International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 5, Issue 2, March - April 2016 **ISSN 2278-6856**

Extraction of key topics from online text reviews

Bhaskarjyoti Das[1] ,Prathima V R[2]

[1]Rajiv Gandhi Institute of Technology, Dept. Of Computer Science and Engineering,
R T Nagar, Bangalore 560032, India

[2] Rajiv Gandhi Institute of Technology, Dept. Of Computer Science and Engineering,
R T Nagar, Bangalore 560032, India

Abstract

Though it has been a subject of active research for a while, extraction of key phrases from unstructured textual data is not a completely mature technology. Apart from the usual challenges in computational linguistics such as synonymy and polysemy, there may be additional domain specific challenges. So, a state-of-art in one domain may not be so in a different domain and there is hardly any universally acceptable and completely accurate solution. In this paper, we evaluate and compare different approaches for key topic extraction from unstructured textual data found in online review and rating portals.

Keywords: Topic, key phrase, co-occurrence, supervised learning, unsupervised learning, dimensionality reduction, graph theory, deep learning

1. INTRODUCTION

The problem of finding the main topics in a document or set of documents involves finding the phrases (consisting of one or multiple words) representing a possible topic and sorting them in order of importance. This is a key problem of Information Retrieval (IR) domain and there are many possible applications such as text summarization, search, online advertisement etc.

Key topic extraction is different from text summarization. The goal of text summarization is to come up with a summary consisting of well-formed sentences that meets the dual criteria of compression and retention. The unit of information in text summarization is sentence and it has its own challenge because of this constraint.

While it is relatively easy to extract candidate topics or phrases, it is hard to rank them in order of importance. Apart from "frequency" of "phrase", the other challenge is to come up with a set of "coherent" features. The significant challenge in detecting key phrases is "morphological and semantic" as described by Su Nam Kim et al [1] i.e. same topic can be described in different words or word permutation and so tends to be disregarded as infrequent if we just focus on frequency of word or phrases in the same sequence. Though determination of semantic similarity or relatedness is important, it is very intuitive and hard to implement in software algorithms. As a result, most solutions are customized to a domain and online companies looking for similar things end up implementing their own customized solutions. In this paper we review the existing researches and present a comparative analysis

2. IMPORTANCE OF DOMAIN

The problem poses different structural challenges in different domains. Eibe Frank et al showed [2] that the process of key phrase extraction performs better when the process incorporates domain specific structural information. In the scientific articles, the keywords will be found either in a keyword section or in the abstract part of the article. But there is no such structure in free format text such as web pages. Similarly, the abstract in a scientific article will have enough information about the key topics. An extracted topic may be considered more important if it is related to the listed keywords or mentioned in the abstract. So, looking at abstract and keywords may be sufficient for such domain while doing key phrase extraction. But a web based document such as customer review may have topics appearing throughout its length and we have to consider the full length of the document. So, for finding importance of extracted topics, different strategies need to be devised depending on the domains.

3. GENERATION OF CANDIDATE KEY PHRASES IN ONLINE REVIEWS

In general, the whole process of keyword extraction can be broken down into two steps i.e. finding the candidate key words or phrases and making a selection based on a criteria. The first step is usually based on heuristics. The goal of this heuristics is to get a good number of meaningful key words or phrases as candidates. Generating the candidate terms is either on linguistic basis such as a combination of part of speech tags or on statistical basis such as n-grams. Using a part of speech tag (POS tag) pattern is an effective heuristic to generate candidates. For example, "adverb-adjective-noun" patterns typically are subjective text fragments and may contain important information in a text. Typically we will be removing stop words and unnecessary punctuation, control character, web URLs and web HTML fragments from the corpus before choosing to generate the candidates either linguistically or statistically. Once the candidate words or phrases are generated, we do the necessary pruning in the second phase for selecting candidates.

The possibilities of heuristic based first phase are limited and we can do a limited frequency based analysis of the generated candidates. Word Cloud and Tag Cloud are two such applications where keywords or tags are shown in proportionately larger font based on the frequency of occurrence in the corpus. But these methods provide an approximate overview, are useful but not very precise.

4. SELECTION OF KEY PHRASES

The second step of selecting key topics and phrases from the heuristically generated candidate set has been attempted using statistical, supervised and unsupervised methods.

