Cross-Reading News

Shahbaz Syed[1] Tim Gollub[1] Marcel Gohsen[1] Nikolay Kolyada[1] Benno Stein[1] Matthias Hagen[2]

[1]Bauhaus-Universität Weimar
<firstname>.<lastname>@uni-weimar.de

[2]Martin-Luther-Universität
Halle-Wittenberg
matthias.hagen@informatik.uni-halle.de

Abstract

Journalists often need to perform multiple actions using different tools, in order to create content for publication. This involves searching the web, curating the result list, choosing relevant entities for the article and writing. We aim to improve this pipeline through Cross-Reading News, a modular, and extendable web application aimed at helping journalists easily search for information and draft articles for publication. It combines information retrieval, natural language processing and deep learning in order to provide a focused work-space by leveraging local collections of news articles maintained by media companies. Specifically, users can search for information in multiple local collections of news articles, gather named entities, extract keyqueries, find semantically related content and obtain title suggestions.

1 Introduction

Journalists from various media organizations can be regarded as one of the key sources for generating new content on the web in the form of articles. However, generating new information often comprises searching, reusing and curating information from different sources. A journalist usually has a vague idea of what he/she wants to write about and often consults an expert (here, a person specialized in the taxonomy of the local collection of articles) to obtain relevant documents. This information need is often hard to formulate as a concrete search query, making this process laborious to iterate. Also, journalists covering specific types of news such as sports or politics, are required to write articles in a relatively short amount of time compared to those covering say editorials or literature. In such cases, a tool which provides faster access to relevant data can significantly speed up the workflow. Specifically, a collection of published articles maintained by most media companies today is an excellent resource which can be leveraged to build intelligent tools by applying natural language processing and information retrieval.

Although querying the web is also a possibility for satisfying an information need, it is often overwhelming as the journalist is bombarded with excessive results which can be both distracting and time consuming to be refined further. Using an existing collection can provide a better focused search environment, and also ease the process of finding similar content already published by fellow journalists. This significantly simplifies author citations, and makes it easier to find duplicate or similar articles to be reused.

In this paper, we present our application that integrates multiple features such as; a local collection of published articles that have been preprocessed to recognize named entities and semantically similar paragraphs, a search engine for querying this collection, an automatic summarization model that provides title suggestions for a piece of text (which can also be used as keyphrases) and finally, a text editor to draft articles.

2 Corpus and preprocessing

We use the Signal Media corpus [CAMM16] as our local data source for the search engine. Before indexing the documents, we perform the following preprocessing steps: First, if a document is longer than 400 words, we split the text into paragraphs using NLTK[1].

Second, we perform named entity recognition using Stanford NER [MSB+14] to annotate entities for each paragraph. Specifically, we extract PERSON, ORGANIZATION, LOCATION, FACILITY and GPE

In: D. Albakour, D. Corney, J. Gonzalo, M. Martinez, B. Poblete, A. Vlachos (eds.): Proceedings of the NewsIR'18 Workshop at ECIR, Grenoble, France, 26-March-2018, published at http://ceur-ws.org

[1]`https://www.nltk.org`

(Geo-Political Entity) entities and add this information to our document store based on RocksDB[2].

Finally, in order to compute paragraph similarity we create a vector (300 dimensional) for each paragraph from an embedding model trained on the whole corpus using fasttext [BGJM17]. A paragraph vector is simply the average of consitituent word vectors. As the number of paragraphs is large (1.5million), we perform the similarity search as part of our preprocessing in order to find similar paragraphs based on cosine similarity. We use Faiss [JDJ17] in order to speed up the similarity search in this large vector space. We empirically chose to store only upto 3 similar paragraphs for each paragraph (if available).

3 Gather relevant entities

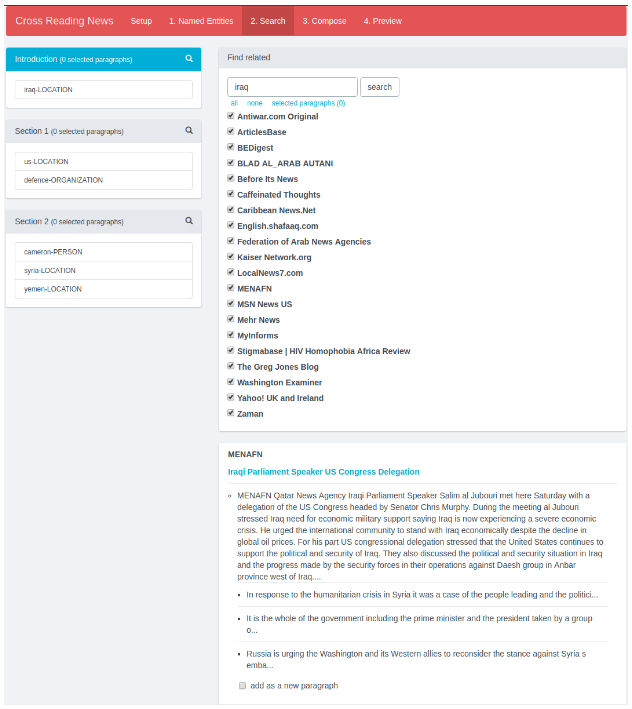
The first stage of writing an article is primarily collecting information about a specific subject or an event. To facilitate looking up documents from a local collection, we built a search engine that uses three standard retrieval models - *tf*, *tf·idf* and *bm-25*. As an intialization step, the user can choose a specific retrieval model to be used by the application.

