Assessing the Quality of Automatic Summarization for Peer Review in Education

Ferry Pramudianto[1], Tarun Chhabra[2], Edward F. Gehringer[3], Christopher Maynard[4]

North Carolina State University
890 Oval Drive, Raleigh, NC 27695
{[1]fferry, [2]tchhabr, [3]efg, [4]ctmaynar}@ncsu.edu

ABSTRACT

Technology supported peer review has drawn many interests from educators and researchers. It encourages active learning, provides timely feedback to students and multiple perspectives on their work. Currently, online peer review systems allow a student's work to be reviewed by a handful of their peers. While this is quite a good way to obtain a high degree of confidence, reading a large amount of feedback could be overwhelming. Our observation shows that the students even ignore some feedback when it gets too large. In this work, we try to automatically summarize the feedback by extracting the similar content that is mentioned by the reviewers, which would capture the strength and weaknesses of the work. We evaluate different auto summarization algorithms and length of the summary with educational peer review dataset, which was rated by a human. In general, the students found that medium-size generated summaries (5-10 sentences) encapsulate the context of the reviews, are able to convey the intent of the reviews, and help them to judge the quality of the work.

Keywords

Automatic Summarization, Peer Assessment, Peer Review, Evaluation.

1. INTRODUCTION

Many pieces of evidence show that peer review could bring benefits for educators and learners [1,2]. It encourages active learning by requiring students to exercise their knowledge when they assess others' work. It also induces students to give extensive written feedback, which is typically more reflective than oral feedback. It can enhance learning outcomes, provide a useful assessment in ill-structured domains, and in general, enhance the formative feedback that students receive. Enhancing formative assessment (feedback) is an important aspect of education as it provides a reflection to improve learning and teaching strategies continuously. Feedback enhances the development of an individuals' metacognitive and critical thinking skills [3]. Giving feedback and adjusting one's own behavior based on feedback received from others is a skill that can, and should, be acquired through practice and training [4]. Peer feedback benefits this interaction more by allowing learners and teachers to experience multiple perspectives on the students' performance, rather than the singular voice of a teacher. In addition, including students in the assessment process also improves the learners' buy-in of the assessment process and helps them to focus their learning strategy better.

Peer assessment is able to provide timely, actionable feedback to learners, which help them to continuously measure how well they have mastered the subjects. Through continuous measurements, learners could adjust the way they learn quickly, in case the method that they use doesn't work quite well. That is essential to the development and execution of self-regulatory skills [5]. Moreover, as a learning tool, assessing their peers can provide students with skills to form judgments about what constitutes high-quality work [6]. In his synthesis of over 500 meta-analyses, Hattie found feedback to be among the top influences on student achievement with an average effect size of .79 [7]. Feedback in the form of cues—such as peer-review comments—had effects sizes approaching 1.0 [7].

While getting feedback from different perspectives is helpful, the extent of these feedbacks could easily overwhelm the students, and may even cancel out the desired effects of helping students to identify their strengths and weaknesses related to the assignment. Our preliminary study shows that within traditional classes that are enrolled in Expertiza, the students get feedback from 3-5 reviewers. In a few cases, students could even get feedback from more than 10 reviewers. Each feedback could be very extensive depending on the given rubric. In Expertiza, we found that the rubric contains on average, 8-9 criteria, with two courses even having 159 criteria. Each student / team receives reviews from 5 reviewers on average, and the highest number of reviews was from 72 reviewers, within multiple rounds. The number of words in the feedback that each student / team receives from multiple reviewers on average is 175 words, and the most extensive feedback reaching 8500 words. When assuming that a sentence in average consist of 15 words, it means that each reviewee gets approximately 11-12 sentences, but it could also reach 566 sentences, which would clearly be overwhelming. It is even worse when the students have to read feedback from multiple reviewers that sounds repetitive.

