|| Volume 3 || Issue 7 || July 2018 || ISO 3297:2007 Certified ISSN (Online) 2456-3293

OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

UNDERSTANDING SHORT TEXT AND EXTRACTION BY USING SEMANTIC KNOWLEDGE

Miss Karishma k. Pardeshi[1],Prof. M. R. Bendre[2]

Department of Computer Engineering, Pravara Rural Engineering College, Loni

Pardeshikarishma14494@gmail.com[1],mininath.bendre@gmail.com[2]

***Abstract*: Seeing short messages is essential to numerous applications, however challenges proliferate. In the first place, short messages don't generally watch the grammar of a composed dialect. Therefore, conventional regular dialect handling apparatuses, extending from grammatical feature labelling to reliance parsing, can't be effectively connected. Second, short messages as a rule don't contain adequate factual signs to help many best in class approaches for content mining, for example, subject demonstrating. Third, short messages are more uncertain and loud, and are produced in a gigantic volume, which additionally expands the trouble to deal with them. We contend that semantic information is required with a specific end goal to better see short messages. In this work, we assemble a model framework for short content understanding which abuses semantic learning gave by an outstanding learning base and consequently reaped from a web corpus. Our insight escalated approaches disturb conventional techniques for undertakings, for example, content division, grammatical feature labelling, and idea naming, as in we concentrate on semantics in every one of these assignments. We direct a far reaching execution assessment on genuine information. The outcomes demonstrate that semantic information is irreplaceable for short content comprehension, and our insight escalated approaches are both compelling and proficient in finding semantics of short messages.**

Keywords: Short text understanding, text segmentation, type detection, concept labelling, semantic knowledge.

I INTRODUCTION

Information explosion highlights the need for machines to better understand natural language texts. In this paper, we focus on short texts which refer to texts with limited context. Many applications, such as web search and micro blogging services etc., need to handle a large amount of short texts. Obviously, a better understanding of short texts will bring tremendous value. One of the most important tasks of text understanding is to discover hidden semantics from texts. Many efforts have been devoted to this field. For instance, named entity recognition (NER) [1] [2] locates named entities in a text and classifies them into predefined categories such as persons, organizations, locations, etc. Topic models [3] [4] attempt to recognize "latent topics", which are represented as probabilistic distributions on words, from a text. Entity linking [5] [6] [7] [8] [9] [10] [11] 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. As stated in Psychologist Gregory Murphy's highly acclaimed book [12], "concepts are the glue that holds our mental world together". Therefore, we define short text understanding as to detect concepts mentioned in a short text.

II LITERATURE REVIEW

Name ambiguity problem has raised urgent demands for efficient, high-quality named entity disambiguation methods. In recent years, the increasing availability of large-scale, rich semantic knowledge sources (such as Wikipedia and WordNet) creates new opportunities to enhance the named entity disambiguation by developing algorithms which can exploit these knowledge sources at best. The problem is that these knowledge sources are heterogeneous and most of the semantic knowledge within them is embedded in complex structures, such as graphs and networks. This paper proposes a knowledge-based method, called Structural Semantic Relatedness (SSR), which can enhance the named entity disambiguation by capturing and leveraging the structural semantic knowledge in multiple knowledge sources.

Empirical results show that, in comparison with the classical BOW based methods and social network based methods, our method can significantly improve the disambiguation performance by respectively 8.7% and 14.7%. [9]

Entity Linking (EL) is the task of linking name mentions in Web text with their referent entities in a knowledge base. Traditional EL methods usually link name mentions in a document by assuming them to be independent. However, there is often additional interdependence between different EL decisions, i.e., the entities in the same document should be semantically related to each other. In these cases, Collective Entity Linking, in which the name mentions in the same document are linked jointly by exploiting the interdependence between them, can improve the entity linking accuracy. This paper proposes a graph-based collective EL method, which can model and exploit the global interdependence between different EL decisions. Specifically, we first propose a graph based representation, called Referent Graph, which can model the global interdependence between different EL decisions. Then we propose a collective inference algorithm, which can jointly infer the referent entities of all name mentions by exploiting the interdependence captured in Referent Graph. The key benefit of our method comes from: 1) The global interdependence model of EL decisions; 2) The purely collective nature of the inference algorithm, in which evidence for related EL decisions can be reinforced into high-probability decisions. Experimental results show that our method can achieve significant performance improvement over the traditional EL methods.[10]

Integrating the extracted facts with an existing knowledge base has raised an urgent need to address the problem of entity linking. Specifically, entity linking is the task to link the entity mention in text with the corresponding real world entity in the existing knowledge base. However, this task is challenging due to name ambiguity, textual inconsistency, and lack of world knowledge in the knowledge base. Several methods have been proposed to tackle this problem, but they are largely based on the co-occurrence statistics of terms between the text around the entity mention and the document associated with the entity. In this paper, we propose LINDEN1, a novel framework to link named entities in text with a knowledge base unifying Wikipedia and WordNet, by leveraging the rich semantic knowledge embedded in the Wikipedia and the taxonomy of the knowledge base. We extensively evaluate the performance of our proposed LINDEN over two public data sets and empirical results show that LINDEN significantly outperforms the state-of-the-art methods in terms of accuracy. [11]

