ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

Literature Review on Fuzzy Score Based Short Text Understanding from Corpus Data Using Semantic Discovery

S.Manimegalai [1], D.Umanandhini [2]

M.Phil Research Scholar, Department of Computer Science, Kovai Kalaimagal College of Arts and Science, Coimbatore, India [1]

Assistant Professor, Department of Computer Applications, Kovai Kalaimagal College of Arts and Science, Coimbatore, India [2]

ABSTARCT: Short text understanding and short text are always more ambiguous. These short texts are produced including Search queries, Tags, Keywords, Conversation or Social posts and containing limited context. Generally short texts do not contain sufficient collection of data to support many state-of-the-art approaches for text mining such as topic modelling. It presents a comprehensive overview of short text understanding. Here we used a novel framework are Text Feature Extraction Algorithm and Fuzzy weighted Vote algorithm First, Text classification based on semantic feature extraction.

Its goal is that use semantic feature extraction to improve the performance of classifier. And second, Fuzzy weighted Vote algorithm is the combination of Fuzzy logic and weighted vote algorithm, which means it generates the fuzzy score and then based on this score the weight is calculated during shortening the text. In experimental results, the novel Feature Extraction and voter has higher safety performance than the previous classification algorithms. This proposed criterion can provide almost accurate safety and also a good range of accessibility. We have proved that in problems where the weighted voting distinguish some alternatives and finds the best alternative. Reduced Computation time comparing to other previous process and schemes.

KEYWORDS: Short text understanding, Text segmentation, Concept labelling, Tagger

I. LITERATURE REVIEW

COMPUTING TERM SIMILARITY BY LARGE PROBABILISTIC IS A KNOWLEDGE

Computing semantic similarity between two terms is essential for a variety of text analytics and understanding applications. However, existing approaches are more suitable for semantic similarity between words rather than the more general multi-word expressions (MWEs), and they do not scale very well. Therefore, we propose a lightweight and effective approach for semantic similarity using a large scale semantic network automatically acquired from billions of web documents. Given two terms, we map them into the concept space, and compare their similarity there. Furthermore, an introduce a clustering approach to orthogonalize the concept space in order to improve the accuracy of the similarity measure.

Corpus-based approaches also face several serious limitations. First, such measures are biased because of the indexing and ranking mechanisms used in search engines. For example when querying the term “date” or “range” on Google, none of the first 100 results has anything to do with fruits (a sense for date) or cooking stoves (a sense for range), because these are rare senses of the two terms. With such search results, it is not surprising that a corpus-based method would think “Asian pear” and “date” share very little commonality. Second, some search-result oriented similarity methods require interaction with the search engine which has high communication overhead and high index costs, and are not suitable for online applications. Third, statistical distribution based on words or n-grams in the

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

context ignores the fact that i) the semantic units can be MWEs and not words, let alone n-grams; and ii) many words or phrases are ambiguous in meaning. Finally, corpus-based methods focus on surrounding context of a term or the co-occurrence of two terms within a neighborhood, both of which are more suitable to the calculation of semantic relatedness rather than similarity. Under this approach, "car" and "journey" would have high semantic relatedness because they co-occur very frequently on web texts.

Extensive studies show that our clustering-based refined algorithm outperforms the state-of-the-art methods as well as our basic algorithm in terms of Pearson correlation coefficient on word pairs and MWE pairs. The method is efficient enough to be applied on large scale data sets.

II. CONTEXT-DEPENDENT CONCEPTUALIZATION

Conceptualization seeks to map a short text to a set of concepts as a mechanism of an 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. An overcome this limitation, propose a framework in which the 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, develop a corpus-based framework for context-dependent conceptualization. A probabilistic topic model, an estimate how words is semantically related based on their general co-occurrence statistics, is a natural candidate for capturing the semantic relationships. A probabilistic knowledge base, Probase, which models the probabilistic concept instance mappings, is a good resource for capturing the conceptual relationships. The propose a two stage approach in which the first an estimate the topical context using LDA and then estimate the most likely concepts given the topic context using Probase.

Sentence level conceptualization using different weighting schemes can be used to match Web search queries and advertisements, as well as queries and URL titles. For both of these tasks, used large real world click through logs to quantitatively evaluate our framework against baseline approaches. Conceptualization is an important and general problem, and showed a simple but effective framework to combine Probase and LDA.

