Volume 4 || Special Issue 6 || ICCEME 2019-2020 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

ANALYZING AND HARVESTING SEMANTIC KNOWLEDGE FOR UNDERSTAND SHORT TEXTS

ANUP ARVIND DANGE[1], ASST. PROF. V. S. KARWANDE[2]

ME STUDENT,CSE Dept., EVEREST EDUCATIONAL SOCIETY'S GROUP OF INSTITUTIONS, AURANGABAD[1],

ASSISTANT PROFESSOR, CSE Dept., EVEREST EDUCATIONAL SOCIETY'S GROUP OF INSTITUTIONS, AURANGABAD[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 labeling 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 labeling, 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: *Text Segmentation, Semantic Knowledge, Type Detection, Concept Labeling, Interpretation*

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) locates named entities in a text and classifies them into predefined categories such as persons, organizations, locations, etc. Topic models attempt to recognize "latent topics", which are represented as probabilistic distributions on words, from a text. 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. As stated in Psychologist Gregory Murphy's highly acclaimed book, "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. Contrasted and records, short messages are created in a much bigger volume. For instance, Google, as the most generally utilized web internet searcher starting at 2014, got more than 3 billion hunt questions day1. Twitter likewise detailed in 2012 that it pulled in more than 100 million clients who posted 340 million tweets for each day 2. Along these lines, an attainable structure for short content comprehension ought to be capable to deal with short messages continuously. Be that as it may, a short content can have many conceivable divisions, a term can be marked with numerous sorts, and a case can allude to a large number of ideas. Consequently, it is amazingly tedious to take out these ambiguities and accomplish the best semantic translation for a short content.

Large number of short text producing every day from various resources such as whatsapp or internet blogging. The short text may be leads to the misunderstanding of meaning of text. So there is need to get actual information and understand short text.

1.1 **Text Segmentation-** divide a short text into a collection of terms (i.e., words and phrases)

contained in a vocabulary (e.g., "book Disneyland hotel California" is segmented as book Disneyland hotel California.

1.2 **Type Detection-** determine the types of terms and recognize instances (e.g., both "Disneyland" and "California" are recognized as instances in Fig. 1.1, while "book" is recognized as a verb and "hotel" a concept).

1.3 **Concept Labeling-** infer the concept of each instance (e.g., "Disneyland" and "California" refer to the concept theme park and state respectively in Although the three steps for short text understanding sound quite simple, challenges still abound and new approaches must be introduced to handle them.

II LITERATURE SURVEY

2.1 Xianpei Han and Jun Zhao, " Structural semantic relatedness : A knowledge based method to named entity disambiguation ".

In this Paper, it is found that 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) created new opportunities to enhance the named entity disambiguation by developing algorithms which coulexploit these knowledge sources at best. Problem was that these knowledge sources were heterogeneous and most of the semantic knowledge within them was embedded in complex structures, such as graphs and networks. The paper proposes a knowledge-based method, called Structural Semantic Relatedness (SSR), which enhances the named entity disambiguation by capturing and leveraging the structural semantic knowledge in multiple knowledge sources. Empirical results showed that, in comparison with the classical

BOW based methods and social network based methods, the method used could significantly improve the disambiguation performance by respectively 8.7% up to 14.7%.

Advantage: The primary advantage is the exploitation of semantic knowledge. One another advantage is the rich meaningful features, which is brought by the multiple semantic knowledge sources.

Limitations : 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.

2.2. X. Han, L. Sun, and J. Zhao, " Collective entity linking in web text : A graph-based method ".

In this paper, its is found that entity linking (EL) was the task of linking name mentions in Web text with their referent entities in a knowledge base. Traditional EL methods usually linked name mentions in the document by assuming them to be independent. However, there was often additional interdependence between different EL decisions, i.e., the entities in the same document would be semantically related to each other. In these cases, Collective Entity Linking, in which the name mentions in the same document were linked jointly by exploiting the interdependence between them, which could improve the entity linking accuracy. The paper proposes a graph-based collective EL method, which could model and exploit the global interdependence between different EL decisions. Specifically, the authors first proposed a graph based representation, called Referent Graph, which could model the global interdependence between different EL decisions. Then they proposed the collective inference algorithm, which could jointly infer the referent entities of all name mentions by exploiting the interdependence captured in Referent Graph.

The key benefit of the method came 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 could be reinforced into high-probability decisions. Experimental results showed that the method could achieve significant performance improvement over the traditional EL methods.

Advantage : It would be possible for the user to identify and explore the background knowledge of the searched item. Compared with the local compatibility based methods, the collective EL methods showed a greater advantage in linking less important name mentions in a document.

Limitations : The graph-based method was only leading by a pseudo NIL entity into the model.

3. W. Shen, J. Wang, P. Luo, and M. Wang, " Linden : Linking named entities with knowledge base via semantic knowledge".

In this paper, it is found that integrating the extracted facts with an existing knowledge base had 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 had 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, the proposed LINDEN, 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. The authors extensively evaluated the performance

of the 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.

