}$ as lower bounds of the actual SynRank scores, we successively and alternatively compute $\mathbf{S}_T^{(k)}$ based on $\mathbf{S}_{TW}^{(k-1)}$ and $\mathbf{S}_{TT}^{(k-1)}$ by Equation (1), and $\mathbf{S}_W^{(k)}$ based on $\mathbf{S}_T^{(k)}$ by Equation (6), respectively. Limited by the space of the paper, we don't show the convergence analysis of SynRank here, which follows similar procedure as that in SimRank [18]. Algorithm 1 summarizes the iterative procedure for computing SynRank. Based on each of the similarity computation procedures in previous

Algorithm 1 SynRank

1: **Input**: tuning parameters C and λ, frequency matrix $\mathbf{D}$.
2: Initialize $\mathbf{S}_W^{(0)} = \mathbf{I}_{|W|}$ and $\mathbf{S}_T^{(0)} = \mathbf{I}_{|T|}$.
3: **for** $k = 1 \rightarrow maxIter$ **do**
4: Compute content-based similarity matrix $\mathbf{S}_{TW}^{(k)}$ based on $\mathbf{S}_W^{(k-1)}$ by Equation (2);
5: Compute context-based similarity matrix $\mathbf{S}_{TT}^{(k)}$ based on $\mathbf{S}_T^{(k-1)}$ by Equation (4);
6: Calculate triplet similarity matrix $\mathbf{S}_T^{(k)}$ based on $\mathbf{S}_{TW}^{(k)}$ and $\mathbf{S}_{TT}^{(k)}$ using Equation (1);
7: Update $\mathbf{S}_W^{(k)}$ based on $\mathbf{S}_T^{(k)}$ by Equation (6) and post-process it following Equation (7).
8: **end for**
9: **Output**: Converged matrices $\mathbf{S}_T^{(\infty)}$ and $\mathbf{S}_W^{(\infty)}$.

sections, computational cost for SynRank is $\mathcal{O}(K \cdot |T|^2 |W|)$, where typically $|T| \gg |W|$, and K is the number of iterations needed for SynRank.

C. Triplet Clustering for Representative Frame

Once we compute triplet similarity $\mathbf{S}_T$ by SynRank, various off-the-shelf clustering algorithms (*e.g.* DBSCAN [11] and Affinity Propagation [9]) can then be applied to group these triplets together into clusters. From each cluster, we select one triplet which best summarizes the cluster and use its corresponding frame as the representative frame. Finally, each document is described by the corresponding representative frames derived from all triplet clusters.

More precisely, given triplet similarity matrix $\mathbf{S}_T(i,j)$, and suppose there is totally K frame clusters $\mathcal{C} = \{\mathcal{C}_1, ..., \mathcal{C}_K\}$ derived from the triplet clustering algorithm, we calculate the K representative frames $\{\hat{f}_1, ..., \hat{f}_K\}$ corresponding to the K clusters as follows:

$$\hat{f}_k = \underset{f_i \in C_k}{\operatorname{argmin}} \sum_{f_j \in C_k} \left(\mathbf{S}_T(i,j)\right)^2, \quad k = 1, ..., K. \qquad (9)$$

Each document d is then summarized by a bag of representative frames $d = \{\hat{f}_1, ..., \hat{f}_{K_d}\}$, where K_d is the total number of clusters involved by frames of d.

We choose to use DBSCAN as our triplet clustering algorithm because it has the notion of noise objects, and does not require the number of clusters as an input. Like many other cluster algorithms, DBSCAN have tuning parameters for a given dataset. The two parameters $MinPts$ and Eps [11] are tuned in our experiments so that each news article has at most 100 different frames, and at most 3 same frames. The assumption is that each news article has at most 100 different statements or facts, and should not repeat to mention the same information more than 3 times because they are well-written articles. These constraints can be relaxed for different types of documents like blog posts.

V. Experiments

In this section, we first explain how we obtain three real-world monolingual comparable corpora. Then, we demonstrate the effectiveness of the frame-based document representation by the event tracking analysis of monolingual comparable corpora. Lastly, we evaluate our information network-based solution to similar triplet grouping, *i.e.*, SynRank, quantitatively on human labled datasets, by comparing to different kinds of baselines methods.