Intuitively, as users of search engines, we go through some of the retrieved documents and identify the entities that might be of interest to our subject. Taking this as a motivation, we designed our first stage to return a list of named entities related to a query instead of the full documents themselves. By providing such a list (all entities from the top 20 documents returned by the selected retrieval model), we can help users to easily decide which of these entities to include in their article. Selected entities can be organized into sections which already provides a structure that serves as a guiding map in the subsequent stages of the application. At any time, the user can navigate back to this step in order to add more entities to their sections.

4 Gather related text

Using the *entity list* from the previous stage, the user can now start searching for information with better focus. It is possible to quickly perform a document search for any section with the click of a button. To facilitate this, we combine all the entities in a given section to form a search query. It is also possible to perform a free search with any other query. Results from this search can be filtered in order to select content from specific publishers and each paragraph in the displayed result list can be quickly added to the draft as shown in Figure 1. This speeds up the process of collecting and reusing the existing content.

[2]`http://rocksdb.org`



Introduction (0 selected paragraphs)
- iraq-LOCATION

Section 1 (0 selected paragraphs)
- us-LOCATION
- defence-ORGANIZATION

Section 2 (0 selected paragraphs)
- cameron-PERSON
- syria-LOCATION
- yemen-LOCATION

Find related

iraq

- ✓ Antiwar.com Original
- ✓ ArticlesBase
- ✓ BEDigest
- ✓ BLAD AL_ARAB AUTANI
- ✓ Before Its News
- ✓ Caffeinated Thoughts
- ✓ Caribbean News.Net
- ✓ English.shafaaq.com
- ✓ Federation of Arab News Agencies
- ✓ Kaiser Network.org
- ✓ LocalNews7.com
- ✓ MENAFN
- ✓ MSN News US
- ✓ Mehr News
- ✓ MyInforms
- ✓ Stigmabase | HIV Homophobia Africa Review
- ✓ The Greg Jones Blog
- ✓ Washington Examiner
- ✓ Yahoo! UK and Ireland
- ✓ Zaman

MENAFN

Iraqi Parliament Speaker US Congress Delegation

- MENAFN Qatar News Agency Iraqi Parliament Speaker Salim al Jubouri met here Saturday with a delegation of the US Congress headed by Senator Chris Murphy. During the meeting al Jubouri stressed Iraq need for economic military support saying Iraq is now experiencing a severe economic crisis. He urged the international community to stand with Iraq economically despite the decline in global oil prices. For his part US congressional delegation stressed that the United States continues to support the political and security of Iraq. They also discussed the political and security situation in Iraq and the progress made by the security forces in their operations against Daesh group in Anbar province west of Iraq....
  - In response to the humanitarian crisis in Syria it was a case of the people leading and the politici...
  - It is the whole of the government including the prime minister and the president taken by a group o...
  - Russia is urging the Washington and its Western allies to reconsider the stance against Syria s emba...

☐ add as a new paragraph

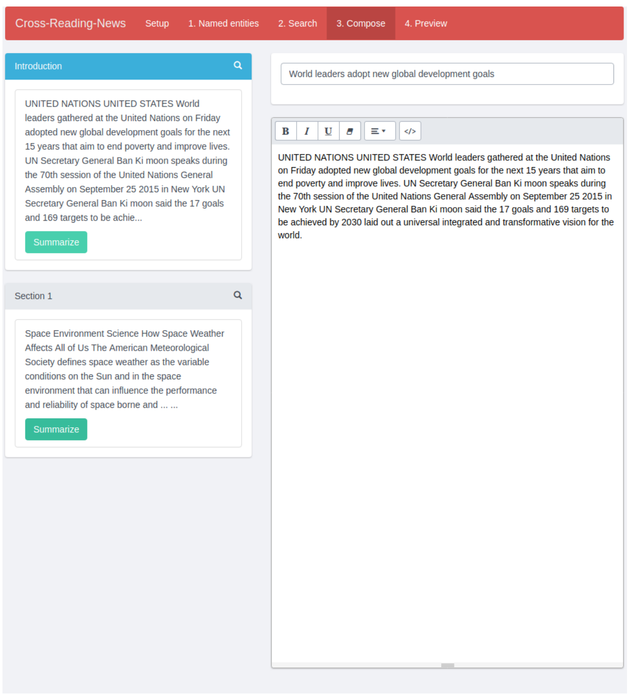
Figure 1: Search and gather related text

Furthermore, for each main paragraph we show the top k similar paragraphs (if available) according to the cosine similariy scores computed for the paragraph vectors as part of our preprocessing (we chose k=3). This helps the user to quickly find related content compared to reading a full document, to then reuse a portion of it.

