Text summarization based on quality phrase

Wei Chen Lin (wclin2), Xi Li (xil2), Yekai Yu (yekaiyu2)

ABSTRACT
Automatic summarization is designed to catch the main idea of a given textual document. With massive corpus, like news, generated every day, it’s necessary to have text summarization to help people mine the information. To quickly and precisely perform text summarization on unknown domain corpus, in this paper, an adopted text ranking algorithm is proposed. The algorithm employs the quality phrases and some sentence-level textual features to improve the quality of the summarization results. The summary is evaluated by using ROUGE to quantify the similarity between selected sentences and article title.

KEYWORDS
Text Summarization, TextRank, Unknown Domain, Large Corpus, Quality Phrase

1 INTRODUCTION
Quick and precise text summarization is useful for extracting main idea from massive text documents. Text summarization is classified into abstractive method [18], [20] and extractive method, where extractive method tries to select the sentences that could best represent the whole document while abstractive usually generate summaries via learning in a neural network. Compared to abstractive method, extractive method is direct and quick, and it usually only relies on the objective document itself. In previous extractive method, the domain knowledge which can provide valuable information for text summarization was not introduced.

In this study, the framework is built on an extractive summarization method called TextRank. TextRank is a graph-based ranking algorithm, scoring each individual sentence by calculating the cosine similarity to other sentences. To help discover domain knowledge and enhance the quality of text summarization, quality phrase is introduced in our modified method. Intuitively, a precise and informative summary contains the most important sentences of the document. Those important sentences usually include outstanding contents: names, locations, and many other types of words/phrases. Therefore, those contents are considered as quality phrases.

The tool, AutoPhrase, was used in our method, which facilitates quality phrase discovery and document clustering. With the assistance of quality phrase, quality phrases are mined and converted to vectors by word embedding to vectorize corresponding document and find background domain knowledge. Moreover, the sentences containing quality phrases will be emphasized during the ranking process.

There are four features of sentence employed in this model:
• Sum of quality phrase score(s) in one sentence
• Length of sentence
• Number of occurrence of each word/phrase within the document
• Cosine similarity between sentences

Based on the features described above, the scores of each sentence are computed for ranking process. The similarity between generated sentence and title is evaluated by ROUGE-1, ROUGE-2.

2 RELATED WORK
In this section, the existing methods that are related to this topic are summarized.

TextRank:
TextRank [1] is a text ranking model for text processing. In previous studies, most of the text summarization methods are supervised, which means that a large number of well-labeled corpus is required during the training phase. TextRank can extract keywords and sentences in an unsupervised way, outperforming over previous works.

As a graph-based ranking algorithm, TextRank is essentially a method to determine the importance of vertex based on global graph recursion. The basic idea of a graph-based ranking model is "vote" or "recommendation". When a vertex links to another vertex, the link is basically also a vote for another vertex. The higher the number of votes casts on the top, the higher the importance of the vertex is. In addition, the importance of voting is also considered in the ranking model. Therefore,
the score associated with a vertex is determined based on the number of votes cast and the number of points in the vote.

\[
WS(V_j) = (1 - d) + d \times \sum_{j \in \text{Out}(V_i)} \frac{1}{|\text{Out}(V_j)|} WS(V_j)
\]

where \(d\) is the damping coefficient, set between 0 and 1. It has the role of integrating into the model the probability of jumping from a given vertex to another random vertex in the graph.

When surfing the Internet, this graph-based ranking algorithm based on the model of random surfing, where a user randomly clicks on links with probability \(d\) and jumps to a new page with probability \(1 - d\). The factor \(d\) usually sets to 0.85 [3].

Assigning arbitrary values to every node within the graph, the process keeps running until convergence. After running the algorithm, a score is associated with each vertex, which represents the “importance” of the vertex within the graph.

