Short Text Conceptualization using a Probabilistic Knowledgebase

Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, Weizhu Chen
Microsoft Research Asia
Beijing, China
{yangqiu.song,haixun.wang,zhy.wang,hongsli,wzchen}@microsoft.com

Abstract

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 knowledgebases (e.g., WordNet, Freebase and Wikipedia), our approach brings significant improvements in short text understanding as reflected by the clustering accuracy.

1 Introduction

Psychologist Gregory Murphy began his highly acclaimed book [Murphy, 2004] with the statement *"Concepts are the glue that holds our mental world together"*. Still, Nature magazine book review calls it an understatement, because *"Without concepts, there would be no mental world in the first place"* [Bloom, 2003]. Doubtless to say, the ability to conceptualize is a defining characteristic of humanity.

We focus on conceptualizing from texts or words. For example, given the word "India," a person will form in his mind concepts such as *country* or *region*. Given two words, "India" and "China," the top concepts may shift to *Asian country* or *developing country*, etc. Given yet another word, "Brazil," the top concepts may change to *BRIC* or *emerging market*, etc. Besides generalizing from instances to concepts, humans also form concepts from descriptions. For example, given words "body," "smell" and "color," the concept of *wine* comes into our mind. Certainly, instances and descriptions may mix, for example, we conceptualize {"apple," "headquarter"} to *company*, but {"apple," "smell," "color"} to *fruit*.

The problem is, can machines do it? Much work has been devoted to topic discovery from text. Statistical approaches such as topic models [Blei and Lafferty, 2009] treat text as a bag of words in a vector space, and discover "latent topics" from the text. But finding latent topics is not tantamount to understanding. A latent topic is represented by a set of words. Machines are not able to grasp the concepts behind these words, nor do they know the properties and relationships associated with the concepts. In particular, using statistical approaches to find topics from short text (search queries, twitter messages, etc.) is often infeasible, as short text does not have enough content from which we can build a statistically meaningful model.

Recent work on explicit semantic analysis [Gabrilovich and Markovitch, 2006; Egozi *et al.*, 2008] focuses on finding *explicit* rather than latent topics. Instead of using a bag of words to represent a topic, the new approach uses a bag of Wikipedia articles, or a distribution over the entire set of Wikipedia articles, to represent a topic. The approach represents a big step forward in promoting semantics in text mining, however, there is still a big difference between a bag of Wikipedia articles and a clear concept in our mental world.

In this paper, we propose a probabilistic framework, which includes a knowledgebase and certain inferencing techniques on top of the knowledgebase, to enable machines to perform human-like conceptualization. The knowledgebase, known as Probase [Wang, 2011; Wu *et al.*, 2011], contains concepts (of worldly facts) that are as rich as those in our mental world. Probase scans billions of documents to obtain millions of concepts, and for each concept, it finds instances and attributes that make the concept concrete. Moreover, Probase scores the concepts, instances, attributes and their relationships. In our work, we use these scores as priors and likelihoods for various statistical inferencing over the text data. Finally, we develop a model which enables us to derive the most likely concepts from a set of words or a short piece of text.

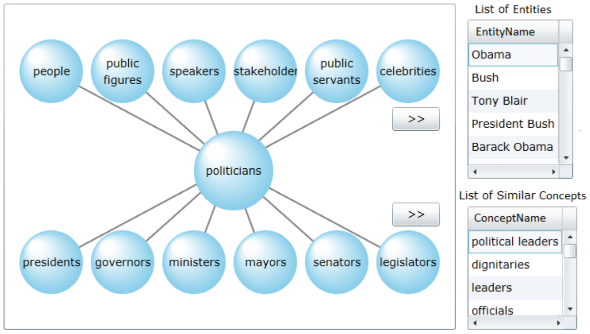
The rest of the paper is organized as follows. Section 2 introduces Probase – a probabilistic knowledgebase that captures concepts in human minds. Section 3 shows how machines can conceptualize. Section 4 uses conceptualization for clustering. We discuss related work in Section 5 and conclude in Section 6.

2 A Knowledgebase of Many Concepts

To enable machines to understand human concepts, we need a knowledgebase. WordNet [Fellbaum, 1998], Wikipedia, Cyc [Lenat and Guha, 1989] and Freebase [Bollacker *et al.*, 2008] are created by human experts or community efforts. Recently, much work has been devoted to building knowledgebases automatically. Representative systems include KnowItAll [Etzioni *et al.*, 2004], TextRunner [Banko *et al.*, 2007], WikiTaxonomy [Ponzetto and Strube, 2007], and YAGO [Suchanek *et al.*, 2007].

