A Survey on Understand Short Texts by Collecting and Evaluating Semantic Knowledge

Roopali Gupta[1], Naresh Kumar Kar[2]
[1]PG scholar, [2]Associate professor
Dept of CSE, Rungta College of Engineering and technology, Raipur, Chhattisgarh 492009.

Abstract **-** Seeing short messages is pivotal to numerous applications, yet challenges flourish. To begin with, short messages don't generally watch the sentence structure of a composed dialect. Accordingly, conventional regular dialect preparing devices, running from grammatical form labeling to reliance parsing, can't be effectively connected. Second, short messages typically don't contain adequate factual signs to help many cutting edge approaches for content mining, for example, point demonstrating. Third, short messages are more equivocal and uproarious, and are created in a tremendous volume, which additionally expands the trouble to deal with them. We contend that semantic learning is required with the end goal to all the more likely see short messages. In this work, we assemble a model framework for short content understanding which abuses semantic learning given by a notable knowledgebase and naturally collected from a web corpus. Our insight escalated approaches upset conventional techniques for undertakings, for example, content division, grammatical feature labeling, and idea marking, as in we center around semantics in every one of these errands. We lead a far reaching execution assessment on genuine information. The outcomes demonstrate that semantic information is basic for short content comprehension, and our insight concentrated methodologies are both viable and effective in finding semantics of short messages.

Keywords - Short text understanding, text segmentation, type detection, concept labeling, semantic knowledge.

I. INTRODUCTION

Info surge highlights the demand for makers to much better comprehend all-natural language messages. In this paper, we concentrate on short messages which describe messages with minimal context. Lots of applications, such as internet search as well as micro-blogging solutions and so on, require taking care of a big quantity of short messages. Clearly, a far better understanding of short messages will certainly bring remarkable worth. Among one of the most essential jobs of message understanding is to uncover covert semiotics from messages. Several initiatives have actually been dedicated to this area. As an example, called entity acknowledgment (NER) [1] [2] finds called entities in a message as well as identifies them right into predefined classifications such as individuals, companies, places, and so on. Subject versions [3] [4] effort to acknowledge "unexposed subjects", which are stood for as probabilistic circulations on words, from a message. Entity connecting [5] [6] concentrates on recovering "explict subjects" revealed as probabilistic circulations on a whole knowledgebase. Nonetheless, classifications, "unexposed subjects", in addition to "specific subjects" still have a semantic space with human beings' psychological globe. As specified in Psycho therapist Gregory Murphy's extremely well-known publication, "principles are the adhesive that holds our psychological globe with each other". As a result, we specify short message understanding regarding discover principles discussed in a short message. Fig. 1 shows a common technique for short message understanding which contains 3 actions: Text Division - separate a short message right into a collection of terms (i.e., words as well as expressions) consisted of in a vocabulary (e.g., "publication Disney land resort california" is fractional as book Disney land resort california); _ Kind Discovery - figure out the sorts of terms as well as acknowledge circumstances (e.g., both "disneyland" as well as "the golden state" are acknowledged as circumstances in Fig. 1, while "publication" is identified as a verb as well as "resort" an idea); _ Principle Identifying - presume the idea of each circumstances (e.g., "disneyland" as well as "the golden state" describe the principle amusement park as well as state specifically in Fig. 1). In general, 3 ideas are found from short message "publication Disneyland resort california" utilizing this approach, particularly amusement park, resort, and also state in Fig. 1.

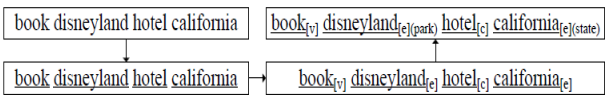

Figure 1: An example of short text understanding

Although the three steps for short text understanding sound quite simple, challenges still abound and new approaches must be introduced to handle them. In the following, we use several examples to illustrate such a need.

Challenge 1 (Ambiguous Segmentation):"april in paris verses" vs. "holiday april in paris" Both a term as well as its sub-terms can be included in the vocabulary, resulting in numerous feasible divisions for a short message. Nonetheless, a legitimate division must preserve semantic comprehensibility. For instance, 2 divisions can be stemmed from "april in paris verses", specifically april in paris lyrics as well as april paris lyrics. Nonetheless, the previous is a far better division according to the understanding that "verses" is a lot more semantically connected with tunes

("april in paris") than months ("april") or cities ("paris"). Typical Longest Cover technique, which is widely-adopted for message division [4] [5], seeks for lengthiest terms consisted of in a vocabulary. It disregards the demand of semantic comprehensibility, as well as hence will certainly cause wrong divisions often. When it comes to "trip april in paris", the Longest Cover technique sections it as vacation april in parisg which is certainly a mute division.