As shown in Figure 1, let \(G = (V, E)\) be a directed graph with a set of vertices \(V\) and a set of edges \(E\). \(\text{In}(V_i)\) is the set of vertices (predecessors) that point to \(V_i\), \(\text{Out}(V_j)\) is the set of vertices (successors) that \(V_i\) points to. The score of \(V_i\) is defined as follows:

\[
S(V_i) = (1 - d) + d \times \sum_{j \in \text{In}(V_i)} \frac{1}{|\text{Out}(V_j)|} S(V_j)
\]

TextRank successfully identifies the most important sentences on the basis of textual information itself. Unlike other supervised systems, they build the collection of other articles for training. TextRank is completely unsupervised, and only depends on the given text to extract the summary. This is a summarization model closed to human when produce summary on a given document. Another important aspect of TextRank is that it presents an overall ranking of the given text, which means it can be easily adapted to extract either long or short summaries. The combination of key phrases and sentence extraction techniques as a way to construct long/short summaries is also explored. Finally, another merit of using TextRank to build up the extraction summary is that it does not require training corpus, which makes it easy to adapt to other languages or domains.

The general steps would be as follow,
1. Identify text units such as sentences, words etc., and add them as vertices in the graph.
2. Identify relations (Cosine similarity, semantic similarity) that connect such text units, and use these relations to draw edges between vertices in the graph. Edges can be either directed or undirected, weighted or unweighted.
3. Iterate the graph-based ranking algorithm until convergence.
4. Sort vertices based on their final score. Use the values attached to each vertex for ranking / selection decisions.

**TextRank (Summa):**
This Python package proposes some changes over the original TextRank algorithm, intending to change the way in which distances between sentences are computed to weight the edges of the graph used for PageRank.
1. **Longest Common Substring** From two sentences we identify the longest common substring and report the similarity to be its length [15].
2. **Cosine Distance** The cosine similarity is a metric widely used to compare texts represented as vectors. We used a classical TF-IDF model to represent the documents as vectors and computed the cosine between vectors as a measure of similarity. Since the vectors are defined to be positive, the cosine results in values in the range [0,1] where a value of 1
represents identical vectors and 0 represents orthogonal vectors [16].

3. BM25 BM25 / Okapi-BM25 is a ranking function widely used as the state of the art for Information Retrieval tasks. BM25 is a variation of the TF-IDF model using a probabilistic model [17].

\[
BM25(R, S) = \sum_{i=1}^{n} IDF(s_i) \cdot \frac{f(s_i, R) \cdot (k_1 + 1)}{f(s_i, R) + k_1 \cdot (1 - b + b \cdot \frac{|R|}{\text{avgDL}})}
\]

where \(k\) and \(b\) are parameters. We used \(k = 1.2\) and \(b = 0.75\). \(\text{avgDL}\) is the average length of the sentences in our collection.

This above function definition implies that if a word appears in more than half the documents of the collection, it will have a negative value.

3 METHOD

3.1. Overview

Given an unknown-domain collection of documents \(D\), a quality-phrase-based clustering process is applied to roughly concentrate similar topic documents into one cluster. In each cluster, quality phrases within the cluster are discovered by AutoPhrase, producing a scored quality phrase rank list. Based on the scores of high-quality phrases, a dictionary \(Dict\) with quality phrase as key and score as value is built. The adopted graph-based ranking algorithm incorporates the scores of high-quality phrases applied to each sentence. More frequent and higher quality phrase scores will produce higher ranking in the final results. Figure 2 shows the pipeline of our work.

3.2. Cluster Construction

Given an unknown-domain collection of documents \(D\), AutoPhrase is applied to \(D\) to locate quality phrases \(Q_i\) in each document \(d_i\). According to the quality phrases \(Q_i\), Word2Vec is used to generate word embedding \(w_i\) for each quality phrase \(q_i\) in document \(d_i\). The weighted mean vector \(W_i\) of vectors \(\{w_{i}\}_0^N\), where \(N\) is the number of vectors in document \(d_i\), represents this document \(d_i\). The weighted mean \(W_i\) incorporates the number of occurrence of quality phrase times the word embedding \(w_i\) and the score of phrases \(s_i\), summing the multiplication up and dividing the total number of word embeddings.

\[
W_i = \frac{\sum_{0}^{N} n_i w_i s_i}{\sum_{0}^{N} n_i}
\]

where \(n_i\) is the number of occurrence of \(w_i\).