Existing knowledgebases fall short of supporting machines to perform human-like conceptualization. There are two major obstacles. First, the scale and scope of the knowledgebases is not big enough. For example, Freebase has about 2,000 categories, and Cyc, after 25 years of efforts, has 120,000 categories. In other words, they are limited in coverage and granularity in representing concepts in our mental world. Second, most of these knowledgebases are deterministic instead of probabilistic. This means, for example, although we can find the concepts that a term may belong to, it is not possible to find which concept is the most typical concept for that term.

We built Probase [Wang, 2011; Wu *et al.*, 2011][1], a taxonomy that contains millions of concepts learned iteratively from 1.68 billion web pages in Bing's web repository. The core taxonomy consists of the *isa* relationships extracted by using syntactic patterns including the Hearst patterns [Hearst, 1992]. For example, we consider "... artists such as Pablo Picasso ..." as a piece of evidence for the claim that "Pablo Picasso" is an instance of the concept *artist*. Next, given a concept C, we use syntactic patterns such as "What is the A of B" to find its attributes (where B is an instance of C, and A is the attribute we are after). For example, sentences such as "What is the capital of China?" and "What is the GDP of Japan?" suggest that "capital" and "GDP" are candidate attributes of concept *country*. Furthermore, every claim in Probase is associated with a few scores that model the consensus, typicality, ambiguity, and other characteristics of the claim. Finally, we expand the taxonomy to include other relationships, and one of the most important relationships is the "similar" relationship between concepts. Figure 1 shows the top super-concepts and sub-concepts of the concept *politicians*, as well as instances of *politicians*, and concepts that are similar to the concept of *politicians*. The current version of the taxonomy contains about 8 million concepts and 17 million instances. We refer the readers to [Wu *et al.*, 2011] for a detailed description of how the taxonomy is constructed.



List of Entities

EntityName
Obama
Bush
Tony Blair
President Bush
Barack Obama

List of Similar Concepts

ConceptName
political leaders
dignitaries
leaders
officials

Figure 1: Browsing the Probase Taxonomy.

In Table 1, we compare Probase with a few other taxonomies, including WordNet, Wikipedia, and Freebase. WordNet specializes in the linguistics of English words. For the word "cat," WordNet has detailed descriptions of its various senses, although many of them are rarely used, or even unknown to many people (e.g., *gossip* and *woman* as concepts for "cat"). Also, it does not contain information for entities such as "IBM," which is not considered as a word. Wikipedia and Freebase, on the other hand, contain limited number of concepts for the word "cat." In fact, the categories there are biased and sometimes inaccurate. For example, Freebase's concept space is biased toward entertainment, media related concepts. More importantly, the categories in WordNet, Wikipedia, and Freebase are not ranked or scored, and users cannot tell the difference in terms of their importance or typicality. In comparison, the concepts in Probase are more consistent with human's common knowledge. Concepts such as *gossip* and *woman* for "cat" are either not included or ranked very low because people rarely associate them with "cat." In addition, for a word such as "language," Probase indicates it can be both an instance on its own or an attribute of some concepts. Thus, Probase provides additional information that is not available from WordNet, Wikipedia, or Freebase.

3 Conceptualization

We first introduce a method to infer concepts from a set of instances, or a set of attributes. Then, we discuss how to handle cases where instances and attributes are mixed. In the rest of the section, we use e to denote an instance, a an attribute, and c a concept. Our problem is to identify candidate concepts $C = \{c_k, k \in 1, ..., K\}$ ranked by their likelihood when we observe a set of instances $E = \{e_i, i \in 1, ..., M\}$, or a set of attributes $A = \{a_j, j \in 1, ..., N\}$, or a set of terms of unknown types $T = \{t_l, l \in 1, \ldots, L\}$.

3.1 Conceptualizing Instances

There are tens of millions of concept-instance pairs in Probase. Given a set of observed instances $E = \{e_i, i \in 1, ..., M\}$, we want to abstract a set of most representative concepts that can best describe the instances. We estimate the probability of concepts using a naive Bayes model:

$$P(c_k|E) = \frac{P(E|c_k)P(c_k)}{P(E)} \propto P(c_k)\prod_{i=1}^{M} P(e_i|c_k). \quad (1)$$

In this case, the concept with the largest posterior probability is ranked as the most possible concept to describe the observed instances. For example, given instances "China," "Russia," "India," and "USA," the posterior suggests *country* as a concept, while given "China," "Indian," and "Russia," it will suggest *emerging market* as the top match.