4.1 Statistical approach

In a statistical approach, some frequency measure is used to choose top n candidates. Gerard Salton and Christopher Buckley [3] discussed the importance of an appropriate term weighting system for an effective information retrieval system. Using an external resource such as Wikipedia to ascertain the importance of the candidate phrase [4] is also another possibility. Additionally, statistical association among candidate key phrases can be used as a possible proxy of semantic coherence. Rapid Automatic Keyword Extraction (RAKE) by M.W.Berry et al. [15] is a popular keyword extraction algorithm for single document that can be extended to multiple documents. Yutaka Matsuo and Mitsuru Ishizuka [5] presented another statistical algorithm to extract keyword from a single document without relying on a corpus and TF IDF measurement. In their proposed algorithm, first frequent terms are determined and then those are clustered based on some similarity measures. The degree of bias of the probability distribution for co-occurrence of any term with those clusters is investigated. If there exists a bias, then it is very likely that the term is a keyword.

In another approach called non-negative matrix factorization (NMF), we can use dimensionality reduction techniques after TF-IDF (term frequency, inverse document frequency) based processing to come up with fewer dimensions consisting of important key words or phrases. The challenge in this is: it is a trial and error process and the matrix factorization output (the reduced dimensions) are not guaranteed to have very meaningful themes.

4.2 Supervised approach

Supervised approach is extremely effective when training data is available. In this method, we train a classifier on documents annotated with key phrases to determine whether a candidate phrase extracted from a test document is a key phrase or not. Witten et al [6] has used this approach in their work on KEA algorithm and many learning algorithms have been tried out. In this approach, the important first step is feature design. Once that is done, the feature set for test and train dataset needs to be generated. The features can be a combination of external, semantic and statistical features. Examples of statistical feature are TF*IDF, term length, the position of the candidate word etc. Wikipedia frequency of the topic can be an external feature. Frequency in the web server search log can be useful in finding advertising keywords and is an example of domain specific feature. The classifier performance can be evaluated using metrics such as precision, recall and F score. Typically, in machine learning approaches like this, we need a baseline and a possible baseline here can be industry standard web service such as Alchemy API.

4.3 Dimensionality reduction methods

Unsupervised approach does not require any training data. This approach can be mainly classified into dimensionality reduction techniques using various methods, graph theoretical approach and sometimes combination of both.

In Latent Semantic Analysis (LSA), the initially formed sparse term document matrix can be reduced in dimension using Singular Value Decomposition (SVD) to come up with an orthogonal set of topics consisting of words. SVD is a least square method that finds the low rank matrix approximation by minimizing the 'Forbenius Norm' of the difference from the original matrix. Finally, we can pick the top sentences based on absolute value of sentence vector in this concept space. However, LSA assumes normal distribution which may not be the case always. Compared to the Linear Algebra grounding of LSA, Probabilistic Latent Semantic Analysis (PLSA) has probabilistic grounding and the topics are allowed to be non-orthogonal with overlap. In continuation of this, David M. Blei et al [7] proposed Latent Dirichlet allocation (LDA) as a topic mining algorithm which visualizes each document as a probabilistic distribution of topic and each topic as a probabilistic distribution of words. LDA is similar to PLSA with dirichlet priors for the document-topic and topic-word distributions. Finding key topics is the reverse of this i.e. from words to topics. However, if we take a look at LDA output, we can see the need of a fair amount of trial and error to come with topics containing words conveying consistent theme.

4.4 Unsupervised clustering

In a clustering approach, a cluster of documents can be viewed as a network of information units such as sentences, phrases or n-grams. Once the candidate "information units" are identified, either clustering is attempted or a graph theoretical analysis is done. Zhiyuan Liu et al [8] used clustering to find the exemplar terms and then used those terms to find key phrases assuming clusters would represent key topics. The candidates can also be clustered using statistical co-occurrence information combined with Wikipedia based semantic clues or semantic distance based on resources such as Wordnet [9]. In clustering, the key challenge is the quality of the cluster which is hard to guarantee.

