

# Matching Long Text Documents via Graph Convolutional Networks

Bang Liu<sup>1</sup>, Ting Zhang<sup>1</sup>, Di Niu<sup>1</sup>, Jinghong Lin<sup>2</sup>, Kunfeng Lai<sup>2</sup>, Yu Xu<sup>2</sup>

<sup>1</sup>University of Alberta, Edmonton, AB, Canada

<sup>2</sup>Mobile Internet Group, Tencent, Shenzhen, China

## ABSTRACT

Identifying the relationship between two text objects is a core research problem underlying many natural language processing tasks. A wide range of deep learning schemes have been proposed for text matching, mainly focusing on sentence matching, question answering or query document matching. We point out that existing approaches do not perform well at matching *long* documents, which is critical, for example, to AI-based news article understanding and event or story formation. The reason is that these methods either omit or fail to fully utilize complicated semantic structures in long documents. In this paper, we propose a *graph* approach to text matching, especially targeting long document matching, such as identifying whether two news articles report the same event in the real world, possibly with different narratives. We propose the *Concept Interaction Graph* to yield a graph representation for a document, with vertices representing different concepts, each being one or a group of coherent keywords in the document, and with edges representing the interactions between different concepts, connected by sentences in the document. Based on the graph representation of document pairs, we further propose a *Siamese Encoded Graph Convolutional Network* that learns vertex representations through a Siamese neural network and aggregates the vertex features through Graph Convolutional Networks to generate the matching result. Extensive evaluation of the proposed approach based on two labeled news article datasets created at Tencent for its intelligent news products show that the proposed graph approach to long document matching significantly outperforms a wide range of state-of-the-art methods.

## 1 INTRODUCTION

Semantic matching, which aims to model the underlying semantic similarity or relationship among different textual elements such as sentences and documents, has been playing a central role in many Natural Language Processing (NLP) applications, including question answering [32], automatic text summarization [21], top-*k* re-ranking in machine translation [4], as well as information organization [13]. Although a wide range of shallow and deep learning techniques [8, 20, 23, 30] have been proposed to match sentence pairs, question-answer pairs, or query-document pairs, yet up to date, it is still challenging to match a pair of (long) text documents—the rich semantic and syntactic structures in text documents have made it an increasingly difficult task, as document lengths increase. For example, news articles from different news agencies may report a same physical incidence in the real world from different perspectives, possibly with different ways of wording and narratives. Yet, accurately identifying the relationship between long documents is a critical capability expected in the next-generation AI-based news systems, which should automatically organize vast amounts of daily

Internet news articles into *events* and *stories* [13]. This capability, if developed, can largely assist or replace the tedious daily routine work performed by human editors at Internet media organizations.

Traditional approaches to text matching represent text document as vectors in terms of the term frequency-inverse document frequency (TF-IDF), LDA [3] and so forth, and estimate the semantic distances between documents via unsupervised metrics. However, such approaches are not sufficient as they do not take the semantic structures of natural language into consideration. In recent years, a wide variety of deep neural network models based on word-vector representations have been proposed for text matching, e.g., [8, 20, 23, 30]. One category of deep network models [23, 30] takes the word embedding sequences of a pair of text objects as the input, and adopts a Siamese convolutional or recurrent neural network to transform the input into intermediate contextual representations, on which the final scoring is performed. Another category of deep models [8, 20] focuses on the interactions between each word in one text object with each word in the other text object, and aggregates all the pairwise interactions, e.g., using convolutional neural networks (CNNs), to yield a matching score. However, this paper shows that existing deep neural network models do not have a satisfactory performance for matching long documents, since the rich structural information inherent in long documents is not taken into account. In other words, existing methods are mainly matching short text snippets on a word-level or word vector level, omitting the complex way of interactions among sentences, key words or phrases present in any long document.

In this paper, we propose a novel graphical approach to text matching. We argue that the appropriate semantic representation of documents plays a central role in matching long text objects. A successful semantic matching algorithm critically depends on a novel document representation, beyond the linear word-vec representations, that can capture the complex interactions among sentences and concepts in an article. We propose a novel graphical document representation named *Concept Interaction Graph*, which is able to represent a document by an undirected weighted graph, with each vertex denoting a *concept* (i.e., a community of highly coherent keywords) in the document, and the sentences closely related to that concept representing the features of the vertex. Moreover, the edge between a pair of vertices indicates the level of interaction/connection between the two concepts (through sentences). By restructuring documents into a Concept Interaction Graphs, we decompose the semantic focuses in each document into interacting concepts. The task of matching two documents is therefore converted into a graph matching problem.

To compare two Concept Interaction Graphs, we propose a new deep neural network model, named *Siamese Encoded Graph Convolutional Network* (SE-GCN), combining the strengths of Siamese

architectures with Graph Convolutional Network (GCN) [5, 11], an emerging variant of CNN that directly operate on graphs. Specifically, we combine the Concept Interaction Graphs of a pair of documents into one unified graph, by including all vertices, and for each vertex in the unified graph, grouping the features from the two graphs, representing a concatenation of sentence subsets related to this concept from both documents. We introduce a Siamese architecture to encode the concatenated features on each vertex into a match vector. The unified graph obtained this way is subsequently passed through multiple layers of GCN to yield a final matching score. This way, our model factorizes the matching process between two pieces of text into the sub-problems of matching corresponding semantic unit pairs in the two documents.

We performed extensive evaluation on two large datasets of long Chinese news article pairs that were collected from major Internet news providers in China, including Tencent, Sina, WeChat, Sohu, etc., in a two-month period from October 1, 2016 to November 30, 2016, covering diverse topics in the open domain. The datasets also contain ground truth labels that indicate whether a pair of news articles talk about the same event and whether they belong to the same story (a notion larger than events). They are created by the editors and product managers at Tencent for algorithm evaluation purposes.<sup>1</sup> Compared with a wide range of state-of-the-art shallow and deep text matching algorithms that do not take the structural interactions of semantic units into account, our proposed algorithms achieve significant improvements through the use of a graphical representation of documents.

To the best of our knowledge, this is not only the first work that provides a graphical approach to long text document matching, but also the first work that novelly adapts the GCN structure to identify the relationship between a pair of graphs, whereas previously, different GCNs have been mainly used for completing missing attributes/links [5, 11] or for node clustering/classification [7], but all within a single graph, e.g., a knowledge graph, citation network or social network.