One way to improve this situation is to provide a summary of feedback when the amount has grown extensive. There are a few different ways that summaries could be used in peer review systems. First, the instructor could benefit from having a summary of the qualitative feedback that his students get from their peers. It allows the instructor to sense the general of the problems of his class in that particular assignment. On the student side, having a summary of the feedback could also help them to get a quick glimpse of their strength and weaknesses. This paper focuses on studying providing the summaries for the students.

A summary for the students could be visualized differently. For instance, when the peer review systems rely on rubrics, a summary could be provided for each piece of feedback given for a criterion such as depicted in Figure 1. Alternatively, the summary could be presented as a holistic narrative that includes the rubric and the feedback. The summary could also be presented as bullet points grouped under the tone polarity that may resemble the pros and cons of the work. For this study, we choose to show the summary as a narrative since it is the most compact form to show the summary.

The remainder of the paper is organized as the following. Section 2 discusses related works. Section 3 discusses approaches to evaluate text summarization, section 4 discusses our study and the results. Section 5 conclude our paper and discuss the future work.

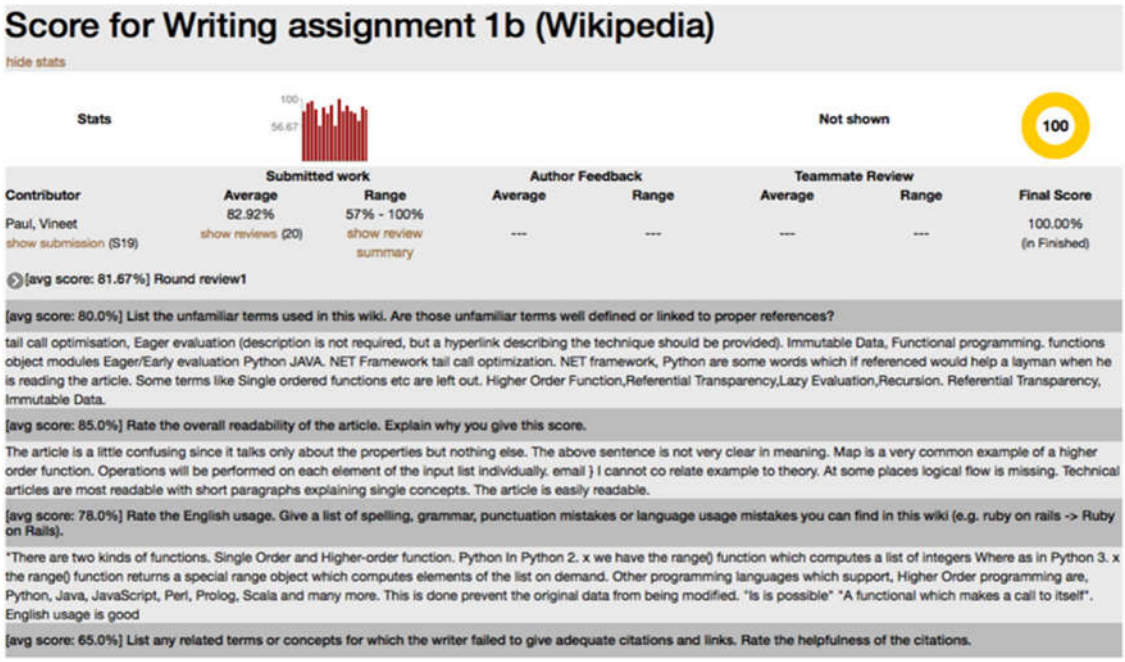

Figure 1. Expertiza summarizes reviews for the students according to the rubric criteria.

2. RELATED WORKS

According to the authors' knowledge, previous work on auto-summarizing peer reviews is not available. However, product reviews and microblogging have taken advantage of the technology. Summarizing product reviews and microblogging have some similarity with summarizing peer review. They deal with multiple short unstructured-free-texts schemes, which may contain some degree of repetitions. Moreover, reviews are usually written casually, which sometimes consist of incomplete sentences or incorrect grammar. Popular products may have hundreds or even thousands of reviews, thus, it is difficult for a user to extract useful information such as product qualities and services [8-11].