According to the vectors of document, K-means clustering method is applied to distribute each document into cluster. In this study, the number of clusters \(k\) is set as 10.

3.3. Quality phrase dictionary construction

In this step, AutoPhrase is performed per cluster to obtain the scored quality phrase rank lists of each cluster. For each list, a dictionary is built with the quality phrase string as key, the quality score as value. When performing later sentence scoring, the corresponding scores will be referred to indicate importance of sentence within the document.
3.4. Ranking Approach
Since the sentences are comprised of multiple phrase(s), understanding the importance of each phrase can be the criteria for sentence selection. Below are four features that can determine the importance of phrases by assigning weights to them.

3.4.1. Domain Knowledge
Based on the intuition that the informative summarization is more likely to contain quality phrases, incorporating quality phrase scores to TextRank can improve the final summaries. Each sentence vector is multiplied by weight $w_j = \exp(\sum_i S(\text{phr}_i) + 1.5)$, where $\text{phr}_i$ represents the high-quality phrases contained in $\text{sentence}_j$, $S(\text{phr})$ is the score of phrase obtained from AutoPhrase. The constant 1.5 is tuned by experiments. The weight of quality phrases is exponentiated to emphasize the importance of high score phrases.

3.4.2. Frequency
To find the summary of articles, the most intuitive way is to look at the word distribution. For example, in the health care article, the word ‘health’ is likely to appear frequently. So, by looking at the occurrences of words, people can have the big picture of the article. Therefore, we count every word within the target article, except the stop words. The higher frequency the word has, the more weight it gets.

3.4.3. Cosine similarity
Cosine similarity evaluates the contextual similarity between two sentences. First, sentences are converted into vectors by using pairwise bag of words. Second, the relation between sentences is found by using cosine similarity. Third, a similarity matrix is built as the input for text ranking algorithm.

3.4.4. Length of sentences
Generally, short sentences may carry fewer information. Therefore, after getting the sentence rank, a penalty for short sentence is introduced to produce more robust results.

4 RESULTS
In this section, we delivered an example of our modified TextRank model and the comparison between the modified model and other popular TextRank packages.