The remainder of this paper is organized as follows. Sec. 2 presents our proposed Concept Interaction Graph for document representation. Sec. 3 presents our proposed Siamese Encoded Graph Convolutional Network for text pair matching based on the derived graphical representation. In Sec. 4, we conduct extensive performance evaluations of the proposed models and algorithms based on two large datasets created at Tencent for its intelligent news products. We review the related literature in Sec. 5 and conclude the paper in Sec. 6.

## 2 CONCEPT INTERACTION GRAPH

In this section, we present our *Concept Interaction Graph* (CIG) to represent a document as a *weighted undirected graph*, which decomposes a document into subsets of sentences, focusing on different sub-topics or *concepts*. Such a graph representation proves to be effective at uncovering the underlying attention structure of a long text document such as a news article, which will help with text matching.

<sup>1</sup>As long text document matching is a relatively new problem and the related datasets are lacking, we are currently under the process of publishing these news article datasets to the public for research purposes.

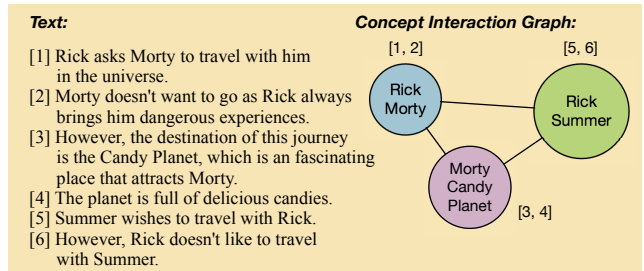


Figure 1: An example to show a piece of text and its corresponding Concept Interaction Graph representation.

We first describe our desired structure for a concept interaction graph before presenting the detailed steps to derive it. Given a document  $\mathcal{D}$ , our objective is to obtain a graph representation  $G_{\mathcal{D}}$  of  $\mathcal{D}$ . Each vertex in  $G_{\mathcal{D}}$  is called a *concept*, which is a community of highly correlated keywords in document  $\mathcal{D}$ . Each sentence in  $\mathcal{D}$  will be assigned onto *one* concept vertex that is the most related to the sentence. We link two vertices by an edge if the similarity (e.g., TF-IDF similarity) of the sentence sets attached to the two vertices, respectively, is above a threshold.

As a toy example, Fig. 1 illustrates how we convert a document into a Concept Interaction Graph. We can extract keywords *Rick*, *Morty*, *Summer*, and *Candy Planet* from the document using standard keyword extraction algorithms [15]. These keywords are further clustered into three concepts, where each concept is a subset of keywords that are highly correlated with each other. After grouping keywords into concepts, we assign each sentence in the document to its most related concept vertex. For example, in Fig. 1, sentences 1 and 2 are mainly talking about the relationship between *Rick* and *Morty*, and are thus assigned to the concept (*Rick*, *Morty*). Other sentences are assigned to sentences in a similar way. The assignment of sentences to concepts naturally leads to multiple sentence subsets. We then connect the concept vertices by weighted edges, where the weight of the edge between a pair of concepts denotes how much the two are related to each other. The edge weights can be determined in various ways, which we will discuss later. This way, we have re-structured the original document into a graph of different focal points, as well as the interaction topology among them.

### 2.1 Construct Concept Interaction Graphs

We now introduce our detailed procedure to transform a document into a desired CIG as described above. The process consists of five steps: 1) document preprocessing, 2) keyword co-occurrence graph construction, 3) concept detection, 4) vertex construction, and 5) edge construction. The entire procedure is shown in Fig. 2.

**Document Preprocessing** Given an input document  $\mathcal{D}$ , our first step is to preprocess the document to acquire its keywords. First, for Chinese text data (which will be used in our evaluation), we need to perform word segmentation using off-the-shelf tools such as Stanford CoreNLP [14]. For English text data, word segmentation is not necessary. Second, we extract named entities from the document. For documents, especially news articles, the named entities

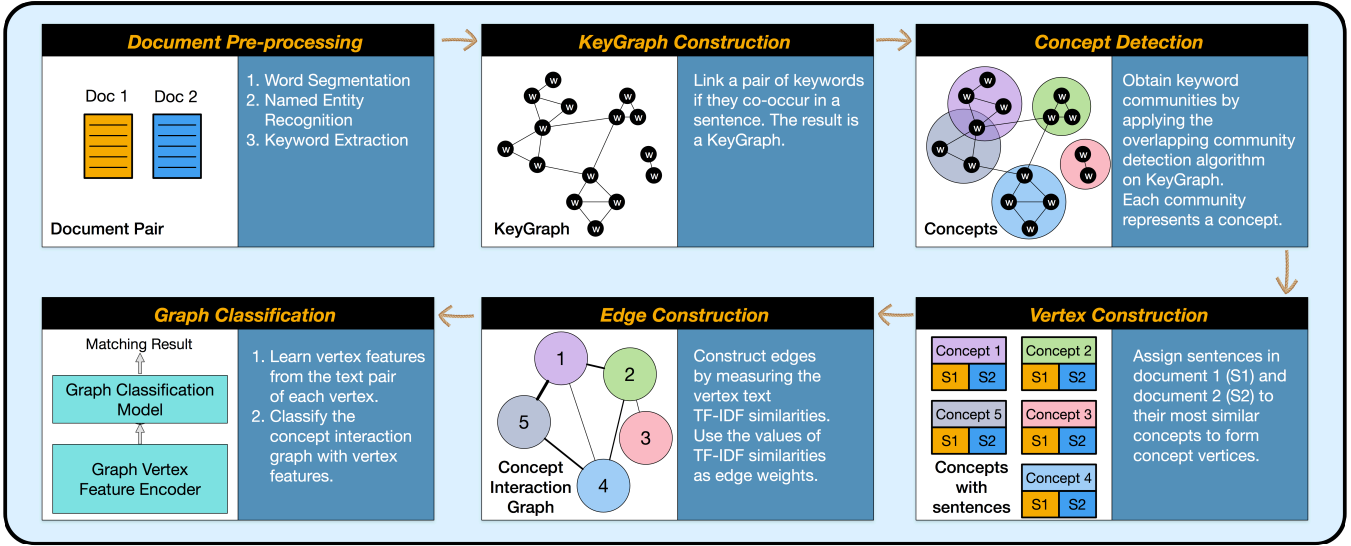


Figure 2: An overview of the procedure of constructing the (joint) *Concept Interaction Graph* (CIG) to match a pair of documents.

are usually critical keywords. Finally, we apply a keyword extraction algorithm to expand the keyword set, as the named entities alone are not enough to cover the main focuses of the document.