Many private and/or public organizations have been reported to create and monitor targeted Twitter streams to collect and understand users' opinions about the organizations. Targeted Twitter stream is usually constructed by filtering tweets with user-defined selection criteria (e.g., tweets published by users from a selected region, or tweets that match one or more predefined keywords). Targeted Twitter stream is then monitored to collect and understand users' opinions about the organizations. There is an emerging need for early crisis detection and response with such target stream. Such applications require a good named entity recognition (NER) system for Twitter, which is able to automatically discover emerging named entities that is potentially linked to the crisis. In this paper, we present a novel 2-step unsupervised NER system for targeted Twitter stream, called TwiNER. In the first step, it leverages on the global context obtained from Wikipedia and Web N-Gram corpus to partition tweets into valid segments (phrases) using a dynamic programming algorithm. Each such tweet segment is a candidate named entity. It is observed that the named entities in the targeted stream usually exhibit a gregarious property, due to the way the targeted stream is constructed. In the second step, TwiNER constructs a random walk model to exploit the gregarious property in the local context derived from the Twitter stream. The highly-ranked segments have a higher chance of being true named entities. We evaluated TwiNER on two sets of real-life tweets simulating two targeted streams. Evaluated using labeled ground truth, TwiNER achieves comparable performance as with conventional approaches in both streams. Various settings of TwiNER have also been examined to verify our global context + local context combo idea. [12]

Microblog platforms such as Twitter are being increasingly adopted by Web users, yielding an important source of data for web search and mining applications. Tasks such as Named Entity Recognition are at the core of many of these applications, but the effectiveness of existing tools is seriously compromised when applied to Twitter data, since messages are terse, poorly worded and posted in many different languages. In this paper, we briefly describe a novel NER approach, called FS-NER (Filter Stream Named Entity Recognition) to deal with Twitter data, and present the results of a preliminary performance evaluation conducted to assess it in the context of the Concept Extraction Challenge proposed by the 2013 Workshop on Making Sense of Microposts - MSM2013. FS-NER is characterized by the use of filters that process unlabeled Twitter messages, being much more practical than existing supervised CRF-based approaches. Such filters can be combined either in sequence or in parallel in a flexible way. Our results show that, despite the simplicity of the filters used, our approach outperformed the baseline with improvements of 4.9% on average, while being much faster. [13]

We designed and implemented Tagme, a system that is able to efficiently and judiciously augment a plain-text with pertinent hyperlinks to Wikipedia pages. The specialty of Tagme with respect to known systems [5, 8] is that it may annotate texts which are short and poorly composed, such as snippets of search-engine results, tweets, news, etc.. This annotation is extremely informative, so any task that is currently addressed using the bag-of-words paradigm could benefit from using this annotation to draw upon (the millions of) Wikipedia pages and their inter-relations. [14]

Most text mining tasks, including clustering and topic detection, are based on statistical methods that treat text as bags of words. Semantics in the text is largely ignored in the mining process, and mining results often have low interpretability. One particular challenge faced by such approaches lies in short text understanding, as short texts lack enough content from which statistical conclusions can be drawn easily. In this paper, we improve text understanding by using a probabilistic knowledgebase that is as rich as our mental world in terms of the concepts (of worldly facts) it contains. We then develop a Bayesian inference mechanism to conceptualize words and short text. We conducted comprehensive experiments on conceptualizing textual terms, and clustering short pieces of text such as Twitter messages. Compared to purely statistical methods such as latent semantic topic modeling or methods that use existing knowledge bases (e.g., WordNet, Freebase and Wikipedia), our approach brings significant improvements in short text understanding as reflected by the clustering accuracy. [15]

Conceptualization seeks to map a short text (i.e., a word or a phrase) to a set of concepts as a mechanism of understanding text. Most of prior research in conceptualization uses human-crafted knowledge bases that map instances to concepts. Such approaches to conceptualization have the limitation that the mappings are not context sensitive. To overcome this limitation, we propose a framework in which we harness the power of a probabilistic topic model which inherently captures the semantic relations between words. By combining latent Dirichlet allocation, a widely used topic model with Probase, a large-scale probabilistic knowledge base, we develop a corpus-based framework for context-dependent conceptualization. Through this simple but powerful framework, we improve conceptualization and enable a wide range of applications that rely on semantic understanding of short texts, including frame element prediction, word similarity in context, ad-query similarity, and query similarity. [16]

III SYSTEM ARCHITECTURE

Fig. 1 illustrates our framework for short text understanding. In the offline part, we construct index on the entire vocabulary and acquire knowledge from web corpus and existing knowledge bases. Then, we pre-calculate semantic coherence between terms which will be used for online short text understanding. In the online part, we perform text segmentation, type detection, and concept labeling, and generate a semantically coherent interpretation for a given short text.