4.1. Sample Document
WASHINGTON — Congressional Republicans have a new fear when it comes to their health care lawsuit against the Obama administration: They might win. The incoming Trump administration could choose to no longer defend the executive branch against the suit, which challenges the administration’s authority to spend billions of dollars on health insurance subsidies for and Americans, handing House Republicans a big victory on issues. But a sudden loss of the disputed subsidies could conceivably cause the health care program to implode, leaving millions of people without access to health insurance before Republicans have prepared a replacement. That could lead to chaos in the insurance market and spur a political backlash just as Republicans gain full control of the government. To stave off that outcome, Republicans could find themselves in the awkward position of appropriating huge sums to temporarily prop up the Obama health care law, angering conservative voters who have been demanding an end to the law for years. In another twist, Donald J. Trump’s administration, worried about preserving executive branch prerogatives, could choose to fight its Republican allies in the House on some central questions in the dispute. Eager to avoid an ugly political pileup, Republicans on Capitol Hill and the Trump transition team are gaming out how to handle the lawsuit, which, after the election, has been put in limbo until at least late February by the United States Court of Appeals for the District of Columbia Circuit. They are not yet ready to divulge their strategy. “Given that this pending litigation involves the Obama administration and Congress, it would be inappropriate to comment,” said Phillip J. Blando, a spokesman for the Trump transition effort. “Upon taking office, the Trump administration will evaluate this case and all related aspects of the Affordable Care Act.” In a potentially decision in 2015, Judge Rosemary M. Collyer ruled that House Republicans had the standing to sue the executive branch over a spending dispute and that the Obama administration had been distributing the health insurance subsidies, in violation of the Constitution, without approval from Congress. The Justice Department, confident that Judge Collyer’s decision would be reversed, quickly appealed, and the subsidies have remained in place during the appeal. In successfully seeking a temporary halt in the proceedings after Mr. Trump won, House Republicans last month told the court that they “and the’s transition team currently are discussing potential options for resolution of this matter, to take effect after the’s inauguration on Jan. 20, 2017.” The suspension of the case, House lawyers said, will provide the and his future administration time to consider whether to continue prosecuting or to otherwise resolve this appeal.” Republican leadership officials in the House acknowledge the possibility of “cascading effects” if the payments, which have totaled an estimated $13 billion, are suddenly stopped. Insurers that receive the subsidies in exchange for paying costs such as deductibles and for eligible consumers could race to drop coverage since they would be losing money. Over all, the loss of the subsidies could destabilize the entire program and cause a lack of confidence that leads other insurers to seek a quick exit as well. Anticipating that the Trump administration
might not be inclined to mount a vigorous fight against the House Republicans given the ‘s dim view of the health care law, a team of lawyers this month sought to intervene in the case on behalf of two participants in the health care program. In their request, the lawyers predicted that a deal between House Republicans and the new administration to dismiss or settle the case “will produce devastating consequences for the individuals who receive these reductions, as well as for the nation’s health insurance and health care systems generally.” No matter what happens, House Republicans say, they want to prevail on two overarching concepts: the congressional power of the purse, and the right of Congress to sue the executive branch if it violates the Constitution regarding that spending power. House Republicans contend that Congress never appropriated the money for the subsidies, as required by the Constitution. In the suit, which was initially championed by John A. Boehner, the House speaker at the time, and later in House committee reports, Republicans asserted that the administration, desperate for the funding, had required the Treasury Department to provide it despite widespread internal skepticism that the spending was proper. The White House said that the spending was a permanent part of the law passed in 2010, and that no annual appropriation was required — even though the administration initially sought one. Just as important to House Republicans, Judge Collyer found that Congress had the standing to sue the White House on this issue — a ruling that many legal experts said was flawed — and they want that precedent to be set to restore congressional leverage over the executive branch. But on spending power and standing, the Trump administration may come under pressure from advocates of presidential authority to fight the House no matter their shared views on health care, since those precedents could have broad repercussions. It is a complicated set of dynamics illustrating how a quick legal victory for the House in the Trump era might come with costs that Republicans never anticipated when they took on the Obama White House.

4.2. Summaries

<table>
<thead>
<tr>
<th>TextRank + Knowledge base</th>
<th>1. The incoming Trump administration could choose to no longer defend the executive branch against the suit which challenges the administrations authority to spend billions of dollars on health insurance subsidies for and American’s handing house republicans a big victory on issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Anticipating that the Trump administration might not be inclined</td>
<td>to mount a vigorous fight against the house republicans given the s dim view of the health care law a team of lawyers this month sought to intervene in the case on behalf of</td>
</tr>
<tr>
<td>3. In a potentially decision in 2015 judge rosemary M’collyer ruled that house republicans had the standing to sue the executive branch over a spending dispute and that the Obama administration had been distributing the health insurance subsidies in violation of the constitution without approval from congress</td>
<td></td>
</tr>
<tr>
<td>4. In their request the lawyers predicted that a deal between house republicans and the new administration to dismiss or settle the case will produce devastating consequences for the individuals who receive these reductions as well as for the nations health insurance and health care systems generally</td>
<td></td>
</tr>
<tr>
<td>5. But on spending power and standing the trump administration may come under pressure from advocates of presidential authority to fight the house no matter their shared views on health care since those precedents could have broad repercussions</td>
<td></td>
</tr>
<tr>
<td>TextRank (Gensim)</td>
<td>1. the incoming trump administration could choose to no longer defend the executive branch against the suit which challenges the administrations authority to spend billions of dollars on health insurance subsidies for and americans handing house republicans a big victory on issues</td>
</tr>
<tr>
<td>2. anticipating that the trump administration might not be inclined to mount a vigorous fight against the house republicans given the s dim view of the health care law a team of lawyers this month sought to intervene in the case on behalf of</td>
<td></td>
</tr>
</tbody>
</table>
two participants in the health care program.