Table I
Description of Three Datasets in Our Experiments

Name	Docs	Sentences	Triplets	Words
Japan's Tsunami	22,108	608,723	402,601	13,114,356
London Riot	6,812	186,394	1,390,960	4,022,380
Egypt Revolution	1,759	70,211	140,348	1,493,745

A. Datasets

We use three different comparable corpora, collected from NewsBank[3], as datasets in the experiments. These corpora consist of news articles published by different news agencies about three news events: Japan's Tsunami (started from 3/11/2011), Egypt Revolution (started from 1/24/2011), and London Riot (started from 8/4/2011), respectively. Overview of the news events are provided as follows

- Japan's Tsunami: A massive 8.9-magnitude earthquake shook Japan on March 11, 2011, causing a devastating tsunami to the coast of Japan. Due to the tsunami, the nuclear power plants in Fukushima were damaged, and one of the reactors in the Fukushima No. 1 nuclear plant partially melted down on the following day. As a result, the nuclear accident caused the exposure of nuclear radiation near the plant.
- Egypt Revolution: Protests started on January 25, 2011, and thousands of people began taking to the streets to protest poverty, rampant unemployment, government corruption, and autocratic governance of President Hosni Mubarak, who has ruled the country for thirty years.
- London Riot: Started from August 6, 2011, thousands of people took to the streets in several London boroughs as well as in cities and towns across England. Resulting chaos generated looting, arson, and mass deployment of police. The disturbances began after a protest in Tottenham, following a death of Mark Duggan, a local who was shot dead by police on August 4, 2011.

We searched news articles in NewsBank with keywords: "Japan Tsunami", "Egypt Revolution", and "London Riot", respectively, and collected articles for 11 days after the corresponding start date of each event. The statistics for the three datasets, and the statistics of semantic text information network constructed from them are shown in Table I.

These datasets are available upon request.

B. Effectiveness of Frame-Based Document Representation

Many of the document representation studies [15] evaluate their proposed representation methods via specific applications like similarity search and document clustering. We choose the event tracking task because it is one of the key applications for monolingual comparable corpora analysis, and it is an interesting task for a collection of news articles.

[3] http://www.newsbank.com

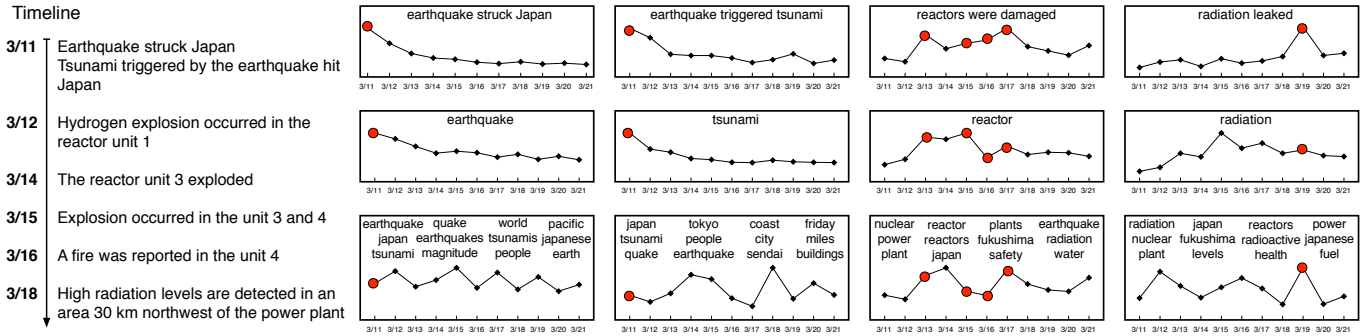


Figure 6. Event Tracking for Japan's Tsunami by Triplets (top), Words (middle), and Topics (bottom)

