# Understand Short Texts by Harvesting and Analyzing Semantic Knowledge

D.Poorani
UG Scholar
Department of CSE
Kongu Engineering College
*poorani.alai@gmail.com*

M.Rajanaveena
UG Scholar
Department of CSE
Kongu Engineering College
*naveepri1996@gmail.com*

R.Priyadharshini
UG Scholar
Department of CSE
Kongu Engineering College
*priyarajen1996@gmail.com*

Vani Rajasekar
Assistant Professor
Department of CSE
Kongu Engineering College
*vanikecit.cse@kongu.edu*

***ABSTRACT:*** Understanding short texts is crucial to many applications, but challenges abound. Here we focus on short texts which refer to texts with limited context. These short texts are produced including Search queries, Tags, Keywords, Conversation or Social posts and containing limited context. Semantic knowledge is required in order to better understand short texts. 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. 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. We conduct a comprehensive performance evaluation on real-life data. The results show that semantic knowledge is indispensable for short text understanding, and our knowledge- intensive approaches are both effective and efficient in discovering semantics of short texts.

*****

## I. Intoduction:

In this paper, we focus on short texts which refer to texts with limited con-text. Many applications, such as web search and microblogging services etc., need to handle a large amount of short texts. A search task represents an atomic information need of a user in web search. Tasks consist of queries and their reformulations, and identifying tasks is important for search engines since they provide valuable information for determining user satisfaction with search results, predicting user search intent, and suggesting queries to the user. Traditional approaches to identify the tasks, exploit either temporal or lexical features of queries. However, many query refinements are topical, which means that a query and its refinements may not be similar on the lexical level. Furthermore, multiple tasks in the same search session may interleave, which means we cannot simply order the searches by their timestamps and divide the session into multiple tasks. Thus, in order to identify tasks correctly, we need to be able to compare queries at the semantic level. In this section we discuss the following factors:

**Text segmentation** gives a short text, find the most semantically coherent segmentation.
**Type detection** detects each term with best type.**Concept labeling** for each ambiguous instance, re-rank its concept clusters according to the context.

## II. Existing Work:

Understanding text to retrieve required content from a huge database is a critical task and more efforts has been devoted to this field. In current available system if a query is processed by user the entire query is considered to be keyword and processing of entire query will takes place. Therefore processing entire query leads to more time consumption and computation power because the machine learning does not understand which word is important or main key to search content. Short text identification is also difficult task in effective retrieval of data.

Entity linking focuses on retrieving “explicit topics” expressed as probabilistic distributions on an entire knowledgebase. However, categories, “latent topics”, as well as “explicit topics” still have a semantic gap with humans’ mental world.

## III. Proposed WORK:

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.

In the proposed system we have stated the semantic knowledge is required in order to better understand short texts. A prototype system for short text understanding which exploits semantic knowledge provided by a well-known knowledge base and automatically harvested from a web corpus is constructed. 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. We introduce three levels of ambiguity, and propose methods to determine ambiguity level by analyzing the hierarchical and overlapping relationships between concept clusters.

Level 0 refers to instances that most people regard as unambiguous. These instances contain only one sense, such as “dog” (animal) and “california” (state);Level 1 refers to instances that both ambiguous and unambiguous make sense. These instances usually contain more than one senses, but all of these senses are related to some extent, such as “google” (company & search engine) and “nike” (brand & company);

Level 2 refers to instances that most people think as ambiguous. These instances contain two or more unrelated senses, such as “apple” (fruit & company) and “jaguar” (animal & company).

