Clickbait Challenge

At SemEval 2023

By Andreas Kruff, Anh Huy Matthias Tran

Agenda

1. Introduction
2. The Data Set
3. Task 1: Spoiler Classification
   a. Ideas & Approaches
   b. Frameworks and Tools
   c. Evaluation
4. Task 2: Spoiler Generation
   a. Ideas & Approaches
   b. Frameworks and Tools
   c. Evaluation
5. Conclusion

1. Introduction

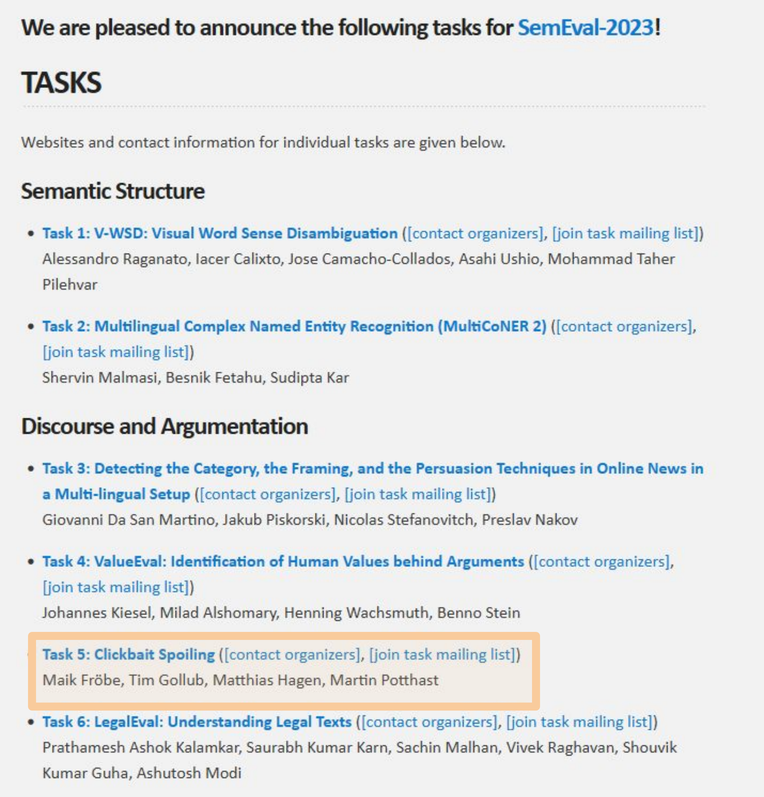
Clickbait Challenge at SemEval 2023 - Clickbait Spoiling

https://semeval.github.io/

SemEval2023

- Series of international NLP research workshops focusing on the evaluation of relevant NLP and computational semantic analysis systems
- Provides high quality annotated data sets
- Organizes and announces shared tasks with various kind of topics

→ **Task 5: Clickbait Spoiling**



We are pleased to announce the following tasks for SemEval-2023!

TASKS

Websites and contact information for individual tasks are given below.

Semantic Structure

- **Task 1: V-WSD: Visual Word Sense Disambiguation** ([contact organizers], [join task mailing list])
  Alessandro Raganato, Iacer Calixto, Jose Camacho-Collados, Asahi Ushio, Mohammad Taher Pilehvar
- **Task 2: Multilingual Complex Named Entity Recognition (MultiCoNER 2)** ([contact organizers], [join task mailing list])
  Shervin Malmasi, Besnik Fetahu, Sudipta Kar

Discourse and Argumentation

- **Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup** ([contact organizers], [join task mailing list])
  Giovanni Da San Martino, Jakub Piskorski, Nicolas Stefanovitch, Preslav Nakov
- **Task 4: ValueEval: Identification of Human Values behind Arguments** ([contact organizers], [join task mailing list])
  Johannes Kiesel, Milad Alshomary, Henning Wachsmuth, Benno Stein
- **Task 5: Clickbait Spoiling** ([contact organizers], [join task mailing list])
  Maik Fröbe, Tim Gollub, Matthias Hagen, Martin Potthast
- **Task 6: LegalEval: Understanding Legal Texts** ([contact organizers], [join task mailing list])
  Prathamesh Ashok Kalamkar, Saurabh Kumar Karn, Sachin Malhan, Vivek Raghavan, Shouvik Kumar Guha, Ashutosh Modi

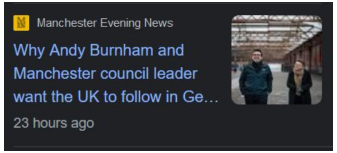
Clickbait Challenge at SemEval 2023 - Clickbait

Clickbait Spoiling

- **Clickbait**
  Posts that generate interest by creating a *curiosity gap*

- **Clickbait Spoiling**
  Generating a short text that answers the *curiosity gap*

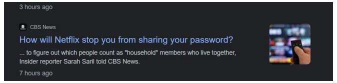
→ **No click required!**



Manchester Evening News

Why Andy Burnham and Manchester council leader want the UK to follow in Ge…

23 hours ago



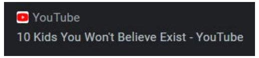
3 hours ago

CBS News

How will Netflix stop you from sharing your password?

... to figure out which people count as "household" members who live together, Insider reporter Sarah Saril told CBS News.

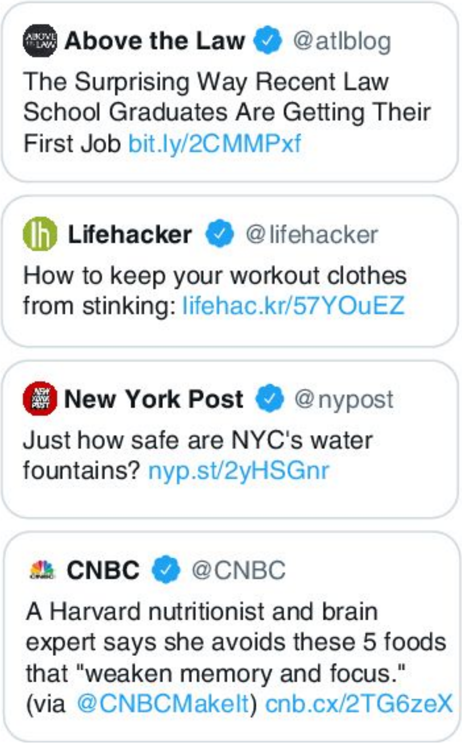
7 hours ago



YouTube

10 Kids You Won't Believe Exist - YouTube

Clickbait tweet



Above the Law @atlblog
The Surprising Way Recent Law School Graduates Are Getting Their First Job bit.ly/2CMMPxf

Lifehacker @lifehacker
How to keep your workout clothes from stinking: lifehac.kr/57YOuEZ

New York Post @nypost
Just how safe are NYC's water fountains? nyp.st/2yHSGnr

CNBC @CNBC
A Harvard nutritionist and brain expert says she avoids these 5 foods that "weaken memory and focus." (via @CNBCMakeIt) cnb.cx/2TG6zeX