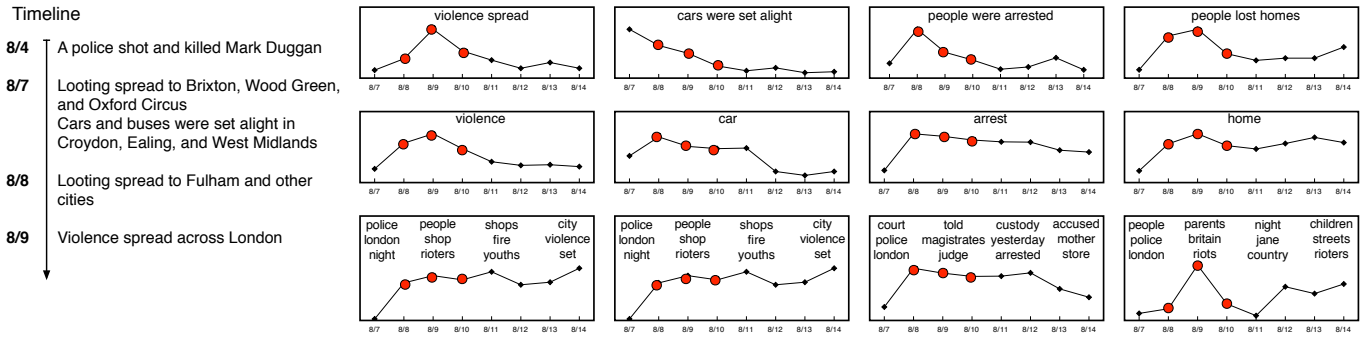


Figure 7. Event Tracking for London's Riot by Frames (top), Words (middle), and Topics (bottom)

We identified four important events from the Japan's Tsunami corpus and London Riot corpus. For each event, we searched for the best triplet clusters, keywords, and topics that describe the event, where topics are from LDA [4] with 20 topics. Then, we plot them by counting their occurrences in the corpus and normalizing by the number of documents in each date. Figure 6 and 7 show the trends in the order of triplets, words, and topics. The bottom row in Figure 6 and 7 show the event tracking by topics. The highest probability words from each topic are listed on each plot.

In Figure 6 and 7, we also indicated the timelines of the two corpora. The red dots in the trend plots indicate the consistent points with the timelines[4]. Thus, those red dots should be higher than other data points.

As shown in the two figures, in general, the frame-based event tracking performs better than the other two baselines. Quantitatively, we can take the average of the rankings of the red dots within the plots as an evaluation measure. For example, in the "reactor" plot, the four red dots ranked 1, 3, 4, and 9. The averages of the rankings of the 19 red dots for frames, words, and topics are respectively 2.33, 2.42, and 3.75. Since lower is better in this measure, the frame-based event tracking is better than the others.

The observation is that if an event cannot be described in a single keyword, it is hard to track events by the keyword. For example, "radiation leaked" cannot be described by a single word. Topic models are designed to model the theme of the words, which are more general concepts than events. It is hard to specify an event using topics. The topic trend plots have many peaks because one topic covers more than one event. These results show that topics are not suitable to specify an event. Increasing the number of topics does not help to specify events. The following survey experiment shows that increasing the number of topics does not make the topics more specific.

[4]Since the timestamps of the news articles are the publication dates, they are off by one from the timeline dates

Table II
TOPIC MODEL EVALUATION SURVEY

The Number of Topics	10	20	50
The Pairwise Agreement	33%	46%	33%

We make multiple choice questions. Each question has a one event description by a sentence and five choices of topics with top ranked words from the word distributions of the topics. Then, participants are asked to pick the most relevant topic for a given event description.

We first generated topics using LDA [4] for the London corpus, and for each of the four events, we selected five most relevant topics by looking at their word distributions and the rankings of several keywords. We repeated this survey for different number of topics (10, 20, and 50). We computed the pairwise agreements for the different number of topics as shown in Table II. The pairwise agreement indirectly measures the specificity of topics for events. When the number of topics is 20, the pairwise agreement is lowest, which means the topics from LDA with 20 topics describe

events better than those from LDA with 10 or 50 topics. Thus, increasing the number of topics does not improve the specificity of topics for events.

C. Effectiveness of SynRank

As we addressed in Section I, it is important to make the document representation space dense by clustering redundant features. In this section, we evaluate our information network-based similarity computation algorithm, SynRank, on labeled datasets. Since better similarity measures lead to better clustering, we demonstrate the effectiveness of SynRank by evaluating semantic similarities using the precision at K measure.