[1] http://research.microsoft.com/probase/

Table 1: Comparison between different knowledgebases.

Term	WordNet Hypernyms	Wikipedia Categories	Freebase Types	Probase Concepts
Cat	Feline; Felid; Adult male; Man; Gossip; Gossiper; Gossipmonger; Rumormonger; Rumourmonger; Newsmonger; Woman; Adult female; Stimulant; Stimulant drug; Excitant; Tracked vehicle; ...	Domesticated animals; Cats; Felines; Invasive animal species; Cosmopolitan species; Sequenced genomes; Animals described in 1758;	TV episode; Creative work; Musical recording; Organism classification; Dated location; Musical release; Book; Musical album; Film character; Publication; Character species; Top level domain; Animal; Domesticated animal; ...	Animal; Pet; Species; Mammal; Small animal; Thing; Mammalian species; Small pet; Animal species; Carnivore; Domesticated animal; Companion animal; Exotic pet; Vertebrate; ...
IBM	N/A	Companies listed on the New York Stock Exchange; IBM; Cloud computing providers; Companies based in Westchester County, New York; Multinational companies; Software companies of the United States; Top 100 US Federal Contractors; ...	Business operation; Issuer; Literature subject; Venture investor; Competitor; Software developer; Architectural structure owner; Website owner; Programming language designer; Computer manufacturer/brand; Customer; Operating system developer; Processor manufacturer; ...	Company; Vendor; Client; Corporation; Organization; Manufacturer; Industry leader; Firm; Brand; Partner; Large company; Fortune 500 company; Technology company; Supplier; Software vendor; Global company; Technology company; ...
Language	Communication; Auditory communication; Word; Higher cognitive process; Faculty; Mental faculty; Module; Text; Textual matter;	Languages; Linguistics; Human communication; Human skills; Wikipedia articles with ASCII art	Employer; Written work; Musical recording; Musical artist; Musical album; Literature subject; Query; Periodical; Type profile; Journal; Quotation subject; Type/domain equivalent topic; Broadcast genre; Periodical subject; Video game content descriptor; ...	**Instance of**: Cognitive function; Knowledge; Cultural factor; Cultural barrier; Cognitive process; Cognitive ability; Cultural difference; Ability; Characteristic; **Attribute of**: Film; Area; Book; Publication; Magazine; Country; Work; Program; Media; City; ...

In Eq. (1), the probability of an instance given a concept is computed as:

$$P(e_i|c_k) = \frac{P(e_i, c_k)}{P(c_k)} \tag{2}$$

where $P(e_i, c_k)$ is proportional to the co-occurrence of instances and concepts, and $P(c_k)$ is approximately proportional to the observed frequency of c_k. In Eq. (2), Laplace smoothing [Lidstone, 1920] is used to filter out noise and introduce concept diversities.

3.2 Conceptualizing Attributes

When observing a set of attributes such as "population," "language," and "currency," we assume with a high probability that they are talking about *country*, even though no specific country names is observed.

To achieve this, we use the same naive Bayes inference method and the similar independence assumption used in Eq.(1) to derive the most probable concepts:

$$P(c_k|A) = \frac{P(A|c_k)P(c_k)}{P(A)} \propto P(c_k) \prod_{j=1}^{N} P(a_j|c_k), \tag{3}$$

It means the concept is determined by the posterior given all the observed attributes.

Unfortunately, Probase does not have direct information about $P(a_j|c_k)$. Probase finds attributes for a concept through its instances. Thus, the inference of relationships between attributes and a concept should be intermediated through the instances of the concept as well. Therefore, by applying the Bayes chain rule, we derive the likelihood of concept in Eq. (3) by:

$$P(a_j|c_k) = \sum_{i:e_i \in E} P(a_j|e_i)P(e_i|c_k), \tag{4}$$

where E is the set of instances that are related to attribute a_j and concept c_k. Similar to Eq. (2), the conditional probability is computed based on the co-occurrence counts of attribute-instance and instance-concept relationships. In addition, smoothing is added to guarantee obtaining more meaningful results.

3.3 Mixture Models

A more common case of conceptualizing occurs when we observe a set of terms, $T = \{t_l, l \in 1, \ldots, L\}$, but do not know which t_l is an instance and which is an attribute. Many terms can be attributes and instances at the same time. For example, "population" can be an attribute of *country*, but it can also be an instance of *geographical data*.