Clickbait Challenge: Task 1

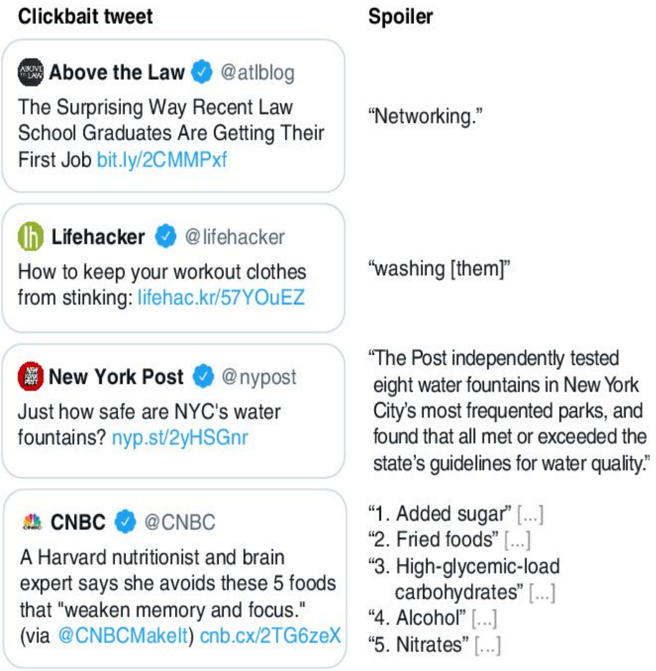
Spoiler Type Classification

Classifying the *spoiler type* of the clickbait post in **three categories**

1. Phrase
2. Passage
3. Multi

Expected Output:

```
{"uuid": "<UUID>", "spoilerType":
"<SPOILER-TYPE>"}
```



Clickbait tweet	Spoiler	
Above the Law @atlblog The Surprising Way Recent Law School Graduates Are Getting Their First Job bit.ly/2CMMPxf	"Networking."	→ *Phrase*
Lifehacker @lifehacker How to keep your workout clothes from stinking: lifehac.kr/57YOuEZ	"washing [them]"	→ *Phrase*
New York Post @nypost Just how safe are NYC's water fountains? nyp.st/2yHSGnr	"The Post independently tested eight water fountains in New York City's most frequented parks, and found that all met or exceeded the state's guidelines for water quality."	→ *Passage*
CNBC @CNBC A Harvard nutritionist and brain expert says she avoids these 5 foods that "weaken memory and focus." (via @CNBCMakeIt) cnb.cx/2TG6zeX	"1. Added sugar" [...] "2. Fried foods" [...] "3. High-glycemic-load carbohydrates" [...] "4. Alcohol" [...] "5. Nitrates" [...]	→ *Multi*

Clickbait Challenge

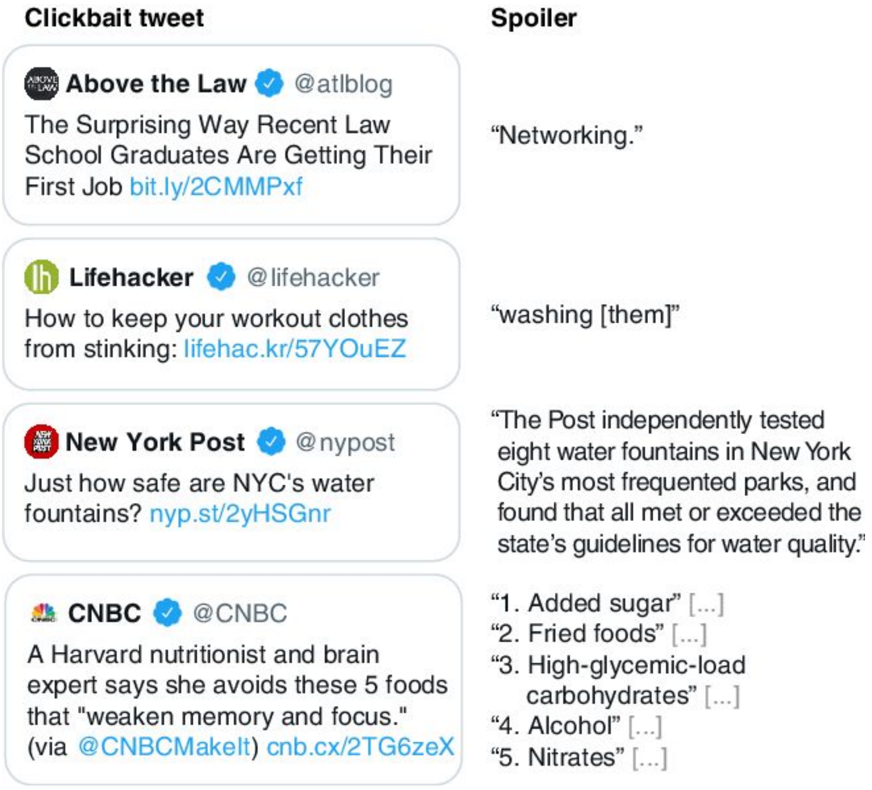
Spoiler Generation

Satisfying *curiosity* via question answering

- Inspect the post and the linked content for relevant passages
- Generate the spoiler for the clickbait post

Expected Output:

{"uuid": "<UUID>", "spoiler": "<SPOILER>"}



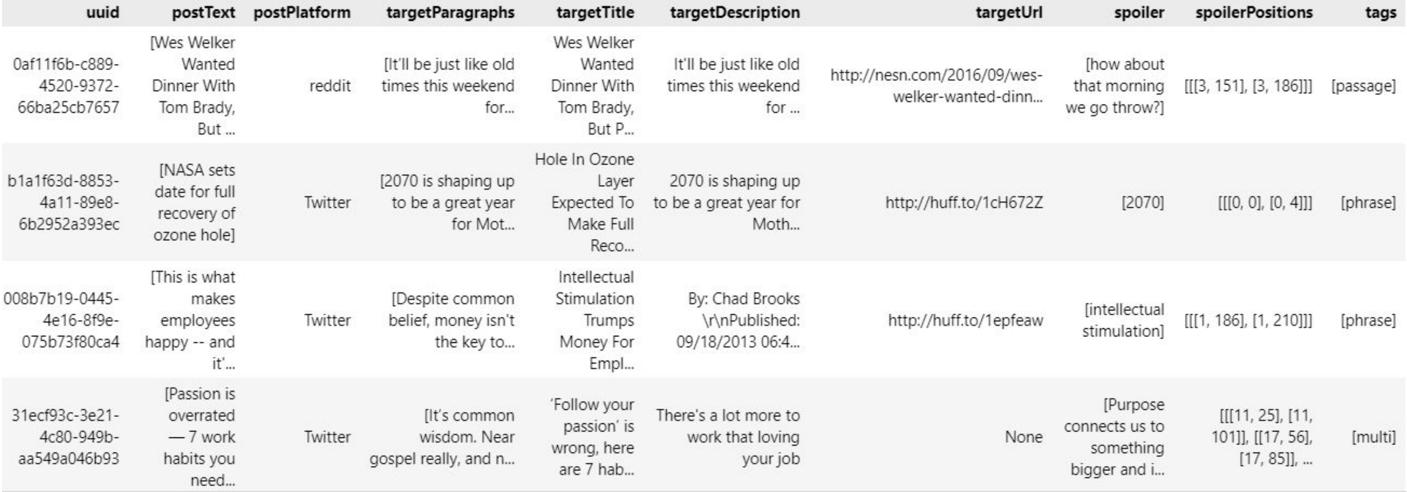
Clickbait tweet	Spoiler
Above the Law @atlblog The Surprising Way Recent Law School Graduates Are Getting Their First Job bit.ly/2CMMPxf	“Networking.”
Lifehacker @lifehacker How to keep your workout clothes from stinking: lifehac.kr/57YOuEZ	“washing [them]”
New York Post @nypost Just how safe are NYC's water fountains? nyp.st/2yHSGnr	“The Post independently tested eight water fountains in New York City’s most frequented parks, and found that all met or exceeded the state’s guidelines for water quality.”
CNBC @CNBC A Harvard nutritionist and brain expert says she avoids these 5 foods that "weaken memory and focus." (via @CNBCMakeIt) cnb.cx/2TG6zeX	“1. Added sugar” [...] “2. Fried foods” [...] “3. High-glycemic-load carbohydrates” [...] “4. Alcohol” [...] “5. Nitrates” [...]

