Assessment of the project report

Student name: David Grönberg, davgr686

I have assessed your report with respect to the criteria defined below. For each criterion, I have assigned one of the scores 0 (below expectation), 3 (meets expectation), or 5 (exceeds expectation).

Clarity Is it clear what was done in this project, why it was done, and how it was done? Is the report well-written and well-structured?

0 There are some important questions about the method, results, or analysis that even expert readers are not able to resolve.

3 Any student who has successfully completed the course should understand what was done in this project, why it was done, and how it was done.

5 The report is well-polished. In regard to clarity of presentation, it would be acceptable for an academic journal or conference proceedings.

Specific comments: Well done!

Soundness and correctness Is the technical approach sound and well-chosen? Are the claims made in the report supported by proper experiments, and are the results of these experiments correctly interpreted?

0 Troublesome. There may be some ideas worth salvaging here, but the work should really have been done or evaluated differently. The claims made in the report have no support in the experimental results.

3 Fairly reasonable work. The approach is not bad, the methods are appropriate, and at least the main claims are probably correct. The report contains a discussion of the possibilities and limitations of the technical approach.

5 The approach is very apt, and the claims are convincingly supported. The report contains a well-developed discussion of the possibilities and limitations of the technical approach, including the reliability and validity of the results.

Specific comments: You could have measured the inter-annotator agreement for your evaluators.

Related work Does the report show awareness and understanding of related work documented in scientific sources? Is it clear where the work done in the project sits with respect to that related work?

0 The report shows little awareness and understanding of related work. References to scientific sources are missing or incomplete. There is no account of how the work done in the project compares to the related work.

3 The report shows some awareness and understanding of related work. Scientific sources are adequately referenced. The relation between the work done in the project and the work documented in the scientific sources is clear.

5 The report features a precise and enlightening comparison with related work. References are complete and consistently formatted. The majority of scientific sources are peer-reviewed research articles.

Specific comments: —

Creativeness How creative is the project? For example: Does the project target a new problem? Does it contribute a new data set? Does it use any machine learning models that were not covered in the course?

0 There are no creative elements in this project. The project is essentially a repetition of one of the lab assignments.

3 The project contains at least one creative element.

5 There are many creative elements in this project. The project goes significantly beyond what has been covered in the course.

Specific comments: Task: ATG. New method for sentence construction. Evaluation scheme.

Substance Based on the report, does this project have enough substance, or would there have been room for more ideas, results, or analysis? (The expected amount of work for the project module is 88 hours.)

0 Seems thin. I (the examiner) would have expected significantly more ideas, results, or analysis for a project with this timeframe.

3 Represents an appropriate amount of work for a project in this course.

5 Contains significantly more ideas, experiments, and analysis than what I (the examiner) would have expected for a project with this timeframe.

Specific comments: —

Grade

For a passing grade, your score *for each criterion* must be at least 3. Your grade is determined by your total score, as specified in the table below.

Total score: 25 Grade: 5

Linköping, 2020-02-05

Marco Kuhlmann, examiner

Total score	15	17	19	21	23
Grade 732A92	E	D	C	B	A
Grade TDDE16	3	3	4	5	5

Generating Headlines for News Articles: A Summarization Approach

David Grönberg
TDDE16 - Text Mining
Linköpings Universitet
davgr686

***Abstract*— This work studies the task of automatically generating titles for a given text. A summarization approach is taken to develop a method that utilize TF-IDF scores, sentence extraction and parse trees. The method is evaluated on the Harvard Dataverse News Articles corpus, containing 3824 news articles. Four judges were presented 20 generated titles and asked to rate them based on grammar, conciseness, relevancy and overall fit. The results show that the presented method can generate perfect titles with a success rate of 20%. A new approach for finding candidate sentences utilizing word similarity is also explored and evaluated and is shown to outperform the baseline approach.**

I. INTRODUCTION

A title for a given document is a compact sentence that can help people capture the central meanings of the text without having to read through the entire document. Automatic title generation (ATG) is a difficult natural language processing problem which requires generating a grammatically correct title that conveys the key concepts of the document, while capturing the readers eye.

