International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

GENERATING WIKIPEDIA BY SUMMARIZING LONG SEQENCES

Ketali Soujanya (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.
G. Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT
In this paper Wikipedia cloning involves designing, writing, and coding a website in a way that helps to improve the volume and quality of traffic to your website from people using informative website. This website will be an informative website which gives any information which is to be needed by the users exactly like a Wikipedia. Informative websites are built for the purpose of providing information. They can include anything like News website, Science Websites, Encyclopaedia etc.

1. INTRODUCTION

The sequence-to-sequence framework has demonstrated success in natural-language sequence transduction tasks such as machine translation. More recently, neural techniques have been applied to do single-document, abstractive (paraphrasing) text summarization of news articles (Rush et al. (2015), Nallapati et al. (2016)). In this prior work, the input to supervised models ranged from the first sentence to the entire text of an article, and they are trained end-to-end to predict reference summaries. Doing this end-to-end requires a significant number of parallel article-summary pairs since language understanding is a pre-requisite to generate fluent summaries. In contrast, we consider the task of multi-document summarization, where the input is a collection of related documents from which a summary is distilled. Prior work has focused on extractive summarization, which select sentences or phrases from the input to form the summaries, rather than generating new text. There has been limited application of abstractive neural methods and one possible reason is the paucity of large, labeled datasets. In this work, we consider English Wikipedia as a supervised machine learning task for multidocument summarization where the input is comprised of a Wikipedia topic (title of article) and a collection of non-Wikipedia reference documents, and the target is the Wikipedia article text. We describe the first attempt to abstractively generate the first section, or lead, of Wikipedia articles conditioned on reference text. In addition to running strong baseline models on the task, we modify the Transformer architecture (Vaswani et al., 2017) to only consist of a decoder, which performs better in the case of longer input sequences compared to recurrent neural network (RNN) and Transformer encoder-decoder models. Finally we show our modeling improvements allow us to generate entire Wikipedia articles.

The existence of an abundance of dynamic and heterogeneous information on the Web has offered many new opportunities for users to advance their knowledge discovery. As the amount of information on the Web has increased substantially in

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
IJARST
ISSN: 2457-0362

the past decade, it is difficult for users to find information through a simple sequential inspection of web pages or recall previously accessed URLs. Consequently, the service from a search engine becomes indispensable for users to navigate around the Web in an effective manner.

PROPOSED SYSTEM

Here we propose to create an exact replica of Wikipedia website which is a Informative website which is very helpful in getting any kind information present in the whole world which is free and more informative for the users. This website are updated by admins or by users as well. Admin only checks the authentication of the website content. This is an effective way to learn for all like students, Businessmen's, Researchers, Politicians, and Actors etc.

ADVANTAGES

anyone can edit

easy to use and learn

Wikis are instantaneous so there is no need to wait for a publisher to create a new edition or update information

people located in different parts of the world can work on the same document

the wiki software keeps track of every edit made and it's a simple process to revert back to a previous version of an article

widens access to the power of web publishing to non-technical users

Wikipedia has no predetermined structure – consequently it is a flexible tool which can be used for a wide range of applications

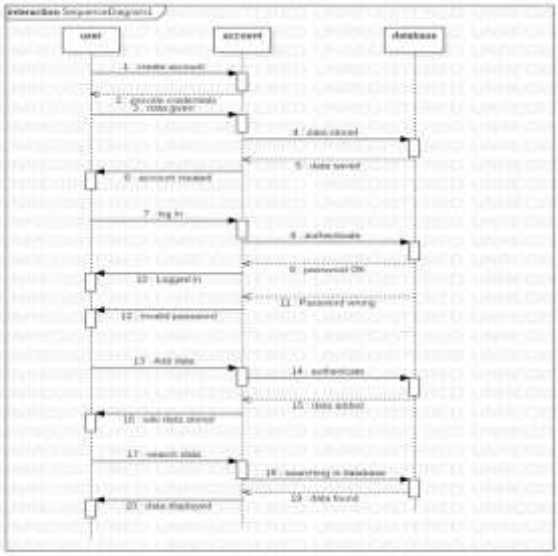
There are a wide range of open source software wiki's to choose from so licensing costs shouldn't be a barrier to installing an institutional Wikipedia.