2. The Data Set

Data Set: Key Facts

Total of 14 fields

➔ Domain Language : Majority English

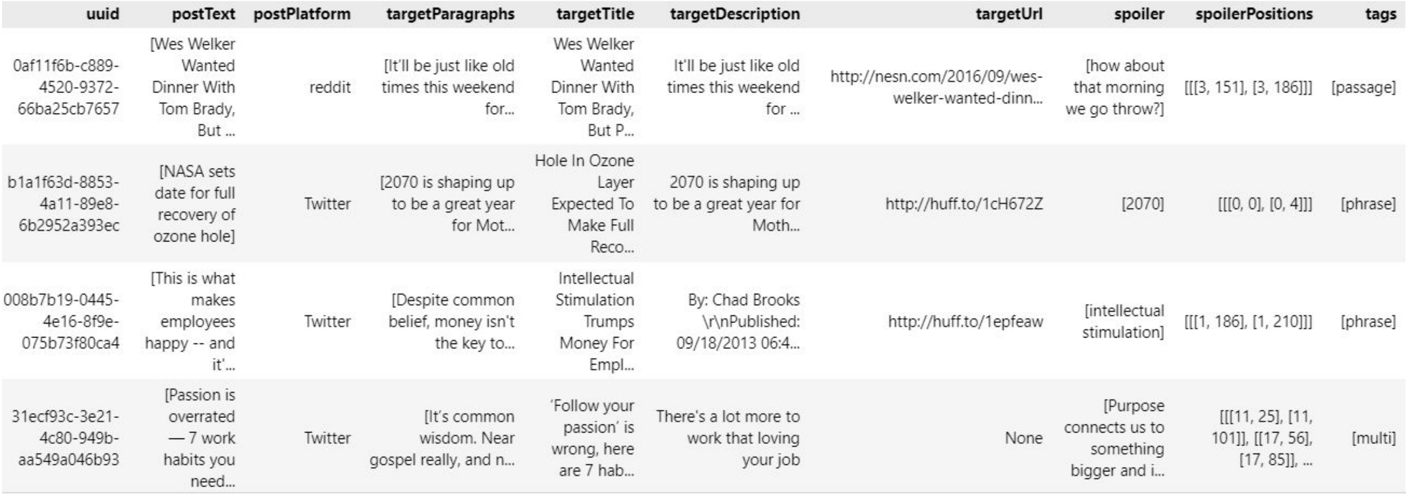


uuid	postText	postPlatform	targetParagraphs	targetTitle	targetDescription	targetUrl	spoiler	spoilerPositions	tags
0af11f6b-c889-4520-9372-66ba25cb7657	[Wes Welker Wanted Dinner With Tom Brady, But ...	reddit	[It'll be just like old times this weekend for...	Wes Welker Wanted Dinner With Tom Brady, But P...	It'll be just like old times this weekend for ...	http://nesn.com/2016/09/wes-welker-wanted-dinn...	[how about that morning we go throw?]	[[[3, 151], [3, 186]]]	[passage]
b1a1f63d-8853-4a11-89e8-6b2952a393ec	[NASA sets date for full recovery of ozone hole]	Twitter	[2070 is shaping up to be a great year for Mot...	Hole In Ozone Layer Expected To Make Full Reco...	2070 is shaping up to be a great year for Moth...	http://huff.to/1cH672Z	[2070]	[[[0, 0], [0, 4]]]	[phrase]
008b7b19-0445-4e16-8f9e-075b73f80ca4	[This is what makes employees happy -- and it'...	Twitter	[Despite common belief, money isn't the key to...	Intellectual Stimulation Trumps Money For Empl...	By: Chad Brooks \r\nPublished: 09/18/2013 06:4...	http://huff.to/1epfeaw	[intellectual stimulation]	[[[1, 186], [1, 210]]]	[phrase]
31ecf93c-3e21-4c80-949b-aa549a046b93	[Passion is overrated — 7 work habits you need...	Twitter	[It's common wisdom. Near gospel really, and n...	'Follow your passion' is wrong, here are 7 hab...	There's a lot more to work that loving your job	None	[Purpose connects us to something bigger and i...	[[[11, 25], [11, 101]], [[17, 56], [17, 85]], ...	[multi]

Data Set: Key Facts

Spoiler Fields mainly contains extractive spoilers

- Extractive (4534)
- Abstrative (88)



	uuid	postText	postPlatform	targetParagraphs	targetTitle	targetDescription	targetUrl	spoiler	spoilerPositions	tags
	0af11f6b-c889-4520-9372-66ba25cb7657	[Wes Welker Wanted Dinner With Tom Brady, But ...	reddit	[It'll be just like old times this weekend for...	Wes Welker Wanted Dinner With Tom Brady, But P...	It'll be just like old times this weekend for ...	http://nesn.com/2016/09/wes-welker-wanted-dinn...	[how about that morning we go throw?]	[[[3, 151], [3, 186]]]	[passage]
	b1a1f63d-8853-4a11-89e8-6b2952a393ec	[NASA sets date for full recovery of ozone hole]	Twitter	[2070 is shaping up to be a great year for Mot...	Hole In Ozone Layer Expected To Make Full Reco...	2070 is shaping up to be a great year for Moth...	http://huff.to/1cH672Z	[2070]	[[[0, 0], [0, 4]]]	[phrase]
	008b7b19-0445-4e16-8f9e-075b73f80ca4	[This is what makes employees happy -- and it'...	Twitter	[Despite common belief, money isn't the key to...	Intellectual Stimulation Trumps Money For Empl...	By: Chad Brooks \r\nPublished: 09/18/2013 06:4...	http://huff.to/1epfeaw	[intellectual stimulation]	[[[1, 186], [1, 210]]]	[phrase]
	31ecf93c-3e21-4c80-949b-aa549a046b93	[Passion is overrated — 7 work habits you need...	Twitter	[It's common wisdom. Near gospel really, and n...	'Follow your passion' is wrong, here are 7 hab...	There's a lot more to work that loving your job	None	[Purpose connects us to something bigger and i...	[[[11, 25], [11, 101]], [[17, 56], [17, 85]], ...	[multi]

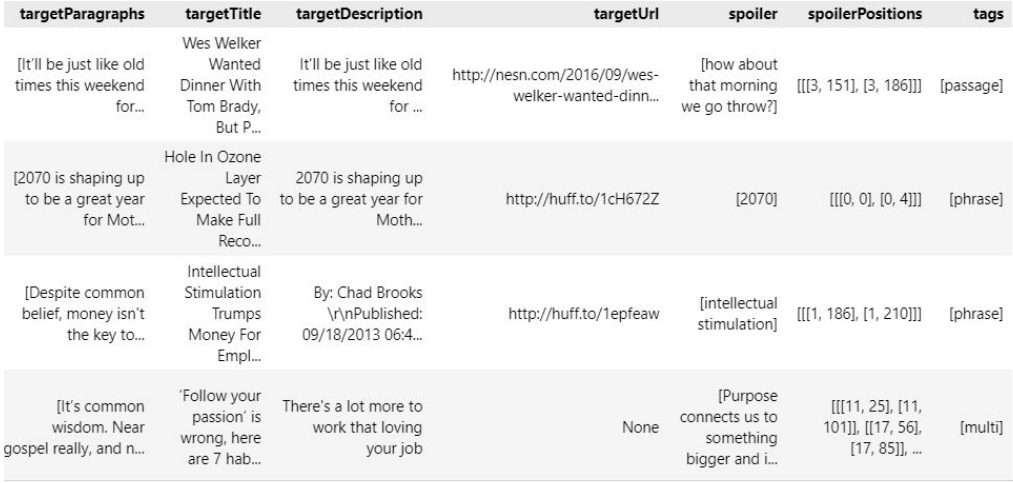
Data Set: Key Facts

Provided data sets:

- → *train.jsonl* with 3200 entries
- → *validation.jsonl* with 800 entries

Not provided:

- → *Test.jsonl* with 1000 entries



targetParagraphs	targetTitle	targetDescription	targetUrl	spoiler	spoilerPositions	tags
[It'll be just like old times this weekend for...	Wes Welker Wanted Dinner With Tom Brady, But P...	It'll be just like old times this weekend for ...	http://nesn.com/2016/09/wes-welker-wanted-dinn...	[how about that morning we go throw?]	[[[3, 151], [3, 186]]]	[passage]
[2070 is shaping up to be a great year for Mot...	Hole In Ozone Layer Expected To Make Full Reco...	2070 is shaping up to be a great year for Moth...	http://huff.to/1cH672Z	[2070]	[[[0, 0], [0, 4]]]	[phrase]
[Despite common belief, money isn't the key to...	Intellectual Stimulation Trumps Money For Empl...	By: Chad Brooks \r\nPublished: 09/18/2013 06:4...	http://huff.to/1epfeaw	[intellectual stimulation]	[[[1, 186], [1, 210]]]	[phrase]
[It's common wisdom. Near gospel really, and n...	'Follow your passion' is wrong, here are 7 hab...	There's a lot more to work that loving your job	None	[Purpose connects us to something bigger and i...	[[[11, 25], [11, 101]], [[17, 56], [17, 85]], ...	[multi]

