Krishna Garg, Jishnu Ray Chowdhury, Cornelia Caragea | University of Illinois Chicago

Keyphrase Generation Beyond the Boundaries of Title and Abstract

EMNLP 2022

Keyphrase Generation Task

Ranking-based Method for News Stance Detection

Qiang Zhang
University College London
London, United Kingdom
qiang.zhang.16@ucl.ac.uk

Emine Yilmaz
University College London
London, United Kingdom
emine.yilmaz@ucl.ac.uk

Shangsong Liang
KAUST
Thuwal, Saudi Arabia
shangsong.liang@kaust.edu.sa

ABSTRACT

A valuable step towards news veracity assessment is to understand stance from different information sources, and the process is known as the stance detection. Specifically, the stance detection is to detect four kinds of stances ("agree", "disagree", "discuss" and "unrelated") of the news towards a claim. Existing methods tried to tackle the stance detection problem by classification-based algorithms. However, classification-based algorithms make a strong assumption that there is clear distinction between any two stances, which may not be held in the context of stance detection. Accordingly, we frame the detection problem as a ranking problem and propose a ranking-based method to improve detection performance. Compared with the classification-based methods, the ranking-based method compare the true stance and false stances and maximize the difference between them. Experimental results demonstrate the effectiveness of our proposed method.

CCS CONCEPTS

• **Computing methodologies → Information extraction**;

KEYWORDS

Fake news; stance detection; learning to rank

Stance detection has been suggested as a crucial first step to detect fake news. [1] Researchers from both academia and industry initiated the Fake News Challenge (FNC) [1] and the first stage of FNC (FNC-1) aimed to accelerate the establishment of automatic systems for evaluating the positions that a news source holds about a particular claim. More specifically, given a news headline as a claim and its article body, FNC-1 tried to develop models to estimate the stance of the article body towards its headline. The stance could be one of the labels: "agree", "disagree", "discuss" and "unrelated". All the news with the "agree", "disagree" and "discuss" are assumed as "related". According to the FNC-1, formal definitions of the four stances are as: "**Agree**" – the body text agrees with the headline; "**Disagree**" – the body text disagrees with the headline; "**Discuss**" – the body text discusses the same claim as the headline, but does not take a position; and "**Unrelated**" – the body text discusses a different claim but not that in the headline.

A GradientBoosting classifier is implemented as the FNC-1 official baseline with a relative score of 75.20%. This classifier makes use of semantic analysis and overlap between headlines and bodies. [1] As for the FNC-1 submissions, an ensemble of convolutional neural network and gradient-boosted decision trees achieves the highest detection performance [2]. Another ensemble of five multi-layer perceptrons (MLP) achieves a little worse accuracy. These two

Applications

- Reviewer matching systems for conferences/ journals
- Mining literature
- Recommendations to readers

Motivation

- Current works are limited to Title and Abstract
- Can we do better by going beyond Title and Abstract?

Contributions

- Leverage information from body of the article
- Present new large-scale scholarly dataset FullTextKP
- Extractive summary sentences are the richest sources
- Comprehensive analysis on four models

Methodology

- Input:
  - Title [SEP] Abstract [SEP] Sent_1 [SEP] Sent_2 [SEP] …. Sent_k

- Output
  - kp_1 [SEP] kp_2 [SEP] … [SEP] kp_n

Methodolody: <Sent$_i$>

Highlighted:
Citation Sentences

Rest:
Non-citation sentences

Either:
Random sentences

1 Introduction

Since the introduction of public key cryptography (Diffie & Hellman 1976, Rivest, Shamir & Adleman 1978) there has been great interest their homomorphic properties. In general terms, this means being able to operate on the encrypted data as you would on unencrypted data. This idea was suggested by (Rivest, Adleman & Dertouzos 1978) under the term *privacy homomorphisms.*

Since then there has been great interest in homomorphic encryption, and the ability to construct a Fully Homomorphic Encryption (FHE) scheme. That is, having a complete set of operations as one does with the unencrypted data. The first provably secure solution was given by (Gentry 2009), which confirmed that such a scheme is possible.