There are several ways to approach ATG tasks. Previous studies have treated it as a summarization, key phrase extraction or machine translation problem [1], [2], [3], [4], [5]. Witbrock and Mittal [1] used a statistical model to produce brief summaries for a document of learnt styles. The Naive Bayes model describes the document word and title word correlation, i.e. the order and likelihood of appearance of the words in the target document in the context of certain words in the source. Witbrock and Mittal tested the model on 1000 news articles, comparing the words in the generated title with the words in the actual title. The model performed surprisingly well considered the simplicity of the approach, but was poor at generating grammatically correct titles.

Kennedy and Hauptmann [2] tackled the ATG problem with adopting a machine translation approach and treated the document title as a short translation of the document. The title generation model used an expectation-maximization algorithm to generate a title for a document by estimating:

$$\begin{aligned}&\text{Argmax}_{title} p(title|document)\\ &= Argmax_{title}(document|title)p(title)\end{aligned} \tag{1}$$

When comparing Kennedy and Hauptmann's approach with Witbrock and Mittal's Naive Bayes approach on a corpus of 40,000 transcripts of news stories, where 100 were used for evaluation, Kennedy and Hauptmann's model yielded a higher F1-score.

Shao and Wang [3] approached the ATG task by extracting central sentences in the document and applied sentence compression to generate a title. This was done by keyword extraction and ranking the sentences based on the number of keywords they contain. A dependency tree was constructed for each *n* sentences chosen for compression and certain branches were pruned based on a set of empirical rules.

In this study, a three-step approach is taken to generate grammatically correct news headlines for given news stories. it is inspired by Shao and Wang's approach. First, relevant keywords of the document are extracted using the TF-IDF measure presented in section II-A. Second, candidate sentences containing the keywords are selected using a new approach described in section III-A. Lastly, the candidate sentences are compressed using a rule-based approach described in section III-B to form the final titles.

II. THEORY

The following chapter presents the collected theory forming the basis of this study.

A. TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a formula used to measure how important a word is in a document within a collection of documents [6]. The formula is divided into two parts; the *Term Frequency* and *Inverse Document Frequency*. The *Term Frequency* calculates the number of times a word appears in a document divided by the total amount of words in the document:

$$tf_{i,j} = \frac{n_{i,j}}{\sum n_{i,j}} \tag{2}$$

where *i* is a word and *j* is a document.

If only the *TF* measure is used, we would end up with stop words that frequently appear in general text (i.e. *him, her, they, am, is, are* and so on) that has no relevancy to the document yielding high scores. The *Inverse Document Frequency* is a measure that describes how common or rare a word is across all documents. It represents the inverse of the share of the documents in which the word can be found.

$$idf_i = \log(\frac{N}{df_i}) \tag{3}$$

Multiplying the measure with the *Term Frequency* yields the TF-IDF measure.

$$w_{i,j} = \frac{n_{i,j}}{\sum n_{i,j}} * \log(\frac{N}{df_i}) \tag{4}$$

These two calculations combined provide a weighting factor (between 0 and 1) of the importance of each word with respect to the document it is in [6].

B. Sentence Compression using Parse Trees

Sentence compression can be described as the task of generating a shorter paraphrase of a sentence, while minimizing information loss. A common approach to sentence compression is operating on parse trees and pruning them [7] to derive a more compact sentence. A parse tree is a tree that represents the syntactic structure of a set of words. Parse trees are usually divided into constituency and dependency parse trees. Constituency parse trees are based on constituency relation, while dependency parse trees are based on dependency relation. Dependency parse trees differ from syntactical parse trees as it uses dependency grammar and is constructed only on the relationships between words rather than sentence structure and relationship. Dependency trees are often more concise than constituency trees because they only display grammar between a governor and is dependents. Figure 1 and 2 visualize the parse tree representation for the sentence: *'Eric eats meat'*.