Data Set: Key Facts

Three types of spoilers:

1. Phrase
   a. *E.g. Organisations, Persons , dates (single n-grams)*
2. Passage
3. Multi
   a. *Listing (Enumerations)*
   b. *Related informations*
   c. *Listing integrated in full text*

Data Set: Key Facts

Types of Fields

Description of field	Related fields
Identifiers	Uuid, postID
Source	postPlatform, targetMedia, targetUrl
context	postText, targetParagraphs, targetTitle, targetDescription, targetKeywords,
Task related field	Spoiler, tags

4. Task 1: Spoiler Type Classification

Task 1: Spoiler Type Classification

Spoiler Type Classification

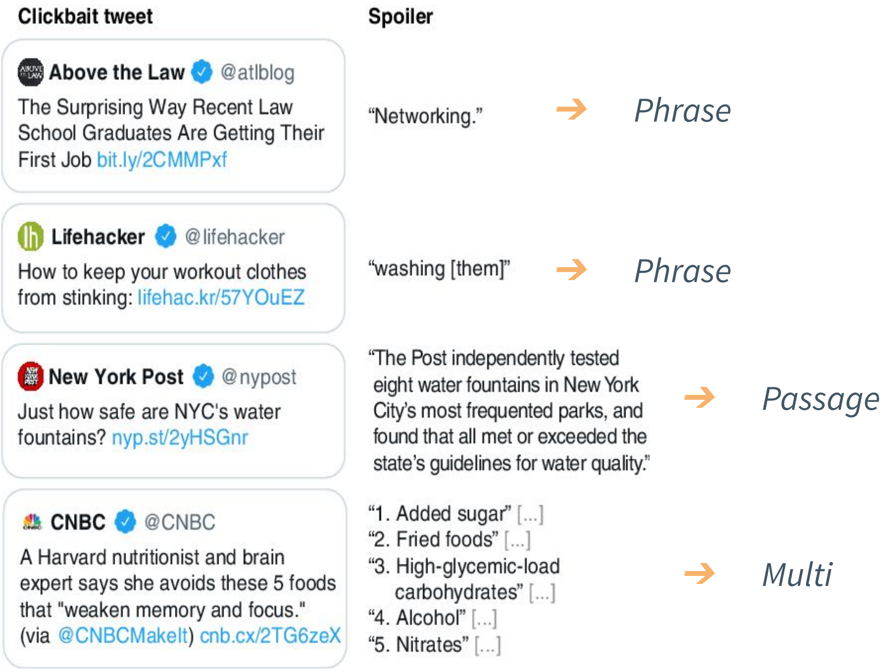
Classifying the *spoiler type* of the clickbait post in **three categories**

1. Phrase
2. Passage
3. Multi

➔ **Multi Class Classification**

Expected Output:

{"uuid": "<UUID>", "spoilerType": "<SPOILER-TYPE>"}



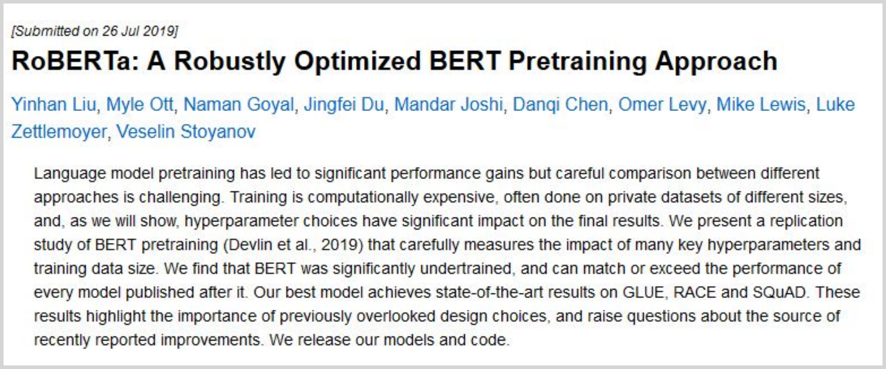
Clickbait tweet	Spoiler	
Above the Law @atlblog The Surprising Way Recent Law School Graduates Are Getting Their First Job bit.ly/2CMMPxf	"Networking."	➔ *Phrase*
Lifehacker @lifehacker How to keep your workout clothes from stinking: lifehac.kr/57YOuEZ	"washing [them]"	➔ *Phrase*
New York Post @nypost Just how safe are NYC's water fountains? nyp.st/2yHSGnr	"The Post independently tested eight water fountains in New York City's most frequented parks, and found that all met or exceeded the state's guidelines for water quality."	➔ *Passage*
CNBC @CNBC A Harvard nutritionist and brain expert says she avoids these 5 foods that "weaken memory and focus." (via @CNBCMakeIt) cnb.cx/2TG6zeX	"1. Added sugar" [...] "2. Fried foods" [...] "3. High-glycemic-load carbohydrates" [...] "4. Alcohol" [...] "5. Nitrates" [...]	➔ *Multi*

Spoiler Type Classification

Classifying clickbait posts into the categories: Phrase, Passage, Multi

Components

- **roBERTa** via simpletransformers
- **NER Recognition** with SpaCy
- **Input Reformulation**
- **Custom Metrics**



[Submitted on 26 Jul 2019]

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov

Language model pretraining has led to significant performance gains but careful comparison between different approaches is challenging. Training is computationally expensive, often done on private datasets of different sizes, and, as we will show, hyperparameter choices have significant impact on the final results. We present a replication study of BERT pretraining (Devlin et al., 2019) that carefully measures the impact of many key hyperparameters and training data size. We find that BERT was significantly undertrained, and can match or exceed the performance of every model published after it. Our best model achieves state-of-the-art results on GLUE, RACE and SQuAD. These results highlight the importance of previously overlooked design choices, and raise questions about the source of recently reported improvements. We release our models and code.

spaCy

https://spacy.io/

Simple Transformers

https://simpletransformers.ai/

Spoiler Type Classification

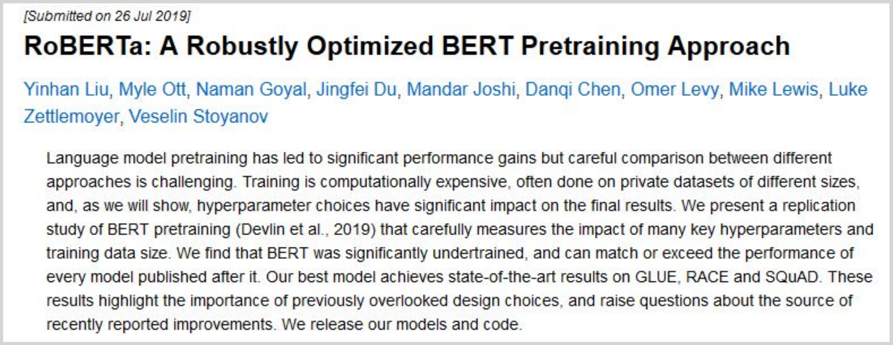
roBERTa

- → Adaptation of *BERT* and BERT’s language masking strategy
- → Modification on pre-training steps, masking and batch sizes

Trained on

A larger and **more task-relevant** union of data than BERT

→ Task 1 deals with social media and news posts

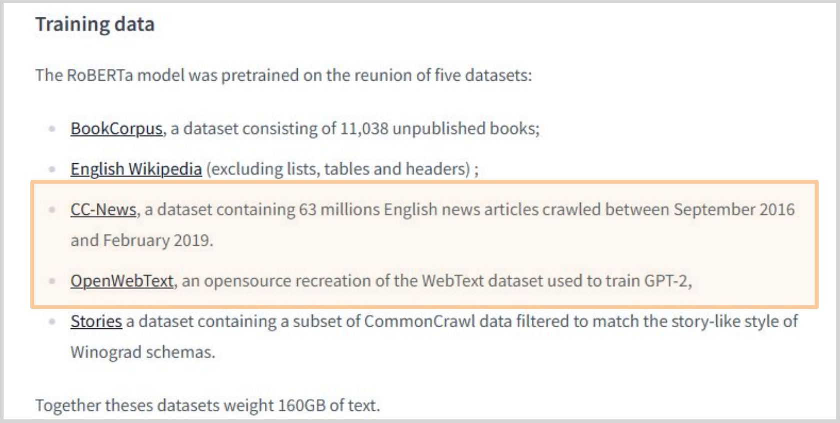


[Submitted on 26 Jul 2019]

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov

Language model pretraining has led to significant performance gains but careful comparison between different approaches is challenging. Training is computationally expensive, often done on private datasets of different sizes, and, as we will show, hyperparameter choices have significant impact on the final results. We present a replication study of BERT pretraining (Devlin et al., 2019) that carefully measures the impact of many key hyperparameters and training data size. We find that BERT was significantly undertrained, and can match or exceed the performance of every model published after it. Our best model achieves state-of-the-art results on GLUE, RACE and SQuAD. These results highlight the importance of previously overlooked design choices, and raise questions about the source of recently reported improvements. We release our models and code.