To efficiently and accurately extract keywords for Chinese news articles, we have constructed a supervised classifier to decide whether a word is a keyword or not for a document. In particular, we have a document-keywords dataset of over 10,000 documents at Tencent, including over 20,000 positive keyword samples and over 350,000 negative samples. Each word is transformed into a multi-view feature vector and classified by a binary classifier which involves a combined use of Gradient Boosting Decision Tree (GBDT) and Logistic Regression (LR) [13]. For English documents, we can use TextRank [15] to get the keywords of each document. Notice that our proposed graphical representation of documents is not language-dependent and can easily be extended to other languages.

**KeyGraph Construction.** Having extracted the keywords of a document  $\mathcal{D}$ , we construct a keyword co-occurrence graph, called *KeyGraph*, based on the set of keywords. Each keyword is a vertex in the *KeyGraph*. We connect a pair of keywords by an edge if they co-occur in at least one sentence.

**Concept Detection.** The structure of *KeyGraph* reveals the connections between keywords. If a subset of keywords are highly correlated with each other, they will form a densely connected sub-graph in the *KeyGraph*, which we call a *concept*.

Concepts can be extracted by applying community detection algorithms on the constructed *KeyGraph*. Community detection is able to split a *KeyGraph*  $G_{key}$  into a set of communities  $C = \{C_1, C_2, \dots, C_{|C|}\}$ , where each community  $C_i$  contains the keywords for a certain concept. By using overlapping community detection, each keyword may appear in multiple concepts.

A lot of existing algorithms can be utilized for community detection. In our case, the number of concepts in different documents varies a lot, and the number of keywords contained in a constructed *KeyGraph* is rather small. Based on these observations, we utilize the *betweenness centrality score* [26] of edges to measure the strength

of each edge in *KeyGraph* to detect keyword communities. An edge’s betweenness score is defined as the number of shortest paths between all pairs of nodes that pass through it. An edge between two communities is expected to achieve a high betweenness score. Edges with high betweenness score will be removed iteratively to extract separated communities. The iterative splitting process will stop until the number of nodes in each sub-graph is smaller than a predefined threshold, or until the maximum betweenness score of all edges in the sub-graph is smaller than a threshold that depends on the sub-graph’s size. We refer interested readers to [26] for more details on community detection over a *KeyGraph*.

**Vertex Construction.** After we obtain the concepts of a document, the next step is to assign each sentence to its most related concepts. We calculate the cosine similarity between each sentence and a concept, where sentences are represented by TF-IDF vectors. As a concept is a bag of keywords, it can also be represented by a TF-IDF vector. We assign each sentence to the concept which is the most similar to the sentence in terms of the TF-IDF vector and whose similarity score is above a predefined threshold. After this step, sentences in the documents are grouped by concepts. For sentences that do not match any concept in the document, we create a special *dummy vertex* that does not contain any keyword and attach all the unmatched sentences to it.

**Edge Construction.** Given the set of extracted concepts with attached sentences, we further organize these concept vertices into a *weighted undirected graph* to reveal the correlations between different concepts. There are various ways to construct edges between vertices and to calculate edge weights. For example, for each vertex, we can combine the sentences attached to it into a long piece of concatenated text, and calculate the edge weight between any two vertices as the TF-IDF similarity between the two pieces of concatenated text on the two vertices, respectively. We also tried multiple alternative methods for weight calculation, such as counting the number of sentences that contain at least one keyword

from each of the two vertices respectively. Our empirical experience shows that constructing edges by TF-IDF similarity generates a good Concept Interaction Graph for NLP tasks, as the resulted graph is more densely connected compared with the graph weight weights determined by other methods.

Until now, we have transformed an input document into a Concept Interaction Graph. Compared with the original document with a sequential structure, CIG discovers the focal point distribution in the document by detecting all the concepts and grouping sentences according to different concepts. Furthermore, the weighted edges represent the strengths of interactions among these concepts. In the next section, we will show how to use such a graphical representation of documents for text matching purposes.

### 3 A GRAPHICAL APPROACH TO DOCUMENT MATCHING

In this section, we exploit the graphical representation of documents provided by concept interaction graphs, and propose the so-called *Siamese Encoded Graph Convolutional Network* (SE-GCN) for text matching. Fig. 3 illustrates the overall architecture of our proposed model, which is trained end-to-end.

#### 3.1 The Joint CIG for a Pair of Documents

Since our goal is to classify the relationship of a pair of input documents  $\mathcal{D}_A$  and  $\mathcal{D}_B$ , we need a mechanism to merge the two corresponding CIGs  $G_A$  and  $G_B$ , which can be eventually aggregated to a final matching score. One straightforward way is to have a ‘‘Siamese GCN’’, where  $G_A$  is encoded into a contextual vector via multiple layers of graph convolutional networks (GCN), and the same procedure is applied to  $G_B$ . Finally, we can match the two contextual vectors to obtain the matching score. However, this approach does not lead to good performance according to our experiments, as the comparison is only done in the final layer between the short encoded vectors, with too much information lost at the initial GCN layers.

Intuitively speaking, a better approach to utilize the concept interaction graph is to compare the sentence subsets on each vertex, and aggregate such fine-grained comparisons on different vertices, possibly weighted by the interaction topology, to get an overall matching result. To preserve the contrast between  $G_A$  and  $G_B$  on a per-vertex level and let such vertex contrasts propagate through multiple GCN layers, we propose a novel procedure to merge a pair of CIGs.

Specifically, for a pair of input documents  $\mathcal{D}_A$  and  $\mathcal{D}_B$ , we can construct a joint Concept Interaction Graph (joint CIG)  $G_{AB}$  by taking the ‘‘union’’ of the respective two CIGs  $G_A$  and  $G_B$  in the following way:

- Include all the concept vertices from  $G_A$  and  $G_B$  into the joint CIG.
- For each vertex  $v$  in the joint CIG, its associated sentence set is given by the union  $\{\mathcal{S}_A(v), \mathcal{S}_B(v)\}$ , where  $\mathcal{S}_A(v)$  (or  $\mathcal{S}_B(v)$ ) is the set of sentences associated with  $v$  in  $G_A$  (or  $G_B$ ).
- The edge weight  $w_{uv}$  for every pair of vertices  $u$  and  $v$  in the joint CIG  $G_{AB}$  is recalculated based on the TF-IDF similarity

between their respective sentence sets,  $\{\mathcal{S}_A(u), \mathcal{S}_B(u)\}$  and  $\{\mathcal{S}_A(v), \mathcal{S}_B(v)\}$ .