1) Data Labeling: In order to quantitatively conduct empirical evaluations, we generated three labeled datasets (subsets of original ones in Table I). Since getting pairwise labels for large datasets is very expensive, we sampled the datasets as follows:

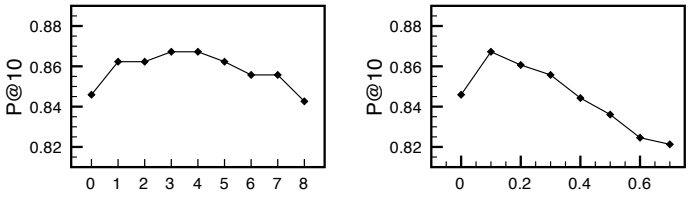
We first chose one specific date from each dataset to increase the chance of having similar documents. Then, we randomly sampled 800 news articles published in the selected date. We performed a labeling procedure as follows: 1) randomly select a triplet t (called a query); 2) from each of our method and our baselines, generate the top 20 similar triplets to t; 3) combine the top 20 triplets of the all methods; 4) label the triplets. Repeating the steps 1-4, we can generate queries with its labeled pairs on which Precision at 20 (P@20) can be calculated.

We asked two participants to label the pairs of triplets with two labels "same" and "different".

After eliminating pairs with different labels from two labelers (the inter-judge agreement rate was 86%), and rejecting queries with all positive and all negative cases, we have 650 queries for Japan corpus, 1,784 queries for London corpus, and 752 queries for Egypt corpus.

2) Quantitative Comparison with Baselines: In this section, we conduct a quantitative comparison between SynRank and other similarity measuring methods to demonstrate the effectiveness of our method on capturing semantic similarity between triplets. Methods based on unstructured text (non-link-based), and based on our semantic text information network are both considered as follows:

- Content (TF-IDF Based Cosine Similarity): This content-based baseline first indexes triplets into tf-idf vectors, and then computes their similarity by cosine similarity measure.
- Corpus (Corpus-Based Distributional Similarity) [21]: This method computes distributional (corpus-based) similarity between words and compose them to get triplet similarity.
- WordNet (Knowledge-Based Similarity) [24]: It computes word similarity based on word synonym information from WordNet and compose them to calculate similarity of triplets.
- SimRank (Homogeneous Link-Based Similarity) [18]: Bipartite SimRank is applied on modified text information network where contextual links are removed since SimRank can only handle homogeneous links.
- P-Rank (Heterogeneous Link-Based Similarity) [34]: P-Rank is applied on our text information network by treating TT and TW relations as in-links and out-links in its framework.



(a) Window Size σ (b) Tradeoff Factor λ

Figure 8. **Parameter Studies of σ and λ by P@10 on Japan's Tsunami Dataset**. Window size σ controls range of contexual information and λ controls the information trade-off between context and content of triplets.

We set all shared parameters between our method and those of baselines the same ($C = 0.8$, $\lambda = 0.1$), and the window size is set as $\sigma = 4$. We ran 20 iterations for SynRank, SimRank and P-Rank. The comparison of SynRank with the other five baseline methods in terms of P@5, P@10 and P@20 are shown in Table III. It shows that SynRank outperforms other methods, demonstrating that leveraging both contextual and content information helps to measure similarities among triplets.

3) Parameter Study: Recall that the two parameters, σ and λ in SynRank formulas control their information gain between context and content. The window size σ controls the range of contextual information, whereas λ in Equation (1) controls the information trade-off between the context and content of triplets. We now study the influence of parameters on SynRank's performance by measuring P@10 on Japan's Tsunami labeled dataset. Parameter study results on other data sets suggest similar trend. In Figure 8(a), SynRank gained best P@10 when $\sigma = 4$, and has relatively low P@10 when σ is small or large. As an extreme case, when $\sigma = 0$ it means no context of triplet is used in the calculation and only content is considered. Low P@10 at small σ indicates that context is useful to enhance similarity measure performance. Also, low P@10 at large σ demonstrates the fact that taking too large range of neighbor triplets as context may introduce too much unrelated and noisy information.

From Figure 8(b) we can examine the appropriate balance between content and contextual information in terms of similarity measuring performance. When $\lambda = 0$, we only make use of content information, which causes low performance gain. On the other hand, when λ goes close to 1, which means only context is leveraged, the performance gain also drops. We found the optimal value for λ is 0.1.

