ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

Short Text Segmentation Using Semantic Knowledge

Saima Jabeen[1], Rekha S[2]

P.G. Student, Department of Computer Engineering, PDA College of Engineering, Aiwan shahi Gulbarga, India[1]

Associate Professor, Department of Computer Engineering, PDA College of Engineering, Aiwan shahi, Gulbarga, India[2]

ABSTRACT. Understanding short text is a crucial task, as the text are more ambiguous it is difficult to understand. Short text are usually produced by search queries, tweets, conversation and tags. Short text are ambiguous and noisy as the text has more than one meaning ,the text does not contain sufficient data it is difficult to handle. In the proposed work we construct a model framework for seeing short content which misuses semantic information conveyed by a notable knowledge base and this can be consequently reaped from a web corpus Our knowledge-intensive methodologies disturb old-style methods for tasks such as part-of-speech tagging, text segmentation and the concept labelling, in the sense that in all these tasks revolve around the semantics. On real-data a comprehensive performance evaluation is being conducted here, the results for this show that to understand the short text semantic knowledge is indispensable. The approaches we are using are both efficient and effective in identifying or discovering the semantics of short texts.

KEYWORDS: Text detection, Segmentation, Concept labelling

I. INTRODUCTION

Understanding short text and classifying them is a difficult and challenging task as short text has more than one meaning. Short text are ambiguous and noisy as they do not contain sufficient data. Most of the applications which may refer to microblogging sites or search engines to handle large volume of short text. For such reason proper understanding of short text will bring immense value to discover hidden semantic from the available text. This is one of the significant tasks of text understanding. Short text give rise to significant amount of ambiguity, new approaches are introduced to handle them. Short text understanding strategy is defined in three steps. I. Text Segmentation 2.Type Detection 3.Concept Labelling

1. Text Segmentation

It is the process of dividing written text into meaningful units such as words, topic. An important topic of short text is to calculate the semantic similarity between short texts. Incorrect segmentation of short text leads to incorrect semantic similarity. The drawback in the existing system is that it only considers the surface features and ignores the requirement of semantic coherence within a segmentation. To overcome an exploit context semantics is been proposed when conducting text segmentation.

Ex. Ambiguity in Text Segmentation
"Paris to Berlin lyrics vs vacation Paris to Berlin"

The text "*vacation Paris to Berlin*" can be segmented in a confused way, as "*vacation Paris to Berlin*" refers the longest term in the vocabulary, it refers only Paris to Berlin ignoring the rest and leading to incorrect segmentation. The "*Paris to Berlin*" lyrics is unambiguous as lyrics is syntactically related to songs. The incorrect segmentation is similar to that of Paris to Berlin lyrics, thus correct segmentation is required for the text.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

2. Type Detection

It determines the purpose of the terms, and determines the lexical types. It taggers include, Rule Based Taggers, Statistical parts-of-speech Taggers and Mainstream Part-of-speech Tagger.

In Rule based Part-of-speech Taggers it assigns Part-of –speech tags to abstruse and unknown words based on large number of verbal rules which are automatically learned. The drawback of this tagger is that it is less accurate.

In statistical Part-of-speech Tagger it forms a collection of written text and unlabelled texts based on the set of information. The drawback of this tagger is that it required vast amount of store information.

In Mainstream Part-of-speech tagger it involves the well-known Markov model which acquires knowledge on both lexical and sequential probabilities from the information given and labels another sentence via hunting down label succession that amplifies the blend of lexical and sequential probabilities. The method ignores semantics and considers only lexical features which causes mistakes. To avoid all the drawbacks pairwise model and chain model is compared, and pairwise is used as it provides all kinds of accuracy measures to the part-of-speech taggers.

Ex. Ambiguity in Type Detection

"Blue Song vs Blue Bag"

This is tagged with the lexical and semantic type, it describes the type of term. In example the Blue song refer to the singer and the Blue Bag refer to the colour of the Bag. As it focuses on surface features, it is difficult to determine the type of term in short text.