Training data

The RoBERTa model was pretrained on the reunion of five datasets:

- BookCorpus, a dataset consisting of 11,038 unpublished books;
- English Wikipedia (excluding lists, tables and headers) ;
- CC-News, a dataset containing 63 millions English news articles crawled between September 2016 and February 2019.
- OpenWebText, an opensource recreation of the WebText dataset used to train GPT-2,
- Stories a dataset containing a subset of CommonCrawl data filtered to match the story-like style of Winograd schemas.

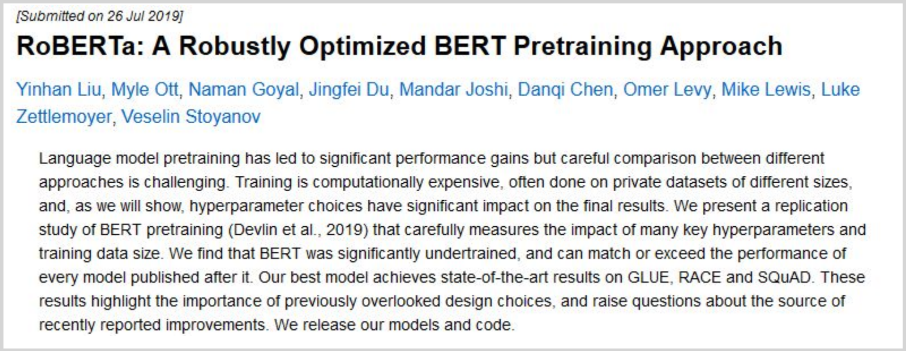
Together theses datasets weight 160GB of text.

https://huggingface.co/roberta-base

Spoiler Type Classification

roBERTa

- → Adaptation of *BERT* and BERT's language masking strategy
- → Modification on pre-training steps, masking and batch sizes



[Submitted on 26 Jul 2019]

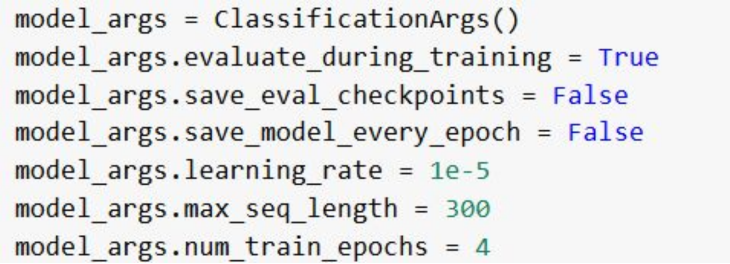
RoBERTa: A Robustly Optimized BERT Pretraining Approach

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov

Language model pretraining has led to significant performance gains but careful comparison between different approaches is challenging. Training is computationally expensive, often done on private datasets of different sizes, and, as we will show, hyperparameter choices have significant impact on the final results. We present a replication study of BERT pretraining (Devlin et al., 2019) that carefully measures the impact of many key hyperparameters and training data size. We find that BERT was significantly undertrained, and can match or exceed the performance of every model published after it. Our best model achieves state-of-the-art results on GLUE, RACE and SQuAD. These results highlight the importance of previously overlooked design choices, and raise questions about the source of recently reported improvements. We release our models and code.

Trained on

A larger and **more relevant** union of data than BERT



```
model_args = ClassificationArgs()
model_args.evaluate_during_training = True
model_args.save_eval_checkpoints = False
model_args.save_model_every_epoch = False
model_args.learning_rate = 1e-5
model_args.max_seq_length = 300
model_args.num_train_epochs = 4
```

→ Task 1 deals with social media and news posts

Custom Metrics

Spoiler-Title Ratio (st-r, range [0,1])

Inspects the Length of the title and the full article

- → Aims to identify entries where **passages** are likely *(st-r → low)*
- → Or **phrases** are likely *(st-r → high)*

Contains-Enumeration (c-e, [0,1])

Inspects the context of the entry for enumerations or lists

- → Aims to identify entries where **multi** is likely

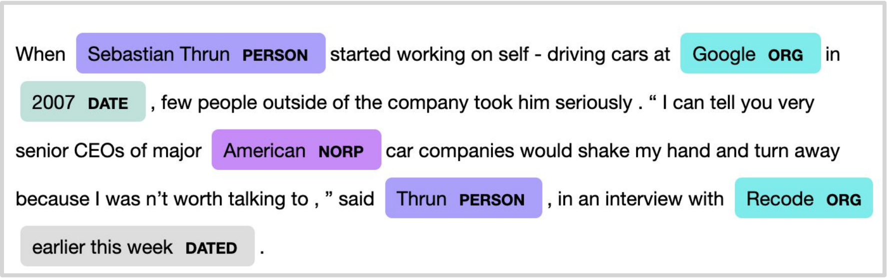
NER Recognition with SpaCy

Approach

Recognizing and emphasizing special

Entities and their Categories

- → Organisations
- → Persons
- → Dates
- → Locations



When Sebastian Thrun PERSON started working on self - driving cars at Google ORG in 2007 DATE , few people outside of the company took him seriously . " I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I was n't worth talking to , " said Thrun PERSON , in an interview with Recode ORG earlier this week DATED .

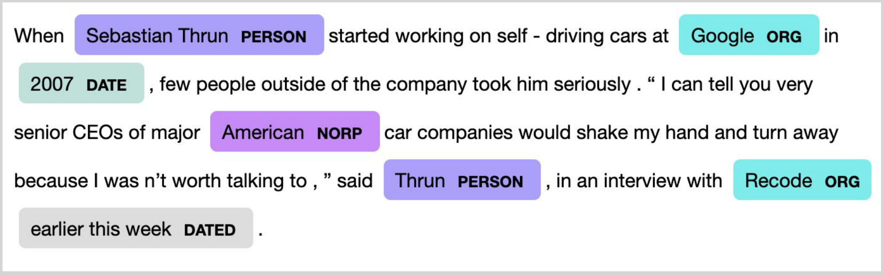
https://spacy.io/

SpaCy & NER Recognition

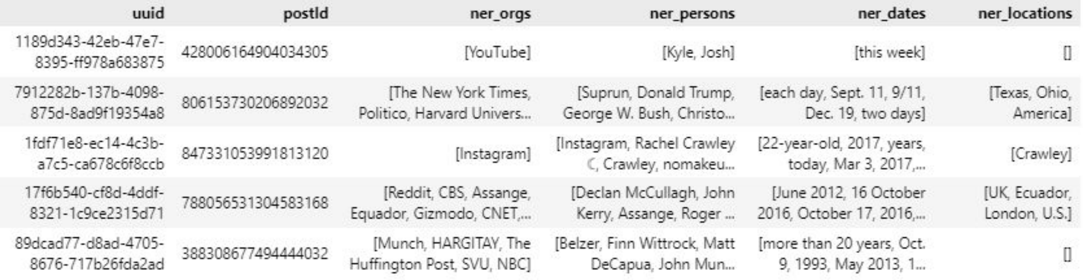
Approach

Recognizing and emphasizing special Entities and their Categories

- Organisations
- Persons
- Dates
- Locations



When Sebastian Thrun **PERSON** started working on self - driving cars at Google **ORG** in 2007 **DATE** , few people outside of the company took him seriously . “ I can tell you very senior CEOs of major American **NORP** car companies would shake my hand and turn away because I was n’t worth talking to , ” said Thrun **PERSON** , in an interview with Recode **ORG** earlier this week **DATED** .