4.5 Unsupervised graph theoretical techniques

In a model that uses bag of words with co-occurrence metric, we come up with N-dimensional vector for each document, where N is the number of all possible words in the corpus. The documents may be represented by a cosine similarity matrix. Now, we can draw a graph where vertices are the words and edges are the calculated similarity. TF*IDF (term frequency, inverse document frequency) is used here to ensure that we take care of the discriminating power of words in addition to frequency.

In a model that uses bag of parts of speech pattern, instead of extracting all possible n-grams, we extract part of speech (POS) patterns and then examine their co-occurrence to build a network. We can examine the degree centrality measures in this network arguing that the significantly similar POS patterns will be connected to each other. We here model the importance of a word in terms of centrality in this graph and an optimum threshold can be chosen to extract the top n nodes. We can also consider other type of centrality metrics to measure global prominence such as Eigen Vector centrality [10] or Pagerank [11] instead of pure degree centrality. In PageRank, the importance of a node is decided by the edges to the neighbouring nodes representing votes for relevance. The ranking score is recursively calculated by considering the weights of these edges and the rank of the neighbouring nodes. In this context, we must mention Textrank [12], a graph theoretical method that can be applied to do both text summarization and keyword extraction. For text summarization, sentences are the nodes of the graph and for keyword extraction, words are the nodes. The undirected edges of the graph have weights that are derived from co-occurrence in the text in case of words and similarity based on sentence vectors in case of sentences. Textrank uses the concept of prestige in the network and uses Pagerank to rank the nodes of the graph. The top n key words or sentences in the graph are the top ranked nodes.

In another approach, we use concept of semantic similarity with graph theoretical models. Methods discussed so far (statistical, graphical theoretical, cluster based, dimensionality reduction etc.) do not attempt to use the semantic meaning of the text and instead use some statistical proxy for the same. Semantic meaning in any language is very intuitive and has a large human component in it. Attempts have been made to incorporate the concept of semantic distance or similarity in the previously discussed methods. For example, the Textrank similarity measure can be modified to incorporate semantic similarity. When we talk of semantic similarity of textual pieces, the basic unit is semantic similarity of a pair of words. This can be done by using a carefully curated lexical corpus such as Wordnet. However, this approach has the limitation that human language is rapidly evolving and every possible word is not part of Wordnet. The sentence similarity can be shown to be a function of word level semantic similarity [13].

Recently, there has been another approach to derive semantic distance using machine learning. In this, semantic distance is derived by training a model with distributional properties of information units in a very large corpus. Word2vec [14] by Tomas Mikolov et al from Google is such an approach inspired by deep learning. It uses methods that are similar to recurrent neural network but at the same time is computationally more efficient by using some specific algorithm. Word2vec maps each word from the corpus vocabulary into some low-dimensional vector space by observing the distributional properties of the words in the corpus it trained with.

4.6 Mixed strategies

We may also have mixed strategies. There are researches such as Topic Rank by Adrien Bougouin et al [17] which is a modification proposed on Textrank by using LDA. CollabRank [18] by Xiaojun Wan et al is again an approach where documents are initially clustered and then key phrases are picked up in each cluster using Textrank. However, just like any clustering effort, the success in this approach also depends on quality of clusters.

5. EXPERIMENTAL RESULTS

We have implemented and evaluated some of these approaches and in this section, we will look at results of these experiments with a sample dataset to have an idea about their relative effectiveness. For our experiment, we have the chosen the review dataset (399 reviews) for a particular restaurant (Carolina's Mexican Food) in Phoenix in USA from Yelp challenge dataset (https://www.yelp.com/dataset_challenge).