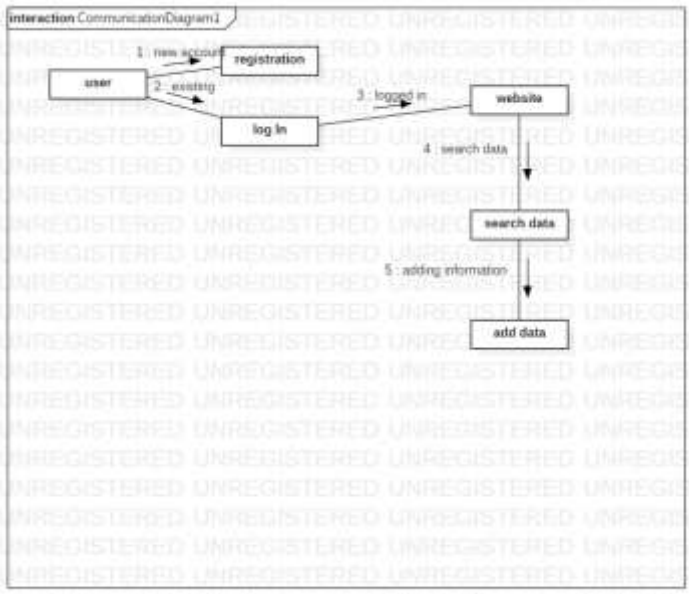
2. RELATED WORK

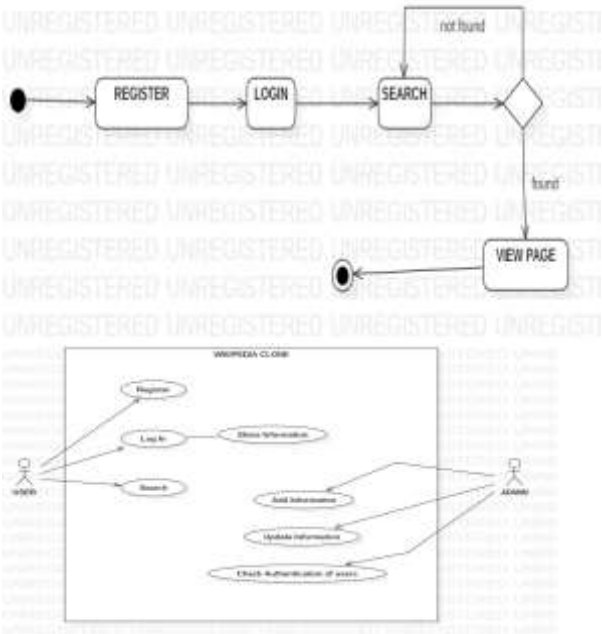
2.1 OTHER DATASETS USED IN NEURAL ABSTRACTIVE SUMMARIZATION Neural abstractive summarization was pioneered in Rush et al. (2015), where they train headline generation models using the English Gigaword corpus (Graff & Cieri, 2003), consisting of news articles from number of publishers. However, the task is more akin to sentence paraphrasing than summarization as only the first sentence of an article is used to predict the headline, another sentence. RNN-based encoder-decoder models with attention (seq2seq) perform very well on this task in both ROUGE (Lin, 2004), an automatic metric often used in summarization, and human evaluation (Chopra et al., 2016). In Nallapati et al. (2016), an abstractive summarization dataset is proposed by modifying a questionanswering dataset of news articles paired with story highlights from Daily Mail and CNN. This task is more difficult than headline-generation because the information used in the highlights may come from many parts of the article and not only the first sentence. One downside of the dataset is that it has an order-of-magnitude fewer parallel examples (310k vs. 3.8M) to learn from. Standard seq2seq models with attention do less well, and a number of techniques are used to augment performance. Another downside is that it is unclear what the guidelines are for creating story highlights and it is obvious that there are significant stylistic differences between the two news publishers. In our work we also train neural abstractive models, but in the multi-document regime with Wikipedia. As can be seen in Table 1, the input and output

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

text are generally much larger, with significant variance depending on the article. The summaries (Wikipedia lead) are multiple sentences and sometimes multiple paragraphs, written in a fairly uniform style as encouraged by the Wikipedia Manual of Style1 . However, the input documents may consist of documents of arbitrary style originating from arbitrary sources. We also show in Table 1 the ROUGE-1 recall scores of the output given the input, which is the proportion of unigrams/words in the output co-occuring in the input. A higher score corresponds to a dataset more amenable to extractive summarization. In particular, if the output is completely embedded somewhere in the input (e.g. a wiki-clone), the score would be 100. Given a score of only 59.2 compared to 76.1 and 78.7 for other summarization datasets shows that ours is the least amenable to purely extractive methods.







3. CONCLUSION

We have shown that generating Wikipedia can be approached as a multi-document summarization problem with a large, parallel dataset, and demonstrated a two-stage extractive-abstractive framework for carrying it out. The coarse extraction method used in the first stage appears to have a significant effect on final performance, suggesting further research on improving it would be fruitful. We introduce a new, decoder-only sequence transduction model for the abstractive stage, capable of handling very long input-

output examples. This model significantly outperforms traditional encoder decoder architectures on long sequences, allowing us to condition on many reference documents and to generate coherent and informative Wikipedia articles

4. REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

Sumit Chopra, Michael Auli, and Alexander M Rush. Abstractive sentence summarization with
attentive recurrent neural networks. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.
93–98, 2016.

Hoa Trang Dang. Overview of duc 2005. In Proceedings of the document understanding conference,
volume 2005, pp. 1–12, 2005.

David Graff and Christopher Cieri. English gigaword 2003. Linguistic Data Consortium, Philadeplhia,
2003.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han,
Matthew Kelcey, and David Berthelot. Wikireading: A novel large-scale language understanding
task over wikipedia. arXiv preprint arXiv:1608.03542, 2016.

R´emi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with
application to the biography domain. In Proceedings of the 2016 Conference on Empirical Methods
in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, pp.
1203–1213, 2016. URL http://aclweb.org/anthology/D/D16/D16-1128.pdf.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes,
Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, S¨oren Auer, et al. Dbpedia–a largescale,
multilingual knowledge base extracted from wikipedia. SemanticWeb, 6(2):167–195, 2015.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Text summarization
branches out: Proceedings of the ACL-04 workshop, volume 8. Barcelona, Spain, 2004.

Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In Proceedings of the 2004
conference on empirical methods in natural language processing, 2004.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, C¸ a glar Gulc¸ehre, and Bing Xiang. Abstractive
text summarization using sequence-to-sequence rnns and beyond. CoNLL 2016, pp. 280, 2016.

Ani Nenkova and Lucy Vanderwende. The impact of frequency on summarization. Microsoft Research,
Redmond, Washington, Tech. Rep. MSR-TR-2005, 101, 2005.

Lawrence Page, Sergey Brin, Rajeev Motwani, and TerryWinograd. The pagerank citation ranking:
Bringing order to the web. Technical report, Stanford InfoLab, 1999.