Short Text Similarity Understanding with Word Embeddings

Rutuja Subhash Gadekar[1], Prof. Bhagwan Kurhe[2]

[1]*M.E. Student, SPCOE, Otur, Pune*
[2] *Assistant Professor, SPCOE, Otur, Pune*

Abstract

Understanding short texts is crucial to many applications, but challenges abound. First, short texts do not always observe the syntax of a written language. As a result, traditional natural language processing tools, ranging from part-of-speech tagging to dependency parsing, cannot be easily applied. Second, short texts usually do not contain sufficient statistical signals to support many state-of-the-art approaches for text mining such as topic modeling. Third, short texts are more ambiguous and noisy, and are generated in an enormous volume, which further increases the difficulty to handle them. We argue that semantic knowledge is required in order to better understand short texts. In this work, we build a prototype system for short text understanding which exploits semantic knowledge provided by a well-known knowledgebase and automatically harvested from a web corpus. Our knowledge-intensive approaches disrupt traditional methods for tasks such as text segmentation, part-of-speech tagging, and concept labeling, in the sense that we focus on semantics in all these tasks. Weconduct a comprehensive performance evaluation on real-life data. The results show that semantic knowledge is indispensable for short textunderstanding, and our knowledge-intensive approaches are both effective and efficient in discovering semanticsof short texts.

Keyword:*Topic Model, Short Texts, Word Embeddings*

INTRODUCTION:

Short texts have become a fashionable form of information on the Internet. Examples include web page snippets, news headlines, text advertisements, tweets, status updates, and questions/answers, to name a few. Given the large volume of short texts available, effective and efficient models to discover the latent topics from short texts become fundamental to many applications that require semantic understanding of textual content, such as user interest profiling [4], topic detection [3], comment summarization [1], content characterizing [2], and classification [3].

Conventional topic modeling techniques, e.g., pLSA and LDA, are widely used to infer latent topical structure from text corpus [2, 12]. In these models, each document is represented as a multinomial distribution over topics and each topic is represented as a multinomial distribution over words. Statistical techniques (e.g., Gibbs sampling) are then employed to identify the underlying topic distribution of each document as well as word distribution of each topic, based on the high-order word co-occurrence patterns [29]. These models and their variants have been studied extensively for various tasks in information retrieval and text mining [14, 36, 42]. Despite their great success on many tasks, conventional topic models experience large performance degradation over short texts because of limited word co-occurrence information in short texts. In other words, data sparsity impedes the generation of discriminative document-topic distributions, and the resultant topics are less semantically coherent.

When a human being interprets a piece of text, the understanding is not solely based on its content, but also her background knowledge, e.g., semantic relatedness between words. It is also natural to exploit external lexical knowledge to guide the topic inference over short texts. Existing works in this line largely rely on either external thesauri (e.g., WordNet) or lexical knowledge derived from documents in a specific domain (e.g., product comments) [5–7]. The availability of such knowledge becomes vital for these models. This calls for a more generic model that can be effectively applied to short texts, without the need of manually constructed thesauri, and not limited to external documents in specific domains.

2. RELATED WORK

We review recent advances on learning better topic representations on short texts. Wethen focus on models with word embeddings because our model uses word embeddingsas external knowledge.

Topic Models for Short Texts.

Conventional topic models such as pLSA and LDAare designed to implicitly capture word co-occurrence patterns at document-level, toreveal topic structures. Thus more word co-occurrences would lead to more reliableand better topic inference. Because of the length of each document, conventional topicmodels suffer a lot from the data sparsity problem in short texts, leading to inferiortopic inferences. Earlier studies focus on exploiting external knowledge to help refinethe topic inference of short texts. Phan et al. [2008] propose to infer topic structure ofshort texts by using the learnt latent topics from Wikipedia.