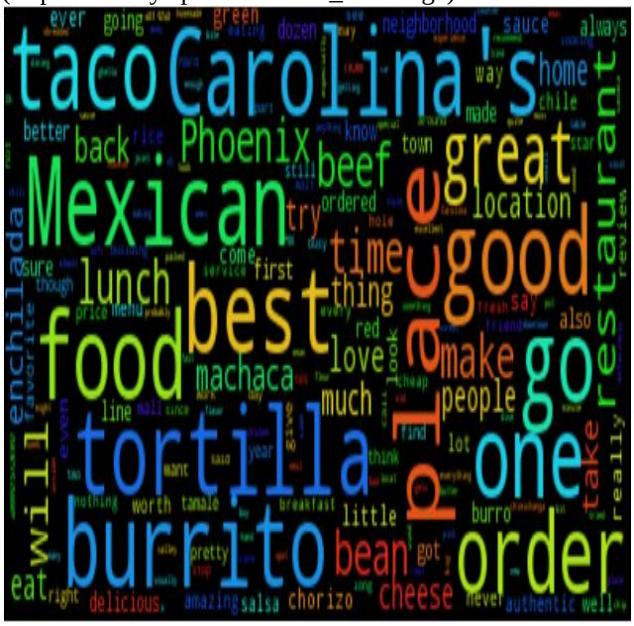

Figure 1 Word cloud drawn in a statistical approach

Figure 1 above shows Word Cloud drawn using simple frequency based statistics and it shows the main words that are popping up in customer reviews.

1. "arrived ... nom nom ... minutes ago ... nom nom ... order"
2. indicating food prices"
3. 'scissors ... choose rocks..cuz thier jynormous tortillas rock'
4. 'several dozen mexican style pseudo fast food eateries'
5. regularly overflowing waste cans hurts carolina 's reputation

Figure 2 Top 5 key phrases picked up by RAKE

Figure 2 above shows the top 5 key phrases selected by RAKE, a statistical keyword extraction algorithm that is extended for key phrase extraction. Clearly its output is not quite meaningful here.

'zone', 'iron', 'helping', 'helpful', 'held', 'heckling', 'heck', 'heavily', 'heavenly', 'healthiest', 'heads', 'headfirst', 'headaches', 'haves', 'hardworking'', 'harbor', 'happiness', 'helpings'

Figure 3 Top key words picked up by TF IDF

Figure 3 shows the top key words picked up by TF-IDF (term frequency, inverse document frequency). Again, it is not very meaningful as a pure statistical approach misses the coherence and semantic aspects.

Topic 0: best town stars place mexican
Topic 1: taco beef got shredded ordered
Topic 2: lunch time buy quick parking
Topic 3: chile green red stuff cash
Topic 4: great prices food homemade moved
Topic 5: good flour thing tortillas beans

Figure 4 Top 5 topics (dimensions) picked up by NMF

Next we have tried the Non Negative Matrix Factorization method which is a dimensionality reduction technique on the same dataset. The Figure 4 above lists the topics (cluster of words) arrived at by matrix factorization. If we look at top 20 topics, few will be conveying some themes but not all.

Topic 0: dozen amazing years friends stop run close spot half bad
Topic 1: location service enchilada minutes eating yummy cactus mohave happy cafeteria
Topic 2: food restaurant ghetto potato ll dirty world flavorful floor clean
Topic 3: phoenix made fresh time isn special price thin eggs top
Topic 4: beans beef rice good shredded plate perfect tasty quality windows

Figure 5 Top 5 key topics picked up by LDA

If we try out Latent Dirichlet allocation (LDA) which has a probabilistic grounding (assumes that a document is a probabilistic distribution of topic and each topic is a probabilistic distribution of words), the key topics detected by it are not guaranteed to show meaningful themes as shown in Figure 5. We got somewhat meaningful themes with 30 topics with each having 10 words.

good steak, good food, chicken, place, food good chicken fried steak, good food good atmosphere, good chicken fried chicken, good delicious food, delicious steak, good portions

Figure 6 Some of the key phrases extracted by Textrank

Figure 6 above lists some of the key phrases picked up by Textrank which is unsupervised graph theoretical method. Note that all key phrases are still not meaningful as the similarity metric of Textrank is still statistical as a weak proxy to semantic similarity. Nevertheless, the results show dramatic improvements.

cool place, location great, full order, full line , full plate, people parking, quality food, quality meat, quality good, food portions, potatoes amazing, great salad, great deal, shrimp excellent, weekend brunch

Figure 7 Some of the key phrases picked up by Textrank modified with Wordnet based similarity

Finally, the figure above shows some of the key phrases that are picked up by the Textrank modified with Wordnet based semantic similarity between words. Now, many of the key phrases start looking like meaningful English. Even our experiment with Word2Vec based similarity yielded similarly encouraging results.

So, we can conclude the followings from our experiments:

(a) For domains such as online reviews, we have to rely on either statistical or graph theoretical techniques for key phrase extraction due to lack of training data.