3. Concept Labelling

It determines the appropriate concepts of an instance with specific context. It focuses on named entities, and classifies them into predefined categories by using statistical techniques like Conditional Random field and Hidden Marcov Model. Named entity recognition cannot be applied to short text directly as it do not always observe the syntax of written language.

Ex. Ambiguity in Concept Labelling

"*Watch Hunger games vs Read Hunger games"*

In this there are multiple concept which have single instance. It can have mapping between instance and concept. When context varies, different concepts can be referred to a single instance. Related term is used to avoid ambiguity of the term.

II.RELATED WORK

T. Gri_ths and M. Rosen-Zvi**,**We show the maker subject model, a generative model for files that expands Latent Dirichlet Allocation (LDA; Blei, Ng, and Jordan, 2003) to join commencement information. Each creator is related with a multinomial dispersion over subjects and every theme is related with a multinomial conveyance over words. A record with different creators is demonstrated as a conveyance over subjects that is a blend of the appropriations related with the creators. Correct induction is immovable for these datasets and we utilize Gibbs inspecting to appraise the subject and creator appropriations. We contrast the execution and two other generative models for reports, which are extraordinary instances of the creator theme show: LDA (a subject model) and a basic creator display in which each creator is related with a circulation over words as opposed to a conveyance over points. We indicate subjects recuperated by the creator point display, and show applications to registering likeness amongst creators and entropy of creator yield. Li and A. McCallumModels for some, common dialect assignments advantage from the adaptability to utilize covering, non-free highlights. For instance, the requirement for named information can be definitely decreased by exploiting area learning as word records, grammatical feature labels, character n-grams, and upper casing designs. While it is hard to catch such between subordinate highlights with a generative probabilistic model, restrictively

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

prepared models, for example, contingent most extreme entropy models, handle them well. There has been noteworthy work with such models for voracious arrangement demonstrating in NLP (Ratnaparkhi, 1996; Borthwick et al., 1998).

J. Su and G. ZhouThis paper proposes a Hidden Markov Model (HMM) and a HMM-based piece tagger, from which a named substance (NE) affirmation (NER) system is attempted to perceive and characterize names, times and numerical amounts. Through the HMM, our framework can apply and incorporate four kinds of inside and outer confirmations: 1) basic deterministic interior element of the words, for example, upper casing and digitalization; 2) inner semantic element of critical triggers; 3) inward gazetteer include; 4) outside full scale setting highlight.Thus, the NER issue can be settled enough. Evaluation of our system on MUC-6 and MUC-7 English NE endeavors achieves F-measures of 96.6% and 94.1% individually. It demonstrates that the execution is essentially superior to anything announced by some other machine-learning framework. In addition, the execution is even reliably superior to those in light of hand-created rules

H. Witten and D. MilneThis paper portrays how to consequently cross-reference reports with Wikipedia: the biggest learning base at any point known. It elucidates how machine learning can be used to recognize basic terms inside unstructured substance, and propel it with associations with the best possible Wikipedia articles. The subsequent connection indicator and disambiguate performs exceptionally well, with review and exactness of just about 75%. This execution is consistent whether the framework is assessed on Wikipedia articles or "genuine" documents. This work has suggestions a long ways past enhancing records with illustrative connections. It can give organized information about any unstructured section of content. Any undertaking that is as of now tended to with packs of words - ordering, grouping, recovery, and rundown to give some examples - could utilize the procedures depicted here to draw on an immense system of ideas and semantic

A. Singh and S. Kulkarni to venture out catchphrase based inquiry toward substance based pursuit, appropriate token ranges ("spots") on archives must be recognized as references to true elements from an element index. A few frameworks have been proposed to interface spots on Web pages to substances in Wikipedia. They are generally in view of nearby similarity between the content around the spot and literary metadata related with the element. Two late frameworks misuse between mark conditions, however in restricted ways. We propose a general aggregate disambiguation approach. Our introduction is that intelligible reports allude to substances from one or a couple of related themes or areas. We give details for the exchange off between nearby spot-to-element similarity and measures of worldwide rationality between substances. Upgrading the general element task is NP-hard. We research viable arrangements in view of neighbourhood slope climbing, adjusting whole number straight projects, and pre-grouping elements took after by nearby advancement inside bunches. In tests including over a hundred physically clarified Web pages and a huge number of spots, our methodologies essentially beat as of late proposed calculations