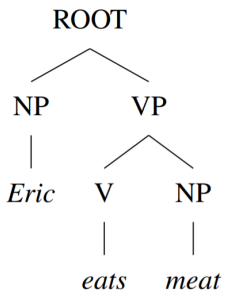

Fig. 1. Example of a constituency parse tree

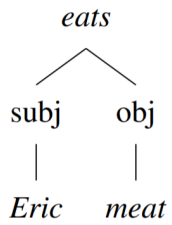

Fig. 2. Example of a dependency parse tree

Marku and Knight [9] presented a decision-based compression model that takes a syntactical parse tree and rewrites it into a smaller tree. Given a syntactical tree, the purpose of the decision-based classifier is to learn what action to choose from a set of 210 possible actions to derive a compressed tree. Jing [8] uses a rule-based approach based on three sources of knowledge to determine if a phrase can be removed. A phrase is removed only if:

- The phrase is not the focus of the original sentence.
- The resulting sentence after removing the phrase is grammatically sound.
- The phrase has a reasonable probability of being removed by a human (based on training data).

The grammatical soundness of a sentence is determined by traversing the parse tree and applying simple linguistic based rules and the probability of a phrase being removed is based on a corpus consisting of sentences reduced by human professionals and their corresponding original sentences.

III. SOLUTION

The following section gives a detailed description of each step in the presented method.

A. Candidate Sentence Extraction

First, the document is divided into sentences. This is easily done with spaCy's *Doc.sents* property. SpaCy is a widely used Natural Language Processing library for Python and Cython and is proven to work well on general-purpose text. It contains a pre-trained statistical models and supports Part-Of-Speech tokenization for 50+ languages. [1]

After sentence segmentation is done, TF-IDF is used to find the *n* most important keywords and each sentence is assigned a score depending on how many keywords occurs in the sentence.

Consider the example sentence: *"The kitten slept in the house"* and the extracted document keywords: *'cat', 'home'*, with the weights: *0.7* and *0.5*, respectively. If a straight forward approach is used, i.e. summing the occurrences of keywords in the sentence, the score would yield 0. However, *'cat'* is similar to *'kitten'*. Likewise, *'home'* is similar to *'house'*. Intuitively, one would say that the sentence is highly relevant in relation to the extracted keywords, thus should be given a high score. This paper propose a scoring measure that takes the similarity between the words in the sentence and the keywords into account.

When determining the score of a sentence, the similarity between each keyword and each word in the sentence is computed using *spaCy's* similarity property. The similarity measure is then multiplied with the keyword's weighting factor and then summed. This can be summarized to the following equation:

$$score = \sum_i \sum_j s_{i,j} \cdot j_w \tag{5}$$

Where *i* represents a word in the sentence, *j* represents a keyword, $s_{i,j}$ represents the similarity between *i* and *j*, and *w* represents the weighting factor.

[1] https://spacy.io/

The measure will allow the model to assign high scores to sentences that contain words that are semantically similar to the document's keywords. When the score has been computed for each sentence, the top n sentences with the highest score is chosen as candidate sentences.

B. Sentence Compression

In order to generate adequate titles that are concise and grammatically correct, the candidate sentences needs to be compressed. This is done by constructing a constituency parse tree for each sentence, and trimming sub-trees and nodes that are considered redundant. The following empirical rules are devised to determine if a sub-tree or node should be trimmed:

- If the node's POS is of type *Determiner*, trim the node.
- If the node's POS is of type *Adposition* or *Auxiliary* and the node's word is not a keyword, trim the node and its sub-tree if no keywords were found in the sub-tree.
- If the node's word is *"According"*, trim the node and its sub-tree.