Methodolody: <Sent_i>

- (Extractive) Summary Sentences
  - Extractive summary of the body
  - PacSum (Zheng and Lapata, 2019) algorithm
  - More central sentences occur on the top
- (Abstractive) Summary Sentences
  - Abstractive summary of the body
  - BirdBirdPegasus pretrained on Arxiv dataset

Methodolody: <Sent$_i$>

- Retrieval-Augmentation Sentences
  - Create a pool of all sentences from T+A of training set (values) and embed using SPECTER[1]
  - Query: embed T+A of a given article using SPECTER
  - Retrieve k most similar sentences using FAISS[2]

1 SPECTER: https://aclanthology.org/2020.acl-main.207.pdf
2 FAISS: https://github.com/facebookresearch/faiss

Concrete Example

Method	Snippets from title and abstracts and beyond
Title	Performance of Power Detector Sensors of DTV Signals in IEEE 802.22 WRANs
Abstract	Sensing is the most important component in any cognitive radio system. The IEEE 802.22 Working Group (WG) is formulating the first worldwide standard for cognitive radios to operate in the television (TV) bands...
Citations	The IEEE 802.22 functional requirements document [2] states that spectrum sensing is required ...
Non-Citations	The keep-out region is a region around the primary user (e.g. DTV transmitter) where... Given these simulation scenarios we can evaluate various spectrum sensing techniques.
Summary	Of course, identification of which TV channels... This paper will describe ... in the IEEE 802.22 WG to evaluate spectrum sensing techniques... Calculating the base station keep-out region ...
Ret-Aug	The Bernoulli nonuniform sampling is further extended to matrix formulation, which allows the application of spectrum sensing for cognitive radio signal detection. *- From Paper: 'Noise Enhancement and SNR Equivalence in Bernoulli Nonuniform Sampling'*
Keyphrases	Cognitive Radio, IEEE 802.22, Spectrum Sensing, Power Detector, Keep-out Region, False alarm probability, Misdetection probability

FullTextKP Dataset

- Existing datasets are either small scale or do not contain body of the articles

Total papers in the corpus	142,844
Training set size	114,271
Validation set size	14,287
Test set size	14,286
% of present keyphrases	55.8
% of absent keyphrases	44.2
Average keyphrases per paper	4.3

Table 2: FULLTEXTKP dataset statistics

Evaluation

- Precision, Recall, F1

- Treat a keyphrase as *Present* if it is verbatim present in **Title [SEP] Abstract**, else *Absent*
  - Recall our Input: Title [SEP] Abstract [SEP] <additional sentences from body>

Results

Results - Present Keyphrase Generation

- Using additional sentences from body helps on all metrics
- Summary sentences give maximum performance boost
- Citation sentences worse than Non-Citation sentences (unlike Keyphrase Extraction Task)
- Retrieval-Augmentation sentences hurt the performance
- Longformer Encoder-Decoder (LED)
    - First time demonstrated for the task
    - Suited for long documents
    - Best among all 4

Model	Method	P@M	R@M	F1@M
catSeq	Title+Abstract	0.333	0.343	0.338
	+ Ret-Aug	0.338	0.343	0.340
	+ Citations	0.360	0.351	0.355
	+ Non-Citations	0.365	0.366	0.366
	+ Random	0.370	0.359	0.364
	+ Summary	**0.379**	**0.380**	**0.380**
One2Set	Title+Abstract	0.300	**0.425**	0.351
	+ Ret-Aug	0.313	0.370	0.339
	+ Citations	0.316	0.396	0.351
	+ Non-Citations	0.316	0.397	0.352
	+ Random	0.314	0.394	0.349
	+ Summary	**0.325**	0.387	**0.353**
ExHiRD	Title+Abstract	0.320	0.382	0.325
	+ Ret-Aug	0.317	0.391	0.325
	+ Citations	0.336	0.403	0.340
	+ Non-Citations	0.353	0.407	0.354
	+ Random	0.352	0.404	0.351
	+ Summary	**0.369**	**0.416**	**0.366**
LED	Title+Abstract	0.333	0.392	0.360
	+ Ret-Aug	0.319	0.394	0.353
	+ Citations	0.338	0.407	0.369
	+ Non-Citations	0.348	0.425	0.382
	+ Random	0.354	0.431	0.389
	+ Summary	**0.365**	**0.435**	**0.397**