We introduce an auxiliary variables z_l to indicate the status of term t_l. Specifically, $z_l = 1$ if t_l is an instance, and $z_l = 0$ if t_l is an attribute. An intuitive way to handle terms with different types is to use a generative mixture model, which assumes that

$$P(t_l|c_k) = P(t_l|z_l = 1, c_k)P(z_l = 1|c_k) + P(t_l|z_l = 0, c_k)P(z_l = 0|c_k) \quad (5)$$

where $P(t_l|z_l = 1, c_k) = P(e_l|c_k)$, i.e., term t_l is regarded as an instance e_l when $z_l = 1$; and $P(t_l|z_l = 0, c_k) = P(a_l|c_k)$, i.e., term t_l is regarded as an attribute a_l when $z_l = 0$, and $P(z_l = 1|c_k)$ and $P(z_l = 0|c_k)$ are the prior probabilities of the type of t_l in concept c_k. Then, the concept posterior is given by:

$$P(c_k|T) = \frac{P(T|c_k)P(c_k)}{P(T)} \propto P(c_k)\prod_{l}^{L} P(t_l|c_k). \quad (6)$$

However, in practice, it is rare that a term is both an instance and an attribute of the same concept. Given that the knowledgebase contains noises, we handle the logic in a discriminative manner. We introduce a noisy-or model to first infer the probability $P(c_k|t_l)$:

$$P(c_k|t_l) = 1 - (1 - P(c_k|t_l, z_l = 1))(1 - P(c_k|t_l, z_l = 0)) \quad (7)$$

Intuitively, it means term t_l invokes concept c_k if it is an instance of c_k or it is an attribute of c_k. Here, we have $P(c_k|t_l, z_l = 1) = P(c_k|e_l) = P(c_k, e_l)/P(e_l)$ where term t_l is regarded as an instance e_l, and $P(c_k|t_l, z_l = 0) = P(c_k|a_l) = \sum_{i:e_i \in E} P(c_k|e_i)P(e_i|a_l)$ where term t_l is regarded as an attribute a_l, and E is the set of instances that are related to attribute a_l and concept c_k. Then, using the naive Bayes rule, we derive the concept posterior given a set of terms by:

$$P(c_k|T) \propto P(c_k)\prod_{l}^{L} P(t_l|c_k) \propto \frac{\prod_l P(c_k|t_l)}{P(c_k)^{L-1}} \quad (8)$$

where $P(c_k|t_l)$ is given by Eq. (7). It can be proved that when all the terms can only have one type, Eq. (8) is identical to Eq. (6).

The generative model and the discriminative model make different assumptions. The generative model assumes that a term can be both instance and attribute of a specific concept, with some probability. Discriminative model assumes that if we observe that the term is either an instance or an attribute of a concept, it will have a high probability of belonging to that concept. More detailed comparison of these models can be found in [Chen, 2006].

3.4 Multiple Concepts

A set of terms may contain multiple classes of unrelated concepts. For instance, the set {China, Brazil, Russia, apple, banana, BBC, New York Time} contains 3 obvious concepts that are not related. To find the 3 concepts, we represent the retrieved concepts and the terms as a bipartite graph, and then we perform a simple co-clustering of concepts and terms by identifying the disjoint cliques. Since we know what are the instances and attributes related to each concept, we use heuristic rules to obtain the disjoint cliques sequentially after ranking the concepts.

3.5 Examples of Conceptualization

We give our system a set of terms. Each term is associated with a type: instance, or attribute, or unknown. Fig. 2 shows the resulting ranked list of concepts. As we can see, similar sets of terms may produce quite different concepts. For {"China," "Russia," "India," "Brazil"}, the top concepts are *emerging market* and *BRIC country*. For {"China," "India," "Japan," "Singapore"}, the results are *asian*-related concepts. This shows that the system captures the subtlety in the input. For a set of attributes {"population," "location," "president"}, the inferred concepts in Fig. 2(c) are consistent with human intuition. For terms of unknown-types, {"California," "Florida," "population"}, the retrieved concepts in Fig. 2(d) also make sense. Fig. 2(e) shows a case where the input contains multiple unrelated concepts.