IV. Evaluation

The following section describes the methodology used when evaluating the proposed approach.

A. Data

The proposed approach is evaluated on the Harvard Dataverse News Articles dataset [10], containing 3824 news articles collected on a three month period from December 2016 to March 2017. The source of the articles are from ABC News, CNN news, The Huffington Post, BBC News, DW News, TASS News, Al Jazeera News, China Daily and RTE News.

B. Experiment Setup

Previous studies evaluates ATGs by comparing the generated titles with existing human-constructed titles [1], [2], [4]. This could be done by counting the number of word co-occurrences in both titles and deriving a F1-score. However, when using this approach, generated titles can be very good (in the sense of conciseness, relevancy and grammar), but penalized because they do not match the human-constructed titles. This study is based on the premise that titles can equally well capture the main topic of a text, while having completely different words and sentence structures.

20 news articles were selected randomly from the Harvard dataset to form a test set. To evaluate the performance of the presented candidate sentence extraction approach utilizing similarity described in III-A, two models were used when generating titles, denoted *Sim* and *Base*. *Sim* implements the candidate sentence extraction approach using similarities, and *Base* simply rates sentences based on the sum of the keyword weights found in the sentence.

For each news article in the test set, the original title and the two generated titles from each model were presented to four English speaking judges. The judges were then asked to rate each title on a three point scale in four different categories:

Grammar

- 3 - The title contains no grammatical flaws.
- 2 - The title contains few grammatical flaws.
- 1 - The title is hard to understand.

Summary

- 3 - The title perfectly reflects the the main topic of the article.
- 2 - The title moderately reflects the main topic of the article.
- 1 - The title is does not reflect the main topic of the article.

Conciseness

- 3 - The title has no redundant words.
- 2 - The title has some redundant words.
- 1 - The title has a lot of redundant words or phrases.

Overall Fit

- 3 - The title fits perfectly as a title.
- 2 - The title could work as a title with minor adjustments.
- 1 - The title does not fit at all.

Each evaluation category represents important aspects when creating sufficient titles. The category *Overall Fit* was introduced because several generated titles could have a high score in all the other categories, but still not be suitable to be a title. For example, if the main topic of a news article is that a man was killed in a car crash, the title *He passed away after car crash on highway 64* is not a suitable title, even though it is concise, grammatically correct and summarizes the article well. However, if the word *'He'* would be changed to *'Man'*, the title would be considered fitting. There are other examples where a generated title scores high on most of the evaluation categories but is not suitable to be a title and it is up to the judges to subjectively define the characteristics of a suitable title.

V. Results

The following section presents the results produced by the evaluation of the proposed method.

A. Mean Rating Score

Figure 3 presents the mean rating for each title, categorized by the evaluation category. As the bar plot indicates, the original titles received the highest score of the three titles in all categories. This was expected due to the original titles being constructed by professional journalists. When comparing *Sim* and *Base*, the two approaches seem to generate titles having roughly the same ratings. Both approaches tend to generate titles having a low amount of grammatical flaws and few redundant words.

B. Candidate Sentence Extraction

To get a better understanding of the performance of the proposed candidate sentence extraction approach incorporating similarity, the two approaches generated titles are further