III. FS-NER: A LIGHTWEIGHT FILTER-STREAM APPROACH TO NAMED ENTITY RECOGNITION ON TWITTER DATA

Micro blog 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. Also, Twitter follows a streaming paradigm, imposing that entities must be recognized in real-time. In view of these challenges and the inappropriateness of existing tools, propose a novel approach for Named Entity Recognition on Twitter data called FS-NER (Filter-Stream Named Entity Recognition). 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. Moreover, because these filters are not language dependent, FS-NER can be applied to different languages without requiring a laborious adaptation. Through a systematic evaluation using three Twitter collections and considering seven types of entity, show that FS-NER performs 3% better than a CRF-based baseline, besides being orders of magnitude faster and much more practical. To evaluate the effectiveness of FS-NER, we used multi-lingual Twitter data obtained from different domains and involving diverse entity types. Our results reveal that FS-NER achieves similar recognition performance when compared to CRF-based approaches. On the other hand, in terms of computational performance, FS-NER surpassed by large the CRF-based approaches, indicating to be more practical to the Twitter environment.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

IV. LINDEN: LINKING NAMED ENTITIES WITH KNOWLEDGE BASE VIA SEMANTIC KNOWLEDGE

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 line 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. The propose LINDEN[9], 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. An 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.

In addition, we define the global coherence for each candidate entity to measure the global document-level topical coherence among the mapping entities in the document. And then we can give a rank to the candidate entity list for each entity mention with the combination of these four measures, link probability, semantic associativity, semantic similarity and global coherence. Furthermore, LINDEN learns how to return NIL for the entity mention which has no matching entity in the knowledge base. To validate the effectiveness of LINDEN, we empirically evaluate it over two public data sets. Since in LINDEN we use the whole information in Wikipedia to generate candidate entities in the Candidate Entities Generation module, we have to add some un-linkable mentions prediction strategies to the module of Un-linkable Mentions Prediction

Entity linking is a very important task for many applications such as Web people search, question answering and knowledge base population. It proposes a LINDEN, a novel framework to link named entities in text with YAGO, a knowledge base unifying Wikipedia and WordNet. By leveraging the rich semantic knowledge derived from the Wikipedia and the taxonomy of YAGO, LINDEN can obtain great results on the entity linking task. A large number of experiments were conducted over two public data sets, i.e., the CZ data set and the TAC-KBP2009 data set. Empirical results show that LINDEN significantly outperforms the state-of-the-art methods in terms of accuracy. Moreover, all features adopted by LINDEN are quite effective for the entity linking task.

V. COLLECTIVE ENTITY LINKING IN WEB TEXT: A GRAPH-BASED METHOD

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.

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. First propose a graph-based representation, called Referent Graph, which can model the global interdependence between different EL decisions as its graph structure. Then we propose a purely collective inference algorithm, which can jointly infer the referent entities of all name mentions in the same document by exploiting both the global interdependence between different EL decisions and the local mention-to-entity compatibility. It have evaluated our method on a standard EL dataset. The experimental results show that our method can achieve significant performance improvement over the traditional EL methods.

In our method, we did not take into account the NIL entity problem of the EL task, i.e., the referent entity of a name mention may not be contained in the given knowledge base. For future work, we will resolve this aspect in our graph-based method by leading a pseudo NIL entity into our model. Furthermore, using the entity linking method, we want to develop a Web entity search and mining system by annotating billions of Web pages with their entity information.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

VI. COLLECTIVE ENTITY LINKING IN WEB TEXT: A GRAPH-BASED METHOD

Entity Linking 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. 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.

It 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.

Then we propose a purely collective inference algorithm, which can jointly infer the referent entities of all name mentions in the same document by exploiting both the global interdependence between different EL decisions and the local mention-to-entity compatibility. It have evaluated our method on a standard EL dataset. The experimental results show that our method can achieve significant performance improvement over the traditional EL methods.

In our method, we did not take into account the NIL entity problem of the EL task, i.e., the referent entity of a name mention may not be contained in the given knowledge base. For future work, we will resolve this aspect in our graph-based method by leading a pseudo NIL entity into our model. Furthermore, using the entity linking method, we want to develop a Web entity search and mining system by annotating billions of Web pages with their entity information.

VII.SHORT TEXT CONCEPTUALIZATION USING A PROBABILISTIC KNOWLEDGEBASE

Most text mining tasks, an 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. The particular challenges faced by such approaches in short text understanding, as short texts lack enough content from which statistical conclusions can be drawn easily. It improved the text understanding by using a probabilistic knowledge base that is as rich as our mental world in terms of the concepts it contains.