4 Clustering Short Texts

We use short text conceptualization to cluster Twitter messages. We collected 605,501 tweet messages. We pre-process the tweets to detect Probase entities first. When multiple entities can be detected from a single piece of text, we choose the longest entity. For example, we obtain "President Barack Obama" as an entity instead of "President", "Barack Obama" or "Obama", although we have these terms in the knowledgebase. For performance, we build a trie to index the terms. Several examples of the tweets and corresponding concepts are shown in Table 2.

4.1 Problem Definition

Because tweets data has no ground-truth labels, to evaluate the effectiveness of conceptualizing, we design clustering problems in the following way. We collect tweets using some hashtag keys, and then we group the tweets into several categories based on the keys. We use different methods to extract features from the tweets (see below for a description of the methods), and use K-means to cluster the tweets based on extracted features. We define two clustering problems.

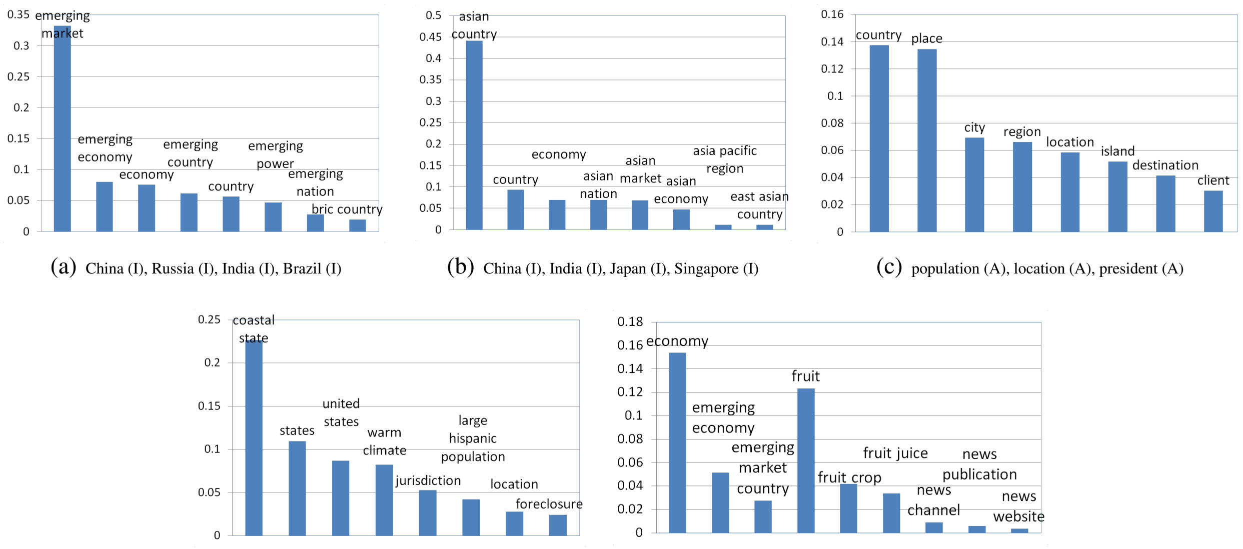
Problem 1 (unique concepts): We use the following hashtag keywords to retrieve tweets in 3 categories:

1. Microsoft, Yahoo, Google, IBM, Facebook
2. cat, dog, fish, pet, bird
3. Brazil, China, Russia, India

We obtained $5,613$ tweets ($2,056$, $1,813$, and $1,744$ tweets in the 3 categories). All together there are $51,743$ tokens after removing URLs and stop-words, and the size of vocabulary is $2,134$.

Problem 2 (concepts with subtle differences): We use the following hashtag keywords to retrieve tweets in 4 categories:

1. United states, American, Canada
2. Malaysia, China, Singapore, India, Thailand, Korea
3. Angola, Egypt, Sudan, Zambia, Chad, Gambia, Congo
4. Belgium, Finland, France, Germany, Greece, Spain, Switzerland



(a) China (I), Russia (I), India (I), Brazil (I)

(b) China (I), India (I), Japan (I), Singapore (I)

(c) population (A), location (A), president (A)

(d) California (U), Florida (U), population (U)

(e) China (U), Brazil (U), Russia (U), apple (U), banana (U), BBC (U), New York Time (U)

Figure 2: Term conceptualization examples. (I: known type as instance. A: known type as attribute. U: unknown type.)

Table 2: Twitter Conceptualization Examples.