Challenge 2 (Noisy Short Text): "New York City city" vs. "nyc" vs. "large apple" In order to discover the most effective division for an offered message by thinking about semantic comprehensibility, we initially require to draw out all prospect terms. It can be quickly as well as effectively done by developing a hash index on the whole vocabulary. Nevertheless, short messages are normally casual and also error-prone, loaded with acronyms, labels, misspellings, and so on. For instance, "New York City city" is generally abbreviated to "nyc" as well as referred to as "huge apple". This asks for the vocabulary to integrate as much info concerning acronyms and also labels as feasible. On the other hand, approximate term removal is additionally needed to deal with misspellings simply put messages.

Challenge 3 (Ambiguous Type):"pink (vocalist) tracks" vs. "pink footwear" We identify terms with lexical kinds (i.e., POS tags) as well as semantic kinds (i.e., quality, idea, as well as circumstances). We will certainly describe why we take into consideration these kinds as well as exactly how they add to short message understanding in text segmentation. A term can come from a number of kinds, and also its ideal enter a short message depends upon context semiotics. For instance, "pink" in "pink tunes" describes a well-known vocalist and also therefore ought to be classified as circumstances, whereas it is an adjective in "pink footwear" explaining the shade of footwear. Typical POS taggers figure out lexical kinds based upon etymological guidelines or lexical as well as consecutive possibilities picked up from identified corpora. Nevertheless, such surface area functions are inapplicable simply put messages, as a result of the truth that short messages do not constantly observe the phrase structure of a composed language. Take into consideration "pink tunes" as an instance. Considering that both the possibility of "pink" as an adjective as well as the likelihood of an adjective coming before a noun are reasonably high, typical POS taggers will wrongly identify "pink" in "pink tunes" as an adjective.

Challenge 4 (Ambiguous Instance):"review harry potter (publication)" vs. "watch harry potter (flick)" vs. "age harry potter (personality)" A circumstances (e.g., "harry potter") can come from numerous ideas (e.g., publication, flick, personality, and so on). We can get such one-to-many mappings in between circumstances as well as ideas straight from existing understanding bases. Nevertheless, circumstances could describe various principles when context differs. Some approaches try to get rid of circumstances uncertainty based upon comparable or associated circumstances, however the variety of circumstances that can be recovered from a short message is normally restricted, making these approaches inapplicable to circumstances disambiguation simply put messages. We observe that terms, such as verbs, adjectives, as well as associates, can additionally aid with circumstances disambiguation. As an example, "harry potter" is a publication in "review harry potter", a motion picture in "watch harry potter", as well as a personality in "age harry potter". Human beings can efficiently acknowledge one of the most suitable ideas for a circumstance within a particular short message, considering that we have the understanding regarding semantic relatedness in between numerous sorts of terms. Nonetheless, it is nontrivial for equipments to disambiguate circumstances without such understanding.

Challenge 5 (Enormous Volume): Compared to papers, short messages are created in a much bigger quantity. As an example, Google, as one of the most commonly utilized internet search engine since 2014, gotten over 3 billion search questions daily1. Twitter likewise reported in 2012 that it brought in greater than 100 million customers that published 340 million tweets per day2. For that reason, a practical structure for short message understanding must have the ability to manage short messages in actual time. Nevertheless, a short message can have 10s of feasible divisions, a term can be identified with numerous kinds, as well as a circumstances can describe numerous ideas. For this reason, it is very taxing to get rid of these obscurities and also accomplish the very best semantic analysis for a short message.

II. LITERATURE WORK

In this section, we discuss related work in three aspects: text segmentation, POS tagging, and semantic labeling.

A. Text segmentation: We think about message division regarding separate a message right into a series of terms. Existing techniques can be identified right into 2 groups: analytical techniques as well as vocabulary based strategies. Analytical strategies, such as N-gram Design, compute the regularities of words co-occurring as next-door neighbors in a training corpus. When the regularity surpasses a predefined limit, the matching bordering words can be dealt with as a term. Vocabulary-based techniques remove terms in a streaming way by looking for presence or regularity of a term in a predefined vocabulary. Particularly, the Longest Cover approach, which is widely-adopted for message division as a result of its simplicity as well as real-time nature, look for lengthiest terms, had in a vocabulary while checking the message. One of the most evident downside of existing approaches for message division is that they just take into consideration surface area functions and also overlook the demand of semantic comprehensibility within division. This will certainly cause wrong divisions in

instances such as "getaway april in paris" defined in Difficulty 1. To this end, we suggest to manipulate context semiotics when performing message division.