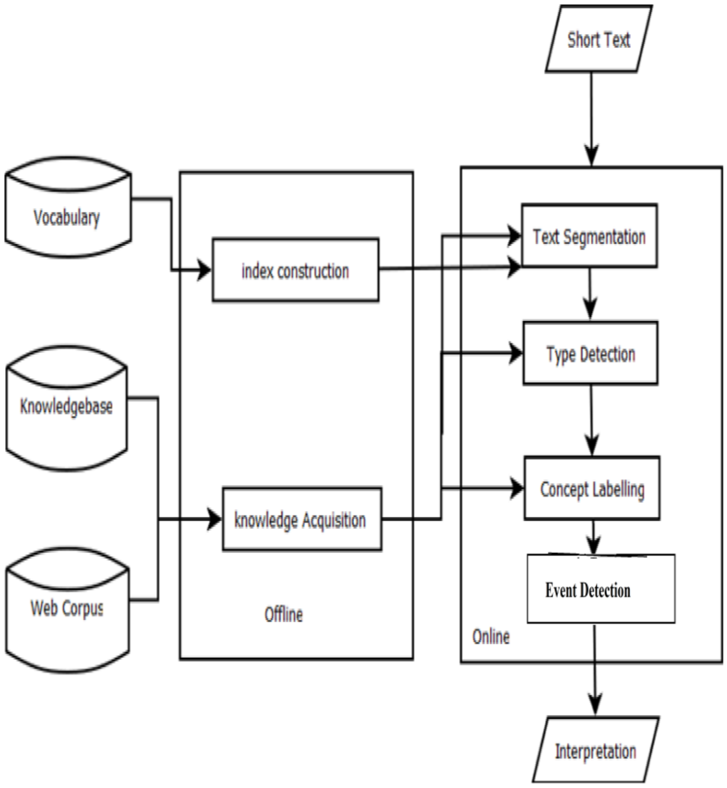


Figure1. System Architecture

Fig. 1 illustrates our framework for short text understanding. In the offline part, we construct index on the entire vocabulary and acquire knowledge from web corpus and existing knowledge bases. Then, we pre-calculate semantic coherence between terms which will be used for online short text understanding. In the online part, we perform text segmentation, type detection, and concept labeling, and generate a semantically coherent interpretation for a given short text.

Methodology

Indexing of vocabulary and knowledge acquisition.

Approximate term extraction aims to locate substrings in a text which are similar to terms contained in a predefined vocabulary. To quantify the similarity between two strings, many similarity functions have been proposed including token-based similarity functions (e.g., jaccard coefficient) and character-based similarity functions (e.g., edit distance). Due to the prevalence of misspellings in short texts, we use edit distance as our similarity function to facilitate approximate term extraction.

- **Text Segmentation.**

We can recognize all possible terms from a short text using the tried-based framework described. But the real question is how to obtain a coherent segmentation from the set of terms. We use two examples to illustrate our approach of text segmentation. Obviously, fapril in paris lyricsg is a better segmentation of "april in paris lyrics" than fapril paris lyricsg, since "lyrics" is more semantically related to songs than two months or cities. Similarly, fvacation april parisg is a better segmentation of "vacation april in paris", due to higher coherence among "vacation", "april", and "paris" than that between "vacation" and "april in paris".

- **Type Detection.**

Recall that we can obtain the collection of typed-terms for a term directly from the vocabulary. For example, term "watch" appears in instance-list, concept-list, as well as verb-list of our vocabulary, thus the possible typed-terms of "watch" are watch[c]; watch[e]; watch[v]g. Analogously, the collections of possible typed-terms for "free" and "movie" are free[ad j]; free[v]g and movie[c]; movie[e]g respectively, as illustrated. For each term derived from a short text, type detection determines the best typed-term from the set of possible typed-terms. In the case of "watch free movie", the best typed-terms for "watch", "free", and "movie" are watch[v], free[ad j], and movie[c] respectively.

Concept Labeling.

The most important task in concept labeling is instance disambiguation, which is the process of eliminating inappropriate semantics behind an ambiguous instance. We accomplish this task by re-ranking concept clusters of the target instance based on context information in a short text (i.e., remaining terms), so that the most appropriate concept clusters are ranked higher and the incorrect ones lower. Our intuition is that a concept cluster is appropriate for an instance only if it is a common semantics of that instance and it achieves support from surrounding context at the same time. Take "hotel california eagles" as an example. Although both animal and music band are popular semantics of "eagles", only music band is semantically coherent (i.e., frequently co-occurs) with the concept song and thus can be kept as the final semantics of "eagles".

IV APPLICATIONS

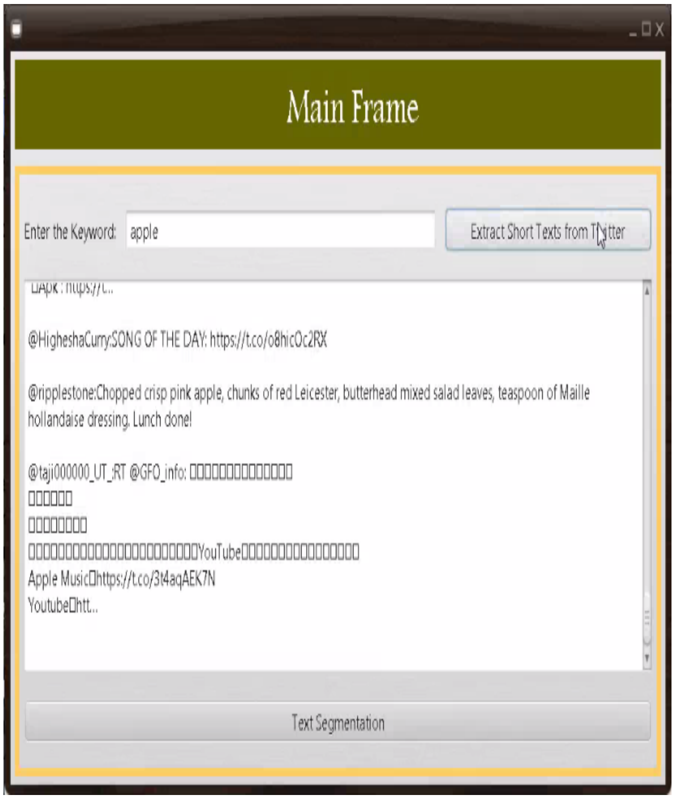
- Use of social media increases rapidly in society also short text increased accordingly.
- The labeling of concept or text which is received on social site is crucial.

