Unsupervised Topic Labeling of Text based on Wikipedia Categorization

Tetyana LOSKUTOVA
Wits Business School, University of the Witwatersrand
Johannesburg, Gauteng, South Africa

ABSTRACT

Defining text topicality is often an expensive problem that requires significant resources for text labeling. Though many packages already exist that provide dictionaries of labeled text, synonyms, and Part-of-Speach tagging, the problem is ongoing as language develops and new meanings of words and phrases emerge. This paper proposes a cheap in human labor solution to topic labeling of any text in the majority of languages. The methodology uses links to the naturally emerging corpus of labeled text – the Wikipedia. Wikipedia categories are processed to extract a weighted set of topic labels for the analyzed text. The approach is evaluated by processing categorized texts and comparing the similarity of the top ranks of topic labels to the text category. The topic labels extracted using this methodology can be used for comparing similarity of texts, for the assessment of the completeness of topic coverage in automated marking of essays, and for coding in qualitative text analysis. The paper contributes to the field of NLP by offering a cheap and organically developing method of topical text labeling. The paper contributes to the work of qualitative analysts by offering a methodology for the analysis of interview transcripts and other unstructured text.

Keywords: Unsupervised Topic Labeling, Context Recognition, Abstractive Labeling.

1. INTRODUCTION

The development of the Internet has made large amounts of text data available for analysis, thus setting in motion significant advances in Natural Language Processing (NLP). Large amounts of text are difficult to analyse due to the hardware memory requirements and the difficulty of comparing texts: very different in word representation texts may mean the same and the meaning is often dependent on the questions of the analysis. The task of recognizing the topic of the text and finding a suitable representation for it, such as labels, is becoming more important as it allows crunching a large body of text into a smaller representation without the loss of meaning. This smaller representation can be for human consumption (summarization) or for computer analysis and comparison (topic labeling). While earlier approaches to text analysis were based on quantitative measures, such as frequency [1], such approaches often failed to preserve the meaning. Later approached attempted to fix the problem by applying predefined context rules, such as removing most commonly used words in a language, assigning different weights to the words based on their location in text, and adjusting for the similarity to the title [2]. These approaches, despite being successful in certain applications lack the ability to define the topic of a random text without prior knowledge of the context. Language has words that have different meaning in different contexts (homonyms) and same phrases can mean different things depending on whether they are used literally, sarcastically, or metaphorically. Additionally, different syntactical structure of texts in different languages prevents from applying same rules across languages. Allahyari et al. [3] pointed that commonly used approaches take context and type of text as inputs in order to produce meaningful outputs.

Despite the significant recent advances in text classification, translation, and text generation, current methods of topic identification depend on the prior knowledge of the higher-level context, labeled, tagged, and classified data, and context-dependent rules. The goal of this paper is to propose and evaluate a methodology for unsupervised topic labeling of text without prior filtering on the text structure or context.

2. LITERATURE REVIEW

Topic labeling is defined as the task of generating a set of words and phrases that capture the topics and subtopics discussed in text. Topic labeling task pursued in this paper is similar to both the tasks of summarization and keyword extraction. The difference from summarization is that topic labeling does not intend to capture the overall meaning expressed in text but rather selects the main topics being discussed. Allahyari et al. [3] suggested using the terms extraction and abstraction to explain how text summaries can be created: extractive summarization is the approach of building summaries from the phrases already used in text while abstractive summarization attempts to create a completely new text that captures the meaning of the text under study. These terms help understanding the difference between keyword extraction and topic labeling: although keywords and topic labels generally aim to capture the main topics, keywords extraction is an extractive method (limited to the words used in text) while topic labeling is an abstractive method that can use completely different words. Topic labeling accomplishes an important task of re-phrasing. The approach is different from classification as text is expected to cover several topics – one main topic and several topics supporting the main point [4] – and topic labels need to capture these topics instead of giving probability of text belonging to a particular class.

Existing methods of unsupervised summarization and keyword extraction are mostly extractive. Among the extractive methods, Latent Semantic Analysis (LSA) has been proposed to produce high-quality summaries from existing sentences. Gong and Liu [4] proposed an iterative approach where the most relevant sentence is selected first, then the terms from the most relevant sentence are eliminated from the text, and the approach is repeated until a desired number of sentences is selected. In this approach, the sentences are ranked by their importance and the most different sentences are extracted. Gong and Liu applied Singular Value Decomposition (SVD) as a solution for sentence ranking. In semantic sense the application of SVD allows assigning of close ranks to semantically similar sentences; those ranks are represented by singular values in the method's matrix.

While Gong and Liu made provision for the use of any suitable function for getting the importance score [4], in the most basic version, term frequency in the general corpora and the term frequency in the text are used for summarization using SVD.