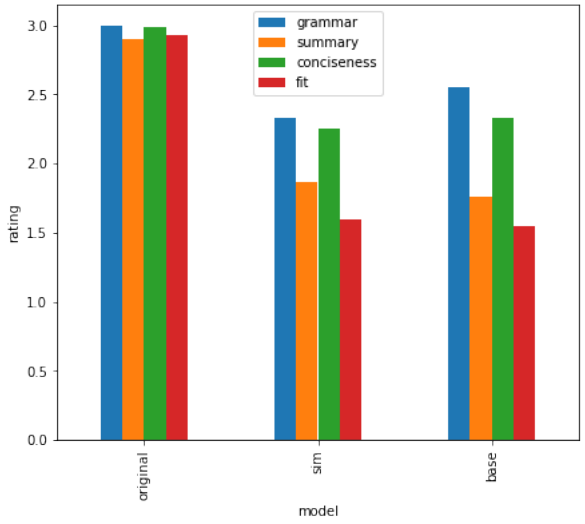

Fig. 3. Mean rating for the different titles

investigated in the summary and fit categories. The *Sim* approach yields a slightly higher mean rating in *fit* and *summary* (3% and 5% respectively), compared to the baseline. Figure 4 and 5 shows a more detailed plot describing the rating distribution for the evaluation measures: *summary* and *overall fit*.

The two approaches are able to generate titles that fits the main topic either perfectly or moderately in more than 50% of the time. Furthermore, the approaches generate perfect titles or near perfect titles in more than 40% of the time. *sim* received a perfect title rating 18% of the time, while *base* received a perfect title rating 16% of the time. This can be seen as the approaches having a success rate of almost 20%. These results will further be discussed in VI. Note that *Sim* and *Base* generated the same titles for a couple articles. A sample of the the original and generated titles are shown in table I.

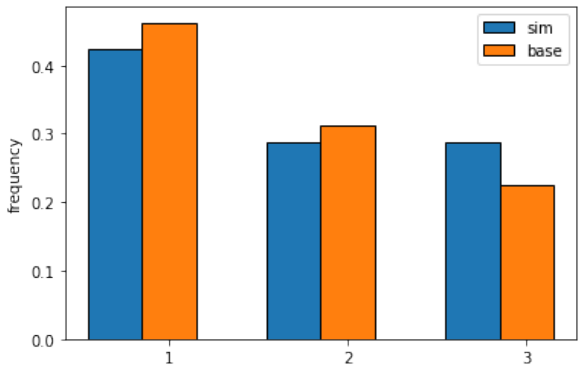

Fig. 4. Distribution of rating score on evaluation measure: *summary*

VI. DISCUSSION

In the following section, the work conducted in this study is discussed and reviewed. First, the produced results

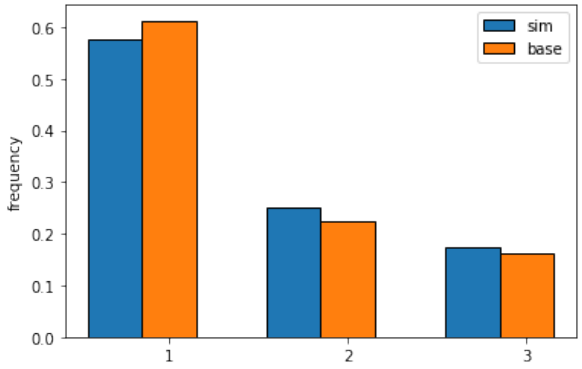

Fig. 5. Distribution of rating score of evaluation measure: *overall fit*

are examined. Secondly, the evaluation method chosen is discussed and areas where further testing would be necessary to strengthen the results produced in this study is presented.

A. Results

The results of this study shows that the presented model has the ability to generate sufficient titles with a frequency of roughly 20%. The models success rate can be more than doubled if the constraints for a title being sufficient is relaxed and allowed to have a small number of flaws. The proposed Candidate Sentence Extraction approach using similarity yield titles having a higher relevancy to the main topic of the article, compared to titles produced by the baseline approach. Although the increase in performance is minor, the results can be seen as a motivation to further develop the approach and evaluate its capacity.

B. Evaluation

The evaluation of the presented model was done by letting four English speakers assess the quality of the generated titles. This evaluation approach introduces a factor of subjectiveness into to evaluation results. The judges might have a different perception of what characteristics make up a good title. This was partly countered with the introduction of evaluation categories, giving the judges a more concrete evaluation scale. However, the evaluation categories *Overall Fit* and *Summary*, can still be highly subjective, as people can have different opinions on what the main topic of a article is. Another possible flaw in the evaluation approach is the disregardance of grade inflation. Grade inflation occurs when judges use different grading scales when evaluating the titles. One judge might only rate titles between 1-2, while another tend to rate titles between 2-3, considered more optimistic in a sense, compared to the former judge. Further evaluation, preferably using more judges and a larger test set, will need to be done to substantiate the obtained results.