Tweets	Concepts
Google: CEO Eric Schmidt Steps Down, Co-Founder Larry Page To Take Over.	(1) "Google": search engine; company; top search engine; competitor ... (2) "CEO": company; position; role; title; senior executive; leader; director ... (3) "Eric Schmidt": top person; speaker; executive; corporate leader; successful person ... (4) "co-founder": top official; executive-level position; leadership; angel investor ... (5) "Larry Page": top executive; person; investor; smart person; successful person ...
Facebook is the place to connect, but Twitter is the place to create new relationships,	(1) "Facebook": social networking site; social media; website; service; social media platform ... (2) "place": circumstance; factor; event; environmental factor; criterion ... (3) "Twitter": social networking site; social media; service; platform; social networking website ... (4) "new relationships": life change; serious issue; sensitive topic; challenge ...
US economy is growing again, but not fast enough, says President Barack Obama.	(1) "US": country; market; currency; nation; region; western country; economy ... (2) "economy": location; field; territory ... (3) "President Barack Obama": leader; democrats; politician; official; democratic leader; federal official; celebrity; national leader ...
House Republicans have repealed Obama's healthcare reform law. Now what?	(1) "Obama": democrats; politician; leader; candidate; president; senator; supporter ... (2) "healthcare reform": issue; legislation; critical issue; healthcare issue; policy initiative; government program ...

We obtained $2,572$ tweets (600, 941, 214, and 817 in the 4 categories). All together there are $22,599$ tokens, and the size of vocabulary is $1,006$.

We evaluate clustering quality using purity [Zhao and Karypis, 2002], adjusted random index (ARI) [Hubert and Arabie, 1985], and normalized mutual information (NMI) [Strehl and Ghosh, 2002]. The purity measure assumes that each cluster is predicted to be the dominant class for that cluster; ARI penalizes both false positive and false negative clustering results; and NMI can be information-theoretically interpreted and has been increasingly used. Larger purity/ARI/NMI scores mean better clustering results.

4.2 Comparison Results

To demonstrate the effectiveness of using a probabilistic knowledgebase for conceptualization, we compare our approach with several other methods, including statistical methods such as LDA, and methods that use other knowledgebases including WordNet, Freebase, and Wikipedia.

- Original Data: We derive TF-IDF vectors from bag-of-words representations of tweets and cluster by cosine distances.
- LDA: We apply LDA to obtain topics, and use the topics as features for clustering. The number of topics we specify for LDA is equal to or twice the number of clusters in our problem sets.
- WordNet: We break each tweet into a set of words, and retrieve their hypernyms in WordNet as additional features for clustering.
- Freebase: Given a tweet, we find terms that correspond to Freebase instances through a trie index. Then, for each term, we use its types (concepts) as extended features.
- Wikipedia (Category-Link): We break each tweet into a set of words, and find their categories on Wikipedia (through Wikipedia's Category-Link). Then, we use anchor texts on the links as extended features.

- Wikipedia (ESA): ESA (explicit semantic analysis) maps text (tweets in our case) to a vector of Wikipedia pages, and we use the vector as features for clustering.
- Probase: We conceptualize tweets, and add concepts as additional features. We varied the retrieved concept number from 10 to 5000 (if it has).

We perform K-means clustering on the extracted features. The purity scores are shown in Table 3; the ARI scores are shown in Table 4; and the NMI scores are shown in Table 5. We draw the following conclusions:

First, statistical approaches (LDA) do not work well for short text. It is difficult to infer topic distributions from text when each document contains approximately ten words. In particular, when LDA uses more topics, the clustering accuracy actually decreases.

Next, we evaluate approaches that use knowledgebases, including WordNet, Wikipedia (Category-Link) and Freebase. WordNet has low accuracy. This is because words in WordNet have many senses, and many of them are rarely used. However, WordNet does not differentiate among those senses by their popularity or typicality. Thus, using WordNet hypernyms actually introduces a lot of noise. Wikipedia (Category-Link) and Freebase showed good improvement on *Problem 1*. This is because the category links in Wikipedia and types in Freebase can easily handle the concepts related to instances in *Problem 1*. However, when using these two knowledgebases in *Problem 2*, both fail to improve clustering results. Freebase has no more than 2,000 types. Thus, the concept space in Freebase is not sufficient to express the tweet content in *Problem 2*. For Wikipedia (ESA), experiments also show that it has better results for *Problem 1* than *Problem 2*. The problem is two-fold: the tweets are too short to create a good mapping, and the concept space of Wikipedia pages is still insufficient.