3. in a potentially decision in 2015 judge rosemary m collyer ruled that house republicans had the standing to sue the executive branch over a spending dispute and that the obama administration had been distributing the health insurance subsidies in violation of the constitution without approval from congress.

4. in their request the lawyers predicted that a deal between house republicans and the new administration to dismiss or settle the case will produce devastating consequences for the individuals who receive these reductions as well as for the nations health insurance and health care systems generally.

5. but on spending power and standing the trump administration may come under pressure from advocates of presidential authority to fight the house no matter their shared views on health care since those precedents could have broad repercussions.

4.3. Evaluation
In this project, precision, recall and F-score and ROUGE are used as criteria for algorithm judgement. We will briefly show the definition of score and show out experiment results compared to the current methods.

Precision: In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the query.

\[
\text{Precision} = \frac{|\text{relevant documents} \cap \text{retrieved documents}|}{|\text{retrieved documents}|}
\]

Recall: In information retrieval, recall is the fraction of the relevant documents that are successfully retrieved.

\[
\text{Recall} = \frac{|\text{relevant documents} \cap \text{retrieved documents}|}{|\text{relevant documents}|}
\]

F-score: This measure is approximately the average of the two when they are close.

\[
F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [13] is one of the most widely used standard metric to evaluate the quality of sentences in text summarization. It compares the generated summary to summaries made by humans and gives scores, which saves great human resource on judgements. In the ROUGE package, ROUGE-N is the most widely applied metric. ROUGE-N, using n-gram co-Occurrence statistics, could be expressed as:

\[
\text{ROUGE - N} = \frac{\sum_{S \in \text{HumanSummary}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \text{HumanSummary}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)}
\]

where \( n \) is the length of n-grams.

ROUGE-N is a recall-related measure. Typically, the summary sharing more common n-gram words with human summaries tend to have higher score, which is straight and intuitive.

There are also other ROUGE metrics that could be applied under various situations. ROUGE-L takes the longest common subsequence into consideration and ROUGE-W add weights to ROUGE-L. ROUGE-S considers skip-bigrams rather than continuous n-grams.

<table>
<thead>
<tr>
<th>TextRank (Summa)</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. trump about to take control of the white house it would seem a dark time for the renewable energy industry.</td>
<td>0.300</td>
<td>0.538</td>
<td>0.381</td>
</tr>
<tr>
<td>2. after all mr.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. trump has mocked the science of global warming as a chinese hoax threatened to kill a global deal on climate change and promised to restore the coal industry to its former glory.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. we do not know for sure that the new york wind farm will get built but we do know this the energy transition is real and mr.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. when mr.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For the comparison, we input the top 5 summaries obtained from TextRank (Gensim), TextRank (Summa) and our algorithm into the PyRouge package (from https://github.com/pyrouge/PyRouge). From the above result, it obviously shows that our algorithm has the highest score on recall which means that our algorithm can provide the most comprehensive summaries and also rank highest on F-score. However, we get the lowest score on the precision. The problem might be caused by taking knowledge base into consideration, since some quality phrases coming from related text may not be appear in the target text. In order to solve the problem, reducing the weight that quality phrases contribute might be an applicable approach. In short, our algorithm outperforms other method on recall and F-score, meanwhile, still maintain acceptable scores on precision.

5 CONCLUSIONS
This work presents two variations to the TextRank algorithm. First, we extract high quality phrases from the articles within the same domain. Second, the quality phrases incorporate with phrase frequencies as weights for sentence ranking. Finally, we get the summary representing the target articles. In short, based on the experiment results, TextRank can be improved by taking advantage of the knowledge base and by considering the textual features.

REFERENCES
[19] GANESAN, Kavita; ZHAI, ChengXiang; HAN, Jiawei. Opinosis: a graph-based approach to abstractive