The LSA SVD approach was improved by Steinberger and Jezek [5] who proposed a salience score for summarization based on the length of each sentence's vector adjusted for its corresponding singular value (singular value is used for ranking in the basic approach). Dokun and Celebi [6] proposed yet another modification of the initial ranking algorithm: in their approach summarization is created from sentences whose term-ranking is average.

As an alternative to LSA approach, Han et al. [7] proposed a Sentence-Level Semantic Graph Model where sentences are vertices in the graph and the relations between the sentences are graph edges. The graph is evaluated using sentences' importance values calculated using PageRank and the values of edges' importance derived using semantic analysis. Several other modifications of the above methods and other methods based on frequencies and rules exist for extractive summarization [3].

Abstractive methods are more difficult because the meaning of the topic needs to be reconstructed anew [3]. Ganesan, Zhai, and Han [8] suggested separating two types of abstractive methods: the first uses preexisting knowledge and the second uses natural language generation. The first approach is generally template-based and abstraction can be perceived as filling in a template or a form. The second approach generates text that is domain dependent [3]. Both these approaches are not useful for achieving unsupervised topic-labeling of unknown domain text.

More relevant approach is Opinosis method [8], which intends to summarize text of unknown domain using abstraction. Opinosis's approach is based on (1) generating graphs of text, (2) evaluating the paths of the graph using the measures of valid paths, redundant paths, and collapsed paths, and (3) using a special algorithm to stitch parts of paths into a valid sentence. While this algorithm allowed achieving reasonable performance on highly-redundant opinion polls (spoken language on a particular topic) [8], its application for an unspecified text structure and other than English languages may require significant additional work.

Other relevant abstractive methods are based on Neural Networks and benefit from the achievements in machine translation. Attention-Based Summarization (ABS) [9] has proven to be effective on paraphrasing short texts, however the outputs bear high similarity with the initial text and are not capable of labeling related subtopics. ABS model also tends to pick correct names and places from keywords while showing lesser accuracy in getting the meaning of the text. A Deep Reinforced Model for abstractive summarization [10] is also based on a neural network and uses a hybrid learning objective to account for the deficiencies of the previous methods; this model demonstrates higher performance of larger texts. Both models require large computational resources and the creation of training sets. Compared to the goal of the current paper, all summarization approaches, including those with paraphrasing ability, are not able to extract related subtopics from text.

Overall, the limitations of the existing methods for text topic labeling include the lack of abstractive methods, the lack of unsupervised methods, the dependence of the existing methods on context-dependent rules, and the lack of language-independent methods.

3. METHODOLOGY

The goal of this paper was to propose and evaluate a methodology for unsupervised topic labeling of text without prior filtering of the text based on structure or context. The following methodology was applied to achieve this goal:

Step 1: Data pre-processing: the extraction of nouns from text.
Step 2: Wikipedia search for topics based on nouns.
Step 3: Extraction of Wikipedia categories and ranking of the categories by the number of appearances in the results.
Step 4: Selection of the top 6 categories or top 3 ranks (the actual number was dependent on the number of categories in a rank). These categories were considered topic labels.

The methodology was evaluated using quantitative and qualitative approaches. The quantitative evaluation was done using Wu-Palmer similarity index to compare the topic labels with the category of the sample. Wu-Palmer similarity index is a measure that defines the similarity of words by comparing their positions in the hierarchy of concepts that define them [11]. The qualitative evaluation was performed on a 10% sample and included reading the text of the article and dividing the generated topic labels into 3 groups: "Topic of text", "Related topic", "Unrelated label".

The details of each step and the evaluation are discussed further.

Implementation

Three sets of data were downloaded from webhose.io: Set 1 was archived articles categorized as "business", Set 2 was archived news articles and news highlights (2-line extract from a news article) categorized as "politics", Set 3 was generated using search term "coffee or tea" and consisted of recent news articles. Out of these datasets, 100 texts were selected randomly for further processing. The texts were minimally processed using Natural Language Toolkit: NLTK 3.4.1. The NLTK processing removed the common stopwords and names using and selected nouns from each sentence, which were then used to search Wikipedia articles. Python package wikipediaapi was used to do the search. If a correspondent page was found (wikipedia page was found for all texts in Set 1 and 3 and for 86% of texts in Set 2), the page categories were added to the list of labels for the text. A list of "stop-labels" was created to remove categorization specific to Wikipedia, such as "pages using dmy dates", "pages using American English", "pages needing attention", and so on. Additionally, nouns related to the time of events described in the text were removed.