Advantage: LINDEN is quite effective for the entity linking task.

Limitations: The semantic knowledge embedded in the taxonomy of concepts cannot be utilized to take advantage by using these methods.

2.4. 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 this paper, it is found out that 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, the authors 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.

Advantage: It utilizes and balances the advantages of both global context and local context.

Limitations: Performance is low as the strategy is only used to identify suitable K value.

2. 5. 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 this paper, it if found that 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, the authors 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. The results show that, despite the simplicity of the filters used, the approach outperformed the baseline with improvements of 4.9% on average, while being much faster.

Advantage : Affix filter can recognize entities that have similar affix to the entities analyzed before.

Limitations : The processing power of the application is slow.

2.6. P. Ferragina and U. Scaiella, " Tagme : On-the-fly annotation of short text fragments ".

In this paper, the authors 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 bagof-words paradigm could benefit from using this annotation to draw upon (the millions of) Wikipedia pages and their inter-relations.

Advantage : The annotations help to improve the effectiveness of the labeling and the clustering phases. Also the explanatory links and the structured knowledge produced allow the efficient and effective resolution of ambiguity issues which often occur when advertiser's keywords are matched against the content of Web pages offering display-ads.

Limitations : The most time consuming step is the calculation of the relatedness score, because anchor detection and other scores require time linear in the

length of the input text T. Its time complexity grows linearly with the number of detected anchors.

2.7. Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen, "Short text conceptualization using a probabilistic knowledge base ".

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.

Advantage :

_ Model short text semantics as concept distribution.

_ Knowledge base can help short text understanding.

Limitations :

_ De-noising Probase.

_ Better taxonomy building and inference

_ More relationships require.

III SYSTEM ARCHITECTURE

Following Fig 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.

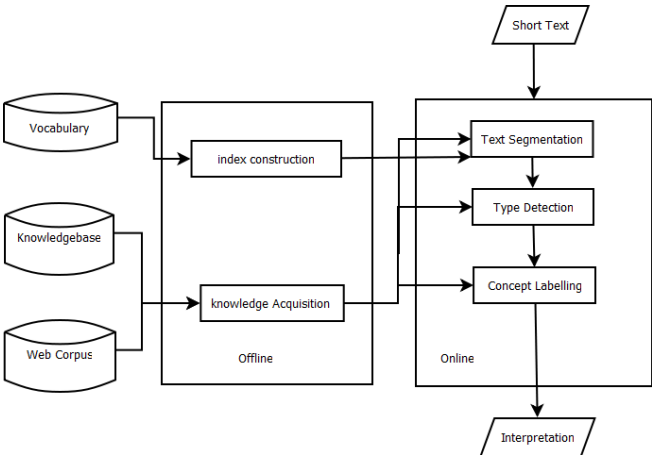

Fig 1: System Architecture

3.1 OVERVIEW OF PROJECT MODULES

3.1.1 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.

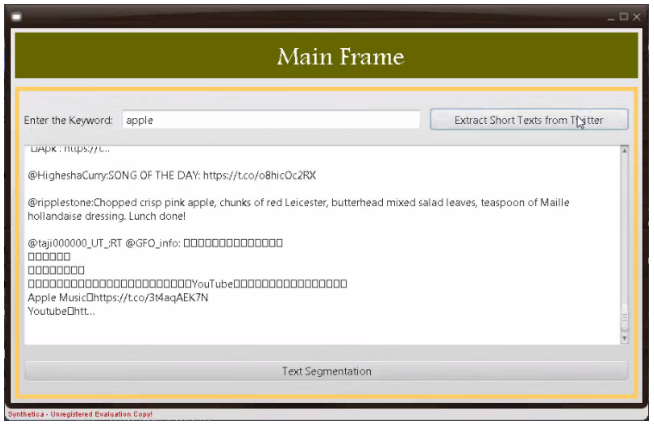

Fig 2: Main Frame

3.1.2. Text Segmentation

We can recognize all possible terms from a short text using the triedbased 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".

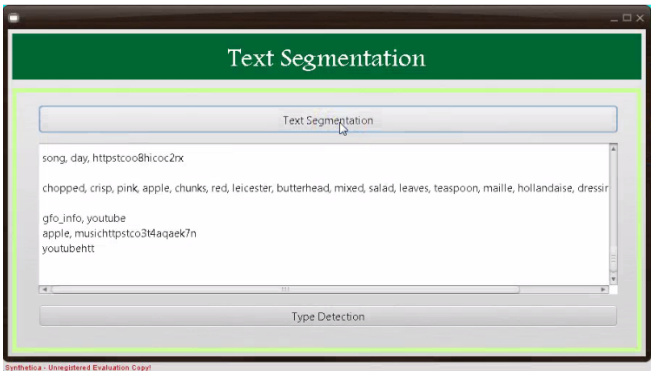

Fig 3: Text Segmentation

3.1.3. 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 instancelist, concept-list, as well as verb-list of our vocabulary, thus the possible typed-terms of "watch" are fwatch[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 fmovie[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.