uuid	postId	ner_orgs	ner_persons	ner_dates	ner_locations
1189d343-42eb-47e7-8395-ff978a683875	428006164904034305	[YouTube]	[Kyle, Josh]	[this week]	[]
7912282b-137b-4098-875d-8ad9f19354a8	806153730206892032	[The New York Times, Politico, Harvard Univers...	[Suprun, Donald Trump, George W. Bush, Christo...	[each day, Sept. 11, 9/11, Dec. 19, two days]	[Texas, Ohio, America]
1fdf71e8-ec14-4c3b-a7c5-ca678c6f8ccb	847331053991813120	[Instagram]	[Instagram, Rachel Crawley ☾, Crawley, nomakeu...	[22-year-old, 2017, years, today, Mar 3, 2017,...	[Crawley]
17f6b540-cf8d-4ddf-8321-1c9ce2315d71	788056531304583168	[Reddit, CBS, Assange, Equador, Gizmodo, CNET,...	[Declan McCullagh, John Kerry, Assange, Roger ...	[June 2012, 16 October 2016, October 17, 2016,...	[UK, Ecuador, London, U.S.]
89dcad77-d8ad-4705-8676-717b26fda2ad	388308677494444032	[Munch, HARGITAY, The Huffington Post, SVU, NBC]	[Belzer, Finn Wittrock, Matt DeCapua, John Mun...	[more than 20 years, Oct. 9, 1993, May 2013, 1...	[]

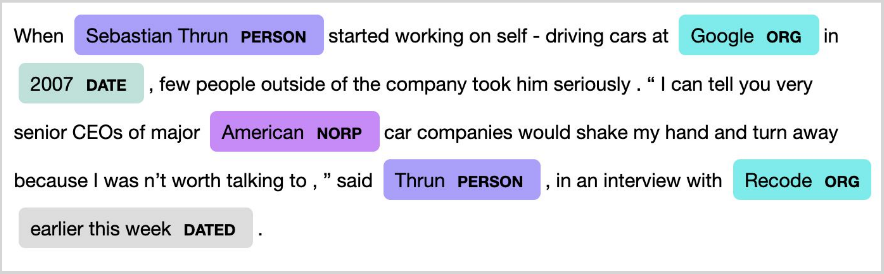
→ extracted from the **whole context** of the post

SpaCy & NER Recognition

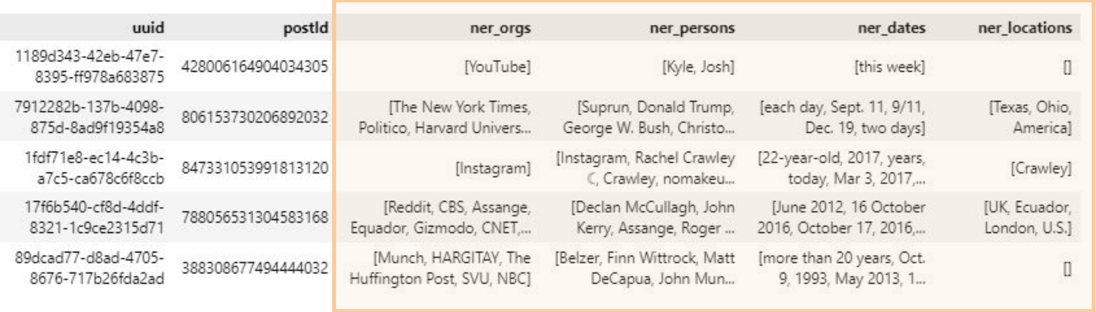
Approach

Recognizing and emphasizing special Entities and their Categories

- Organisations
- Persons
- Dates
- Locations



When Sebastian Thrun PERSON started working on self - driving cars at Google ORG in 2007 DATE , few people outside of the company took him seriously . " I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I was n't worth talking to , " said Thrun PERSON , in an interview with Recode ORG earlier this week DATED .



uuid	postId	ner_orgs	ner_persons	ner_dates	ner_locations
1189d343-42eb-47e7-8395-ff978a683875	428006164904034305	[YouTube]	[Kyle, Josh]	[this week]	[]
7912282b-137b-4098-875d-8ad9f19354a8	806153730206892032	[The New York Times, Politico, Harvard Univers...	[Suprun, Donald Trump, George W. Bush, Christo...	[each day, Sept. 11, 9/11, Dec. 19, two days]	[Texas, Ohio, America]
1fdf71e8-ec14-4c3b-a7c5-ca678c6f8ccb	847331053991813120	[Instagram]	[Instagram, Rachel Crawley ☾, Crawley, nomakeu...	[22-year-old, 2017, years, today, Mar 3, 2017,...	[Crawley]
17f6b540-cf8d-4ddf-8321-1c9ce2315d71	788056531304583168	[Reddit, CBS, Assange, Equador, Gizmodo, CNET,...	[Declan McCullagh, John Kerry, Assange, Roger ...	[June 2012, 16 October 2016, October 17, 2016,...	[UK, Ecuador, London, U.S.]
89dcad77-d8ad-4705-8676-717b26fda2ad	388308677494444032	[Munch, HARGITAY, The Huffington Post, SVU, NBC]	[Belzer, Finn Wittrock, Matt DeCapua, John Mun...	[more than 20 years, Oct. 9, 1993, May 2013, 1...	[]

Motivation

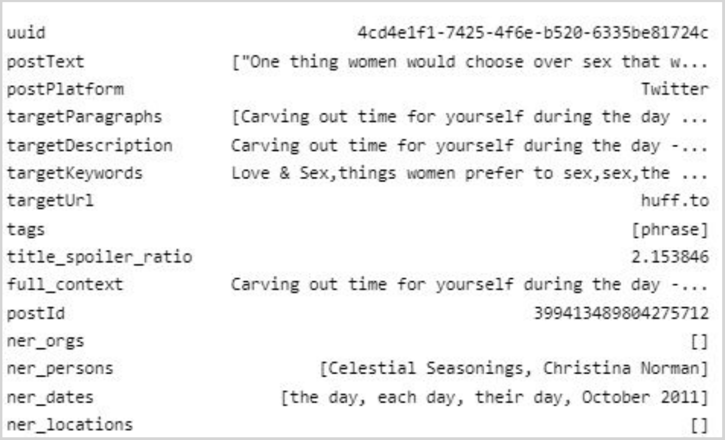
Allow the LM to recognize the feature difference between normal text and special entities

→ extracted from the **whole context** of the post

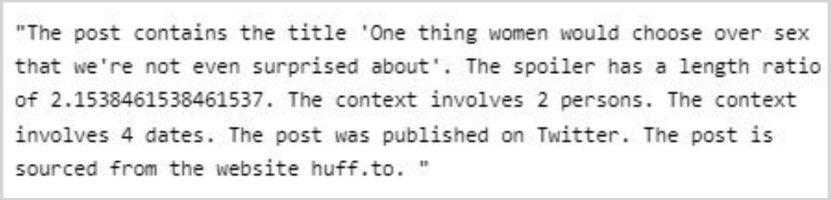
Input Reformulation (Long)

Approach

Transform the data into natural language that is parseable for **roBERTa**



```
uuid                            4cd4e1f1-7425-4f6e-b520-6335be81724c
postText               ["One thing women would choose over sex that w...
postPlatform                                                  Twitter
targetParagraphs       [Carving out time for yourself during the day ...
targetDescription      Carving out time for yourself during the day -...
targetKeywords         Love & Sex,things women prefer to sex,sex,the ...
targetUrl                                                     huff.to
tags                                                         [phrase]
title_spoiler_ratio                                          2.153846
full_context           Carving out time for yourself during the day -...
postId                                             399413489804275712
ner_orgs                                                           []
ner_persons                     [Celestial Seasonings, Christina Norman]
ner_dates                   [the day, each day, their day, October 2011]
ner_locations                                                      []
```