V ALGORITHM

Algorithm 1 Maximal Clique by Monte Carlo (MaxCMC)

```
Input:
    $G = (V, E); W(E) = \{w(e) | e \in E\}$
Output:
    $G' = (V', E'); s(G')$
$V' = \emptyset; E' = \emptyset$
while $E \neq \emptyset$ do
    randomly select $e = (u, v)$ from $E$ with probability proportional to its weight
    $V' = V' \cup \{u, v\}; E' = E' \cup \{e\}$
    $V = V - \{u, v\}; E = E - \{e\}$
    for each $t \in V$ do
        if $e' = (u, t) \notin E$ or $e' = (v, t) \notin E$ then
            $V = V - \{t\}$
            remove edges linked to $t$ from $E$: $E = E - \{e' = (t, *)\}$
        end if
    end for
end while
calculate average edge weight: $s(G') = \frac{\sum_{e \in E'} w(e)}{|E'|}$
```

VI RESULTS



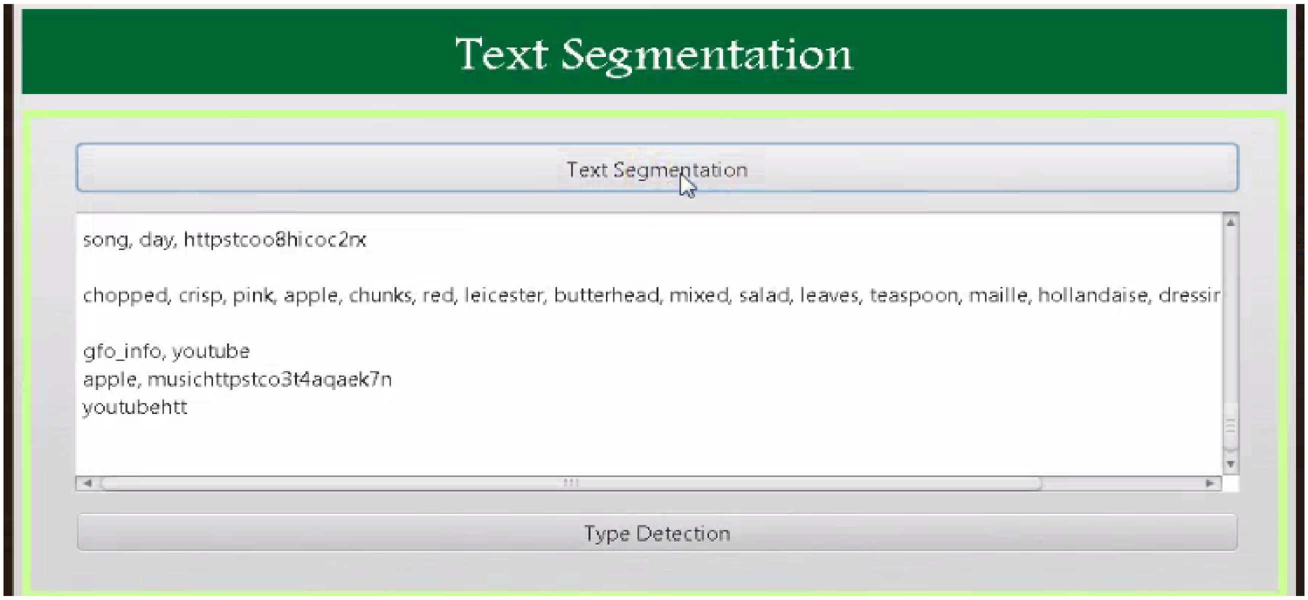
Enter the Keyword: apple

□Apk : https://t...

@HigheshaCurry:SONG OF THE DAY: https://t.co/o8hicOc2RX

@ripplestone:Chopped crisp pink apple, chunks of red Leicester, butterhead mixed salad leaves, teaspoon of Maille hollandaise dressing. Lunch done!

@taji000000_UT_:RT @GFO_info: □□□□□□□□□□□□□□
□□□□□□
□□□□□□□□
□□□□□□□□□□□□□□□□□□□□□□□□□YouTube□□□□□□□□□□□□□□□□□
Apple Music□https://t.co/3t4aqAEK7N
Youtube□htt...