The quantitative evaluation of the relevance of the extracted labels consisted of the following steps:

Step 1. The categories extracted from the Wikipedia were ranked by the number of appearances in a text with the aim of extracting top 3 ranks or top 6+ categories. The following algorithm was used for the selection: (a) categories from the first rank were selected; (b) if the selection resulted in 6 or more categories, the process was completed and the categories from rank 1 were used as topic labels; if not, the next rank was selected. The process was repeated until at least 6 categories or

3 ranks were selected. The extracted categories were considered topic labels and were used for comparison with the initial dataset category

Step 2. NLTK synonyms (synsets) were collected for each of the labels: *allsyns_labels*. NLTK synsets are based on WordNet database where synsets are defined as hierarchies of words connected by their lexical and conceptual-semantic relationships [12].

Step 3. NLTK synonyms were collected for the initial categories: *allsyns_categories*.

Step 4. Wu-Palmer Similarity score was computed between *allsyns_labels* and *allsyns_categories*. WordNet's Wu-Palmer similarity implementation was used. In this particular implementation, Wu-Palmer similarity is a measure between 0 (not similar at all) and 1 (the same) of two words based on their hierarchical position in WordNet's synsets.

The qualitative assessment was performed on 10% of the texts in all analyzed datasets to determine the goodness of fit of all the top labels.

4. RESULTS

The results for all three sets of data are presented separately and then summarized below.

Set 1 – Business news: Total articles: 14794, sample 100. The quantitative assessment showed average best similarity of synonymous terms 0.78 and the worst of 0.14.

An example of the analyzed text (for readability, special symbols are replaced with spaces):

Cameroon: Crime Prevention - UN Discusses Social, Economic Challenges. BRICS countries hold 40 per cent of the world's population and a quarter of all economic output. The development bank is expected to be operational by the end of the year from its headquarters in Shanghai and predicts it will have up to 100 billion dollars in capital to finance infrastructure projects in developing countries. Dumisani Hlophe, the Director of the Kunjalo Centre for Development Research in Johannesburg, told RFI that the new bank may be met with suspicion from South Africans. He is sceptical that South Africa needs to belong to another foreign development investment bank. "What remains to be seen is whether it is going to remain one of those diplomatic points that the country scores without necessarily speaking to the needs of society," he said. For President Putin, hosting of the seventh BRICS summit comes at a great time. The sanctions imposed by the European Union and the United States over the country's involvement in the conflict in Ukraine continue to bite, and Russia is looking to strengthen its economic ties elsewhere. Andrew Foxall, the director of the Russia Studies Centre at London-based thinktank the Henry Jackson Society, believes Putin is using the opportunity to show the West that the country isn't isolated economically. Foxall told RFI that "the dynamics within the BRICS has changed a lot" over the period that the bloc has been in existence, and the "shining star" of Russia has been waning. The summit runs Thursday and Friday. South Africa

The text above received the following topic labels:

'Countries'
'Human geography'
'Banking'
'Banks'
'Economic history of Italy'
'Italian inventions'
'Legal entities'

The results show that topic capture the context better than the original category 'business' as the article is concerned with countries, geopolitics ('human geography'), and banking. The labels 'Economic history of Italy' and 'Italian inventions' are less clearly related to the topic of the article and their presence is explained by the fact that banking in its modern form (deposit banking) is considered to be an Italian invention and an important event in Italian economic history. While first 4 labels and the last label capture the topic of the article, the 5th and 6th represent related topics.

Set 2 – Political news and highlights: Total articles: 87156, sample 100, out of which 86 got labeled. The quantitative assessment showed average best similarity of synonymous terms 0.68 and the worst of 0.12.

An example of the analyzed text:

Privacy Policy. More Newsletters AP FILE - In this Aug. 24, 2015, file photo, former Arkansas Governor and Republican presidential candidate Mike Huckabee speaks to reporters in Little Rock, Ark. Republican presidential candidate Mike Huckabee came under fire for a tweet he sent while the Democratic candidates took to the debate stage in Las Vegas, Nevada, on Tuesday, NBC News reported. The former Arkansas governor likened his trust of Vermont senator Bernie Sanders to "a North Korean chef with my labrador!" I trust @BernieSanders with my tax dollars like I trust a North Korean chef with my labrador! #DemDebate — Gov. Mike Huckabee (@GovMikeHuckabee) October 14, 2015 The response on Twitter to Huckabee was swift with many calling his comments racist. Huckabee followed up his tweet hours later, writing, "Poor liberals think it's racist to deplore a brutal dictatorship."

The text above received the following topic labels:

'Elections'
'Chefs'
'Culinary terminology'
'Occupations'
'Restaurant staff'
'Restaurant terminology'
'Skills'
'Internal territorial disputes of Canada'
'Labrador'
'Discrimination'
'Hatred'

'Politics and race'
'Racism'

The results show lesser accuracy than with Set 1, which is likely explained by the prevalence of shorter texts with incomplete sentences. Potentially, improvement may be achieved by considering other than nouns parts of speech, stable phrases and colloquialisms, and supplementing the Wikipedia with dictionaries or other encyclopedias to account for terms unavailable in the Wikipedia.