Named entity recognition using an hmm-based chunk tagger

This proposes a Hidden Markov Model (HMM) and an HMM-based chunk tagger, from which a named entity (NE) recognition (NER) system is built to recognize and classify names, times and numerical quantities. Through the HMM, our system is able to apply and integrate four types of internal and external evidences: 1) simple deterministic internal feature of the words, such as capitalization and digitalization; 2) internal semantic feature of important triggers; 3) internal gazetteer feature; 4) external macro context feature. In this way, the NER problem can be resolved effectively.

3. EXISTING SYSTEM

Existing semantic visualization models are not designed for short texts. For example, PLSV represents documents as bags of words, and topic distributions are inferred from word co-occurrences in documents. This assumes sufficiency in word co-occurrences to discover meaningful topics. This may be valid for regular-length documents, but not for short texts, due to the extreme sparsity of words in such documents. Methods based on tf-idf vectors, such as SSE would also suffer, because tf-idf vectors are not efficient for short text analysis. Many words appear only once in a short document, and may appear in only a few documents. Consequently tf and idf are not very distinguishable in short texts.

The Existing system is a generalized framework to understand short texts effectively and efficiently. More specifically, it divide the task of short text understanding into three subtasks: text segmentation, type detection, and concept labeling. It formulate text segmentation as a weighted Maximal Clique problem, and propose a randomized approximation algorithm to maintain accuracy and improve efficiency at the same time. It introduce a Chain Model and a Pairwise Model which combine lexical and semantic features to conduct type detection. They achieve better accuracy than traditional POS taggers on the labeled benchmark. It employ a Weighted Vote algorithm to determine the most appropriate semantics for an instance when ambiguity is detected. The experimental results demonstrate that framework outperforms existing state-of-the-art approaches in the field of short text understanding. It unable to analyze and incorporate the impact of spatial-temporal features into framework for short text understanding.

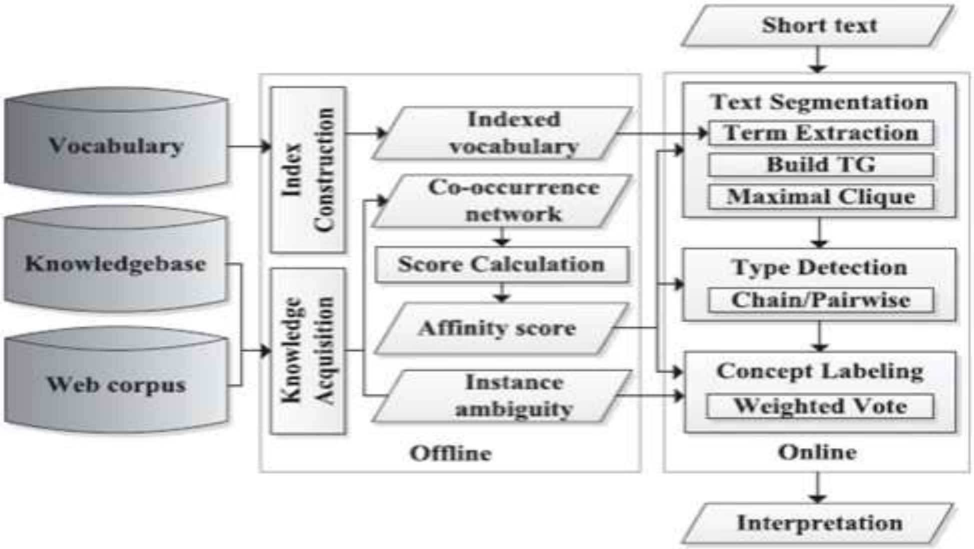


Fig 1. Existing System

4. PROPOSED ARCHITECTURE:

Effective learning of general word semantic relations is now feasible and practical with recent developments in neural network techniques, which have contributed improvements in many tasks in Information Retrieval (IR) and Natural Language Processing (NLP). Specifically, neural network language models, e.g., Continuous Bag-of-Words (CBOW), Continuous Skip-gram model, and Glove model], learn word embeddings (or word vectors) with the aim of fully retaining the contextual information for each word, including both semantic and syntactic relations. Such general word semantic relations can be efficiently learned from a very large text corpus, in any language. In fact, there are many pre-trained word embeddings learned from resources like Wikipedia, Twitter, and Freebase, publicly available on the Web. Because of its good performance, in this paper, we propose to extend the DMM model for topic modeling over short texts by addressing its two limitations.