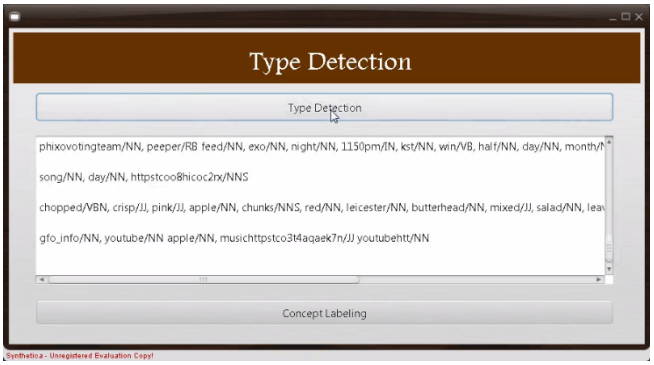

Fig 4: Type Detection

3.1.4. 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 cooccurs) with the concept song and thus can be kept as the final semantics of "eagles".

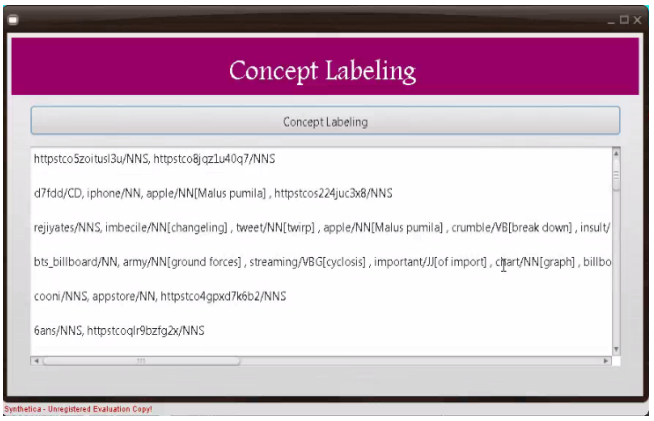

Fig 5: Concept Labeling

IV ALGORITHM

Maximal Clique by Monte Carlo (MaxCMC)

- **Input:**
  $G = (V, E)$; $W(E) = \{ w(e) | e \in E \}$
- **Output:**
  $G' = (V', E')$; $S(G')$

1 : $V' = \emptyset$; $E' = \emptyset$
2 : **while** $E \neq \emptyset$ **do**
3 : randomly select $e = (u, v)$ from E with probability proportional to it's weight.
4 : $V' = V' \cup \{u, v\}$; $E' = E' \cup \{e\}$
5 : $V = V - \{u, v\}$; $E = E - \{e\}$
6 : for each $t \in V$ **do**
7 : **if** $e' = (u,t)$ E *or* $e' = (v, t)$ E **then**
8 : $V = V - \{t\}$
9 : remove edges linked to t from E : $E = E - \{e' = (t, *)\}$
10 : **end if**
11 : **end for**
12 : **end while**
13 : calculate average edge weight : $S(G') = \frac{\sum W(e)}{|E'|}$

Algorithm runs as follows: First, it randomly selects an edge e = (u; v) with probability proportional to its weight. In other words, the larger the edge weight, the higher the probability to be selected. After picking an edge, it removes all nodes that are disconnected (namely mutually exclusive) with the picked nodes u or v. At the same time, it removes all edges that are linked to the deleted nodes. This process is repeated until no edges can be selected. The obtained sub-graph G0 is obviously a maximal clique of the original TG. Finally, it evaluates G0 and assigns it with a score representing the average edge weight. In order to improve the accuracy of the above algorithm, we repeat it for k times, and choose the maximal clique with the highest score as the final segmentation.In Algorithm, the while loop will be repeated for at most e times, since each time the algorithm removes at least one edge from the original TG. Here, ne is the total number of edges in TG. Similarly, the for loop in each while loop will be repeated for at most nV times. Therefore, the total time complexity of this randomized algorithm is O (k. ne. nV) or O (k. n3v). Our experimental results in Sec. 5 verify the effectiveness and efficiency of this randomized algorithm.

V CONCLUSION

In this Paper, we have 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 HLTNAACL 2003.

[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 , in 2002.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, " Latent dirichlet allocation ", J. Mach. Learn. Res., vol. 3, pp. 9931022, 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 , 2004.

[5] R. Mihalcea and A. Csomai, " Wikify! Linking documents to encyclopedic knowledge ", ACM Trans. Inf. in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, 2007.

[6] D. Milne and I. H. Witten, " Learning to link with Wikipedia ", in Proceedings of the 17th ACM conference on Information and knowledge management, 2008.

[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, 2009 .

[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, 2009.

[9] X. Han, L. Sun, and J. Zhao, " Collective entity linking in web text: A graphbased 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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. IJCAI11, 2011.

[15] D. Kim, H.Wang, " Context-dependent conceptualization ", in Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, ser. IJCAI13, 2013.

[16] Xianpei Han and Jun Zhao, " Structural semantic relatedness: A knowledgebased 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.