Cohen's κ ?

The corpus used in this study was collected on a three month time span from several major news websites. Because of the short time span, there could be a risk that a few topics dominate the news feed, thus the topics were not diversified enough for the keyword extraction to perform well. If a large portion of the news articles had topics concerning for

Original title	Man arrested after fatal stabbing in Dublin
Sim title	Man in 40s stabbed in house last night.
Base title	Man in 40s stabbed in house last night.
Original title	Sikh Temples Open Their Doors To Oroville Dam Evacuees
Sim title	Sikh Temples in Sacramento region are open for people evacuated.
Base title	More than 180,000 people in Northern California ordered to evacuate late due to erosion of emergency spillway in nation's tallest dam.
Original title	This Year The LA Pride Parade Is Being Replaced By A 'Resist March'
Sim title	Organizers of Los Angeles Pride announced very different plan.
Base title	We resist forces would divide us
Original title	North Korea 'conducts high-thrust engine test'
Sim title	In South Korea on Friday, he said.
Base title	North Korea's state media says military tested new high-performance rocket engine.

TABLE I

SUBSET OF THE EVALUATED TITLES

example Donald Trump, then TF-IDF might fail to assign a high score to *'Donald Trump'*, when he is the main topic of a article. It would be interesting to test the presented model on other data sets and investigate the potential change in accuracy.

C. CONCLUSION

This study has presented a Automatic Title Generator utilizing TF-IDF scores, candidate sentence extraction techniques and semantic parse trees, to generate feasible titles for a given text. The presented candidate sentence extraction approach utilizing word similarity outperformed the baseline approach by a fraction. There are areas in the model that have room for improvement and some shortcomings that needs to be addressed. Future work could be to further develop the rule-based sentence compression phase and introduce more sophisticated rules. Another improvement could be introducing a step in the final stage that evaluates the grammar in each generated title and tries to improve it before outputting the final title. This could also be done with semantic parse trees. Considering the present results, The proposed ATG could be used for tasks as automatically assigning titles to a large set of documents, where there is room for some error.

REFERENCES

[1] Witbrock, Michael and Mittal, Vibhu. "Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries". (1999), pp. 315-316.

[2] Kennedy, Paul E. and Hauptmann, Alexander G. "Automatic Title Generation for EM". In DL '00 (2000), pp. 230-231.

[3] Liqun Shao and Jie Wang "DTATG: An Automatic Title Generator Based on Dependency Trees". In Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (2016), pp. 166–173.

[4] Jin, Rong and Hauptmann, Alexander G. "Automatic Title Generation for Spoken Broadcast News". In Proceedings of the First International Conference on Human Language Technology Research (2001).

[5] Filippavo et al. "Sentence Compression by Deletion with LSTMs" in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (2015), pp 360-368.

[6] Wen Zhang, Taketoshi Yoshida and Xijin Tang. "A comparative study of TF*IDF, LSI and multi-words for text classification". In Expert Systems with Applications (2011), pp. 2758-2765.

[7] Filippova, Katja and Altun, Y. "Overcoming the lack of parallel data in sentence compression" in Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference (2013), pp. 1481-1491.

[8] Jing, Hongyan. "Sentence Reduction for Automatic Text Summarization" In Proceedings of the Sixth Conference on Applied Natural Language Processing (2000), pp.310–315.

[9] Knight, Kevin and Marcu, Daniel. "Summarization beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression". In Artificial Intelligence (2002) pp. 91-107.

[10] Dai, Tianru. News Articles (2017). Harvard Dataverse.