Newsjam

A Multilingual Summarization Tool for News Articles

Joseph Keenan
joseph-alexander.keenan@loria.fr

Shane Kaszefski-Yaschuk Adelia Khasanova Maxime Méloux
{kaszefsk1u, khasanov2u, meloux1u}@etu.univ-lorraine.fr

eKnow 2022 - June 26, 2022 to June 30, 2022

Presenter Biography

- Joseph Keenan
  - B.A. Cognitive Science from University of Southern California
  - Master's Student in Natural Language Processing at Université de Lorraine

Motivation

Overload of COVID News

- Many people feeling overwhelmed about all of the constant COVID-19 updates [Savage, 2020]
- One study found that more than 26 million coronavirus related articles have been posted since the beginning of the pandemic [Krawczyk et al., 2021]
- Concise, summarized news about the COVID-19 pandemic keeps readers informed while reducing this sort of 'news fatigue'
- Our solution: create a model which automatically scrapes, classifies, and summarizes articles relevant to particular regions then posts them to Twitter

Motivation

Overload of COVID News

- Many people feeling overwhelmed about all of the constant COVID-19 updates [Savage, 2020]
- One study found that more than 26 million coronavirus related articles have been posted since the beginning of the pandemic [Krawczyk et al., 2021]
- Concise, summarized news about the COVID-19 pandemic keeps readers informed while reducing this sort of 'news fatigue'
- Our solution: create a model which automatically scrapes, classifies, and summarizes articles relevant to particular regions then posts them to Twitter

Motivation

Overload of COVID News

- Many people feeling overwhelmed about all of the constant COVID-19 updates [Savage, 2020]
- One study found that more than 26 million coronavirus related articles have been posted since the beginning of the pandemic [Krawczyk et al., 2021]
- Concise, summarized news about the COVID-19 pandemic keeps readers informed while reducing this sort of 'news fatigue'
- Our solution: create a model which automatically scrapes, classifies, and summarizes articles relevant to particular regions then posts them to Twitter

Motivation

Overload of COVID News

- Many people feeling overwhelmed about all of the constant COVID-19 updates [Savage, 2020]
- One study found that more than 26 million coronavirus related articles have been posted since the beginning of the pandemic [Krawczyk et al., 2021]
- Concise, summarized news about the COVID-19 pandemic keeps readers informed while reducing this sort of 'news fatigue'
- Our solution: create a model which automatically scrapes, classifies, and summarizes articles relevant to particular regions then posts them to Twitter

Datasets

French Datasets

- French version of the MultiLingual SUMmarization corpus (MLSUM) [Scialom et al., 2020]
  - contains over 400,000 articles from *Le Monde*
- Custom corpora of French COVID-19 news articles
  - 895 articles scraped from *Actu*[1]
  - 1,703 scraped from *L'Est Républicain*[2]

English Datasets

- CNN/Daily Mail Corpus [Hermann et al., 2015]
  - 11,490 English articles selected
- Custom corpus of English COVID-19 news articles
  - 2,010 articles scraped from *The Guardian*[3]

[1] `https://www.actu.fr`
[2] `https://www.estrepublicain.fr/`
[3] `https://theguardian.com`

Datasets

French Datasets

- French version of the MultiLingual SUMmarization corpus (MLSUM) [Scialom et al., 2020]
  - contains over 400,000 articles from *Le Monde*
- Custom corpora of French COVID-19 news articles
  - 895 articles scraped from *Actu*[1]
  - 1,703 scraped from *L'Est Républicain*[2]

English Datasets

- CNN/Daily Mail Corpus [Hermann et al., 2015]
  - 11,490 English articles selected
- Custom corpus of English COVID-19 news articles
  - 2,010 articles scraped from *The Guardian*[3]

[1] `https://www.actu.fr`
[2] `https://www.estrepublicain.fr/`
[3] `https://theguardian.com`

Breakdown of Custom Corpora

Table: Geographical Breakdown of Custom corpora

French corpus		English corpus			
French	**Global**	**North America**	**British Isles**	**Oceania**	**Global**
58%	42%	46%	24%	16%	11%

Table: Inter-annotator agreement between annotators *A*, *B*, and *C*

Dataset	*Actu*	*Guardian*			*L'Est Républicain*		
Metric	**A-B**	**A-B**	**A-C**	**B-C**	**A-B**	**A-C**	**B-C**
A_o	0.995	0.987	0.917	0.913	0.966	0.976	0.978
S	0.990	0.974	0.834	0.827	0.958	0.952	0.955
π	0.956	0.961	0.761	0.747	0.949	0.879	0.888
κ	0.956	0.961	0.761	0.748	0.949	0.879	0.888