Results – Absent Keyphrase Generation

Model	Method	P@M	R@M	F1@M
catSeq	Title+Abstract	0.030	0.017	0.021
	+ Ret-Aug	0.037	0.021	0.027
	+ Citations	0.052	0.031	0.039
	+ Non-Citations	0.068	0.046	0.055
	+ Random	0.073	0.049	0.059
	+ Summary	**0.093**	**0.069**	**0.079**
One2Set	Title+Abstract	0.043	0.042	0.043
	+ Ret-Aug	0.044	0.039	0.041
	+ Citations	0.053	0.047	0.049
	+ Non-Citations	0.054	**0.053**	0.053
	+ Random	**0.056**	**0.053**	**0.054**
	+ Summary	0.023	0.050	0.032
ExHiRD	Title+Abstract	0.042	0.026	0.030
	+ Ret-Aug	0.035	0.022	0.025
	+ Citations	0.060	0.041	0.045
	+ Non-Citations	0.071	0.051	0.055
	+ Random	0.076	0.054	0.059
	+ Summary	**0.100**	**0.074**	**0.080**
LED	Title+Abstract	0.061	0.063	0.061
	+ Ret-Aug	0.053	0.034	0.041
	+ Citations	0.073	0.058	0.065
	+ Non-Citations	0.088	0.071	0.079
	+ Random	0.093	0.078	0.085
	+ Summary	**0.112**	**0.100**	**0.106**