B. POS Tagging: POS tagging establishes lexical kinds (i.e., POS tags) of words in a message. Mainstream POS labeling formulas fall under 2 groups: rule-based methods as well as analytical techniques. Rule-based POS taggers try to appoint POS tags to unidentified or unclear words based upon a lot of handmade [8] [9] or immediately found out etymological regulations. Analytical POS taggers stay clear of the price of building identifying regulations by constructing an analytical version immediately from a corpora as well as labeling untagged messages based upon those discovered analytical info. The majority of the widely-adopted analytical methods use the popular Markov Design which discovers both lexical likelihoods (P(tag jword)) and also consecutive chances (P(tagijtagi1; tagi2,..., tagin)) from an identified corpora and also tags a brand-new sentence by looking for tag series that makes the most of the mix of lexical and also consecutive likelihoods. Keep in mind that both rule-based as well as analytical strategies to POS tagging count on the presumption that messages are appropriately structured. Simply put, messages ought to please identifying guidelines or consecutive relationships in between successive tags. Nevertheless, this is not constantly the instance for short messages. Extra notably, every one of the abovementioned job just thinks about lexical attributes as well as overlooks word semiotics. This will certainly result in blunders occasionally, as highlighted when it comes to "pink tracks" explained in Obstacle 3. Our job tries to develop a tagger which takes into consideration both lexical attributes as well as underlying semiotics for kind discovery.

C. Semantic Labeling: Semantic labeling finds concealed semiotics from an all-natural language message. According to the depiction of semiotics, existing service semantic labeling can be about categorized right into 3 classifications, particularly called entity acknowledgment (NER), subject modeling, and also entity connecting. NER situates called entities in a message as well as identifies them right into predefined groups (e.g., individuals, companies, places, times, amounts and also percents, and so on) making use of etymological grammar-based strategies along with analytical designs like CRF [1] and also HMM [2] Subject designs [3] [4] effort to acknowledge "unrealized subjects", which are stood for as probabilistic circulations on words, based upon evident analytical relationships in between messages and also words. Entity connecting utilizes existing understanding bases and also concentrates on getting "explict subjects" revealed as probabilistic circulations on the whole knowledgebase. In spite of the high precision that has actually been accomplished by existing deal with semantic labeling, there are still some constraints. Initially, groups, "concealed subjects", in addition to "specific subjects" are various from human-understandable ideas. Second, short messages do not constantly observe the phrase structure of a composed language which, nevertheless, is an important attribute made use of in mainstream NER devices. Third, short messages normally do not include adequate web content to sustain analytical versions like subject designs. The job most pertaining to ours are performed by Tune et alia as well as Kim et al. [8] specifically, which additionally stand for semiotics as principles. Utilizes the Bayesian Reasoning system to conceive circumstances and also short messages, and also gets rid of circumstances uncertainty based upon uniform circumstances. Catches semantic relatedness in between circumstances making use of a probabilistic subject version (i.e., LDA), as well as disambiguates circumstances based upon associated circumstances. In this job, we observe that terms, such as verbs, adjectives, as well as connects, can likewise assist with circumstances disambiguation. As a result, we include kind discovery right into our structure for short message understanding and also carry out circumstances disambiguation based upon different sorts of context details.

III. STATISTICAL MODEL FOR TEXT SEGMENTATION

We recommend an analytical approach that locates the maximum-probability division of a provided message. This technique does not need training information since it approximates likelihoods from the offered message. As a result, it can be put on any type of message in any type of domain name. An experiment revealed that the approach is a lot more precise than or a minimum of as exact as an advanced message division system. Papers normally consist of different subjects. Recognizing as well as separating subjects by separating records, which is called message division, is very important for several all-natural language handling jobs, consisting of details access (Hearst and also Plaunt, 1993; Salton et al., 1996) and also summarization (Kan et al., 1998; Nakao, 2000). In details access, customers are frequently curious about specific subjects (components) of gotten files, as opposed to the records themselves. To fulfill such demands, files ought to be fractional right into systematic subjects. Summarization is usually utilized for a lengthy file that consists of several subjects. A recap of such a paper can be made up of recaps of the element subjects. Recognition of subjects is the job of message division. A great deal of study has actually been done on message division (Kozima, 1993; Hearst, 1994; Okumura and also Honda, 1994; Salton et al., 1996; Yaari, 1997; Kan et al., 1998; Choi, 2000; Nakao, 2000). A significant feature of the approaches utilized in this study is that they do not call for training information to sector offered messages. Hearst (1994), as an example, made use of just the resemblance of word circulations in an offered message to section the message. Subsequently, these approaches can be put on any type of message in any type of domain name, also if training information do not exist. This residential or commercial property is necessary when message division is related to details access or summarization, since both jobs handle