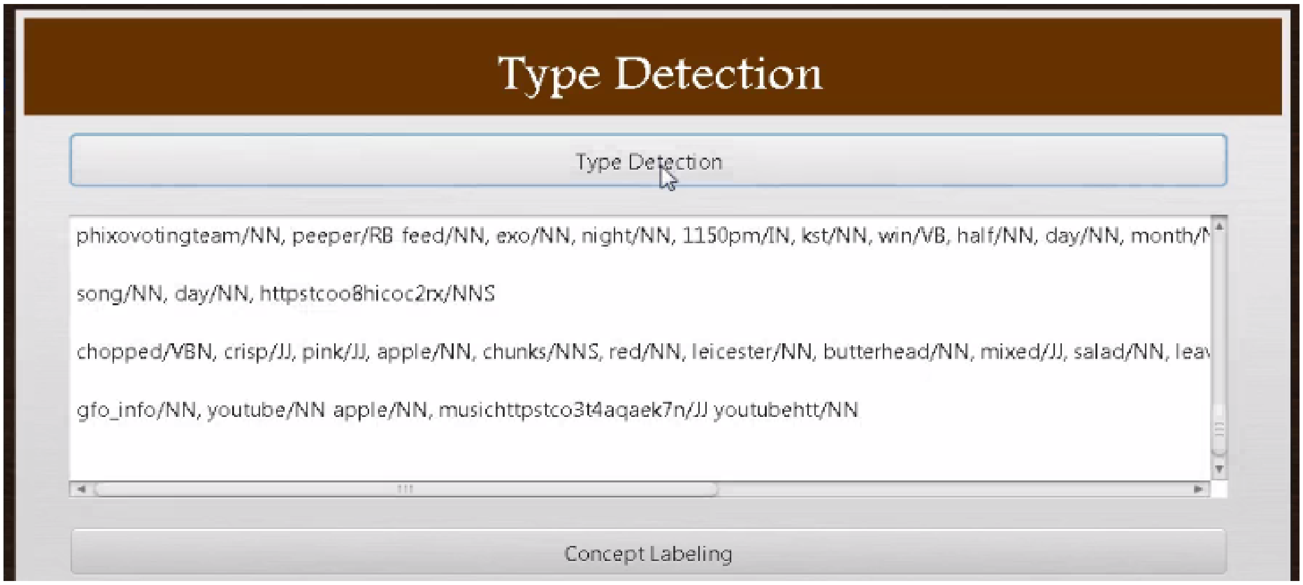


Text Segmentation

song, day, httpstcoo8hicoc2rx

chopped, crisp, pink, apple, chunks, red, leicester, butterhead, mixed, salad, leaves, teaspoon, maille, hollandaise, dressir

