Learning to Generate Correct Numeric Values in News Headlines

Jui Chu[1], Chung-Chi Chen[1], Hen-Hsen Huang[2,3], Hsin-Hsi Chen[1,3]

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taiwan

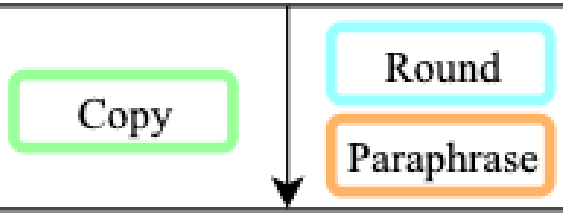
[2]Department of Computer Science, National Chengchi University, Taiwan

[3]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

The Process of Numeric Values from News Articles to Headlines

NLP Lab
National Taiwan University

(Article1) FED raises the rates, showing the confidence for economic recovery. This also stimulates the European stock market, especially the export stocks, in which the Germany ones boosts by 2.6%.
(Article2) The numbers of manufacturing are in bad situation, dragging the US stocks to pullback for 24 days. The NASDAQ composite index of technology stocks drops down by 1.67%.
(Article3) Ride-hailing giant Uber released its first safety report on Thursday, disclosing that it received a total of 5,981 reports of sexual assault incidents during its ride-hailing trips in the U.S. in 2017 and 2018.



(Headline1) European Stock Takes a Rides of US Rates Raise. The Germany Stock Rises by 2.6%.
(Headline2) US Manufacturing Stocks Struggle. NASDAQ Composite Collapses by 1.7%.
(Headline3) Uber Discloses Nearly 6K Reports of Sexual Assaults

Methodology

Base Architecture: Character-based Transformer

Two Different Forms:

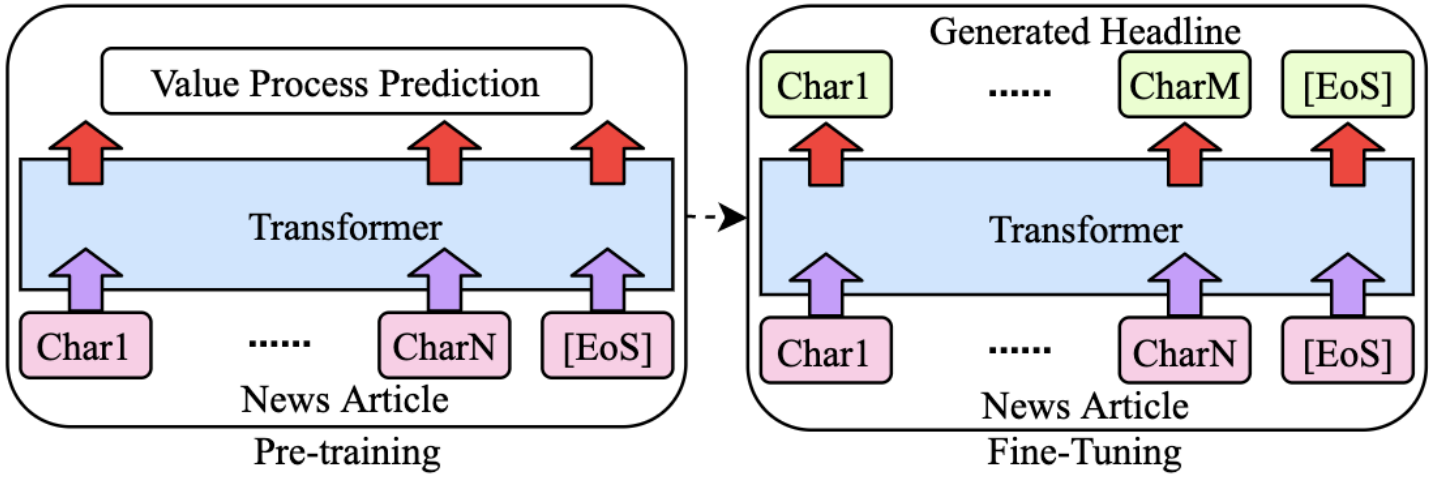
(1) Token: whole value as a token

(2) Digit: represent the value digit by digit.

Two Different Settings:

(1) Numeral Category: year, month, date, percentage, price, and population

(2) Pre-train by Value Process Prediction:



F1 Scores of Numeric Correctness (%)

NLP Lab
National Taiwan University

Method	From Article	New to Article	Weighted Average
Base Model	45.23	21.28	23.81
w/ Token	**50.76**	18.06	35.71
w/ Token w/o category	49.71	18.74	35.46
w/ Digit	41.37	21.45	32.21
w/ Digit w/o category	39.07	21.63	31.05
w/ Digit + pre-train	45.71	**24.25**	**35.83**

(1) For the numerals copied directly from the news articles, the token-based model performs the best.

(2) For the numerals rounded or paraphrased from the news articles, the proposed pre-train process performs the best.

(3) We also find that add the category information is helpful for the overall results.

Conclusion & Future Work

NLP Lab
National Taiwan University

- We experiment with various methods to guide the generator to process values from articles for headlines.

- **The proposed pre-train task leads the model to correctly generate numbers** even in the extreme case of target values in headlines are newly generated by rounding and paraphrasing.

- The results show the challenge of these processes, implying the importance of numerical reasoning.

- **Exploring numerical reasoning methods for generative models** is our future work.

Related Events and Datasets

NLP Lab
National Taiwan University

FinNum-2 in NTCIR-2020

Semantic Understanding

Numeral Attachment:
$NE OK NE, last time oil was over $65 you were close to $8. Giddy-up…

NTCIR

http://finnum.nlpfin.com

FinNLP Workshop in IJCAI-2020

- Submission Deadline: **April 24, 2020**
- The **Best Paper Award** winner: USD$500
- Published at **ACL Anthology**

http://finnlp.nlpfin.com

Related Datasets

Numeracy-600K:
600K Market Comments from Reuters
600K Article Titles from the Web

http://numeracy600k.nlpfin.com

Numeral Understanding:
8,868 Annotations on Financial Tweets
Fine-grained Taxonomy for Numerals

http://numeralunderstanding.nlpfin.com

Please feel free to contact us if you have any questions.

Jui Chu: jchu@nlg.csie.ntu.edu.tw

Chung-Chi Chen: cjchen@nlg.csie.ntu.edu.tw