domain-independent records. We initially specify the likelihood of a division of a provided message in this area. In the following area, we after that define the formula for picking one of the most likely divisions.

Let W= w1 w2,….. w_n be a text consisting of n words, and let S= s1 s2,….. s_m be a segmentation of W consisting of m segments. Then the probability of the segmentation S is defined by,

$$Pr(S|W) = \frac{Pr(W|S)\,|Pr(S)}{Pr(w)}$$

The most likely segmentation S is given S= args max Pr (W|S) and Pr(S), because Pr (W) is a constant for a given text W.

The definitions of Pr (W|S) and Pr(S) are given below, in that order:

Definition of Pr (W|S) - We characterize a point by the dissemination of words in that theme. We accept that distinctive points have diverse word appropriations. We additionally accept that distinctive subjects are factually free of one another. We likewise expect that the words inside the extent of a theme are measurably autonomous of one another given the subject.

Definition of Pr (S) - The definition of Pr (S) can change contingent upon our earlier data about the likelihood of division S. Forinstance, we may know the normal length of the sections and need to join into Pr (S). Our suspicion, in any case, is that we don't have such earlier data. Along these lines, we need to utilize some uninformative earlier likelihood.

Algorithm for Finding the Maximum-Probability Segmentation - This section describes an algorithm for finding the minimum-cost segmentation. First, we define the terms and symbols used to describe the algorithm.

Given a text W= $w_1,w_2,\ldots.w_n$ consisting of n words, we define g_i as the position between w_i and w_{i+1}, so that g_0 is just before w1 and g_n is just after w_n.

Next, we define a graph G= (V, E), where V is a set of nodes and E is a set of edges. V is defined as

V= {gi|0≤i≤n}

And E is defined as

E= {eij|0≤i<j≤n} where the edges are ordered; the initial vertex and terminal vertex of eij are giand g_j, respectively. We say that e_{ij} covers w_{i+1}, $w_{i+2},\ldots.w_j$. This means that e_{ij} is showing a segment w_{i+1}, $w_{i+2},\ldots.w_j$. Thus we define the cost c_{ij} of edge e_{ij}.

Given these definitions, we describe the algorithm to find the minimum-cost segmentation or maximum-probability segmentation as follows:

Step 1: Calculate the cost C_{ij} of edge e_{ij} for $0 \le i < j \le n$

Step 2: locate the minimal price course from g0 to gn. Policies for locating the minimal expense course in a graph are popular. A formula that can provide a solution for step 2 will certainly be a much less intricate adjustment of the estimation made use of to find one of the most severe probability plan in Japanese morphological evaluation (Nagata, 1994). By doing this, a response can be gotten by using vibrant shows (DP) formula. DP computations have actually in addition been made use of for material department by various researchers (Ponte and also Croft, 1997; Heinonen, 1998). The method by doing this obtained talks with the minimal expense sections in when sides connect with sectors. The computation subsequently chooses the variety of sectors. However, the variety of areas can also be suggested specifically by identifying the variety of sides in the minimal expense rub. The formula makes it possible for the material to be split at any type of area in between words; i.e., each of the scenarios in between words is opportunity for division borders. It is easy, nevertheless, to change the estimation so the web content have to be split at particular settings, for instance, the final thought of sentences or flows. We make use of simply the sides whose hidden and also incurable vertices are prospect borders that satisfy certain problems, for instance, being the closures of sentences or areas. We then obtain the minimal expense course by doing steps 1 and also 2. The minimal price division by doing this obtained satisfies the limit problems. In this paper, we anticipate that the sector limits go to the closures of sentences.