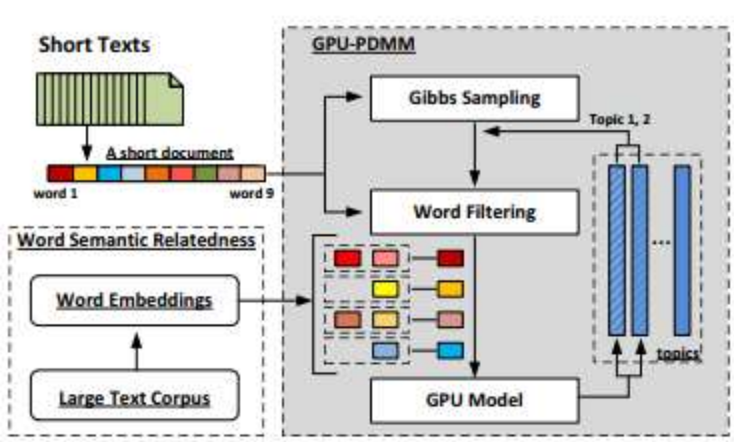


Fig 2. Proposed Architecture

5. POISSON-BASED DIRICHLET MIXTURE MODEL

As its name suggests, the proposed PDMM is an extended DMM model. Given a shortdocument, PDMM first samples a topic number for it based on a Poisson distribution.The specific topics are then sampled based on global topic distribution as well asthe related topic-word distributions. Next, we review the DMM model and detail theproposed PDMM.

6. DIRICHLET MIXTURE MODEL

The Dirichlet Mixture Model is a generative probabilistic model with the assumptionthat a document is generated from a single topic. That is, all the words within a document are generated by the same topic distribution.

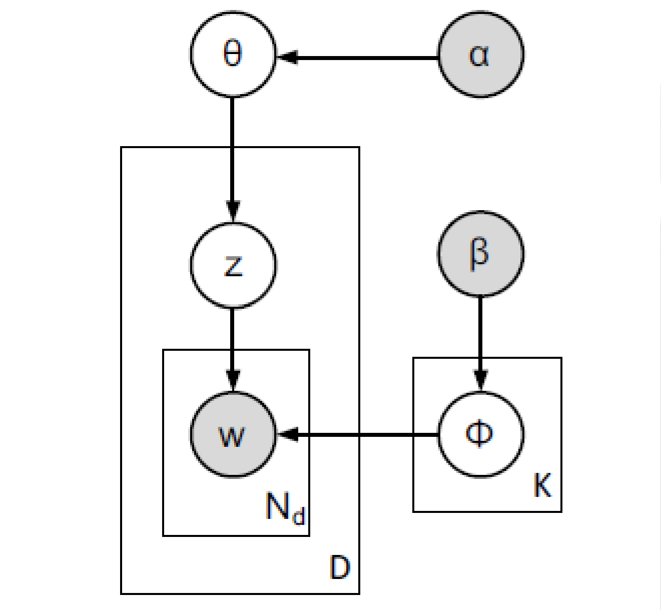
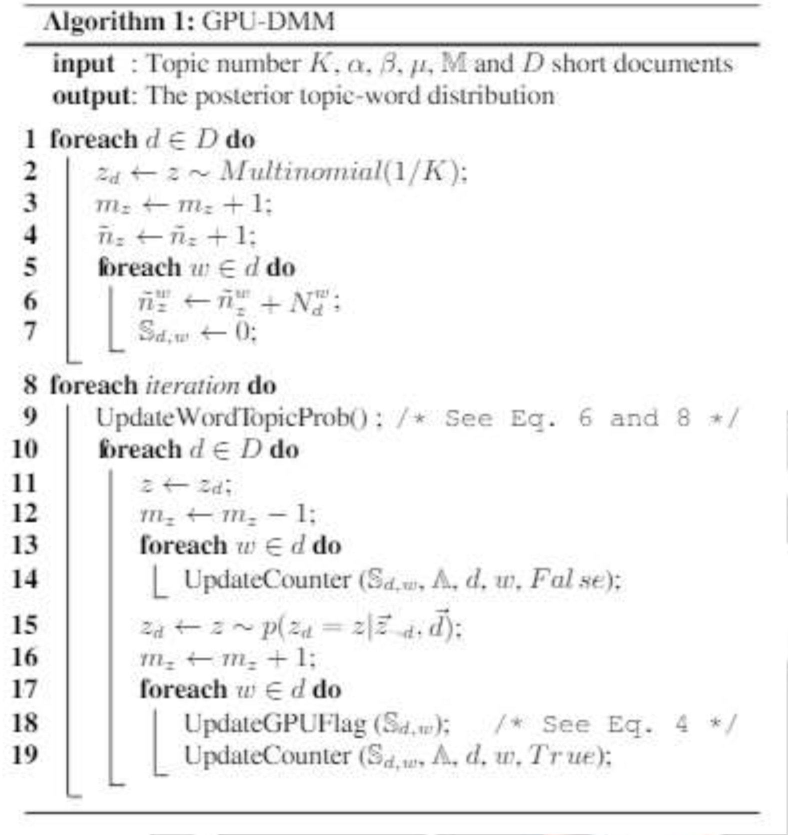