#### 3.2 A Siamese Document Pair Encoder

Given the joint CIG  $G_{AB}$ , our next step is find an appropriate feature vector of a fixed length for each vertex  $v \in G_{AB}$  to express the semantic similarity and divergence between  $\mathcal{S}_A(v)$  and  $\mathcal{S}_B(v)$ , which represents the difference between documents  $\mathcal{D}_A$  and  $\mathcal{D}_B$  on the focal point  $v$ . A natural idea is manually extract various features to compare  $\mathcal{S}_A(v)$  and  $\mathcal{S}_B(v)$ , e.g., in terms of TF-IDF similarity, distance between mean word vectors. However, the performance of such a method is limited and will be highly dependent on feature engineering. To reduce the impact of human judgment in feature engineering, we resort to the power of a neural encoder applied onto every vertex in a distributed manner. As illustrated by Fig. 3 (a), we apply a same Siamese neural network encoder [18] onto each vertex  $v \in G_{AB}$  to convert the word embeddings (e.g., provided by *Word2Vec* [16]) of  $\{\mathcal{S}_A(v), \mathcal{S}_B(v)\}$  into a fixed-sized hidden feature vector  $\mathbf{m}_{AB}(v)$ , which we call the *match vector*.

In particular, the Siamese encoder takes the sequences of word embeddings of  $\mathcal{S}_A(v)$  and  $\mathcal{S}_B(v)$  as two inputs, encode them into two context vectors through the context layers that share weights on both sides, and compare the two context vectors through an aggregation layer to get the match vector  $\mathbf{m}_{AB}(v)$ . The context layer usually contains one or multiple layers of LSTM, bi-directional LSTM (BiLSTM), or CNN with max pooling layers, aiming to capture the contextual information in each text sequence. In a Siamese network, every text sequence is encoded by the same context representation layer. The obtained context vectors are concatenated in the aggregation layer, and can be further transformed by more layers to get a fixed length  $\mathbf{m}_{AB}(v)$ .

In our experiments, the context layer contains a single layer of 1-D CNN that consists of 200 kernels and a max pooling layer. Denote the context vectors of sentences  $\mathcal{S}_A(v)$  and sentences  $\mathcal{S}_B(v)$  as  $\mathbf{c}_A(v)$  and  $\mathbf{c}_B(v)$ . Then, in the aggregation layer, the match vector  $\mathbf{m}_{AB}(v)$  is given by concatenating the element-wise absolute difference and the element-wise multiplication of the two context vectors, i.e.,

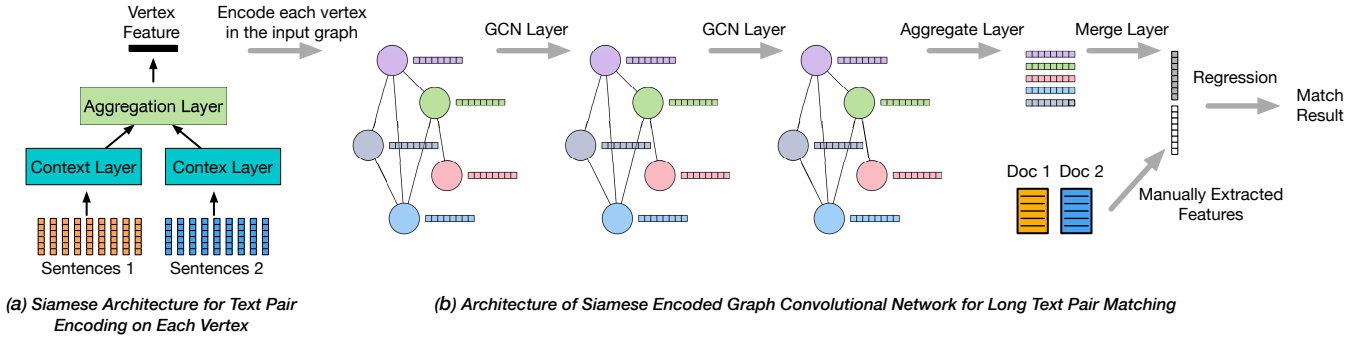
$$\mathbf{m}_{AB}(v) = (|\mathbf{c}_A(v) - \mathbf{c}_B(v)|, \mathbf{c}_A(v) \circ \mathbf{c}_B(v)), \quad (1)$$

where  $\circ$  denotes Hadamard (or element-wise) product.

#### 3.3 Siamese Encoded GCN

Finally, we utilize the ability of Graph Convolutional Network (GCN) [11] to capture the interactions between vertices and get an overall matching score between two documents. GCNs generalize the CNN from low-dimensional regular grids to high-dimensional irregular graph domains. In general, the input to the GCN [11] is a graph  $G = (\mathcal{V}, E)$  with  $N$  vertices  $v_i \in \mathcal{V}$ , and edges  $e_{ij} = (v_i, v_j) \in E$  with weights  $w_{ij}$ . The input also contains a vertex feature matrix denoted by  $X = \{\mathbf{x}_i\}_{i=1}^N$ , where  $\mathbf{x}_i$  is the *feature vector* of vertex  $v_i$ .

For a pair of documents  $\mathcal{D}_A$  and  $\mathcal{D}_B$ , we will input the joint concept interaction graph  $G_{AB}$  (assuming it has  $N$  vertices) with match vectors, as obtained according to the previous subsection, into the GCN, such that  $\mathbf{x}_i = \mathbf{m}_{AB}(v_i)$ , i.e., the match vector obtained for



**Figure 3: An overview of the proposed Siamese Encoded Graph Convolutional Network (SE-GCN) for matching a pair of long text documents. a) The architecture of the Siamese Text Pair Encoder on each vertex of the joint concept interaction graph (CIG) of the two documents, for vertex feature generation. b) The GCN layers to map the initial vertex features in the joint CIG into a final matching score.**

each  $v_i$  from the Siamese encoder will serve as the feature vector for vertex  $v_i$  in GCN.

Now let us briefly describe the GCN propagation layers, as shown in Fig. 3 (b). Interested readers are referred to [11] for details. Denote the weighted adjacency matrix of the graph as  $A \in \mathbb{R}^{N \times N}$  where  $A_{ij} = w_{ij}$ . Let  $D$  be a diagonal matrix such that  $D_{ii} = \sum_j A_{ij}$ . We will utilize a multi-layer GCN with the following layer-wise propagation rule [11]:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}), \quad (2)$$

where  $\tilde{A} = A + I_N$  and  $\tilde{D}$  is a diagonal matrix such that  $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$  are the adjacency matrix and the degree matrix of graph  $G$ , respectively, with added self-connections, and  $I_N$  is the identity matrix.