4) Corpus-based Word Similarity: In order to show the performance gain from corpus-based word similarity matrix updating, we plot the curve in Fig. 9 which shows the change of P@10 as SynRank iteration goes, i.e., S_W is updated iteratively. In the Figure, we show the P@10 with and without updating S_W (i.e., fix the word similarity matrix as $S_W = I_{|W|}$). Even though learning the word similarity from corpus leads to worse performance at first, it eventually

Table III
PRECISION EVALUATIONS OF DIFFERENT COMPARED METHODS ON THREE LABELED DATASETS

Method	Japan's Tsunami			London's Riot			Egypt Revolution		
	P@5	P@10	P@20	P@5	P@10	P@20	P@5	P@10	P@20
Content	0.767	0.698	0.653	0.853	0.787	0.719	0.848	0.773	0.689
Corpus	0.756	0.694	0.650	0.859	0.794	0.723	0.853	0.770	0.681
WordNet	0.770	0.711	0.664	0.854	0.791	0.722	0.850	0.767	0.683
SimRank	0.747	0.683	0.641	0.798	0.737	0.695	0.745	0.722	0.679
P-Rank	0.783	0.726	0.681	0.868	0.803	0.728	0.817	0.746	0.677
SynRank	**0.856**	**0.864**	**0.854**	**0.883**	**0.848**	**0.807**	**0.905**	**0.843**	**0.739**

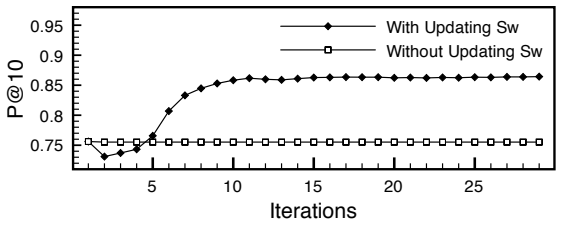


Figure 9. **Performance Gain from Learning Corpus-based Word Similarity Jointly**. P@10 over iterations is plotted, with or without updating the corpus-based word similarity matrix S_W, respectively, on Japan's Tsunami labeled dataset.

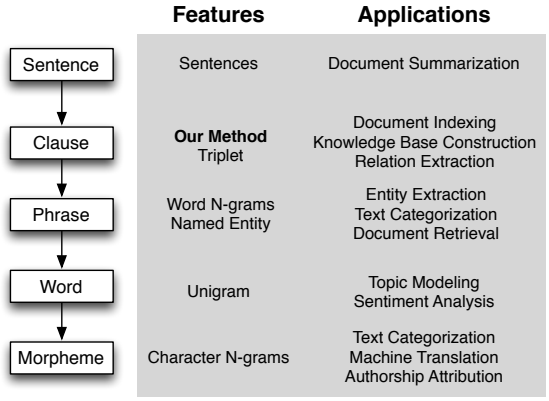


Figure 10. Hierarchy of Grammatical Units and Related Works

enhances it and gets to a stable point, demonstrating that word similarity updating by the corpus bring usefulness.

VI. RELATED WORK

A comprehensive and concise document representation is useful to help the reader quickly obtain an overview of a document or corpus, which is also fundamental for different text processing tasks. In the last few decades, a great number of methods have been proposed for document representation [23], [28].

For its simplicity, *bag-of-words* approaches, e.g., *tf-idf*, are prevalent representations in all kinds of text processing tasks. *Character-level n-gram* representation which considers information from multiple consecutive characters is effectively used for text categorization in [5]. However, this type of approaches are known for the disadvantages that word ordering and grammar structures are missed.

Document summarization [32] can be seen as a sentence-level representation for a single document or multiple documents. Phrase based approaches for document representation model a document using ordered sequences of words, which can be divided into statistical phrase based approaches [8] and syntactic phrase based approaches [10], [28]. However, discouraging results [10] show that using statistical phrase representation degrades some text processing tasks like text categorization and in general syntactic phrase based approaches cannot improve the performance compared to a simple unigram feature based approach.

In addition, specific syntactic and semantic units in documents such as *named entity* [19], POS-tag, word-senses, synonym, and hypernym relations from WordNet [12] have been investigated in different applications [28], [26]. Topic analysis or topic modeling, such as PLSA [16] and LDA [4], which discover main themes that pervade a collection of documents, can be seen as topic-level document representation. Despite of the effectiveness and efficiency of PLSA and LDA, the topics are described by word distributions over the corpus, which sometimes are difficult to interpret and cause ambiguity.

To summarize, different features used for document representation can be understood by a hierarchy of grammatical units in linguistics (See Figure 10), where higher ranked units are composed of and can be analyzed by lower ranked units. In this paper, we explore frame-based features at the clause level from each sentence, which can be viewed as clauses with semantic role labeling.