III.PROBLEM STATEMENT

The short text understanding have different challenges that needs to be understood d is
i. A complete vocabulary
ii. Mapping between instance and concept
iii. Semantic Coherence between terms

PROBLEM DEFINATION

Given a query "*Book dream world motel queens land*" we want to know that the uses is searching for motel near by dream world in Queensland. These are the steps which are taken place.
The important terms are identified using a vocabulary that appears in a short text.
We get two different segmentation

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

"Book dream world motel Queensland"

&

"*Book dream world motel Queensland*"

As *boo*k has multiple types {book (V), book(c)}. By analysing the sentence the book can be stated as a verb within a concept, and motel as a concept and dream world and Queensland as instance. The dream world can be determined as a theme park within the short text more than the company, as the concept is semantically more related to the theme park than the company.

IV.METHODOLOGY

1. Offline Processing
2. Constructing co-occurrence network
3. Scoring semantic coherence
4. Determining instance ambiguity

1. Offline Processing

In this module, we develop the system environment entities i. Admin ii. User

In the admin module, the admin has to login by using valid user name and password. After login successfully, the admin can perform operations such as add data, view user, user search history and most likes details. In this module we develop the entity with the following functionality:

i. View all user information
ii.Add new data into data base
iii.Admin to see the ranking module part in admin page after using searching.

In user module, n number of users are present. Users should register before performing some operations. Registered user details are stored in admin data base. After registering successfully, the user has to login by authorized user name and password. After login, user can give an input like short text keyword, after completing this the step results will be shown.

2. Constructing Co-occurrence Network

The construction of a co-occurrence network is to model semantic relatedness between the terms Co-occurrence systems are the aggregate interconnection of terms in light of their combined nearness inside a particular unit of content. As noticed terms of various sorts happen in various settings. Harry potter as a verb co-occurs with movie while Harry potter as a instance co-occurs with buy and price, thus the construction of co-occurrence network must be between typed terms instead of terms. The more frequently the two typed terms occur the higher the relatedness will be.

3. Scoring Semantic Coherence

In includes two type of coherence similarity and relatedness. It is believed that two typed terms are coherent if they are semantically similar or they often co-occur on the web. Surmised term extraction intends to find substrings in a content which are similar terms contained in a predefined vocabulary To quantify the similarity have been two strings, many similarity functions have been proposed including token-based similarity function and character based , is the way toward separating a surge of content into words, phrases or other significant components called tokens. In token-based function the work will estimating distance between shorter strings that differ largely at character level, they become too computationally expensive and accurate for larger strings. In character based function it characterizes the separation between two strings as the base number of character additions, erasures or substitutions required to change one string into another. Due to the prevalence of incorrect spellings in short messages, we utilize alter separate as our likeness capacity to encourage surmised term extraction.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

4. Determining Instance Ambiguity

The focus of concept labelling instance disambiguation making it important to determine whether an instance is ambiguous or not. Handling unambiguous cause over filtering. A straight forward approach to determine instance ambiguity is to check the number of concepts it belongs to. However the accuracy of such a method is highly dependent on the granularity of concept space in a knowledge base. A course grained knowledge will miss some ambiguous instance while a fined grained knowledge base might lead to false positive i.e. an unambiguous instance is incorrectly recognised as ambiguous. In this work we adopt a fine grained knowledge base.

V.SYSTEM ARCHITECTURE

In the fig1 it shows text segmentation, where the text is divided into sub text. In the construction of co-occurrence network, user wants the meaning of the particular term to display the meaning of occurrence as the simple word has various meaning.

Short text are usually regulated format and error based nicknames, abbreviations and misspellings. Using approximate term extraction, it identifies the meaning of short text. As per user requirement to find out meaning of the word and particular meaning of word provided by user.

In type detection it describes the type of term to contribute short text understanding. The short text does not always observe syntax of the written language. To provide the exact meaning of the short text and to determine the similarity between two strings, it finds all the possible terms for related text.