gfo_info, youtube
apple, musichttpstco3t4aqaek7n
youtubehtt



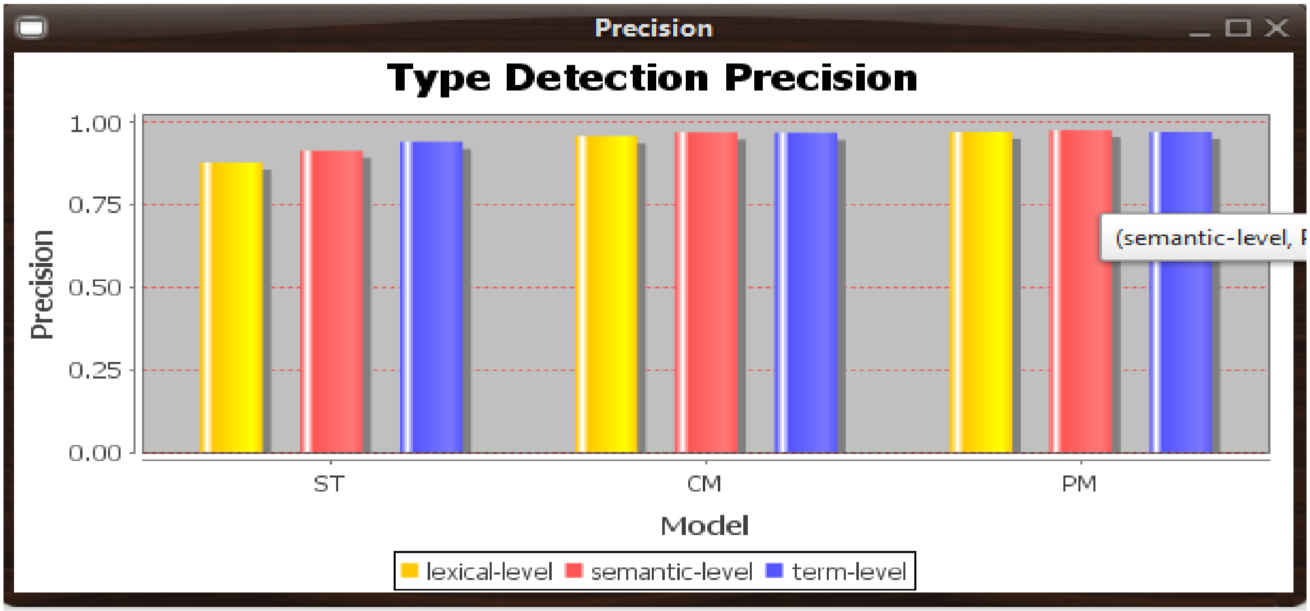
Type Detection

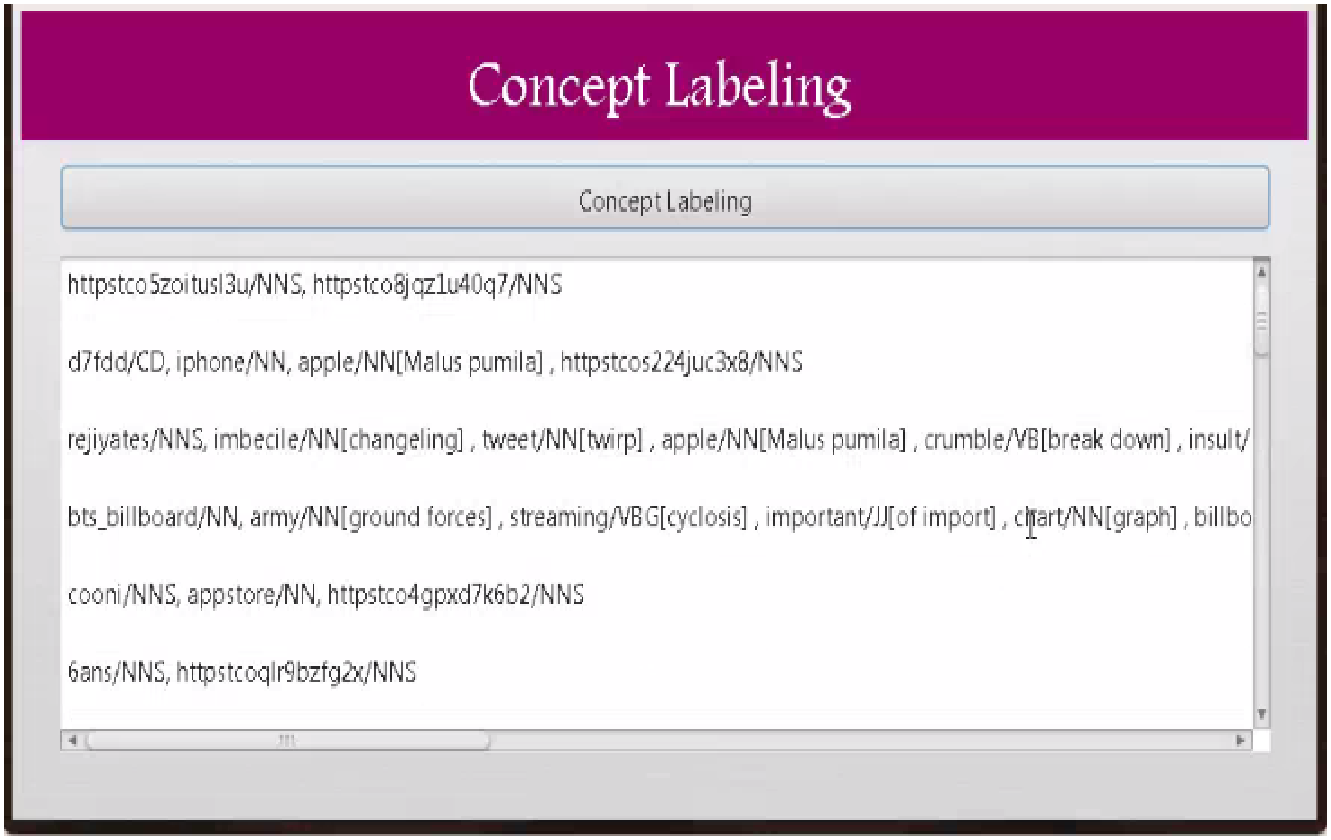
phixovotingteam/NN, peeper/RB feed/NN, exo/NN, night/NN, 1150pm/IN, kst/NN, win/VB, half/NN, day/NN, month/N