The input layer to GCN is  $H^{(0)} = X$ , which contains original vertex features, and  $H^{(l)} \in \mathbb{R}^{N \times M_l}$  is the matrix of activation, containing *hidden vectors* of the vertices in the  $l^{\text{th}}$  layer.  $W^{(l)}$  is the trainable weight matrix in the  $l^{\text{th}}$  layer.  $\sigma(\cdot)$  denotes an activation function such as Sigmoid or ReLU. Such a form of propagation rules is motivated by a first-order approximation of localized spectral filters on graphs, and can be considered as differentiable generalization of the Weisfeiler-Lehman algorithm, as described in [11].

In summary, as shown in Fig. 3, the combination of a Siamese encoder applied to each vertex and multiple layers of GCN leads to the proposed Siamese Encoded GCN (SE-GCN), which takes a joint CIG representation  $G_{AB}$  of a pair of documents  $\mathcal{D}_A$  and  $\mathcal{D}_B$  as the input, pass the original sentences  $\{\mathcal{S}_A(v), \mathcal{S}_B(v)\}$  associated with each vertex  $v$  into the same Siamese encoder in a distributed fashion to get the match vector  $\mathbf{m}_{AB}(v)$ . Next, the concept interaction graph  $G_{AB}$ , together with the match vectors  $\mathbf{m}_{AB}(v)$  serving as vertex features, are fed into multiple layers of GCNs. Finally, the hidden vectors in the last GCN layer is merged into a single vector of a fixed length. Note that these hidden vectors of vertices preserve the structural properties of the entire Concept Interaction Graph with minimum information loss. We use the mean of the hidden vectors of all vertices in the last layer as the merged representation, based on which the final matching score is computed. All the components in the entire proposed SE-GCN model can be jointly trained in an *end-to-end* manner with back-propagation.

**Discussion.** To further improve the performance of our model, we can also manually construct a feature vector for the pair of documents in question, and concatenate the final mean vector representation from the GCN with the manual feature vector for classification. In our experiment, we pass such a concatenated vector to a regression layer, such as a multi-layer feed forward neural network, to get the final matching result.

We can see that SE-GCN solves the problem of long text document matching in a “divide-and-conquer” manner. The matching of two documents is divided into the matching of pairs of text snippets (sentence subsets) on each vertex of the constructed Concept Interaction Graph. Then, the distributed vertex matching results are aggregated and merged through graph convolutional network layers. SE-GCN overcomes the limitation of previous text matching algorithms, by extending text representation from a sequential or grid point of view to graphs, and can therefore better capture the rich intrinsic semantic structures in long text objects.

Finally, it is worth noting that our proposed SE-GCN is highly flexible. Different components in the architecture may be replaced by different neural network modules. Besides, it is not limited to text matching problems and can be applied to a variety of natural language processing tasks, especially those related to the modelling of long text objects, such as document classification, sentiment analysis and so on.

## 4 EVALUATION

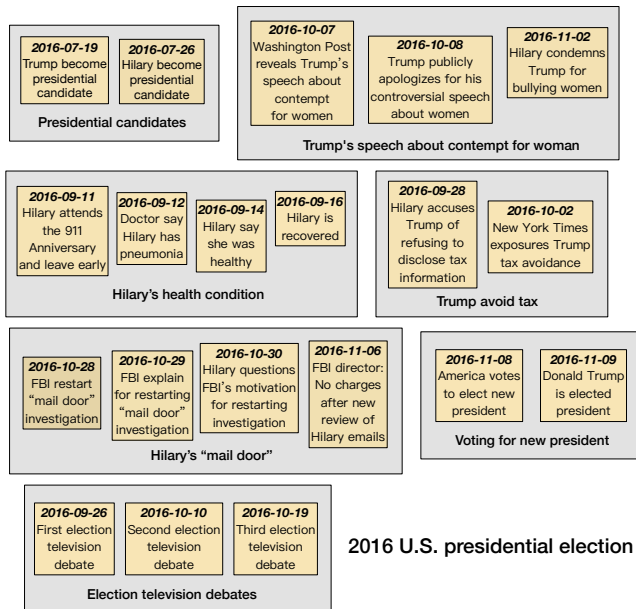
In this section, we evaluate the performance of our proposed SE-GCN model on the document pair matching task. We will first describe the task of semantic relationship classification for news articles, and then introduce two Chinese news datasets we collected specifically for this task at Tencent. After that, to evaluate our model’s efficiency, we will compare our model with a wide variety of existing text matching approaches.

### 4.1 Description of Tasks and Datasets

Most of existing research work on text matching mainly focuses on short text pairs. And there are few research work and publicly available datasets for long document pair matching tasks. However, the problem of matching two documents, such as news articles,

**Table 1: Description of evaluation datasets.**

Dataset	Pos Samples	Neg Samples	Train	Dev	Test
CNSE	12865	16198	17438	5813	5812
CNSS	16887	16616	20102	6701	6700



**Figure 4: The events contained in the story “2016 U.S. presidential election”.**

will be of great value to real-world applications, such as intelligent news systems.

Specifically, we will study the problem of matching a pair of news articles to classify whether they are talking about the same physical event or whether they belong to the same story in the real world. The concepts of *event* and *story* are defined as [13]:

*Definition 4.1. Event:* an event is a set of news documents that contains semantically identical information revolving around a real-world incident. An event always has a specific time of occurrence. It may involve a group of participating persons, organizations or other types of entities, the actions performed by them, and one or several locations.

*Definition 4.2. Story:* a story consists of a set of semantically related or similar events.

To give readers more intuition on what the stories or events look like, here we use an example to clarify the concept of story and event. Fig. 4 shows the events contained in the story *2016 U.S. presidential election*. As we can see, there are multiple sets of sub-events, such as events about *Hillary’s health condition*, *Trump avoid tax*, *Hillary’s “mail door”* and so on, which belong to the same story *2016 U.S. presidential election*. For each event subset, there are multiple events occurred at different time. For example, the event set *Election television debates* contains three events that correspond to the three television debates during the presidential election, respectively. Let us consider the following 4 events under the story

*2016 U.S. presidential election*: 1) *Trump and Hilary’s first television debate*; 2) *Trump and Hilary’s second television debate*; 3) *FBI restarts “mail door” investigation*; 4) *America votes to elect the new president*. Intuitively, these 4 events should have no overlap between them. A news article about *Trump and Hilary’s first television debate* is conceptually separate from *Trump and Hilary’s second television debate*. For news articles, different events from the same story should be clearly distinguishable, because they usually follow the progressing timeline of real-world affairs.