Finally, our approach outperform 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. Furthermore, the number of concepts we use also makes a difference (as shown in Tables 3, 4 and 5). For *Problem 1*, the optimal number of concepts is 500, while for *Problem 2*, smaller numbers give better results. The reason is that, if the content contains a unique class of concepts, more concepts can capture the content in a more comprehensive way. If the content contains multiple classes of concepts, sometimes we end up conceptualizing into very general, vague concepts (such as *topic*, *factor*, etc.) which make the features not discriminative.

Table 3: Clustering purity scores on Twitter data.

Method	@Problem1	@Problem2
Original Data	0.492±0.004	0.592±0.009
LDA (1×Cluster Num)	0.561±0.060	0.497±0.034
LDA (2×Cluster Num)	0.451±0.034	0.464±0.024
WordNet	0.563±0.044	0.439±0.023
Freebase	0.722±0.147	0.551±0.035
Wikipedia (Category-Link)	0.748±0.081	0.515±0.008
Wikipedia (ESA)	0.620±0.096	0.622±0.060
Probase (Top 10)	0.636±0.065	0.512±0.018
Probase (Top 20)	0.728±0.094	**0.635±0.128**
Probase (Top 50)	0.825±0.096	0.631±0.142
Probase (Top 500)	**0.911±0.109**	0.474±0.014
Probase (Top 5000)	0.876±0.144	0.421±0.012

Table 4: Clustering ARI scores on Twitter data.

Method	@Problem1	@Problem2
Original Data	0.0615±0.003	0.262±0.068
LDA (1×Cluster Num)	0.176±0.044	0.091±0.034
LDA (2×Cluster Num)	0.059±0.050	0.036±0.019
WordNet	0.177±0.077	0.069±0.031
Freebase	0.526±0.198	0.149±0.015
Wikipedia (Category-Link)	0.536±0.092	0.140±0.088
Wikipedia (ESA)	0.209±0.145	0.250±0.088
Probase (Top 10)	0.313±0.089	0.193±0.085
Probase (Top 20)	0.504±0.135	**0.324±0.176**
Probase (Top 50)	0.611±0.157	0.299±0.181
Probase (Top 500)	**0.878±0.045**	0.124±0.005
Probase (Top 5000)	0.695±0.249	0.083±0.011

5 Related Works

Analyzing short text is important. A lot of interests lie in understanding user intent from search queries, or mining twitter messages for business insight. Recent work [Phan *et al.*, 2008; Ritter *et al.*, 2010; Ramage *et al.*, 2010; Karandikar, 2010] that applies clustering and topic modeling to Twitter text confirms that the difficulty comes from the fact that highly related Twitter messages often have very little overlapping on the word level. It has been shown that traditional topic analysis methods should consider topic segments with tens of hundreds of words [Hearst, 1997]. Statistical topic modeling [Blei *et al.*, 2003; Blei and Lafferty, 2009] also requires sufficient words in a document to infer the document topic distribution.

A knowledgebase can be used to enrich features derived from bag-of-words representations, and help text understanding. By resolving synonyms and introducing WordNet concepts, the quality of document clustering can be improved [Hotho *et al.*, 2003]. Other research also showed that using knowledgebases such as ODP [Gabrilovich and Markovitch, 2005] and Wikipedia [Gabrilovich and Markovitch, 2006; Egozi *et al.*, 2008; Hu *et al.*, 2008; 2009] helps text categorization and information retrieval. Compared with traditional latent semantic analysis (LSA) [Deerwester *et al.*, 1990] and topic modeling such as latent Dirichlet allocation (LDA) [Blei *et al.*, 2003], explicit semantic analysis (ESA) has the advantage of providing semantics that are interpretable by human beings.

6 Conclusion

We introduce a method of conceptualizing short text using a probabilistic knowledgebase. We detect and map terms in short text to instances and attributes in the knowledgebase. Then we derive the most likely concepts using Bayesian inference. The conceptualization techniques is applied to clustering Twitter messages. Results showed that our approach is highly effective compared to traditional bag-of-words based statistical methods.

Table 5: Clustering NMI scores on Twitter data.

Method	@Problem1	@Problem2
Original Data	0.215±0.010	0.452±0.076
LDA (1×Cluster Num)	0.161±0.065	0.114±0.037
LDA (2×Cluster Num)	0.067±0.022	0.069±0.024
WordNet	0.195±0.070	0.074±0.074
Freebase	0.531±0.164	0.204±0.037
Wikipedia (Category-Link)	0.540±0.077	0.336±0.089
Wikipedia (ESA)	0.351±0.132	0.340±0.800
Probase (Top 10)	0.318±0.110	0.490±0.029
Probase (Top 20)	0.479±0.111	0.555±0.019
Probase (Top 50)	0.559±0.123	**0.632±0.066**
Probase (Top 500)	**0.826±0.062**	0.301±0.189
Probase (Top 5000)	0.690±0.176	0.095±0.084

References

[Banko *et al.*, 2007] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In *IJCAI*, pages 2670–2676, 2007.