It develops Bayesian inference mechanism to the conceptualize words and short text. The 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 modelling or methods that use existing knowledgebase our approach brings significant improvements in short text understanding as reflected by the clustering accuracy.

In comparison, the concepts in Probase are more consistent with humans common knowledge. Finally, our approach out performs all other approaches on both problems. In Probase, the concept space has different granularities, and it is also much larger. This enables Probase to capture short content as expressed by tweets.

VIII. STRUCTURAL SEMANTIC RELATEDNESS: A KNOWLEDGE-BASED METHOD TO NAMED ENTITY DISAMBIGUATION

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. 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.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

To overcome the deficiencies of previous methods, this 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 from multiple knowledge sources.

In particular, first extract the semantic relations between two concepts from a variety of knowledge sources and represent them using a graph-based model, semantic-graph. Then based on the principle that "two concepts are semantic related if they are both semantic related to the neighbour concepts of each other", we construct our Structural Semantic Relatedness measure. In the end, leverage the structural semantic relatedness measure for named entity disambiguation and evaluate the performance on the standard WePS data sets. The experimental results show that our SSR method can significantly outperform the traditional methods.

In particular, propose a semantic relatedness measure, Structural Semantic Relatedness, which can capture both the explicit semantic relations and the implicit structural semantic knowledge. The experimental results on the WePS data sets demonstrate the efficiency of the proposed method.

IX. TAGME: ON-THE-FLY ANNOTATION OF SHORT TEXT FRAGMENTS (BY WIKIPEDIA ENTITIES)

It 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 is that it may annotate texts which are short and poorly composed, such as snippets of search-engine results, tweets, news, etc.

Tagme a software system that, on-the-fly and with high precision/recall, annotates short fragments of text with pertinent hyperlinks to Wikipedia articles. Preliminary experiments show that Tagme outperforms the best known systems when they are adapted to work on short texts, and surprisingly, it results competitive on long texts too. The system by Milne &Witten performed poorly, because many features used by their pruning method are not effective when dealing with short texts. In fact they consider features like location and frequency of anchors (which may be "undefined" or even misleading on short texts), as well as they consider only the un-ambiguous anchors to compute a coherence-score.

It currently investigating the impact of Tagme's annotation onto the performance of our past system Snake T for the on-the-fly labelled clustering of search-engine results. In fact Snake T, as most of its competitors is based only on syntactic and statistical features and thus, believe that, it could benefit from Tagme's annotation to improve the effectiveness of the labelling and the clustering phases. Another promising context of application for Tagme could be Web Advertising. The explanatory links and the structured knowledge produced by Tagme could allow the efficient and effective resolution of ambiguity and polysemy issues which often occur when advertiser's keywords are matched against the content of Web pages offering display-ads.

X. TWINER: NAMED ENTITY RECOGNITION IN TARGETED TWITTER STREAM

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. Targeted Twitter stream is then monitored to collect and understand users opinions about the organizations. The 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. An 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. An evaluated TwiNER on two sets of real life tweets simulating two targeted streams.

This gregarious property of named entities in Twitter motivates us to design a "recursive" algorithm to compute the score of a segment being a named entity. Because a segment's confidence is affected by its neighbors in the graph, which only depends on the tweets themselves, we call the segment graph as the local context of a segment in the

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

tweets. Note that, not only has the local context been considered in this model, but also the global context is integrated to overcome the limitation of random walk model.

XI. CONCLUSION

It presented an effective approach for semantic similarity between terms with any multi-word expression. Using a large scale semantic network automatically acquired from billions of web documents.

And used Fuzzy weighted Vote algorithm is the combination of Fuzzy logic and weighted vote algorithm, which means it generates the fuzzy score and then based on this score the weight is calculated during shortening the text. This method can collectively infer the referent entities of all name mentions in the same document. By modelling and exploiting the global interdependence between different EL decisions, the proposed method can achieve competitive performance over the traditional methods. More importantly, the model enables probabilistic inference between concepts and instances which will benefit a wide range of applications that require text understanding.

The results revealed that the novel voter has higher safety performance than the Majority and fuzzy weighted voting algorithms. This voting criterion can provide almost accurate safety and also a good range of accessibility. It have proved that in problems where the weighted voting distinguish some alternatives and finds the best alternative. Reduced Computation time comparing to other previous process and schemes.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.
[2] 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.
[3] 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.
[4] 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.
[5] 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.
[6] 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.
[7] "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.
[8] 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.
[9] 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.