Extracting events and stories accurately from vast news corpora is critical for online news feed apps and search engines to organize news information collected from the Internet and present it to users in sensible forms. The key problem for such kind of applications is classify whether two news articles are talking about the same event or the same story. However, to our best knowledge, we are the first to study this problem. As there is no publicly available dataset for such task, here we propose two datasets: Chinese News Same Event dataset (CNSE), and Chinese News Same Story dataset (CNSS).

The two datasets contain long Chinese news articles that were collected from major Internet news providers in China, including Tencent, Sina, WeChat, Sohu, etc., in a two-month period from October 1, 2016 to November 30, 2016, covering diverse topics in the open domain. For the Chinese News Same Event dataset, it contains 29063 pairs of news articles with labels that represent whether a pair of news articles are talking about the same event. The labels are created by the editors and product managers of Tencent. Similarly, the number of the Chinese News Same Story dataset is 33503, and the labels are representing whether two documents are talking about the same story. For each document in these two datasets, it also has a publication timestamp, and a topic category, such as “Society”, “Entertainment” and so on. Notice that the negative samples in the two datasets are not randomly generated: we select document pairs that contain similar keywords, and filter out samples with TF-IDF similarity lower than a threshold.

Table 1 shows a detailed breakdown of the datasets used in the evaluation. For both of the two datasets, we use 60% of samples as training set, 20% of samples as development set, and the remaining 20% of as test set. We conduct the experiments on the two datasets. We use training sets to train the models, development set to tune the hyper-parameters and each test set is only used once in the final evaluation. The metrics we used to evaluate the performance of our proposed models on the text matching tasks are the accuracy and the F1 score of classification results. For each model, we carry out training for 10 epochs. We then choose the model with the best validation performance to be evaluated on the test set.

## 4.2 Compared Algorithms

In the following, We briefly describe the baseline methods:

- **Support Vector Machine with Manually Extracted Document Pair Features (Feature + SVM):** this is the most classical approach for classification tasks. In this approach, we extract features for a pair of documents, and train a support vector machine to classify the relationship between two documents. The extracted features include: the TF-IDF cosine similarity and the TF cosine similarity between two

**Table 2: Accuracy and F1-score results of different algorithms on CNSE dataset.**

Algorithm	Dev		Test	
	Accuracy	F1-score	Accuracy	F1-score
ARC-I	0.5308	0.4898	0.5384	0.4868
ARC-II	0.5488	0.3833	0.5437	0.3677
DUET	0.5625	0.5237	0.5563	0.5194
DSSM	0.5837	0.6457	0.5808	0.6468
C-DSSM	0.5895	0.4741	0.6017	0.4857
MatchPyramid	0.6560	0.5299	0.6636	0.5401
SVM	0.7566	0.7299	0.7581	0.7361
SE-GCN	<b>0.7800</b>	<b>0.7785</b>	<b>0.7901</b>	<b>0.7893</b>

documents, the TF-IDF cosine similarity and the TF similarity between the first sentence of two documents, the topic categories of the two documents, and the absolute gap value of the publication time of the two documents.

- **Deep Structured Semantic Models (DSSM)** [9]: it utilizes a deep neural network (DNN) to map high-dimensional sparse features into low-dimensional dense features, and calculates the semantic similarity of the text pair.
- **Convolutional Deep Structured Semantic Models (C-DSSM)** [29]: learning low-dimensional semantic vectors for input text by convolutional neural network (CNN).
- **Multiple Positional Semantic Matching (MV-LSTM)** [30]: matching two text with multiple positional text representations, and aggregating interactions between different positional representations to give a matching score.
- **Match by Local and Distributed Representations (DUET)** [17]: matching two text using both a local representation and learned distributed representations.
- **Convolutional Matching Architecture-I (ARC-I)** [8]: encoding text pairs by CNN, and comparing the encoded representations of each text with a multi-layer perceptron (MLP).
- **Convolutional Matching Architecture-II (ARC-II)** [8]: built directly on the interaction space between two text, and model all the possible combinations of them with 1-D and 2-D convolution.
- **MatchPyramid** [20]: calculating pairwise word matching matrix, and modeling text matching as image recognition, by taking the matching matrix as an image.
- **K-NRM** [31]: using a translation matrix to model word-level similarities and a new kernel-pooling technique to extract multi-level match features, and a learning-to-rank layer that combines those features into the final ranking score.

We utilize the implementation of MatchZoo [6] for the evaluation of above deep text matching models.

### 4.3 Performance Analysis

Table 2 and Table 3 compare the performance of different models in terms of classification accuracy and F1 score, based on the Chinese News Same Event dataset and the Chinese News Same Story dataset. We can see that the results of our Siamese Encoded Graph Convolutional Network achieves the best performance on both two datasets in terms of accuracy and F1 score. This can be attributed

**Table 3: Accuracy and F1-score results of different algorithms on CNSS dataset.**

Algorithm	Dev		Test	
	Accuracy	F1-score	Accuracy	F1-score
ARC-I	0.5267	0.5979	0.5010	0.6658
ARC-II	0.4946	0.5144	0.5200	0.5383
K-NRM	0.4952	0.6609	0.5021	0.6642
MV-LSTM	0.4954	0.6574	0.5021	0.6642
DUET	0.5307	0.6125	0.5233	0.6067
DSSM	0.6063	0.7015	0.6109	0.7058
C-DSSM	0.5368	0.5747	0.5296	0.5675
MatchPyramid	0.6213	0.6479	0.6252	0.6456
SVM	0.7715	0.7531	0.7672	0.7484
SE-GCN	<b>0.8138</b>	<b>0.8203</b>	<b>0.8060</b>	<b>0.8122</b>

to the two characteristics of our model. First, the input of long document pairs are re-organized into Concept Interaction Graphs. Therefore, corresponding semantic units in the two documents will be roughly aligned. Second, our model learns the match vector of each aligned semantic unit through a siamese encoder network, and aggregate the match vectors of all units, or concept vertices, via Graph Convolutional Network to take semantic topology structure of two documents into consideration. Therefore, it solves the problem of matching documents in a “divide-and-conquer” manner to cope with the long length of documents, and fully utilize the connections between semantic units to give an overall matching score or label.