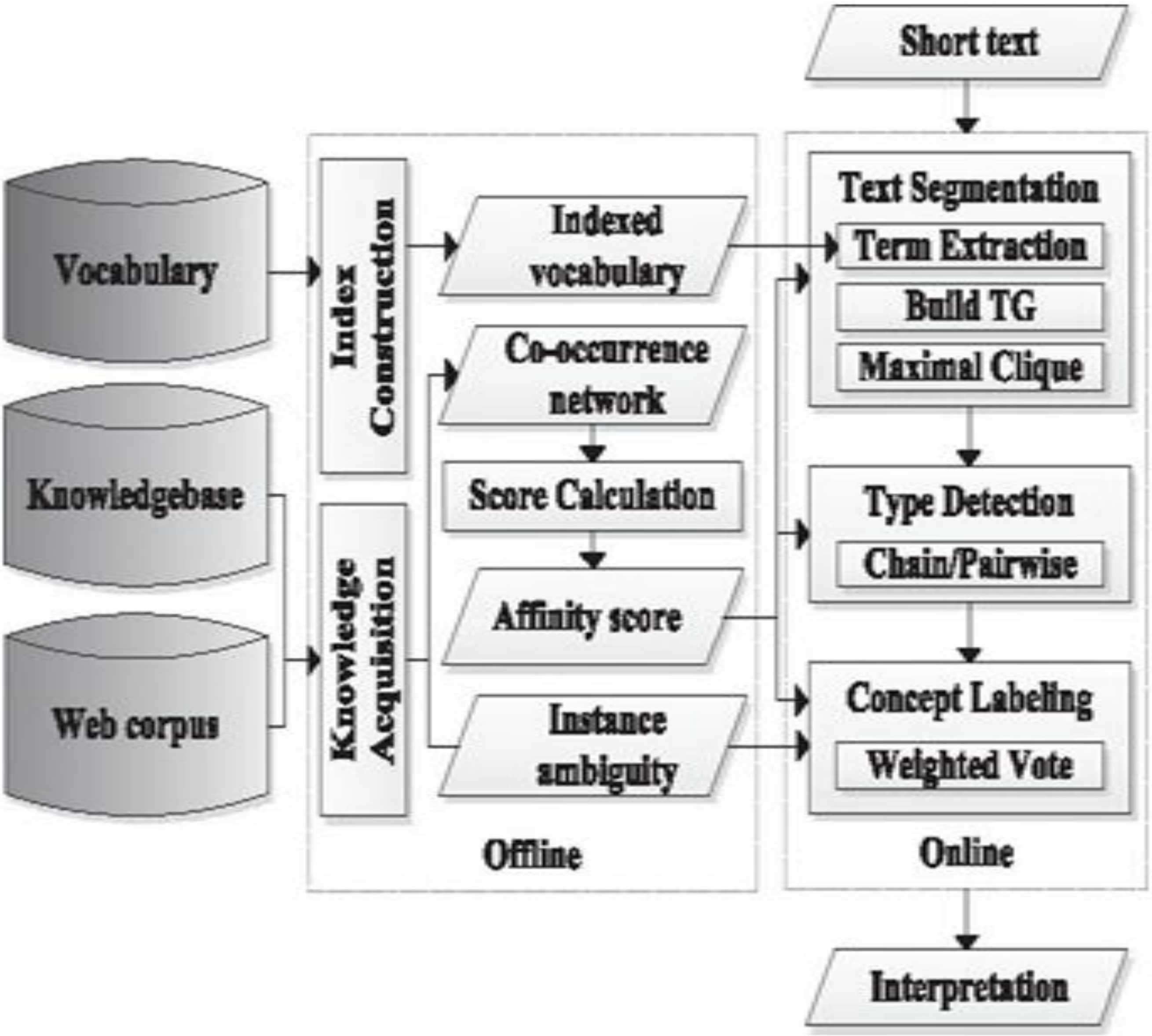

FIG.1: FRAMEWORK OVERVIEW

## Vi. Methodology

### OFFLINE PROCESSING

A prerequisite to short text understanding is the knowledge about semantic relatedness between terms. We describe how we construct the co- occurrence network and quantify semantic coherence. After that, we introduce the indexing strategy to allow for approximate term extraction on the vocabulary, as well as the approach to determine instance ambiguity.

### ONLINE PROCESSING

There are basically three tasks in online processing of short texts, namely text segmentation, type detection, and concept labeling.

### A. Text segmentation

Divide a short text into a collection of terms contained in a vocabulary (e.g., "book dis-neyland hotel california" is segmented as fbookdisneyland hotel californiag);

### B. Type Detection

Determine the types of terms and recognize instances (e.g., "disneyland" and "california" are recognized as instances, while "book" is a verb and "hotel" a concept);

### C. Concept Labeling

Infer the concept of each instance (e.g., "disneyland" and "california" refer t the concept theme park and state respectively). Overall, three concepts are detected from short text "book disneyland hotel california" using this strategy, namely theme park, hotel, and state in FIG: 2.

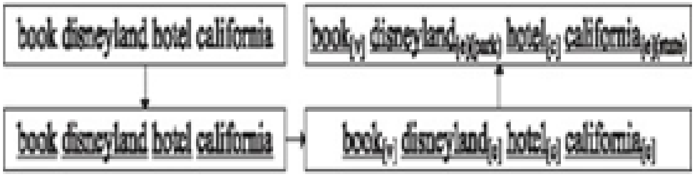
book disneyland hotel california
book(v) disneyland(e)(park) hotel(c) california(e)(state)
book disneyland hotel california
book(v) disneyland(e) hotel(c) california(e)

FIG.2

## IV. Conclusion:

In this work, we propose a generalized framework to understand short texts effectively and efficiently. More specifically, we divide the task of short text understanding into three subtasks: text segmentation, type detection, and concept labeling. The experimental results demonstrate that our proposed framework outperforms existing state-of-the-art approaches in the field of short text understanding. As a future work, we attempt to analyze and incorporate the impact of spatial-temporal features into our framework for short text understanding.

## REFERENCES

[1] A. McCallum and W. Li, "Early results for named entity recogni-tion with conditional random fields, feature induction and web-enhanced lexicons," in Proc. 7th Conf. Natural Language Learn., 2003, pp. 188–191.

[2] G. Zhou and J. Su, "Named entity recognition using an hmm-based chunk tagger," in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics, 2002, pp. 473–480.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[4] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in Proc. 20th Conf. Uncertainty Artif. Intell., 2004, pp. 487–494.

[5] R. Mihalcea and A. Csomai, "Wikify! linking documents to ency-clopedic knowledge," in Proc. 16th ACM Conf. Inf. Knowl. Manage., 2007, pp. 233– 242.

[6] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti, "Collective annotation of wikipedia entities in web text," in Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2009,
pp. 457–466.

[7] X. Han, L. Sun, and J. Zhao, "Collective entity linking in web text: A graph- based method," in Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2011, pp. 765–774.

[8] W. Shen, J. Wang, P. Luo, and M. Wang, "Linden: Linking named entities with knowledge base via semantic knowledge," in Proc. 21st Int. Conf. World Wide Web, 2012, pp. 449–458.

[9] G. L. Murphy, The Big Book of Concepts. Cambridge, MA, USA: MIT press, 2004.

[10] S. Klein and R. F. Simmons, "A computational approach to gram-matical coding of english words," J. ACM, vol. 10, no. 3, pp. 334– 347, 1963.