Results

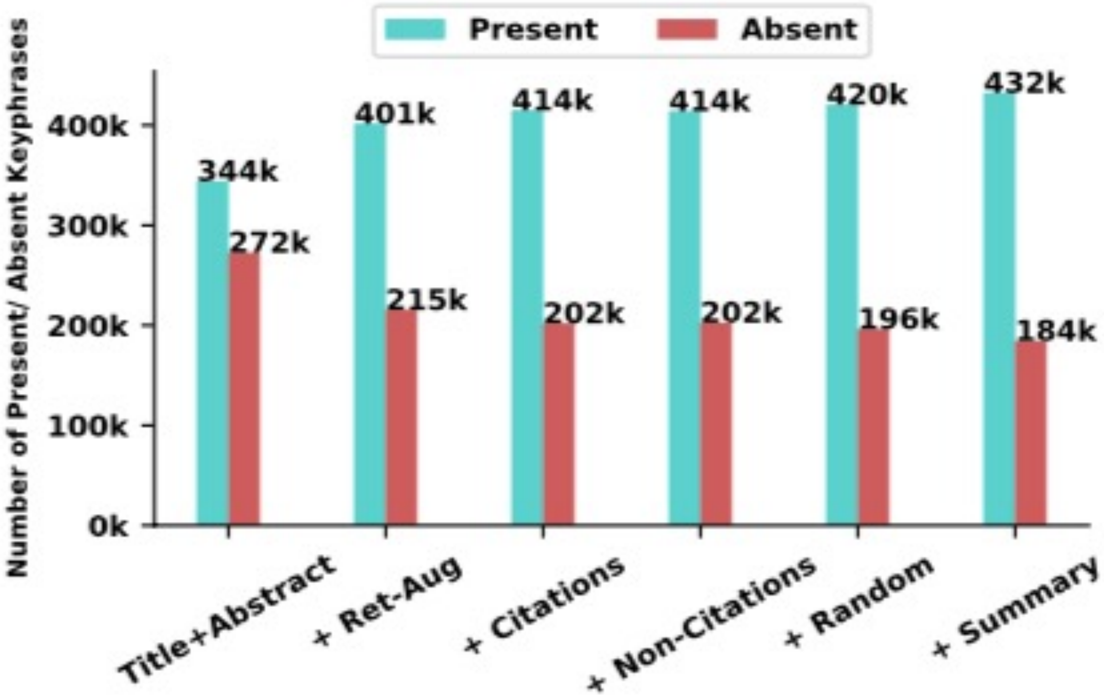

Figure shows number of present and absent keyphrases in the full text (of various methods)

- Summary $>_{88000}$ T+A
- Summary is the richest source of present keyphrases and model learns to generalize better when trained on such sentences

Results – Analysis on Summarization Methods

Present KP	P@M	R@M	F1@M
Title+Abstract	0.333	0.392	0.360
+ Abs_Summary	0.326	0.409	0.363
+ Ext_Summary	**0.365**	**0.435**	**0.397**
Absent KP	P@M	R@M	F1@M
Title+Abstract	0.061	0.063	0.061
+ Abs_Summary	0.062	0.048	0.054
+ Ext_Summary	**0.112**	**0.100**	**0.106**

Keyphrase Generation:
Longformer Encoder-Decoder (LED)

Abstractive Summary:

- BigBirdPegasus pretrained on Arxiv dataset
- Performs at par/ worse than Title+Abstract
- 2/3rd sentences from Introduction, 10-15 % repetitions

Extractive Summary:

- PacSum algorithm
- Bring complimentary information from the entire body

Results – Exploring longer sequence lengths

Type	Method	Seq-Len	P@M	R@M	F1@M
Present Keyphrases	T+A+Summary	800	0.365	**0.435**	0.397
	T+A+Random	800	0.354	0.431	0.389
		1500	0.370	0.427	0.397
		2000	0.372	0.431	0.399
		2500	**0.389**	0.427	**0.407**
Absent Keyphrases	T+A+Summary	800	0.112	0.100	0.106
	T+A+Random	800	0.093	0.078	0.085
		1500	0.110	0.097	0.103
		2000	0.114	0.100	0.107
		2500	**0.126**	**0.108**	**0.116**

- Summary-800 comparable to Random-2000 -> Summary is indeed useful
- Performance increases as we increase sequence length
- For very long sequence lengths, Summary converges to Random

Prediction Samples

Snippets from title and abstracts and summary with model predictions	Gold Keyphrases
Segmentation of brain MR images using intuitionistic fuzzy clustering algorithm <sep> A new algorithm ... using intuitionistic fuzzy clustering (IFCM), is proposed in this paper.	segmentation, brain mr image, intuitionistic fuzzy set
GPU -based simulation of wireless body area network <sep> ... in a local workstation to perform model simulations of a ... the goal of this project is to gain an understanding of the gpu -based performance of wban model simulation ...	gpu, wireless body area network, model, matlab
Spectrum sharing for directional systems <sep> Dynamic spectrum access systems will need to ... An example involving a satellite downlink antenna and a broadband wireless access system using directional antennas is presented ...	dynamic spectrum access, directionality
Automatic classification of anuran sounds using convolutional neural networks <sep> ... networks with mel-frequency cepstral coefficients (mfccs) as input for the task of ... <sep> Different machine learning approaches for anuran classification ...	anuran, convolutional neural networks, wireless sensor networks, mfcc, machine learning

cyan Predicted keyphrases by LED

Conclusion

- New direction for keyphrase generation task by using body of the articles
- New large-scale dataset FullTextKP containing full length of the articles
- Extractive summary sentences contain the most topical information
- Comprehensive analysis with four models, including LED

Thank you

FLOWCODE

PRIVACY.FLOWCODE.COM