Table 2 and Table 3 indicate that the deep text matching models in Matchzoo lead to bad performance in the long document text matching. The main reasons are the following. First, existing deep text matching models are hard to capture meaningful semantic relations between the long document pair. When the input text pairs are long, it is hard to get an appropriate context vector representation to match text pairs. For interaction-focused models, most of the interactions between words in two long documents will be meaningless, therefore it is not easy to extract useful interaction features for further matching steps. Our model effectively solves the above challenges by representing documents as Concept Interaction Graphs to split and align long text pairs, and utilize the semantic structure of long documents through Graph Convolution Network for semantic matching.

Moreover, Fig. 5(a) and Fig. 5(b) show that our SE-GCN performs better than SVM according to ROC and AUC, indicating the higher precision of our model. We also notice that the performance given by the classical “Manual features + SVM” model is relatively not bad compared to other models. Actually that is reasonable, as the extracted features such as the publication time of news articles, topic categories of news articles and so on are quite critical to judge whether two news articles are talking about the same event or story. However, our model provides a method to match a pair of long documents without manually designed features and achieves significant improvement compared to existing deep text matching models. Besides, we can easily incorporate manually designed features into our model by concatenating them with our learned matching vector for two documents.

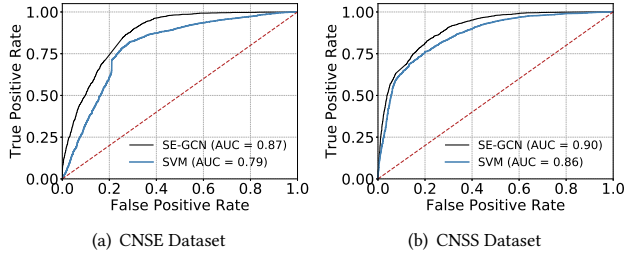


Figure 5: Compare the ROC curves of our model and the SVM baseline model on two datasets.

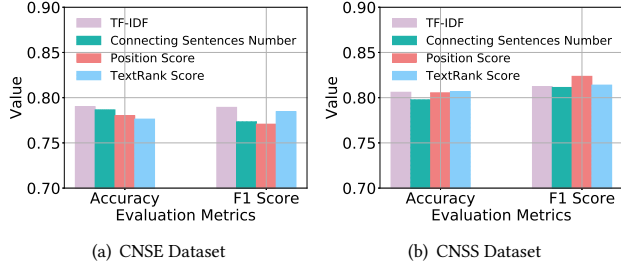


Figure 6: Compare the performance of our model on two datasets using different weight calculation strategies.

Overall, the experimental results demonstrate the superior applicability and generalizability of our proposed model.

**Impact of global feature concatenation.** Compare our model with the version that doesn’t contain global feature concatenation in the last layer. It is not surprising that the performance is worse when we do not feed global feature vectors into our model. However, we can see that our model without global feature concatenation still achieves much better performance than existing deep text matching models. The reason is that existing text matching models are not able to characterize the semantic similarities between long text pairs. Without utilizing the intrinsic semantic structures in long documents, neither representation-focused deep neural models nor interaction-focused models are able to get meaningful comparisons between long documents. In our model, we represent documents by Concept Interaction Graphs so that it is able to align document pairs and match long documents with their semantic structures.

**Impact of different edge weight calculation strategies.** Given a pair of Concept Interaction Graph vertices  $v_i$  with sentence index lists  $S_{iA} = [i_{a1}, i_{a2}, \dots, i_{a|S_{iA}|}]$  and  $S_{iB} = [i_{b1}, i_{b2}, \dots, i_{b|S_{iB}|}]$ , and  $v_j$  with sentence index lists  $S_{jA} = [j_{a1}, j_{a2}, \dots, j_{a|S_{jA}|}]$  and  $S_{jB} = [j_{b1}, j_{b2}, \dots, j_{b|S_{jB}|}]$ . The indices indicate the position of attached sentences in document  $\mathcal{D}_A$  and  $\mathcal{D}_B$ . We tried different strategies to assign weights to the edges:

- **TF-IDF:** for each vertex, concatenating all the sentences from both documents to get a single text snippet, and calculating the TF-IDF similarity between the two text snippets belonging to a pair of vertices.
- **Number of connecting sentences:** counting how many sentences in  $\mathcal{D}_A$  and  $\mathcal{D}_B$  contain at least one keyword in  $v_i$  and one keyword in  $v_j$  (we call them connecting sentences), and use the total number of sentences as weight  $w_{ij}$ .
- **Position of connecting sentences:** counting how many sentences in  $\mathcal{D}_A$  and  $\mathcal{D}_B$  contain at least one keyword in

$v_i$  and one keyword in  $v_j$ . For each sentence, suppose its position in the document is at the  $i_p$ -th paragraph and the  $i_s$ -th sentence in that paragraph. We assign a position score  $score_p$  to it which is calculated as:

$$score_p = e^{-\alpha i_p - \beta i_s}, \quad (3)$$

where  $\alpha$  and  $\beta$  are two hyper parameters ( $\alpha = 0.1$  and  $\beta = 0.3$  for our experiments). We then sum up the position scores of connecting sentences as  $w_{ij}$ .

- **TextRank score of connecting sentence:** similar with above approach, but we use TextRank algorithm to assign scores for sentences. We then sum up the TextRank scores of connecting sentences as  $w_{ij}$ .

Fig. 6 compares the effects of our SE-GAN model on the test sets of Chinese News Same Event dataset and Chinese News Same Story dataset, with different weight calculation strategies. As we can see, for different cases, choosing appropriate edge weight assignment strategies can influence the performance. The TF-IDF strategy achieves slightly better performance than other methods on the event dataset, and the strategies considering sentence positions and sentence TextRank scores can improve the performance over the story dataset. In overall, TF-IDF weight strategy is enough to give us promising performance.

## 5 RELATED WORK

There are mainly two research lines that are highly related to our work: Document Graph Representation and Text Matching.

### 5.1 Document Graph Representation

A various of graph representations have been proposed for document modeling. Based on the different types of graph nodes, a majority of existing works can be generalized into four categories: word graph, text graph, concept graph, and hybrid graph.

For word graphs, the vertices represent different non-stop words in a document, and the edges are constructed based on syntactic analysis [12], co-occurrences [25] or preceding relation [27]. For text graphs, they use sentences, paragraphs or documents as vertices, and establish edges by word co-occurrence, location [15], text similarities [22], or hyperlinks between documents [19].