Early works on product reviews summarization focus on extracting overall customer sentiments (e.g., positive, negative, or neutral), but it has developed further and able to identify topics based on their salience [12], or extract features of a product and identify opinions that associate with them [13]. Most of these approaches depend on the co-occurrence of words. Another interesting approach tries to identify the pros and cons of a product from the reviews by using Maximum Entropy classification [14].

2.1 Text Summarization Approaches

Text summarization is covered under the umbrella of Human Language Technologies(HLT). The goal has been to enable communication with machines using natural skills. HLT also encompasses areas such as retrieval, sentiment analysis, and text classification. Most of these techniques have emerged out of a need to condense content by selection or generalization. With the information explosion, new text summarization techniques have emerged in the last decade. In this section, we will try to broadly define this domain and various methods that have been used for text summarizations

Research on text summarization has existed from the late '50s [15]. Since then other approaches have been proposed, however, there has not been an approach that is able to determine the quality of summaries produced by a human. This shows that producing summaries is a complex problem. The algorithm must be able to detect redundancy, sentence ordering, temporal dimension, identify relevant content and merge it into new fragments of information. Das and Martins [16] emphasize 3 important aspects that characterize research on automatic summarization:

- Summaries may be produced from a single document or multiple documents
- Summaries should preserve important information
- Summaries should be short.

Text summary can be done by extraction, which combines the most important subset of the original document, or abstraction that introduces new concepts that abstract the original content. Condensing large amounts of text into relevant and necessary summary can be challenging. With over 50 years of research in this area, different types of summarization techniques have emerged.

2.1.1 Statistical-based approach

Luhn algorithm [15] is one of the oldest known works in this field. This approach is fundamentally based on a finding of most frequently used words. However, words that do not carry semantic value are ignored while computing summaries eg. "a" or "the". This approach is also known as Term Frequency- Inverse Document Frequency (TF-IDF). Any content is treated like a bag of words and sentence frequency is calculated using weighted term frequencies and inverse sentence frequencies. Using these frequencies sentence vectors are created and scored, the highest scoring ones become part of the summary. Filatova and Hatzivassiloglou [17] have claimed that these methods may not prove to be generating high- quality and relevant summaries. Another approach called SumBasic [18] works based on an observation that the relative frequency of a non-stop word in a document set is a good predictor of a word appearing in a human summary. It assigns scores to each sentence based on how many high-frequency word appears in it. KLSum extract sentences from the source documents which have the highest similarity to the overall distribution of words in the entire document cluster. It measures the similarity across word distributions on KL-Divergence [19].

2.1.2 Topic-based approaches

This idea was developed based off of the cue-based method. It is based on the hypothesis that relevance is computed based on presence or absence of certain words in a dictionary. Edmundson [20] was one of the first attempts in this area. Sentences such as "in conclusion" or "the aim of this" are treated as good identifiers of relevant sentences. This technique also comprises of title method and location method. Title method assigns higher weights to words appearing in title and subtitles while location method looks for higher weighted words based on their location in the start of paragraphs. Algorithms such as Edmundson require a set of sentences as input in order to condense text based on such criteria. Hence, this also limits the use of this approach wherein the structure of sentences may not be known ahead of time and may vary vastly with each context.

2.1.3 Graph/cluster-based approaches

LexRank [21] and TextRank [22] are some of the algorithms which use this approach. The node of each graph is used to represent the text element. Edges define a semantic relationship between each node. These are taken as inputs to the connectivity of a graph and help in exploring various topologies that may be possible. In TextRank, vertices are representative for the units to be ranked. For sentence extraction, a vertex is added to the graph for each sentence in the text. The edges represent a "similarity" relation between sentences, where "similarity" is measured as a function of their content overlap, which can be determined as the number of common tokens between the lexical representations of the two sentences. The content overlap is normalized by dividing with the length of each sentence.