VII. CONCLUSIONS

In this paper, we study the document representation problem, a fundamental problem for many information retrieval and text mining tasks. We propose to use frames to represent documents, which can capture the semantics of documents and represent documents in a comprehensive and concise way. In order to solve the sparsity issue caused by frame variations, we use an information network approach that treats the corpus as a gigantic semantic information network. Then, a link-based similarity function called SynRank is proposed to capture the similarity between frames in an iterative way. Experiments on real-world datasets have shown the power of the frame-based document representation methods.

VIII. ACKNOWLEDGEMENTS

The work was supported in part by the U.S. Army Research Laboratory under Cooperative Agreement No.

W911NF-09-2-0053 (NS-CTA) and W911NF-11-2-0086 (Cyber-Security), the U.S. Army Research Office under Cooperative Agreement No. W911NF-13-1-0193, DTRA, and U.S. National Science Foundation grants CNS-0931975, IIS-1017362, IIS-1320617, IIS-1354329. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press, 1999.

[2] R. Barzilay and N. Elhadad. Sentence alignment for monolingual comparable corpora. In *EMNLP*, 2003.

[3] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter. Distributional word clusters vs. words for text categorization. *JMLR*, 3:1183–1208, 2003.

[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[5] W. Cavnar, J. Trenkle, et al. N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175.

[6] J. Christensen, S. Soderland, O. Etzioni, et al. Semantic role labeling for open information extraction. In *ACL HLT*, 2010.

[7] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *JMLR*, 2011.

[8] W. Croft and J. Lafferty. *Language modeling for information retrieval*, volume 13. Springer, 2003.

[9] D. Dueck and B. J. Frey. Non-metric affinity propagation for unsupervised image categorization. In *ICCV*, 2007.

[10] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *CIKM*, 1998.

[11] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *SIGKDD*, 1996.

[12] C. Fellbaum. Wordnet. *Theory and Applications of Ontology: Computer Applications*, pages 231–243, 2010.

[13] C. Fillmore. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32, 1976.

[14] D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

[15] X. He, D. Cai, H. Liu, and W.-Y. Ma. Locality preserving indexing for document representation. In *SIGIR*, 2004.

[16] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.

[17] A. Ibrahim, B. Katz, and J. Lin. Extracting structural paraphrases from aligned monolingual corpora. In *ACL*, 2003.

[18] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *SIGKDD*, 2002.

[19] G. Kumaran and J. Allan. Text classification and named entities for new event detection. In *SIGIR*, 2004.

[20] C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 1998.

[21] D. Lin and X. Wu. Phrase clustering for discriminative learning. In *ACL*, 2009.

[22] C. Manning and P. Raghavan. *Introduction to information retrieval*, volume 1. Cambridge university press.

[23] Y. Miao, V. Kešelj, and E. Milios. Document clustering using character n-grams: a comparative evaluation with term-based and word-based clustering. In *CIKM*, 2005.

[24] R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, 2006.

[25] M. Mitra, C. Buckley, A. Singhal, C. Cardie, et al. An analysis of statistical and syntactic phrases. In *RIAO*, 1997.

[26] A. Moschitti and R. Basili. Complex linguistic features for text classification: A comprehensive study. *Advances in Information Retrieval*, pages 181–196, 2004.

[27] M. Palmer, D. Gildea, and P. Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005.

[28] S. Scott and S. Matwin. Feature engineering for text classification. In *ICML*, 1999.

[29] F. Sebastiani. Machine learning in automated text categorization. *CSUR*, 34(1):1–47, 2002.

[30] D. Shen and M. Lapata. Using semantic roles to improve question answering. In *EMNLP*, 2007.

[31] L.-A. Tang, Q. Gu, X. Yu, J. Han, T. La Porta, A. Leung, T. Abdelzaher, and L. Kaplan. Intrumine: mining intruders in untrustworthy data of cyber-physical systems. In *SDM*, 2012.

[32] D. Wang, S. Zhu, T. Li, and Y. Gong. Multi-document summarization using sentence-based topic models. In *ACL-IJCNLP*, 2009.

[33] B. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *VLDB*, 5(6):550–561, 2012.

[34] P. Zhao, J. Han, and Y. Sun. P-rank: a comprehensive structural similarity measure over information networks. In *CIKM*, 2009.