5 Generate keyphrases

Keyphrases for unlabeled documents are an effective source to perform clustering, topic search and summarization [FPW+99]. Motivated by the idea of using keyqueries as document descriptions [GHMS13], we use deep learning to automatically generate keyphrases for both existing paragraphs as well as any new content written by the user (Figure 2). We believe that a keyphrase can serve both as a title suggestion for the text, and also be transformed into a keyquery in order to find additional semantically related documents.

We establish the task of generating keyphrase for a text as an abstractive summarization task using neural generative models. This allows the model to not only gain a semantic understanding of the text, but also generate novel words for the title conveying important information selected from the semantic space. A sample of generated summaries is presented in Table 1. While a similar approach for generating keyphrases for scientific articles was adopted by [MZH+17], we only use the title of each document as the target.



World leaders adopt new global development goals

UNITED NATIONS UNITED STATES World leaders gathered at the United Nations on Friday adopted new global development goals for the next 15 years that aim to end poverty and improve lives. UN Secretary General Ban Ki moon speaks during the 70th session of the United Nations General Assembly on September 25 2015 in New York UN Secretary General Ban Ki moon said the 17 goals and 169 targets to be achie...

Space Environment Science How Space Weather Affects All of Us The American Meteorological Society defines space weather as the variable conditions on the Sun and in the space environment that can influence the performance and reliability of space borne and ... ...

UNITED NATIONS UNITED STATES World leaders gathered at the United Nations on Friday adopted new global development goals for the next 15 years that aim to end poverty and improve lives. UN Secretary General Ban Ki moon speaks during the 70th session of the United Nations General Assembly on September 25 2015 in New York UN Secretary General Ban Ki moon said the 17 goals and 169 targets to be achieved by 2030 laid out a universal integrated and transformative vision for the world.

Figure 2: Title generation using summarization

We use a paragraph as the source text to be summarized, and the title of the article to which it belongs as the target summary to train our model. Our training set consists of 1.29million pairs, and the validation set has 68,000 pairs. We use a bidirectional RNN encoder with global dot attention [LPM15] in the decoder, stochastic gradient descent as optimizer and a dropout value of 0.3 as suggested by [GG16].

6 Conclusion and Outlook

We plan to evaluate the usefulness of our tool with a user study in the near future. As a next step, we aim to use Elasticsearch[3] to index and search additional document collections. It is also possible to enhance the experience by adding additional features using deep learning such as; automatic POS tagging using sequence models, locating important parts of the text using attention networks, question answering, text comprehension for visual inspection and dynamic interaction with very long documents. We propose that by leveraging big data and deep learning, we can greatly improve the collaboration between the content creators and consumers in today's age of information.

[3]https://www.elastic.co/products/elasticsearch

Table 1: Gold (**G**) vs predicted (**P**) summaries

G:	Kzoo apartment fire prompts large emergency response
P:	Crews respond to fire at Kalamazoo apartment complex
G:	Man accused of rape refused bail
P:	Bundaberg man accused of sexually abusing women
G:	Third Ukrainian policeman dies from injuries after clashes in Kiev , more than 140 people in hospital
P:	Ukrainian policeman dies in Kiev battle
G:	Duquesne beats Bucknell 26-7
P:	Duquesne Rolls Past Bucknell
G:	Taylor ton helps England keep series alive
P:	Taylor century helps England beat Australia

References

[BGJM17] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *TACL*, 5:135–146, 2017.

[CAMM16] David Corney, Dyaa Albakour, Miguel Martinez, and Samir Moussa. What do a million news articles look like? *ECIR 2016, Padua, Italy, March 20, 2016.*, pages 42–47, 2016.

[FPW+99] Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. Domain-specific keyphrase extraction. *IJCAI '99*, pages 668–673, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[GG16] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. *NIPS 2016, December 5-10, Barcelona, Spain*, pages 1019–1027, 2016.

[GHMS13] Tim Gollub, Matthias Hagen, Maximilian Michel, and Benno Stein. From keywords to keyqueries: Content descriptors for the web. *SIGIR '13*, pages 981–984, New York, NY, USA, 2013. ACM.

[JDJ17] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *CoRR*, abs/1702.08734, 2017.

[LPM15] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421, 2015.

[MSB+14] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. *ACL System Demonstrations*, pages 55–60, 2014.

[MZH+17] Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. Deep keyphrase generation. *ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 582–592, 2017.