Walking through these graphs result into different summaries, could be just a random traversal of graph or even based on some weights. Computing shortest in such graphs may even lead to an abridged version of the larger text. These techniques have proved to be effective in multiple domains such as biomedical, image captions and social media summarizations. Our experiment uses both TextRank to condense student and reviews and see if they could be used effectively in this domain.

2.1.4 Discourse-based approaches

Approaches mentioned in the previous sections do face problems from the linguistic point of view. The sentences formed by these algorithms do not have a rhetorical structure. Hence, a summary produced by such algorithms often not cohesive nor coherent in terms of language. As a result of which new approaches were explored which could combine both statistical as well as linguistic techniques.

Latent Semantic Analysis (LSA) uses Singular Value Decomposition (SVD) technique to find orthogonal dimensions of a multi-dimensional data [23]. This technique involves creating a matrix of document and words that are semantically related to each other. Each column represents the weighted term frequency vector of a sentence in the set of documents. Then singular value decomposition (SVD) is used on the matrix to derive the latent semantic structure. Words that occur in related contexts are placed closer in the matrix which helps in establishing semantics just the way they occur in a human brain. LSA uses word vectors to find relations between different sentences based on their mutual orthogonal positions. If a word combination pattern is salient and recurring in document, this pattern will be captured and represented by one of the singular vectors. Each singular vector represents a salient topic/concept of the document, and the magnitude of its corresponding singular value represents the degree of importance of the salient topic/concept. The summarization process chooses the most informative sentence for each topic by selecting the largest index value in K-th right singular vector in the matrix. This form of chaining of sentences improves the quality of summaries and also helps capture the context of the text.

2.1.5 Machine learning based approaches

Hidden Markov Models and Bayesian methods are some the well-known machine learning techniques. However, text summarization is not just limited to using only these techniques. Some algorithms such as NetSum [24] and RankNet [25] use a learning algorithm in order to score sentences and then extract the most relevant portions out of a text. However just like topic based approach, this seems to perform well in classified knowledge domains. When exposed to a scenario where each text may vary from another the learning aspect of the algorithm may not be well trained with vacillating texts. It could be identified as an interesting problem in the coming years.

3. SUMMARY EVALUATION

Text summarization could be evaluated using diverse approaches as depicted in Figure 2. When evaluating a large scale of text, the automatic approach is usually chosen since it requires significantly less effort than the manual approach. However, the manual approach usually yields more reliable results, relative to text that is natural to human language. The evaluation measures of the text summary can be determined by intrinsic content evaluation which compares its content with an ideal summary [26]. It usually calculates the co-selection metric such as precision, recall, and F-score to measure how many ideal sentences exist in the summary. Another approach evaluates the content by comparing the words in the sentences, instead of the whole sentences. This approach allows comparing the automatic extracts with the summaries generated by human, although they contain newly written sentences.

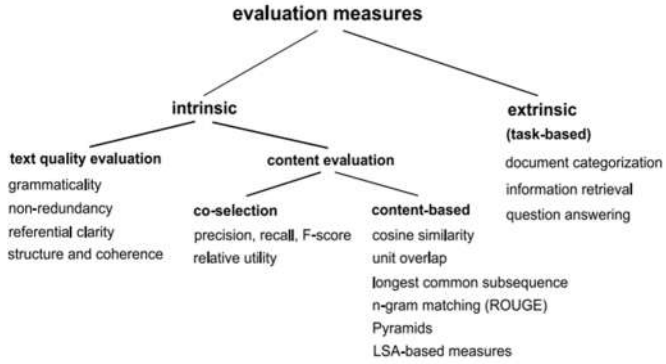

Figure 2. The taxonomy of summary evaluation measures [26].