IV. POS TAGGING

Automatic Tagging: We presently swing to the Automatic Tagging programs which frame the core of the undertaking, and comprise its primary commitment to look into. The process of automatic tagging can itself be separated into three sensibly distinguishable procedures:

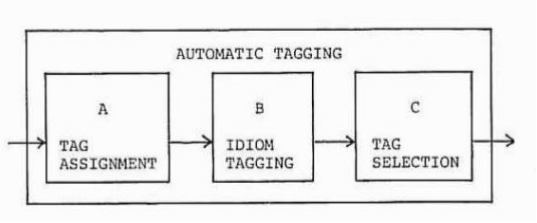

Figure 2: Automatic tagging

For improvement purposes, it was helpful to compose a different program for every one of these three processes; [8] it would be simple enough on a basic level to consolidate them all into a solitary program. Consistently, the Automatic Tagging partitions into Tag Assignment (whereby each word in the corpus is allotted at least one conceivable labels), and Tag Selection (whereby a solitary tag is chosen as the right one in setting, from the at least one choice created by Tag Assignment). It was as something of a reconsideration that we added to the Tag Assignment program (WORDTAG) and the Tag Selection program (CHAINPROBS) a third, middle of the road program (IDIOMTAG) to manage different syntactically atypical word-successions which, without aiming any specialized use of the term, we may call "parts of speech".

Tag Assignment: The least complex sort of Tag Assignment strategy would be only a query in a WORDLIST or lexicon indicating the tag(s) related with

each word. Notwithstanding such a Wordlist, the Brown Tagging Program TAGGIT has a SUFFIXLIST, or rundown of pairings of word-endings and labels (for instance, the closure - NESS is related with things). We pursue Brown in this, utilizing a Wordlist of more than 7000 words, and a Suffix list of around 660 word-endings. Further, the LOB Assignment Program contains various techniques for managing words containing, hyphens, words starting with a capital letter, words finishing with - X, with 'S, and so on. The upsides of having a SUFFIXLIST are that (a) the WORDLIST can be abbreviated, since words whose word class is unsurprising from their closure can be excluded from it; and (b) the arrangement of words acknowledged by the program would be able to open-finished, and can even incorporate neologisms, uncommon words, jabber words, and so on. These focal points likewise apply to the systems for managing hyphenated and uppercase words.

Tag Selection: On the off chance that one a player in the undertaking can be said to have made a specific commitment to programmed dialect preparing, it is the Tag Selection Program (CHAINPROBS, the structure of which is depicted in more prominent detail in Marshall (1982). This program works on a rule very unique in relation to that of the Tag Selection part of the program utilized on the Brown Corpus. The Brown program utilized an arrangement of CONTEXT FRAME RULES, which killed labels on the current word in the event that they were contrary with tags on the words inside a range of two to one side or two to one side of the current word (W). Along these lines expecting a succession of words - 2, - 1, W, +l, +2, an endeavor was made to disambiguate W on the proof of labels as of now unambiguously appointed to words - 2, - 1, +l, or +2. The principles worked just in the event that at least one of these words were unambiguously labeled, and thusly regularly bombed on groupings of vague words. Also, the same number of as 80% of the uses of the Context Frame Rules made utilization of just a single word to one side or to one side of W. These observations, made by running the Brown Program over piece of the LOB Corpus, driven us to create, as a model of the LOB Tag-Selection Program, a program which registers transitional probabilities between one tag and the following for all mixes or conceivable labels, end picks the in all likelihood path through an arrangement of vague labels on this premise.