Fig 3. DMM

Given a short text corpus of D documents, with a vocabulary of size V, and K predefinedlatent topics, each document d is associated with one specific topic k. Thenthe N_d words (wd;1, wd;2, …., wd;N_d) in document d are generated by the topic-wordmultinomial distribution p(w|z = k) = $ø_k$ assuming independence of the words. More formally, with Dirichlet priors α and β, the generative process ofDMM is described as follows:

(1) Sample a topic proportionϴ~ Dirichlet(α)
(2) For each topic k € {1,……….,K}
Draw a topic-word distribution $ø_k$~Dirichlet(β)
(3) For each document d € { 1,. . . . . . ,D}
(a) Sample a topic Z^d~ Multinomial(ϴ)
(b) For each word w € {wd,1,. . . . . . wd;Nd}
Sample a word w~ Multinomial(øzd)



```
Algorithm 1: GPU-DMM
  input  : Topic number $K$, $\alpha$, $\beta$, $\mu$, $\mathbb{M}$ and $D$ short documents
  output: The posterior topic-word distribution
1  foreach $d \in D$ do
2      $z_d \leftarrow z \sim Multinomial(1/K)$;
3      $m_z \leftarrow m_z + 1$;
4      $\tilde{n}_z \leftarrow \tilde{n}_z + 1$;
5      foreach $w \in d$ do
6          $\tilde{n}_z^w \leftarrow \tilde{n}_z^w + N_d^w$;
7          $\mathbb{S}_{d,w} \leftarrow 0$;
8  foreach iteration do
9      UpdateWordTopicProb() ; /* See Eq. 6 and 8 */
10     foreach $d \in D$ do
11         $z \leftarrow z_d$;
12         $m_z \leftarrow m_z - 1$;
13         foreach $w \in d$ do
14             UpdateCounter ($\mathbb{S}_{d,w}$, $\mathbb{A}$, $d$, $w$, $False$);
15         $z_d \leftarrow z \sim p(z_d = z|\vec{z}_{\neg d}, \vec{d})$;
16         $m_z \leftarrow m_z + 1$;
17         foreach $w \in d$ do
18             UpdateGPUFlag ($\mathbb{S}_{d,w}$);    /* See Eq. 4 */
19             UpdateCounter ($\mathbb{S}_{d,w}$, $\mathbb{A}$, $d$, $w$, $True$);
```

7. GIBBS SAMPLING

The Gibbs sampling process of DMM is detailed in Algorithm 1. We firstly sampleeach zd;w within document d conditioned on each possible Zd by using (Line13). Then, the likely Zd is sampled conditioned on all the corresponding zd;w values byusing Equation 4 (Line 14).Afterwards, all the values of zd;w are set to the updated Zd'scorresponding values sampled in the first step (Lines 15-19). The posterior distributionis also calculated by using Equation.