3.1 OUR STUDY

Although previous studies on these algorithms exist, they focus on the intrinsic content evaluation exists, which compare the automatically generated summary with human generated summary in order to obtain the recall, precision, and F-Score. We believe that even human can generate summaries in many different ways, and comparing the automatically generated summaries with only one summary that is generated by human, would not be able to help us determining their usefulness for our use case. Thus, instead of focusing on an intrinsic content evaluation, we design our study similar to usability study, in which we presented the generated summaries to our potential users and have them rate their quality based on a set of rubric criteria.

Additionally, the study was designed to find out whether some of the existing text summarization algorithms are yield different perceived qualities as well as finding out the preferred length of the review summary for the educational peer review. After reviewing the literature, we choose 3 algorithm samples that are created using different approaches. First, based on the Graph analysis category, we choose TextRank since it is quite simple to implement and performs quite fast to extract summaries from a large amount of reviews. Second, we choose LSA (Latent Semantic Analysis), which is based on Discourse since it is the most popular algorithm in this category. Third, we choose KLSum since it a significant improvement of algorithms in this category that is mostly used. Since we cannot know in advance the application domain of the assignments, it would be impractical for our use case to use algorithms that require domain knowledge such as the topic based and machine learning based. Therefore, decided to exclude them in our study.

To perform our study, we use an open source implementation of these algorithms in python that can be found at https://pypi.python.org/pypi/sumy.

We took two samples of peer reviews from a writing assignment in the Object-Oriented Languages and Systems graduate course. We used these three algorithms to generate summaries which have a length of 3, 5 and 10 sentences, which result in nine combinations as shown in table 1. As an initial attempt, we simply present the summary as a narrative that combines feedback based on different criteria.

To administer the survey, we use several google forms that contain the 9 summaries, each of which is followed by the survey questions. The participants received the summaries in different order to cancel out the survey fatigue affecting the survey results of any particular algorithm. In order to be sure that each order is performed by approximately the same number of participants, the order was calculated using the Latin square technique. After reading each summary, the participants had to answer the survey that contains five Likert scale questions as follows:

1. How would you rate the structure of the given summary?
2. Does this summary encapsulate the context of all or most of the reviews?
3. Does the summary convey the intent of the reviewers?
4. Please rate the readability of language for the given summary?
5. Does the summary help you to judge the quality of the work?

We announced the survey to the students who took part in the OSS course in fall semester 2015 and twenty-eight people volunteered to participate in the survey.

3.2 Results

We calculated the Cronbach alpha for the survey items based on the results and we found that in average they are highly reliable (α=0.81).

As depicted in Figure 3A, the result of the survey shows that in average, the participants rated the usefulness of the generated summary 3.6, on the scale of 1 being the worst to 5 being excellent. TextRank with the 10 length to be the best (M=3.9, SD = 1) and KL Sum with the length of 3 to be the worst (M=3.4, SD = 0.9). however, a paired T-test does not show any significant difference between the two means.

When considering only the length of the summary, the participants rated the longer the summary higher than the shorter ones in almost all questions except for the structure. We suspect that a summary of three sentences is potentially too short for the reader to grasp the context of the reviews. However, within a very short summary, the readers expect direct important points instead of good flow between the sentences, unlike the longer summaries where the flow between sentences is more noticeable to the readers. Despite of these trends, unfortunately, a paired T-Test there are no significant differences among the results and therefore, we cannot be sure if this result applies for the future cases.