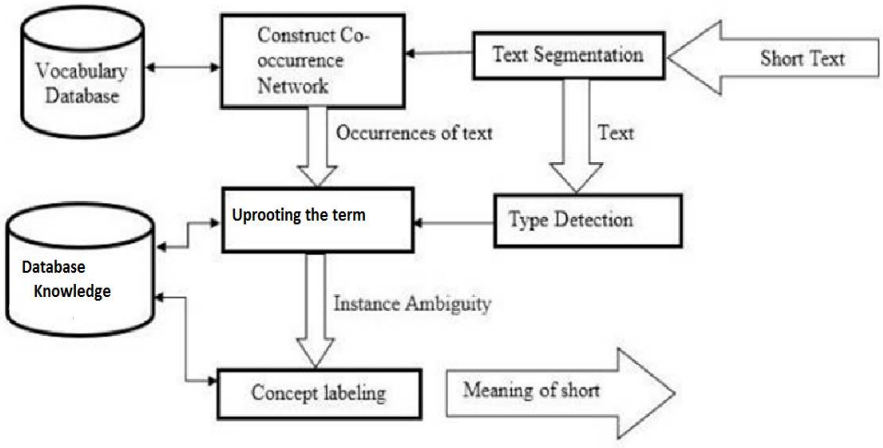

Fig1 –System architecture of exploring semantics of short text messages

The short text is an input of concept labelling. In the type detection, obtain the collection typed of the term from the vocabulary. Concept labelling is used to overcome the ambiguity of the term, same name with different meaning is to be identified by specifying a label. So related term is used to avoid ambiguity

.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

VI. EXPERIMENTATION AND RESU;TS

The contribution of our work is as follows:
By proposing a generalized framework it recognises segmentation, type detection and eliminates instance ambiguity on the bases of various type of context information.
As words are noisy, a large vocabulary is created which includes terms and abbreviation. The efficiency and accuracy of the words are improved by using segmentation of words. Short text is regarded as an online task of many other text mining applications like classification and segmentation. As millions of short text need to be handled by an application, which makes the efficiency of short text understanding critical. In the frame work short text, segmentation uses maximal clique, and type detection uses pairwise model and instance ambiguity uses weighted vote algorithm.
In the text segmentation as it focuses only on the surface features such as length, it uses maximal clique approach to segment the text which gives better performance than the other approaches.

Algorithm-1
1: **procedur**e MaxClique (*G= (V, E)*)
2: **for** i: 1 to *n* do
3**:** **if** d (v_i) $\geq$*max***then**
4: *U* ←Ø
5: **for** each $_{Vj}$ ϵ N(v_i)**do**
6: **if** d (vj) $\geq$*max***then**
7: U←U ∪ {Vj}
8: CLIQUE (G, U, 1)

The Algorithm considers only *one* neighbour with *maximum degree* at each step instead of recursively considering *all* neighbours from the set U, and, thus, is much faster. The vertex with maximum degree is chosen for this intuitive reason: in a relatively fairly connected subgraph, a vertex with the maximum degree is more likely to be a member of the largest clique containing that vertex than to any other

	Accuracy	Efficiency	Optimization
Maximal clique algorithm	✓	✓	×
Randomized approximation algorithm	×	×	✓
Pairwise Model	✓	✓	×
Chain model	✓	✓	×
Weighted vote algorithm	✓	×	×

Table-1: comparison table showing the features of algorithm

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

Table-1 showing different algorithms with different features, the algorithms include different features which helps in segmenting, tagging and better understanding of short text. Each algorithm acquires different features that helps in obtaining better accuracy and efficiency for short texts.