V. SEMANTIC LABELING

In this area we think about the issue of displaying content corpora and different accumulations of discrete information. The objective is to discover short portrayals of the individuals from an accumulation that empower effective handling of huge accumulations while protecting the fundamental measurable connections that are helpful for essential assignments, for example, order, oddity location, synopsis, and comparability and importance judgments. Huge advancement has been made on this issue by analysts in the field of data recovery (IR) (Baeza-Yates and Ribeiro-Neto, 1999). The fundamental strategy proposed by IR analysts for content corpora a technique effectively sent in current Internet web crawlers decreases each archive in the corpus to a vector of genuine numbers, every one of which speaks to proportions of tallies. In the mainstream tf-idf plot (Salton and McGill, 1983), an essential vocabulary of "words" or "terms" is picked, and, for each report in the corpus, a tally is framed of the quantity of events of each word. After appropriate standardization, this term recurrence check is contrasted with a backwards record recurrence tally, which estimates the quantity of events of a word in the whole corpus (by and large on a log scale, and again reasonably standardized). The final product is a term-by-report grid X whose sections contain the tf-idf esteems for every one of the records in the corpus. In this way the tf-idf conspire decreases reports of discretionary length to settled length arrangements of numbers. While the tf-idf decrease makes them bid includes prominently in its fundamental ID of sets of words that are discriminative for records in the gathering the methodology likewise gives a generally little measure of decrease in depiction length and uncovers little in the method for between or intradocument factual structure. To address these inadequacies, IR analysts have proposed a few other dimensionality decrease procedures, most prominently inactive semantic ordering (LSI) (Deerwester et al., 1990). LSI utilizes a solitary esteem decay of the X lattice to distinguish a direct subspace in the space of tf-idf includes that catches the vast majority of the difference in the gathering. This methodology can accomplish noteworthy pressure in vast accumulations. Moreover, Deerwester et al. contend that the determined highlights of LSI, which are straight blends of the first tf-idf highlights, can catch a few parts of fundamental etymological thoughts, for example, synonymy and polysemy. To substantiate the cases with respect to LSI, and to consider its relative qualities and shortcomings, it is helpful to build up a generative probabilistic model of content corpora and to examine the capacity of LSI to recuperate parts of the generative model from information (Papadimitriou et al., 1998). Given a generative model of content, be that as it may, it isn't clear why one ought to embrace the LSI technique one can endeavor to continue all the more specifically, fitting the model to information utilizing most extreme probability or Bayesian strategies. A huge advance forward in such manner was made by Hofmann (1999), who displayed the probabilistic LSI (pLSI) demonstrate, otherwise called the perspective model, as an option in contrast to LSI. In the pLSI approach models each word in a report as an example from a blend demonstrate, where the blend segments are multinomial arbitrary factors that can be seen as portrayals of "subjects." Thus each word is created from a solitary point, and distinctive words in a record might be produced from various themes. Each report is spoken to as a rundown of blending extents for these blend segments and in this way diminished to a likelihood circulation on a settled arrangement of themes. This dissemination is the "lessened depiction" related with the archive.

VI. CONCLUSION

In this script, we suggest a generalized structure to comprehend short messages properly as well as effectively. A lot more particularly, we split the job of short message understanding right into 3 subtasks: text segmentation, type detection, and concept labeling. We develop text segmentation as a heavy Ultimate Inner circle trouble, as well as recommend a randomized estimation formula to keep precision and also boost effectiveness at the very same time. We present a Chain Version as well as a set smart Version which incorporate lexical as well as semantic attributes to perform type detection. They accomplish much better precision than standard POS taggers on the identified criteria. We use a Heavy Ballot formula to identify one of the most ideal semiotics for circumstances when obscurity is discovered. The speculative outcomes show that our suggested structure outshines existing modern techniques in the area of short message understanding. As a future job, we try to assess as well as include the influence of spatial-temporal attributes right into our structure for short message understanding.

VII. REFERENCES

[1]. 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.

[2]. N. Mishra, R. Saha Roy, N. Ganguly, S. Laxman, and M. Choudhury,"Unsupervised query segmentation using only query logs," in Proceedings of the 20th International Conference Companion on World Wide Web, ser. WWW '11, 2011, pp. 91–92.

[3]. M. Hagen, M. Potthast, B. Stein, and C. Br¨autigam, "Query segmentation revisited," in Proceedings of the 20th International Conference on World Wide Web, ser. WWW '11, New York, NY, USA, 2011, pp. 97–106.

[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 Proceedings of the35th International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '12, New York, NY, USA, 2012, pp.721–730.

[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 Proceedings of the 22nd International Conference on World Wide Web, ser. WWW '13 Companion, Republic and Canton of Geneva, Switzerland, 2013, pp. 597–604.

[6]. 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.

[7]. 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.

[8]. 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.

[9]. G. Zhou and J. Su, "Named entity recognition using an hmm-based chunktagger," in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ser. ACL '02, Stroudsburg, PA, USA, 2002,pp. 473–480.

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

[11]. 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.

[12]. 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, NewYork, NY, USA, 2007, pp. 233–242.

[13]. 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.

[14]. K. W. Church, "A stochastic parts program and noun phrase parser for unrestricted text," in Proceedings of the second conference on Applied natural language processing, ser. ANLC '88, Stroudsburg, PA, USA,1988, pp. 136–143.

[15]. S. J. DeRose, "Grammatical category disambiguation by statistical optimization,"Comput. Linguist, vol. 14, no. 1, pp. 31–39, 1988.