When considering only the algorithms, the participants in average rated TextRank higher than LSA and KL Sum in almost all questions except for the readability, where LSA gets the highest rate and the structure of the summaries, where all algorithms get in average almost the same scores. TextRank does produce a summary with more overlapping content as it relies on how often the sentences are "recommended" by the other sentences and calculate recursively to draw information from the entire text. However, there are no significant differences between these algorithms since the generated

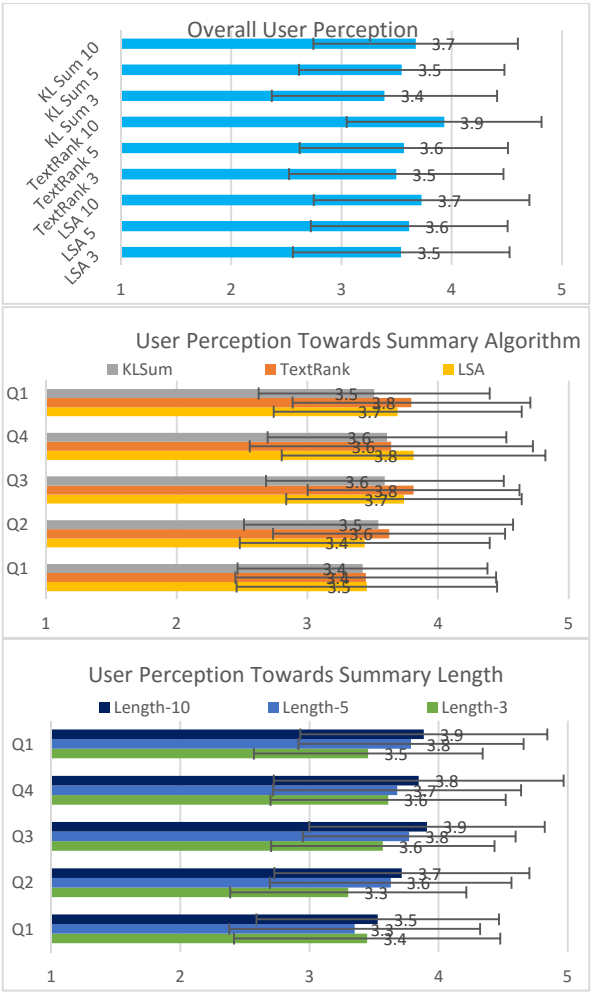

Figure 3. From the scale of 1 (worst) to 5 (best), (A) User perception to each summary. (B) User perception to the summary algorithms. (C) User perception to the summary lengths.

4. CONCLUSION AND FUTURE WORK

Providing summaries to the students help them to quickly scan feedback that they get from their peers. The result of the study shows that the students did not notice any quality differences among the summaries generated by different algorithms. However, although the results are not statistically significant, we see a

tendency that those medium sized summaries were preferable than the shorter ones. Therefore, it is worthwhile to perform a larger study to confirm this.

In the future, we would like to compare the remaining two approaches that we did not include in our current study Topic-based and Machine Learning based approaches. In addition, we would also like to study different visualization approaches to present the summary. For instance, comparing the narrative form with the more concise pros cons bullet points or highlighting feedback that is mentioned multiple times by the reviewers. Visualizing the feedback with some degree of importance could help the students focusing on improving their learning strategies to overcome their weaknesses.

5. ACKNOWLEDGMENTS

This study is partially funded by the PeerLogic project under the National Science Foundation grants 1432347, 1431856, 1432580, 1432690, and 1431975

6. REFERENCES

[1] Topping, K., and Ehly, S.: 'Peer-assisted learning' (Routledge, 1998. 1998)

[2] Topping, K.J.: 'Trends in peer learning', Educational psychology, 2005, 25, (6), pp. 631-645

[3] Wang, S., and Wu, P.: 'The role of motivation on computer-supported learning behaviors and achievement', in Editor 'Book The role of motivation on computer-supported learning behaviors and achievement' (2002).