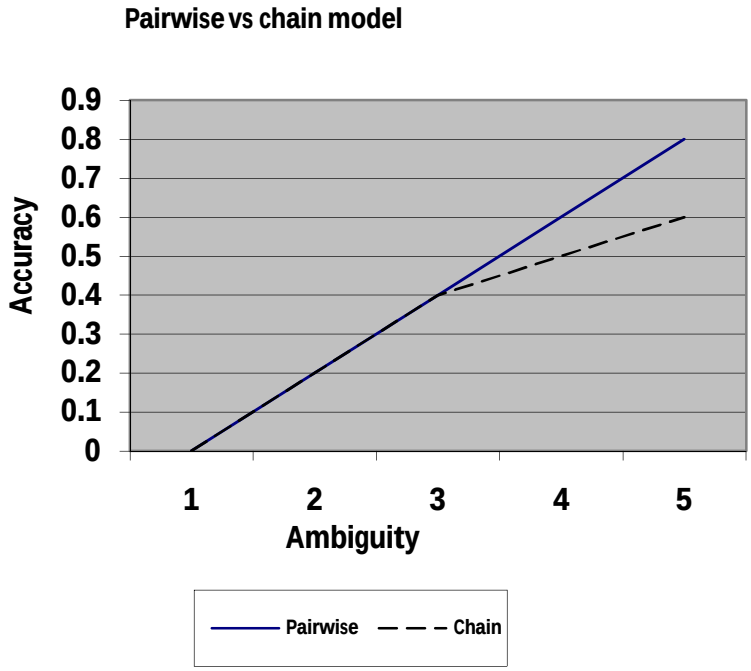

Fig – 2 Graph showing compatibility of pair and chain wise model

In the type detection it use different taggers, we compare the chain and pairwise model. As pairwise model is better than chain model in all kinds of accuracy measures which in turn provides better accuracies than the part of speech taggers.
As seen in fig-2 the graph shows that the pairwise model is more compatible than the chain model, as both are compared the graph shows that the pairwise model provides more accuracy measures than the chain model, thus pairwise is used

As short text understanding is regarded as an online task of many text mining application, such as classification and segmentation. The efficiency of short text is critical as the application needs to handle millions of short text, thus the framework is used to verify its efficiency.

VII.CONCLUSION

A generalized framework is proposed to understand short text efficiently and effectively. The task of short text is to divide the short text inti sub tasks, text segmentation, type detection and concept labelling. In text segmentation a weighted clique problem is formulated and proposed a randomized approximation to maintain accuracy and improve efficiency. In the type detection for lexical and semantic features, chain and pairwise model has been introduced to achieve better accuracy than part-of-speech tagger. In concept labelling a weighted vote algorithm is applied and employed to determine the concept for an instance when ambiguity is detected. The disambiguated sentence improve the accuracy of measuring semantic coherence between terms which helps in the performance of text segmentation and type detection.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

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. Gri_ths, 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] M. Utiyama and H. Isahara, "A statistical model for domain-independent text segmentation," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ser. ACL '01, Stroudsburg, PA, USA, 2001, pp. 499–506.
[6] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: a probabilistic taxonomy for text understanding," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '12, New York, NY, USA, 2012, pp. 481–492
[7] 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.
[8] 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.
[9] D. Deng, G. Li, and J. Feng, "An efficient trie-based method for approximate entity extraction with edit-distance constraints," in *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering*, ser. ICDE '12, Washington, DC, USA, 2012, pp. 762–773
[10] E. BRILL, "Some advances in transformation-based part of speech tagging," in *National Conference on Artificial Intelligence*, 1994, pp. 722–727.
[11] E. Brill, "Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging," *Comput. Linguist.*, vol. 21, no. 4, pp. 543–565, 1995
[12] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "A practical part-of-speech tagger," in *Proceedings of the third conference on Applied natural language processing*, ser. ANLC '92, Stroudsburg, PA, USA, 1992, pp. 133–140.
[13] R. Weischedel, R. Schwartz, J. Palmucci, M. Meteer, and L. Ramshaw, "Coping with ambiguity and unknown words through probabilistic models," *Comput. Linguist.*, vol. 19, no. 2, pp. 361–382, 1993.
[14] B. Merialdo, "Tagging english text with a probabilistic model," *Comput. Linguist.*, vol. 20, no. 2, pp. 155–171, 1994.
[15] H. Schutze and Y. Singer, "Part-of-speech tagging using a variable¨ memory markov model," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ser. ACL '94, Stroudsburg, PA, USA, 1994, pp. 181–187.