(b) Graph theoretical methods tend to fare better than pure statistical methods as it is hard to proxy semantic meaning and coherence with statistical correlation. Dimensionality reduction techniques require many trial and error to yield meaningful results.

(c) Even graph theoretical methods work better if the similarity metrics is replaced by semantic similarity.

So, an approach which uses graph theoretical method with semantic similarity is expected to yield best results.

6. CONCLUSION

In spite of all the state of the art researches, the keyword extraction remains a challenging problem. The challenge here is to select top keywords or phrases which are frequent, coherent, have discriminating power and do justice to the semantic aspects at the same time. State of the art researches are also not outright deployable across domains. It is because, in a specific research, the state of the art techniques are tested on a specific dataset of a particular domain. What comes out as a winning method in one domain's dataset, does not necessarily perform so well on datasets in other domains. The heuristics to generate candidate key phrases is well-understood. In the subsequent step of making a choice amongst those candidates, graph theoretical approaches are found empirically more appealing than methods based on clustering and pure statistical or probabilistic grounding. So, because of this challenge, in spite of all the researches, we tend to fall back to the lowest common denominator which has been proved beyond doubt. Kazi Saidul Hasan and Vincent Ng in their evaluation [16] of state of the art techniques with diverse dataset validate that TF*IDF remains a robust performer across diverse domains though domain specific techniques may do better in the specific domains they are built for.

References

[1] Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010b. SemEval-2010 Task 5: Automatic key phrase extraction from scientific articles. In Proceedings of the 5th International Workshop on Semantic Evaluation, pages 21–26.

[2] Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Domain-specific key phrase extraction. In Proceedings of 16th International Joint Conference on Artificial Intelligence, pages 668–673.

[3] Gerard Salton and Christopher Buckley. 1988. Term weighting approaches in automatic text retrieval. Information Processing and Management, 24(5):513–523.

[4] Extracting key terms from noisy and multi-theme documents, Maria Grineva, Maxim Grinev, and Dmitry Lizorkin, 2009, In Proceedings of the 18th International Conference on World Wide Web, pages 661–670.

[5] Yutaka Matsuo and Mitsuru Ishizuka. 2004. Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools, 13.

[6] Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. 1999. KEA: Practical automatic key phrase extraction. In Proceedings of the 4th ACM Conference on Digital Libraries, pages 254–255.

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

[8] Zhiyuan Liu, Peng Li, Yabin Zheng, and Maosong Sun. 2009b. Clustering to find exemplar terms for key phrase extraction. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 257–266.

[9] Wordnet, a lexical database for English, https://wordnet.princeton.edu/

[10] Centrality and network flow, Stephen P Borgatti, Social Networks 27 (2005) 55–71

[11] The anatomy of a large-scale hypertextual Web search engine. Sergey Brin and Lawrence Page. 1998. Computer Networks, 30(1–7):107–117

[12] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 404–411.

[13] Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett, "Sentence Similarity Based on Semantic Nets and Corpus Statistics", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 8, AUGUST 2006

[14] Distributed representation of words and phrases and their compositionality by Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean of Google Inc. in Advances of Neural Information Processing Systems 26 (NIPS) 2013

[15] Automatic Keyword Extraction from Individual Documents. In M. W. Berry & J. Kogan (Eds.), Text Mining: Theory and Applications: John Wiley & Sons, Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010).

[16] Conundrums in unsupervised key phrase extraction: Making sense of the state-of-the-art by Kazi Saidul Hasan and Vincent Ng. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 365–373. 2010.

[17] Topicrank: Graph-based topic ranking for keyphrase extraction by Adrien Bougouin, Florian Boudin, and B´eatrice Daille. In Proceedings of the 6th International Joint Conference on Natural Language Processing, pages 543–551, 2013

[18] CollabRank: Towards a collaborative approach to single-document keyphrase extraction by Xiaojun Wan and Jianguo Xiao. In Proceedings of the 22nd International conference on Computational Linguistics, pages 969–976. 2008a.

AUTHOR

Bhaskarjyoti Das received B.E.T.C.E in Electronics and Telecommunications Engineering from Jadavpur University, Calcutta in 1985. During 1985-2014, he worked in various capacities in government R&D as well as multinational product engineering organizations. Currently, he is pursuing academics and hence doing his M. Tech in Computer Science and Engineering.