[4] Sluijsmans*, D.M., Brand-Gruwel, S., van Merriënboer, J.J., and Martens, R.L.: 'Training teachers in peer-assessment skills: effects on performance and perceptions', Innovations in Education and Teaching International, 2004, 41, (1), pp. 59-78

[5] Topping, K.J.: 'Peer Assessment', Theory Into Practice, 2009, 48, (1), pp. 20-27

[6] Topping, K.: 'Peer assessment between students in colleges and universities', Review of educational Research, 1998, 68, (3), pp. 249-276

[7] Hattie, J., and Timperley, H.: 'The power of feedback', Review of educational research, 2007, 77, (1), pp. 81-112

[8] Cheung, K.-W., Kwok, J.T., Law, M.H., and Tsui, K.-C.: 'Mining customer product ratings for personalized marketing', Decision Support Systems, 2003, 35, (2), pp. 231-243

[9] Pang, B., Lee, L., and Vaithyanathan, S.: 'Thumbs up?: sentiment classification using machine learning techniques', in Editor (Ed.)^(Eds.): 'Book Thumbs up?: sentiment classification using machine learning techniques' (Association for Computational Linguistics, 2002, edn.), pp. 79-86

[10] Popescu, A.-M., Nguyen, B., and Etzioni, O.: 'OPINE: Extracting product features and opinions from reviews', in Editor (Ed.)^(Eds.): 'Book OPINE: Extracting product features and opinions from reviews' (Association for Computational Linguistics, 2005, edn.), pp. 32-33

[11] Tellis, G. J., & Johnson, J. (2007). The value of quality. Marketing Science,26(6), 758-773.

[12] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. ACM Trans. Program. Lang. Syst. 15, 5 (Nov. 1993), 795-825. DOI= http://doi.acm.org/10.1145/161468.16147.

[13] Ding, W. and Marchionini, G. 1997. A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[14] Kim, S.-M., and Hovy, E.: 'Automatic identification of pro and con reasons in online reviews', in Editor (Ed.)^(Eds.): 'Book Automatic identification of pro and con reasons in online reviews' (Association for Computational Linguistics, 2006, edn.), pp. 483-490

[15] Luhn, H.P.: 'The automatic creation of literature abstracts', IBM Journal of research and development, 1958, 2, (2), pp. 159-165

[16] Das, D., and Martins, A.F.: 'A survey on automatic text summarization', Literature Survey for the Language and Statistics II course at CMU, 2007, 4, pp. 192-195

[17] Filatova, E., and Hatzivassiloglou, V.: 'Event-based extractive summarization', 2004

[18] Nenkova, A., and Vanderwende, L.: 'The impact of frequency on summarization', Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101, 2005

[19] Kullback, S., and Leibler, R.A.: 'On information and sufficiency', The annals of mathematical statistics, 1951, 22, (1), pp. 79-86

[20] Edmundson, H.P.: 'New methods in automatic extracting', Journal of the ACM (JACM), 1969, 16, (2), pp. 264-285

[21] Erkan, G., and Radev, D.R.: 'LexRank: Graph-based lexical centrality as salience in text summarization', Journal of Artificial Intelligence Research, 2004, pp. 457-479

[22] Mihalcea, R., and Tarau, P.: 'TextRank: Bringing order into texts', in Editor (Ed.)^(Eds.): 'Book TextRank: Bringing order into texts' (Association for Computational Linguistics, 2004, edn.), pp.

[23] Gong, Y., and Liu, X.: 'Generic text summarization using relevance measure and latent semantic analysis', in Editor (Ed.)^(Eds.): 'Book Generic text summarization using relevance measure and latent semantic analysis' (ACM, 2001, edn.), pp. 19-25

[24] Svore, K.M., Vanderwende, L., and Burges, C.J.: 'Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources', in Editor (Ed.)^(Eds.): 'Book Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources' (2007, edn.), pp. 448-457

[25] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G.: 'Learning to rank using gradient descent', in Editor (Ed.)^(Eds.): 'Book Learning to rank using gradient descent' (ACM, 2005, edn.), pp. 89-96

[26] Steinberger, J., and Ježek, K.: 'Evaluation measures for text summarization', Computing and Informatics, 2012, 28, (2), pp. 251–2