Article Classification

Methods used:

- Logistic Regression
- Multinomial Naive Bayes
- Support Vector Machine

Classification Results

Classification results for French

Method	**Accuracy**	**Precision**	**Recall**	**F1**
Multinomial Naive Bayes	0.842	0.711	0.842	0.775
MNB (tuned)	0.845	0.732	0.845	0.781
Logistic Regression	0.934	0.934	0.934	0.934
LR (tuned)	**0.943**	**0.943**	**0.943**	**0.943**
Support Vector Machine	0.934	0.934	0.934	0.934
SVM (tuned)	0.943	0.943	0.943	0.943

Classification results for English

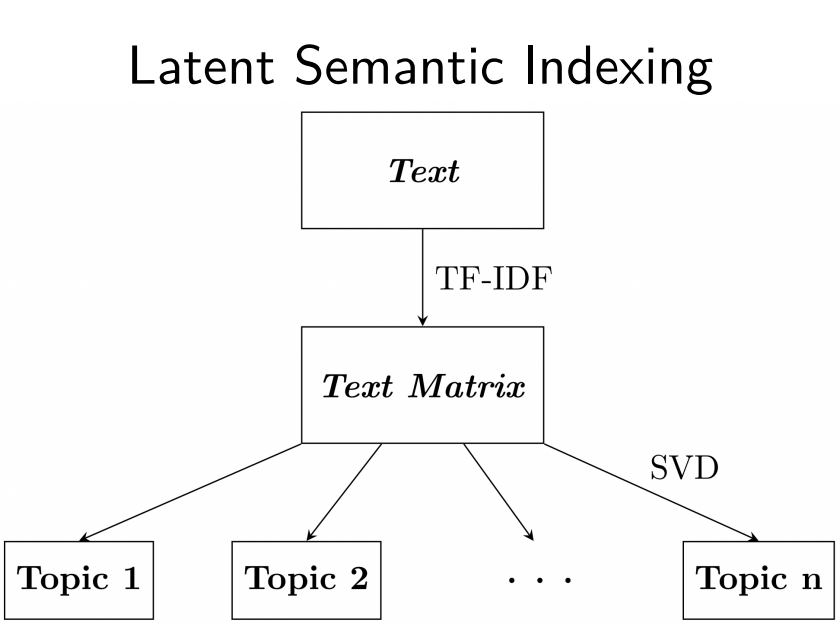
Method	**Accuracy**	**Precision**	**Recall**	**F1**
Multinomial Naive Bayes	0.610	0.505	0.610	0.505
MNB (tuned)	0.628	0.505	0.628	0.505
Logistic Regression	0.934	0.934	0.934	0.934
LR (tuned)	**0.935**	**0.935**	**0.935**	**0.935**
Support Vector Machine	0.921	0.921	0.921	0.918
SVM (tuned)	0.932	0.932	0.932	0.932

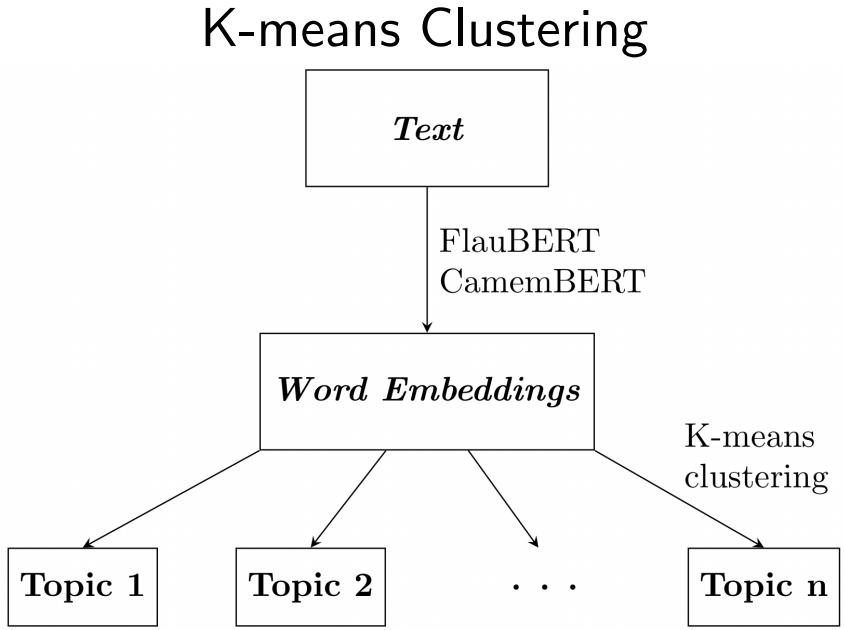
Summarization

Two extractive methods of summarization were implemented for this project:

- Latent Semantic Indexing (LSI)
- K-Means Clustering on Contextual Word Embeddings

Visualization of Summarization Methods





Summarization Results

Summarization evaluation for French (average scores for WMD, average F1-scores for other metrics)

Method	ROUGE-L	Keyword ROUGE-L	BERTScore	Keyword BERTScore	WMD	Keyword WMD
MLSUM corpus, test set (15,828 articles)						
LSI	**0.0652**	**0.0627**	0.5894	0.5885	**0.2517**	0.2652
FlauBERT + k-means	0.0564	0.0571	0.5905	**0.5909**	0.2558	0.2629
CamemBERT + k-means	0.0591	0.0598	**0.5907**	0.5904	0.2528	**0.2602**
Built corpus (787 + 1,803 = 2,560 articles)						
LSI	**0.1536**	**0.1538**	**0.6267**	**0.6238**	**0.2355**	**0.1959**
FlauBERT + k-means	0.1040	0.1075	0.6137	0.6134	0.2471	0.2080
CamemBERT + k-means	0.1093	0.1123	0.6153	0.6143	0.2452	0.2057

Summarization evaluation for English (average scores for WMD, average F1-scores for other metrics)

Method	ROUGE-L	Keyword ROUGE-L	BERTScore	Keyword BERTScore	WMD	Keyword WMD
CNN/Daily Mail corpus, test set (11,490 articles)						
LSI	**0.1207**	**0.1894**	**0.4947**	**0.4807**	**0.2178**	**0.1709**
RoBERTa + k-means	0.0839	0.1513	0.4680	0.4640	0.2342	0.1807
Built corpus (2,010 articles)						
LSI	**0.0748**	**0.1162**	**0.4822**	**0.4663**	**0.2297**	**0.2241**
RoBERTa + k-means	0.0533	0.0953	0.4702	0.4650	0.2390	0.2331

Generated Summary Example

Quality Summary | Poor Score (MLSUM Test Dataset Article 54[4]):

- *Reference Summary*: "The main suspect, a city worker, fired "blindly". He also died."
- *Generated Summary*: "On Friday, May 31st, twelve people were killed by a shooter in a municipal building in Virginia Beach (Virginia), a seaside resort on the east coast of the USA."

ROUGE-L F1-Score: 0.151

BERTScore F1-Score: 0.157

[4] `https://bit.ly/3n9qRuK`

Pipeline

- *Tuned Logistic Regression* was selected as the classification method and *Latent Semantic Indexing* was selected as the default summarization method
- A full pipeline was implemented
- It works as follows:
  1. Choose summarizer and x number of articles
  2. Scraper grabs the newest x articles and feeds them into the classifier
  3. Valid articles in French or English that are classified as 'local' are fed into the summarizer
  4. Generated summary & URL are posted to Twitter

Example Tweets

Links

- The NewsJam twitter page can be found at: `https://www.twitter.com/NewsjamOfficial`
- The NewsJam GitHub repository can be found at: `https://github.com/pie3636/newsjam`

Thank you

Bibliography

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015).

Teaching machines to read and comprehend.

Advances in neural information processing systems, 28.

Krawczyk, K., Chelkowski, T., Laydon, D. J., Mishra, S., Xifara, D., Gibert, B., Flaxman, S., Mellan, T., Schwämmle, V., Röttger, R., Hadsund, J. T., and Bhatt, S. (2021).

Quantifying Online News Media Coverage of the COVID-19 Pandemic: Text Mining Study and Resource.

Journal of Medical Internet Research, 23(6):e28253.

Savage, M. (2020).

Coronavirus: How much news is too much?

Scialom, T., Dray, P.-A., Lamprier, S., Piwowarski, B., and Staiano, J. (2020).

Mlsum: The multilingual summarization corpus.

arXiv preprint arXiv:2004.14900.