[Blei and Lafferty, 2009] David M. Blei and John D. Lafferty. *Topic Models*, chapter: Topic Models. Taylor and Francis, 2009. (in Press).

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Bloom, 2003] Paul Bloom. Glue for the mental world. *Nature*, (421):212–213, Jan 2003.

[Bollacker *et al.*, 2008] Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, 2008.

[Chen, 2006] David Chen. Sisterhood of classifiers: A comparative study of naive Bayes and noisy-or networks, 2006. Master Thesis.

[Deerwester *et al.*, 1990] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[Egozi *et al.*, 2008] Ofer Egozi, Evgeniy Gabrilovich, and Shaul Markovitch. Concept-based feature generation and selection for information retrieval. In *AAAI*, pages 1132–1137, 2008.

[Etzioni *et al.*, 2004] Oren Etzioni, Michael Cafarella, and Doug Downey. Webscale information extraction in knowitall (preliminary results). In *WWW*, 2004.

[Fellbaum, 1998] Christiane Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998.

[Gabrilovich and Markovitch, 2005] Evgeniy Gabrilovich and Shaul Markovitch. Feature generation for text categorization using world knowledge. In *IJCAI*, pages 1048–1053, 2005.

[Gabrilovich and Markovitch, 2006] Evgeniy Gabrilovich and Shaul Markovitch. Overcoming the brittleness bottleneck using Wikipedia: enhancing text categorization with encyclopedic knowledge. In *AAAI*, pages 1301–1306, 2006.

[Hearst, 1992] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING*, pages 539–545, 1992.

[Hearst, 1997] Marti A. Hearst. Texttiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23:33–64, March 1997.

[Hotho *et al.*, 2003] Andreas Hotho, Steffen Staab, and Gerd Stumme. WordNet improves text document clustering. In *SIGIR Workshop on Semantic Web*, 2003.

[Hu *et al.*, 2008] Jian Hu, Lujun Fang, Yang Cao, Hua-Jun Zeng, Hua Li, Qiang Yang, and Zheng Chen. Enhancing text clustering by leveraging Wikipedia semantics. In *SIGIR*, pages 179–186, 2008.

[Hu *et al.*, 2009] Xiaohua Hu, Xiaodan Zhang, Caimei Lu, E. K. Park, and Xiaohua Zhou. Exploiting wikipedia as external knowledge for document clustering. In *KDD*, pages 389–396, 2009.

[Hubert and Arabie, 1985] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[Karandikar, 2010] Anand Karandikar. Clustering short status messages: A topic model based approach, 2010. Master Thesis.

[Lenat and Guha, 1989] Douglas B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley, 1989.

[Lidstone, 1920] George James Lidstone. Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8:182–192, 1920.

[Murphy, 2004] George L. Murphy. *The big book of concepts*. The MIT Press, 2004.

[Phan *et al.*, 2008] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text and web with hidden topics from large-scale data collections. In *WWW*, pages 91–100, 2008.

[Ponzetto and Strube, 2007] Simone Paolo Ponzetto and Michael Strube. Deriving a large-scale taxonomy from wikipedia. In *AAAI*, pages 1440–1445, 2007.

[Ramage *et al.*, 2010] Daniel Ramage, Susan Dumais, and Dan Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.

[Ritter *et al.*, 2010] Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised modeling of twitter conversations. In *HLT*, pages 172–180, 2010.

[Strehl and Ghosh, 2002] Alexander Strehl and Joydeep Ghosh. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research*, 3:583–617, 2002.

[Suchanek *et al.*, 2007] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *WWW*, pages 697–706, 2007.

[Wang, 2011] Haixun Wang. Knowledgebase and our mental world. Technical report, Microsoft Research, 2011.

[Wu *et al.*, 2011] Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Zhu. Towards a probabilistic taxonomy of many concepts. Technical Report MSR-TR-2011-25, Microsoft Research, 2011.

[Zhao and Karypis, 2002] Ying Zhao and George Karypis. Criterion functions for document clustering: Experiments and analysis. Technical report, Department of Computer Science, University of Minnesota, 2002.