[Documentation]

. , .

version1 . Bitbucket .

-
- v1 engine documentation
- 2021.01 PPS
- 2021.02 PPS

- (dacon.io) (kaggle) "AI, " .
- . .
- . . , . task ( ) . .

- . .
  - 2018 FastText Embedding CNN . JIPS | Github
  - 2019 BOAZ | Github
  - 2019 KAIST Deep Hierarchical Encoder . AAAI-19 | Github | Paper
- 2018 JIPS AAAI-19 github repository . AAAI-19 github . 2018 JIPS github .
- AAAI-2019 . .
  -

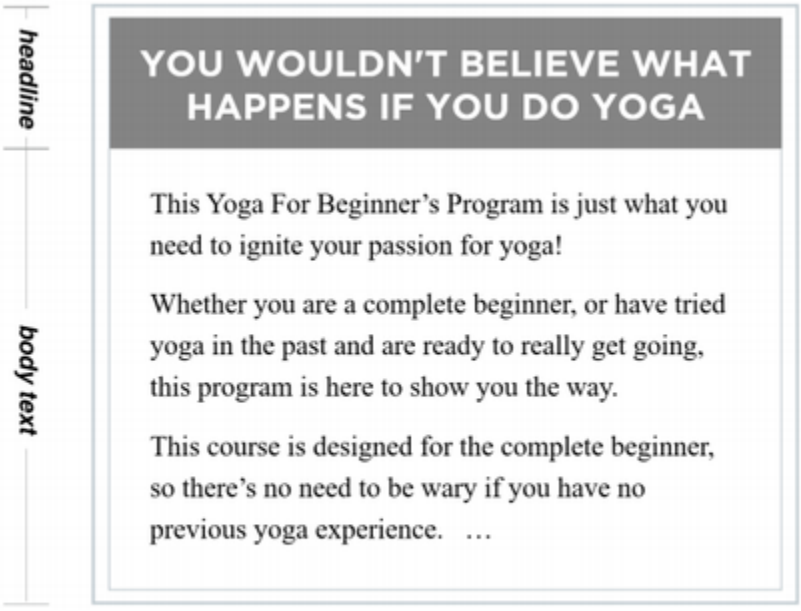

Figure 1: A news article example of the incongruent headline.

  - .
- 2018 JIPS . 1, 2 1 . 2 AAAI-19 . 1 inconsistent 2 irrelevant . Inconsistent Irrelevant 2021 NLP pps . . 1 mission1 2 mission2 .
- 2018 JIPS (inconsistent irrelevant), 2018 JIPS . KAIST (———).
- : , .
- , 2018 JIPS . Inconsistent . Irrelevant ( ). github issue KAIST . . inconsistent . .
- label 2018 JIPS .

2018 JIPS

- Train dataset
  - 1 (inconsistent) 31000
  - 2 (irrelevant) 68000
  - 0 , 1
    - 1 2 5:5 well balanced .
- Test dataset
  - 1 (inconsistent) 100
  - 2 (irrelevant) 100

—

2018 JIPS train dataset test dataset .

- meta feature (,) TF-IDF CatBoost Mecab Bidirectional LSTM . CatBoost Bidirectional LSTM 1 2 test dataset 50% .
- BERT pretrain BERT HanBERT KoELECTRA . HanBERT 54GB , KoELECTRA 34GB .
- monologg NLP pytorch . Tensorflow 2.0 pytorch .
- KoELECTRA HanBERT . ELECTRA replaced token detection pretraining BERT KoELECTRA .
- KoELECTRA "[CLS] title [SEP] content [SEP]" input id . 512 .
- how to fine tune BERT for text classification . classification 129 383 . Head+Tail truncation .
- Head+Tail truncation .
  - Inconsistent Irrelevant 1 0 KoELECTRA
  - Inconsistent Irrelevant
- test dataset .
- 2018 JIPS CNN AUROC (.. ??). Head+Tail KoELECTRA FastText + BCNN AUROC .
-

Model	Inconsistent Type	Irrelevant Type
FastText + BCNN (2018)	0.528	0.726
ELECTRA head+tail, ALL	0.872	**0.811**
ELECTRA head+tail, Inconsistent	**0.875**	--
ELECTRA head+tail, Irrelevant	--	0.654

- Irrelevant inconsistent irrelevant AUROC , inconsistent AUROC .
- Head+tail truncation Irrelevant Irrelevant Inconsistent .
- FastText+BCNN KoELECTRA AUROC .
- AUROC :
  - Inconsistent irrelevant Head+Tail KoELECTRA
    - Inconsistent: 81%, Irrelevant: 74%
  - Inconsistent Head+Tail KoELECTRA
    - Inconsistent: 84%
  - Irrelevant Head+Tail KoELECTRA
    - Irrelevant : 65%

Head+Tail truncation

- Irrelevant Head+Tail truncation .
- . 2018 JIPS document . Naive 1 1 0 0 . KoELECTRA ( )
- (1 labeling!) Head+Tail Head+Tail truncation (test set 60%).
- Hierarchical Transformers for long document classification Naive BERT (pretrain) , chunking feature extraction extracted feature RNN (GRU, LSTM) .
- .

- KoELECTRA . chunking fine-tune KoELECTRA .

  prepare_model.py

- , sliding window chunking .

  chunking.py

- representation stacking

  create_feature_dataframe.py

-
  -

	features	Label
0	[[-0.39957574, 0.052018672, 0.048946593, 0.315...	1
1	[[-0.43301085, -0.013179632, -0.018207176, 0.3...	1
2	[[-0.21780677, 0.20347388, 0.030457448, 0.3164...	1
3	[[-0.55885184, 0.07147807, 0.08166136, 0.37081...	1
4	[[-0.3625559, 0.027224315, 0.14107037, 0.33482...	1

- GRU .

  GRU.py

- GRU LSTM Temporal CNN . feature .
- Irrelevant dataset Head+Tail truncation AUROC , .
- inference time (test feature GRU predict) .

—

- head+tail Head+Tail , .
- Irrelevant,Inconsistent advertisement mismatch . Advertisement mismatch Irrelevant , irrelevant .
- none: 0, inconsistent: 1, irrelevant: 2, mismatch: 3, advertisement: 4 softmax mismatch advertisement . imbalance . Head+Tail irrelevant .
- . .
- head+tail . 512 () . predict .
- train set 400 80 . inconsistent ( ), ( ).
- KoELECTRA . ( )

  bert_checkpoint.py

- V100 1 fine-tune . 1000 step . 1 3-4 . .

- 1 97% . KAIST 2018 JIPS irrelevant irrelevant 79% AUROC 0.851 . (2018 JIPS).

Furture Work

- , .
- NLP .
  - 18GB 34GB KoELECTRA 54GB HanBERT . pretrain .
- .
- 2018 JIPS .
- task . ..