song/NN, day/NN, httpstcoo8hicoc2rx/NNS

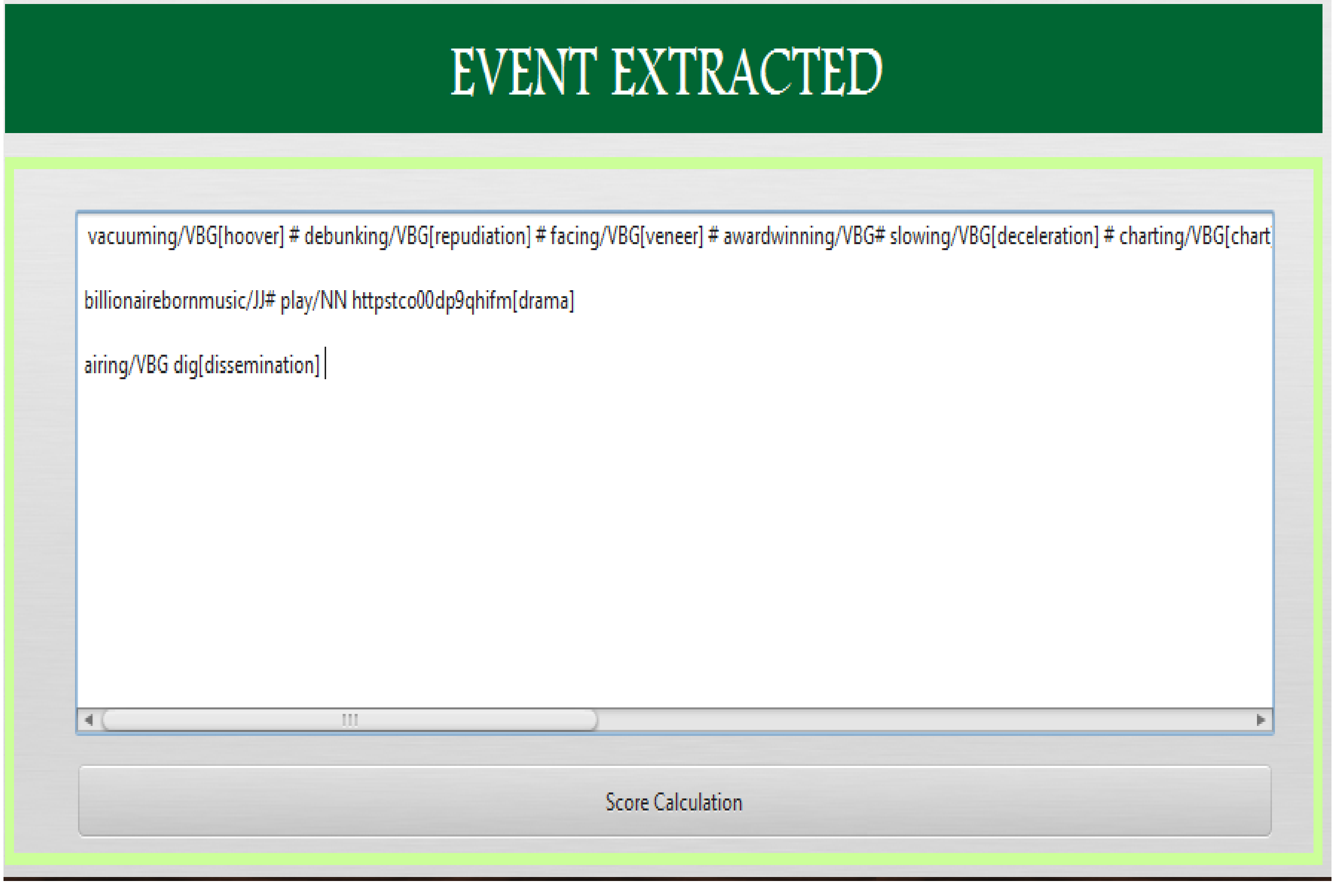
chopped/VBN, crisp/JJ, pink/JJ, apple/NN, chunks/NNS, red/NN, leicester/NN, butterhead/NN, mixed/JJ, salad/NN, leav

gfo_info/NN, youtube/NN apple/NN, musichttpstco3t4aqaek7n/JJ youtubehtt/NN

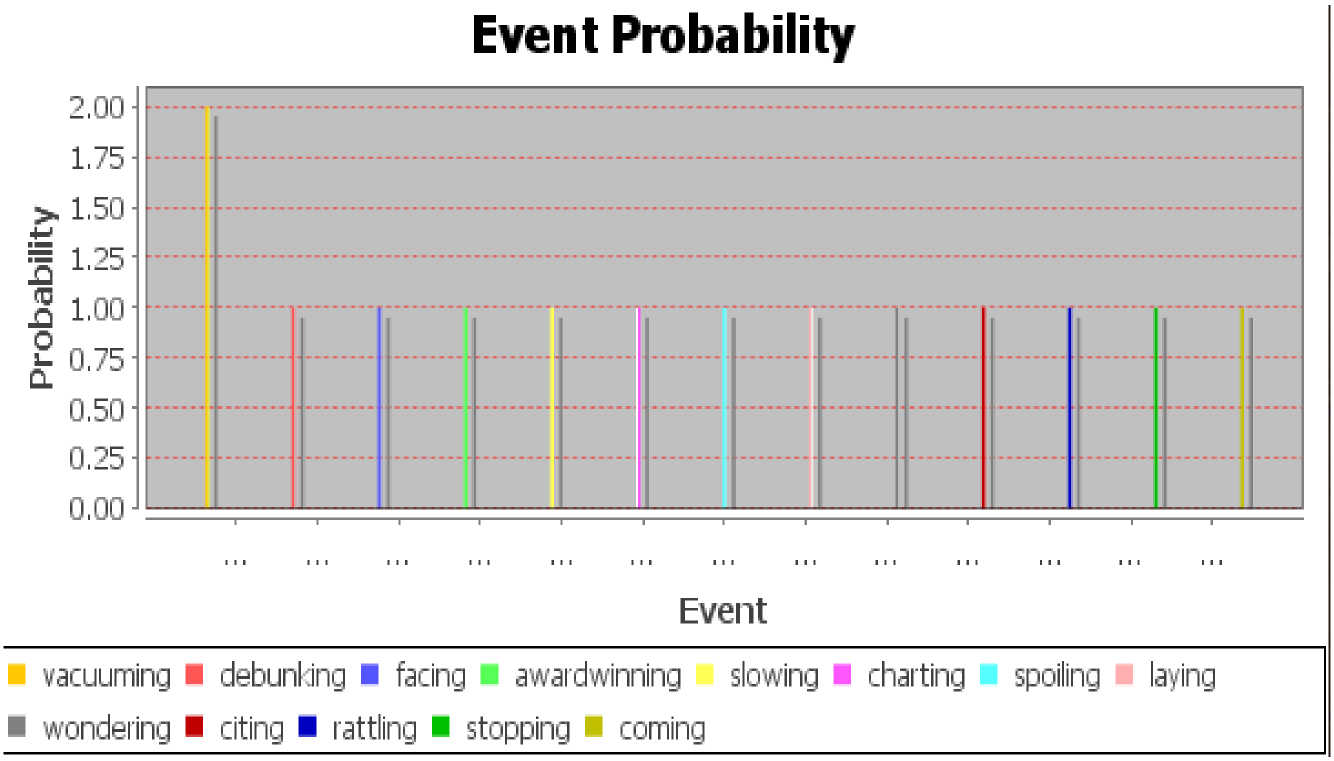




```
httpstco5zoitusl3u/NNS, httpstco8jqz1u40q7/NNS

d7fdd/CD, iphone/NN, apple/NN[Malus pumila] , httpstcos224juc3x8/NNS

rejiyates/NNS, imbecile/NN[changeling] , tweet/NN[twirp] , apple/NN[Malus pumila] , crumble/VB[break down] , insult/

bts_billboard/NN, army/NN[ground forces] , streaming/VBG[cyclosis] , important/JJ[of import] , chart/NN[graph] , billbo

cooni/NNS, appstore/NN, httpstco4gpxd7k6b2/NNS

6ans/NNS, httpstcoqlr9bzfg2x/NNS
```



```
vacuuming/VBG[hoover] # debunking/VBG[repudiation] # facing/VBG[veneer] # awardwinning/VBG# slowing/VBG[deceleration] # charting/VBG[chart

billionairebornmusic/JJ# play/NN httpstco00dp9qhifm[drama]

airing/VBG dig[dissemination]
```