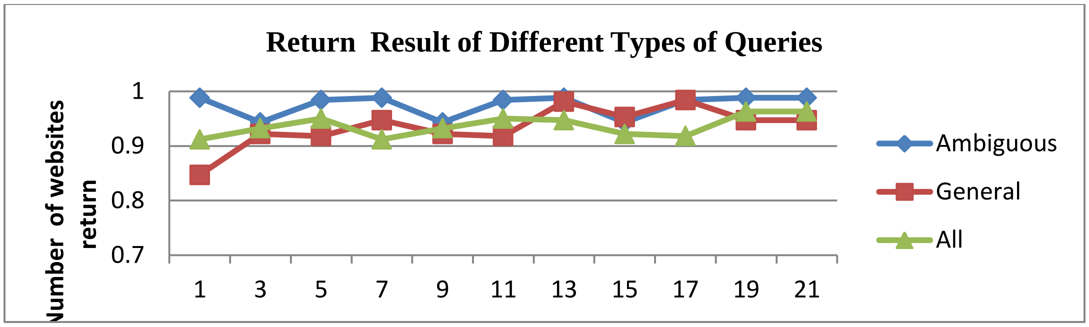
EXPERIMENTAL SETUP

We conducted comprehensive experiments on real-world datasetsto evaluate the performance of our approach to short text understanding.All the algorithms were implemented in JAVA/J2EE, and all theexperiments were conducted on a Windows XP/7 with Pentium IV 2.4 GHz
E5-2690 CPU and 40 GB memory.

RESULT ANALYSIS:

It's a framework for short text understanding whichcan recognize best segmentations, conduct type detection, and eliminate instance ambiguity explicitly based on various typesof context information. Therefore, we manually picked 11 ambiguousterms. Ambiguous in the sense for example "apple" it could be a fruit or a technically it could be a company. So general POS gives you the according to its understanding but to overcome this we proposed this model.

Foreach term derived from a short text, type detection determines thebest typed-term from the set of possible typed-terms. In the caseof "watch free movie", the best typed-terms for "watch", "free",and "movie" are watch as a verb, free as a adjective], and movie as a conceptual respectively.



8. REFERENCES

[1] C. Quirk, C. Brockett, and W. B. Dolan. Monolingual machine translation for paraphrase generation. In EMNLP 2004, 2004.

[2] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In ICML 2008, 2008.

[3] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391–407, 1990.

[5] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain. Neural probabilistic language models. In Innovations in Machine Learning. Springer, 2006

[6] B. Hu, Z. Lu, H. Li, and Q. Chen. Convolutional neural network architectures for matching natural language sentences. In NIPS 2014, 2014

[7] A. Islam and D. Inkpen. Semantic text similarity using corpus-based word similarity and string similarity. TKDD, 2008.

[8] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In Proceedings of the 13st International Conference on Machine Learning, 2014

[9] Y. Kim. Convolutional neural networks for sentence classification. In EMNLP 2014, 2014.

[10] P. Shrestha. Corpus-based methods for short text similarity. Rencontre des Étudiants Chercheurs en Informatique pour le Traitement automatique des Langues, 2011

[11] S. Fernando and M. Stevenson. A semantic similarity approach to paraphrase detection. CLUK 2008, 2008.

[12] R. Ferreira, R. D. Lins, F. Freitas, S. J. Simske, and M. Riss. A new sentence similarity assessment measure based on a three-layer sentence representation. In DocEng 2014, 2014

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