Set 3 – 'Coffee or tea' news: Total: 100 automatically sampled from webhose.io (537 total). The quantitative assessment showed average best similarity of synonymous terms 0.80 and the worst of 0.21.

An example of the analyzed text:

> *(Extract) World Coffee Producers Forum declares need for action on coffee price Posted on Wednesday 27th, March 2019. Share World Coffee Producers Forum organisers have released an official declaration calling for serious and immediate action to be taken on the historically low international coffee price. Thirteen coffee producers' groups, including the Federación Nacional de Cafeteros de Colombia (FNC), are listed on the declaration....Producers who stay in coffee will not be able to afford the proper care of their farms and their coffee which leads to improper fertilization and care of the trees, affects quality and deprives consumers the diversity that they enjoy today. Adaptation and mitigation of the effects of climate change are other burdens that falls on the shoulders of producers.*

The text above received the following topic labels:

'Crops'
'Coffee'
'Herbal and fungal stimulants'
'Hot drinks'
'Non-alcoholic drinks'
'Turkish words and phrases'
'Pricing'

This category was much better defined than the previous two. This clarity also resulted in better results. The quantitative assessment showed best results of 0.80 and the worst of 0.21. For comparison the results of Wu-Palmer similarity for clearly related words 'dog' and 'animal' are: best=0.88, worst=0.27.

Qualitative evaluation: The Wu-Palmer similarity score gives a very rough estimate of the appropriateness of the topic labeling. Qualitative assessment is a much better evaluation method for the goal of the topic labeling method. In this first version of the method, to complement the results of the quantitative (Wu-Palmer similarity) assessment, a qualitative assessment was performed on 10% of texts in all three sets. The results are presented in Table 1.

Table 1. Qualitative evaluation of topic labels.

Data set	Topic of the text, %	Related topic, %	Unrelated label, %
Set 1	65,9	30,8	3,3
Set 2	75	22,2	1,4
Set 3	60	22,9	17,1

Overall, the short twitter-type texts in Set 2 showed the lowest labeling performance, which can be explained by the lack of information that the method receives to evaluate the relevance of a particular topic. This problem can be addressed using additional topic filtering methods, for example, measuring the similarity between the selected labels and the analyzed text.

5. CONCLUSIONS

The method proposed in this paper performs unsupervised topic labeling of texts of unknown context and structure. This functionality is similar to finding synonyms, however synonyms of the whole text as opposed to the synonyms of a word. The method is useful for generating short representation of text that can be used in text comparisons. The topic labels give a good sense of the range of the topics discussed, which can be used practically for essay scoring to ensure the completeness of the topic coverage.

The method is potentially language independent and can be applied to any language that is used in the Wikipedia and can be tagged by parts of speech.

Future development of the method should focus on the improvement of the labeling ability of the method for shorter texts. Another important development is the ability to distinguish the main topic label, subtopic labels, and related topic labels. To improve the reliability of the qualitative assessment of the method, the topic extraction ability of the method should be verified using human-coding of the same texts.

6. REFERENCES

[1] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[2] H. P. Edmundson, "New methods in automatic extracting," *Journal of the ACM (JACM)*, vol. 16, no. 2, pp. 264–285, 1969.

[3] M. Allahyari *et al.*, "Text summarization techniques: a brief survey," *arXiv preprint arXiv:1707.02268*, 2017.

[4] Y. Gong and X. Liu, *Test summarization using relevance measures and latent semantic analysis*. Google Patents, 2009.

[5] J. Steinberger and K. Jezek, "Using latent semantic analysis in text summarization and summary evaluation," *Proc. ISIM*, vol. 4, pp. 93–100, 2004.

[6] O. Dokun and E. Celebi, "Single-Document summarization using Latent Semantic Analysis," *International Journal of Scientific Research in*

Information Systems and Engineering (IJSRISE), vol. 1, no. 2, pp. 57–64, 2015.

[7] X. Han, T. Lv, Q. Jiang, X. Wang, and C. Wang, "Text summarization using sentence-level semantic graph model," in *2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, 2016, pp. 171–176.

[8] K. Ganesan, C. Zhai, and J. Han, "Opinosis: A graph based approach to abstractive summarization of highly redundant opinions," in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 2010, pp. 340–348.

[9] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," *arXiv preprint arXiv:1509.00685*, 2015.

[10] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," *arXiv preprint arXiv:1705.04304*, 2017.

[11] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, 1994, pp. 133–138.

[12] Princeton University, "WordNet. A lexical database for English," *Princeton University*, 01-Jun-2019. [Online]. Available: https://wordnet.princeton.edu/