Concept graphs link terms in a document to real world entities or concepts based on knowledge bases such as DBpedia [1]. After detected concepts in a document as graph vertices, they can be connected by edges based on syntactic/semantic rules. Besides, using these concepts as initial seeds, a concept graph can be expanded by performing a depth-first search along the DBpedia with a maximum depth of two, and adds all outgoing relational edges and concepts along the paths [28].

Hybrid graphs consists of different types of vertices and edges. [24] builds a graph representation of sentences that encodes lexical, syntactic, and semantic relations. [10] extract tokens, syntactic structure nodes, part of speech nodes, and semantic nodes from each sentence, and link them by different types of edges that representing different relationships. [2] combines Frame Semantics and Construction Grammar to construct a Frame Semantic Graph of a sentence.



## 5.2 Text Matching

Most existing works on text matching can be generalized into three categories: unsupervised metrics, representation-focused deep neural models, and interaction-focused deep neural models [6].

Traditional methods represent a text document as vectors of bag of words (BOW), term frequency inverse document frequency (TF-IDF), LDA [3] and so forth, and calculate the distance between vectors. However, they cannot capture the semantic distance and usually cannot achieve good performance.

In recent years, different neural network architectures have been proposed for text pair matching tasks. For representation-focused models, they usually transform the word embedding sequences of text pairs into context representation vectors through a Siamese architectural multi-layer Long Short-Term Memory (LSTM) network or Convolutional Neural Networks (CNN), followed by a fully connected network or score function which gives the matching score or label based on the context representation vectors [23, 30]. For interaction-focused models, they extract the features of all pairwise interactions between words in text pairs, and aggregate the interaction features by deep networks to give a matching result [8, 20]. However, the intrinsic structural properties of long text documents are not fully utilized by these neural models. Therefore, they cannot achieve good performance for long text pair matching.

## 6 CONCLUSION

In this paper, we propose a novel graphical approach to text matching. We propose the *Concept Interaction Graph* to transform one or a pair of documents into a weighted undirected graph, with each vertex representing a concept of tightly correlated keywords and edges indicating their interaction levels. Based on the graph representation of documents, we further propose the *Siamese Encoded Graph Convolutional Network*, a novel deep neural network architecture, which takes graphical representations of documents as the input and matches two documents by learning hidden document representations through the combined use of a distributed Siamese network applied to each vertex in the graph and multiple Graph Convolutional Network layers. We apply our techniques to the task of relationship classification between a pair of long documents, i.e., whether they belong to the same event (or story), based on two newly created Chinese datasets containing news articles. Our extensive experiments show that the proposed approach can achieve significant improvement for long document matching, compared with multiple existing approaches.

## REFERENCES

- [1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. *The semantic web* (2007), 722–735.
- [2] Collin Baker and Michael Ellsworth. 2017. Graph Methods for Multilingual FrameNets. In *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing*. 45–50.
- [3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
- [4] Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics* 19, 2 (1993), 263–311.
- [5] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*. 3844–3852.
- [6] Yixing Fan, Liang Pang, JianPeng Hou, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2017. MatchZoo: A Toolkit for Deep Text Matching. *arXiv preprint arXiv:1707.07270* (2017).
- [7] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation Learning on Graphs: Methods and Applications. *arXiv preprint arXiv:1709.05584* (2017).
- [8] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems*. 2042–2050.
- [9] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2333–2338.
- [10] Chuntao Jiang, Frans Coenen, Robert Sanderson, and Michele Zito. 2010. Text classification using graph mining-based feature extraction. *Knowledge-Based Systems* 23, 4 (2010), 302–308.
- [11] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [12] Jure Leskovec, Marko Grobelnik, and Natasa Milic-Frayling. 2004. Learning substructures of document semantic graphs for document summarization. (2004).
- [13] Bang Liu, Di Niu, Kunfeng Lai, Linglong Kong, and Yu Xu. 2017. Growing Story Forest Online from Massive Breaking News. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 777–785.
- [14] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 55–60.
- [15] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.
- [16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
- [17] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1291–1299.
- [18] Paul Neculoiu, Maarten Versteegh, Mihai Rotaru, and Textkernel BV Amsterdam. 2016. Learning Text Similarity with Siamese Recurrent Networks. *ACL 2016* (2016), 148.
- [19] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web*. Technical Report. Stanford InfoLab.
- [20] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text Matching as Image Recognition. In *AAAI*. 2793–2799.
- [21] Luca Ponzanelli, Andrea Mocchi, and Michele Lanza. 2015. Summarizing complex development artifacts by mining heterogeneous data. In *Proceedings of the 12th Working Conference on Mining Software Repositories*. IEEE Press, 401–405.
- [22] Jan Wira Gotama Putra and Takenobu Tokunaga. 2017. Evaluating text coherence based on semantic similarity graph. In *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing*. 76–85.
- [23] Xipeng Qiu and Xuanjing Huang. 2015. Convolutional Neural Tensor Network Architecture for Community-Based Question Answering. In *IJCAI*. 1305–1311.
- [24] Bryan Rink, Cosmin Adrian Bejan, and Sanda M Harabagiu. 2010. Learning Textual Graph Patterns to Detect Causal Event Relations. In *FLAIRS Conference*.
- [25] François Rousseau and Michalis Vazirgiannis. 2013. Graph-of-word and TW-IDF: new approach to ad hoc IR. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 59–68.
- [26] Hassan Sayyadi and Louiqa Raschid. 2013. A graph analytical approach for topic detection. *ACM Transactions on Internet Technology (TOIT)* 13, 2 (2013), 4.
- [27] Adam Schenker, Mark Last, Horst Bunke, and Abraham Kandel. 2003. Clustering of web documents using a graph model. *SERIES IN MACHINE PERCEPTION AND ARTIFICIAL INTELLIGENCE* 55 (2003), 3–18.
- [28] Michael Schuhmacher and Simone Paolo Ponzetto. 2014. Knowledge-based graph document modeling. In *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 543–552.
- [29] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 373–374.
- [30] Shengxian Wan, Yanyan Lan, Jiafeng Guo, Jun Xu, Liang Pang, and Xueqi Cheng. 2016. A Deep Architecture for Semantic Matching with Multiple Positional Sentence Representations. In *AAAI*, Vol. 16. 2835–2841.
- [31] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 55–64.
- [32] Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632* (2014).