IV CONCLUSION AND FUTURE WORK

We propose a summed up structure to see short messages viably and proficiently. All the more particularly, we separate the undertaking of short content comprehension into three subtasks: content division, sort discovery, and idea marking. We detail content division as a weighted Maximal Clique issue, and propose a randomized estimation calculation to keep up exactness and enhance proficiency in the meantime. We present a Chain Model and a Pair astute Model which join lexical and semantic highlights to lead sort location. They accomplish preferable exactness over customary POS taggers on the named benchmark.

We employ a Weighted Vote algorithm to determine the most appropriate semantics for an instance when ambiguity is detected. The experimental results demonstrate that our proposed framework outperforms existing state-of-the-art approaches in the field of short text understanding.

REFERENCES

1. A. McCallum and W. Li, “Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons,” in Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, ser. CONLL ’03, Stroudsburg, PA, USA, 2003, pp. 188–191.
2. G. Zhou and J. Su, “Named entity recognition using an hmm-based chunk tagger,” in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics , ser. ACL ’02, Stroudsburg, PA, USA, 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 Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence , ser. UAI ’04, Arlington, Virginia, United States, 2004, pp. 487–494.
5. R. Mihalcea and A. Csomai, “Wikify! Linking documents to encyclopedic knowledge,” in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, ser. CIKM ’07, New York, NY, USA, 2007, pp. 233–242.
6. D. Milne and I. H. Witten, “Learning to link with Wikipedia,” in Proceedings of the 17th ACM conference on Information and knowledge management , ser. CIKM ’08, New York, NY, USA, 2008, pp. 509–518.
7. S. Kulkarni, A. Singh, G. Ramakrishna, and S. Chakrabarti, “Collective annotation of Wikipedia entities in web text,” in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining , ser. KDD ’09, New York, NY, USA, 2009, pp. 457–466.
8. X. Han and J. Zhao, “Named entity disambiguation by leveraging wikipedia semantic knowledge,” in Proceedings of the 18th ACM conference on Information and knowledge management, ser. CIKM ’09, New York, NY, USA, 2009, pp. 215–224.

9. "Structural semantic relatedness: A knowledge-based method to named entity disambiguation," in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics , ser. ACL '10, Stroudsburg, PA, USA, 2010, pp. 50–59.
10. X. Han, L. Sun, and J. Zhao, "Collective entity linking in web text: A graph-based method," in Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '11, New York, NY, USA, 2011, pp. 765–774.
11. W. Shen, J.Wang, P. Luo, and M.Wang, "Linden: Linking named entities with knowledge base via semantic knowledge," in Proceedings of the 21st International Conference on World Wide Web, ser. WWW '12, New York, NY, USA, 2012, pp. 449–458.
12. C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee, "Twiner: Named entity recognition in targeted twitter stream," in Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '12, New York, NY, USA, 2012, pp. 721–730.
13. D. M. de Oliveira, A. H. Laender, A. Veloso, and A. S. da Silva, "Fsner: A lightweight filter-stream approach to named entity recognition on twitter data," in Proceedings of the 22nd International Conference on World Wide Web, ser. WWW '13 Companion, Republic and Canton of Geneva, Switzerland, 2013, pp. 597–604.
14. P. Ferragina and U. Scaiella, "Tagme: On-the-fly annotation of short text fragments (by wikipedia entities)," in Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ser. CIKM '10, New York, NY, USA, 2010, pp. 1625–1628.
15. Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen, "Short text conceptualization using a probabilistic knowledgebase," in Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three, ser. IJCAI'11, 2011, pp. 2330–2336.
16. D. Kim, H. Wang, and A. Oh, "Context-dependent conceptualization," in Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, ser. IJCAI'13, 2013, pp. 2654–2661.