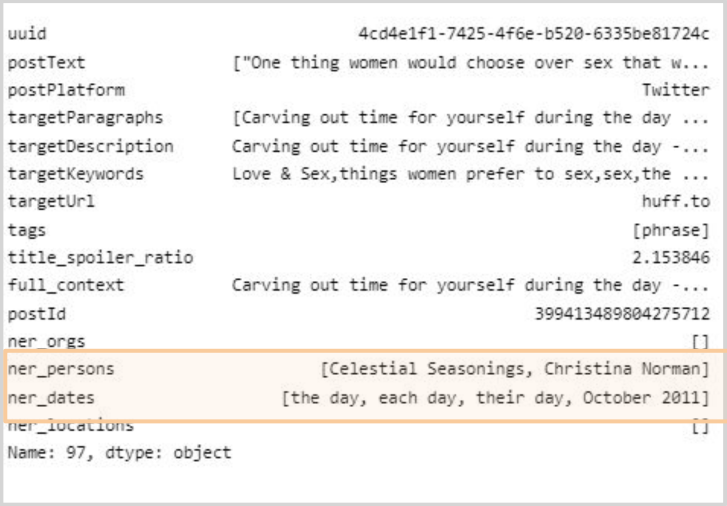


```
"The post contains the title 'One thing women would choose over sex
that we're not even surprised about'. The spoiler has a length ratio
of 2.1538461538461537. The context involves 2 persons. The context
involves 4 dates. The post was published on Twitter. The post is
sourced from the website huff.to. "
```

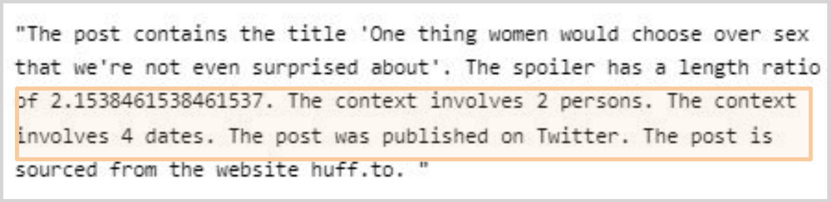
Input Reformulation (Long)

Approach

Transform the data into natural language that is parseable for **roBERTa**



```
uuid                             4cd4e1f1-7425-4f6e-b520-6335be81724c
postText               ["One thing women would choose over sex that w...
postPlatform                                                   Twitter
targetParagraphs       [Carving out time for yourself during the day ...
targetDescription      Carving out time for yourself during the day -...
targetKeywords         Love & Sex,things women prefer to sex,sex,the ...
targetUrl                                                      huff.to
tags                                                          [phrase]
title_spoiler_ratio                                           2.153846
full_context           Carving out time for yourself during the day -...
postId                                              399413489804275712
ner_orgs                                                            []
ner_persons                     [Celestial Seasonings, Christina Norman]
ner_dates                   [the day, each day, their day, October 2011]
ner_locations                                                       []
Name: 97, dtype: object
```

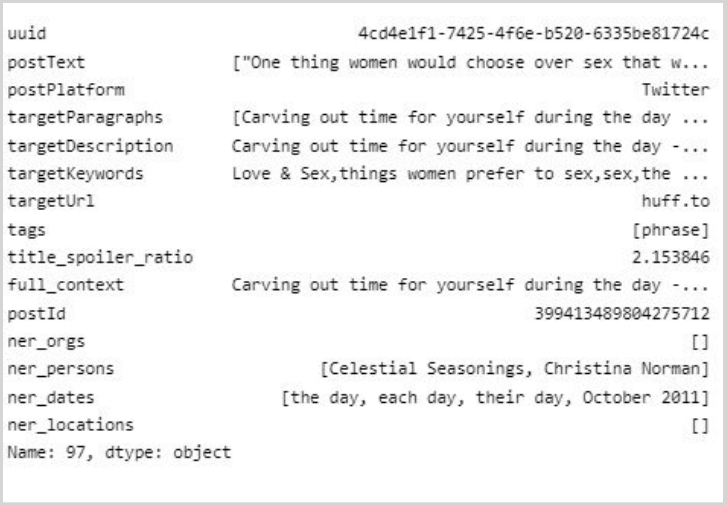


```
"The post contains the title 'One thing women would choose over sex
that we're not even surprised about'. The spoiler has a length ratio
of 2.1538461538461537. The context involves 2 persons. The context
involves 4 dates. The post was published on Twitter. The post is
sourced from the website huff.to. "
```

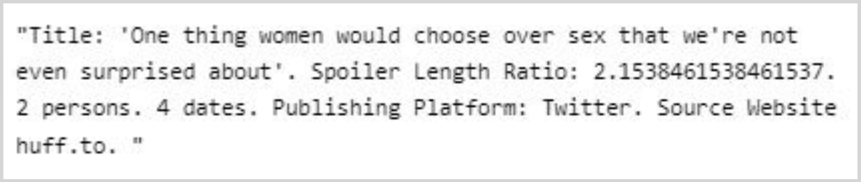
Input Reformulation (Short)

Approach

Transform the data into short language that is parseable for **roBERTa**



```
uuid                             4cd4e1f1-7425-4f6e-b520-6335be81724c
postText             ["One thing women would choose over sex that w...
postPlatform                                                   Twitter
targetParagraphs     [Carving out time for yourself during the day ...
targetDescription    Carving out time for yourself during the day -...
targetKeywords       Love & Sex,things women prefer to sex,sex,the ...
targetUrl                                                      huff.to
tags                                                          [phrase]
title_spoiler_ratio                                           2.153846
full_context         Carving out time for yourself during the day -...
postId                                              399413489804275712
ner_orgs                                                            []
ner_persons                      [Celestial Seasonings, Christina Norman]
ner_dates                   [the day, each day, their day, October 2011]
ner_locations                                                       []
Name: 97, dtype: object
```



```
"Title: 'One thing women would choose over sex that we're not
even surprised about'. Spoiler Length Ratio: 2.1538461538461537.
2 persons. 4 dates. Publishing Platform: Twitter. Source Website
huff.to. "
```

Results on the Validation Set

Submission to tira.io

Against the official **validation data set**

Model	**Balanced Accuracy (in %)**
Naive (Baseline)	33.3
Transformer (Baseline)	73.4
roBERTa with NER	58.87

→ Outperforms Naive,
Outperformed by Transformer Baseline

Results on the Test Set

Submission to tira.io

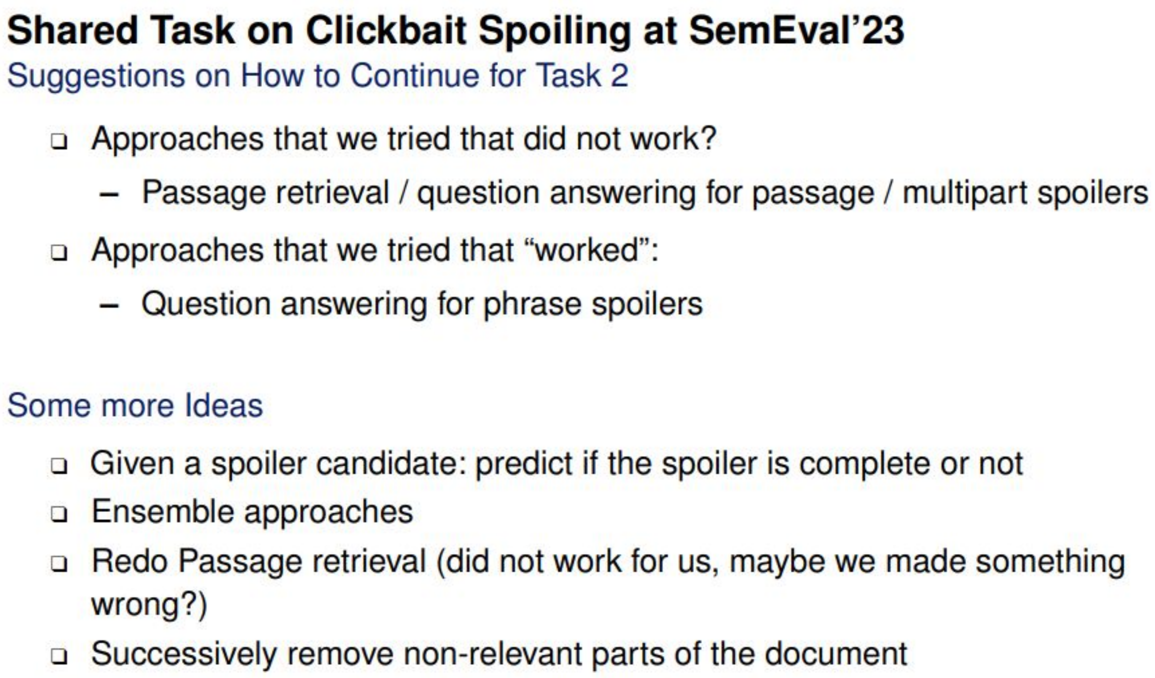
Against the official **test data set**

Model	**Balanced Accuracy (in %)**
roBERTa with NER	59.36

5. Task 2: Spoiler Generation

Via Question Answering

Problem:



Shared Task on Clickbait Spoiling at SemEval’23

Suggestions on How to Continue for Task 2

- ❑ Approaches that we tried that did not work?
  - – Passage retrieval / question answering for passage / multipart spoilers
- ❑ Approaches that we tried that “worked”:
  - – Question answering for phrase spoilers

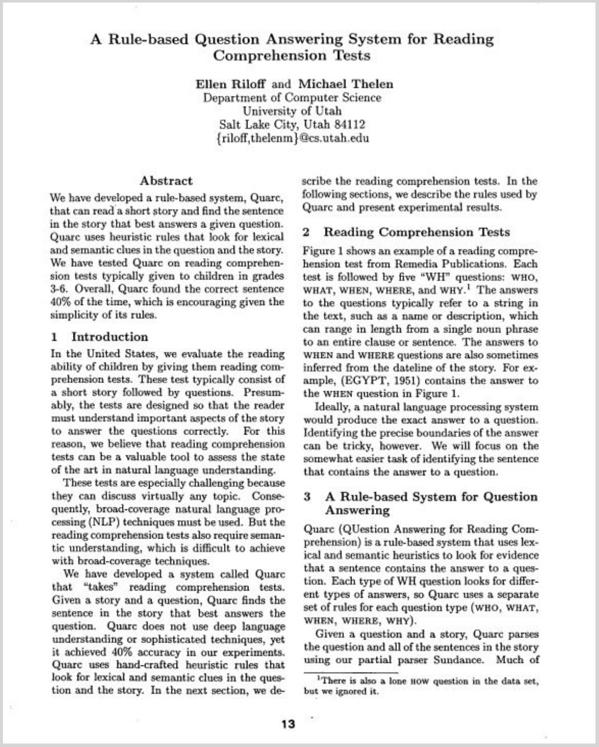
Some more Ideas

- ❑ Given a spoiler candidate: predict if the spoiler is complete or not
- ❑ Ensemble approaches
- ❑ Redo Passage retrieval (did not work for us, maybe we made something wrong?)
- ❑ Successively remove non-relevant parts of the document

First Idea: Rule Based Approach

Inspired by Quarc

- Developed in the year 2000
- Uses NER and Pattern Matching
- Goal: Identifying the context of a sentence by Wh-rules



A Rule-based Question Answering System for Reading Comprehension Tests

Ellen Riloff and Michael Thelen
Department of Computer Science
University of Utah
Salt Lake City, Utah 84112
{riloff,thelenm}@cs.utah.edu

Abstract

We have developed a rule-based system, Quarc, that can read a short story and find the sentence in the story that best answers a given question. Quarc uses heuristic rules that look for lexical and semantic clues in the question and the story. We have tested Quarc on reading comprehension tests typically given to children in grades 3-6. Overall, Quarc found the correct sentence 40% of the time, which is encouraging given the simplicity of its rules.

1 Introduction

In the United States, we evaluate the reading ability of children by giving them reading comprehension tests. These test typically consist of a short story followed by questions. Presumably, the tests are designed so that the reader must understand important aspects of the story to answer the questions correctly. For this reason, we believe that reading comprehension tests can be a valuable tool to assess the state of the art in natural language understanding.

These tests are especially challenging because they can discuss virtually any topic. Consequently, broad-coverage natural language processing (NLP) techniques must be used. But the reading comprehension tests also require semantic understanding, which is difficult to achieve with broad-coverage techniques.

We have developed a system called Quarc that "takes" reading comprehension tests. Given a story and a question, Quarc finds the sentence in the story that best answers the question. Quarc does not use deep language understanding or sophisticated techniques, yet it achieved 40% accuracy in our experiments. Quarc uses hand-crafted heuristic rules that look for lexical and semantic clues in the question and the story. In the next section, we describe the reading comprehension tests. In the following sections, we describe the rules used by Quarc and present experimental results.

2 Reading Comprehension Tests

Figure 1 shows an example of a reading comprehension test from Remedia Publications. Each test is followed by five "WH" questions: WHO, WHAT, WHEN, WHERE, and WHY.[1] The answers to the questions typically refer to a string in the text, such as a name or description, which can range in length from a single noun phrase to an entire clause or sentence. The answers to WHEN and WHERE questions are also sometimes inferred from the dateline of the story. For example, (EGYPT, 1951) contains the answer to the WHEN question in Figure 1.

Ideally, a natural language processing system would produce the exact answer to a question. Identifying the precise boundaries of the answer can be tricky, however. We will focus on the somewhat easier task of identifying the sentence that contains the answer to a question.

3 A Rule-based System for Question Answering

Quarc (QUestion Answering for Reading Comprehension) is a rule-based system that uses lexical and semantic heuristics to look for evidence that a sentence contains the answer to a question. Each type of WH question looks for different types of answers, so Quarc uses a separate set of rules for each question type (WHO, WHAT, WHEN, WHERE, WHY).

Given a question and a story, Quarc parses the question and all of the sentences in the story using our partial parser Sundance. Much of

[1] There is also a lone HOW question in the data set, but we ignored it.

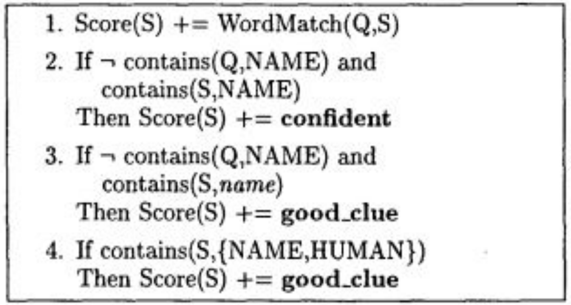
13

Source: https://aclanthology.org/W00-0603.pdf

First Idea: Rule Based Approach

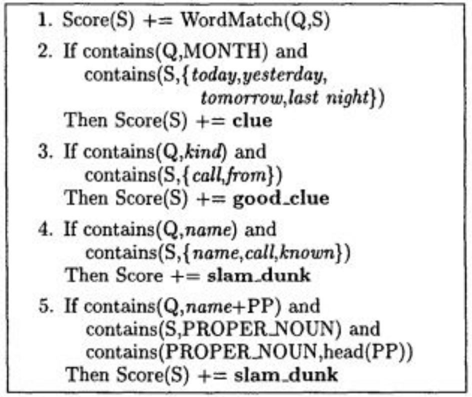
Calculating Scores:

- Clue **(+ 3)**
- Good_clue **(+ 4)**
- Confident **(+ 6)**
- slam_dunk **(+ 20)**



```
1. Score(S) += WordMatch(Q,S)
2. If ¬ contains(Q,NAME) and
      contains(S,NAME)
   Then Score(S) += confident
3. If ¬ contains(Q,NAME) and
      contains(S,name)
   Then Score(S) += good_clue
4. If contains(S,{NAME,HUMAN})
   Then Score(S) += good_clue
```

Figure 2: WHO Rules



```
1. Score(S) += WordMatch(Q,S)
2. If contains(Q,MONTH) and
      contains(S,{today,yesterday,
                  tomorrow,last night})
   Then Score(S) += clue
3. If contains(Q,kind) and
      contains(S,{call,from})
   Then Score(S) += good_clue
4. If contains(Q,name) and
      contains(S,{name,call,known})
   Then Score += slam_dunk
5. If contains(Q,name+PP) and
      contains(S,PROPER_NOUN) and
      contains(PROPER_NOUN,head(PP))
   Then Score(S) += slam_dunk
```

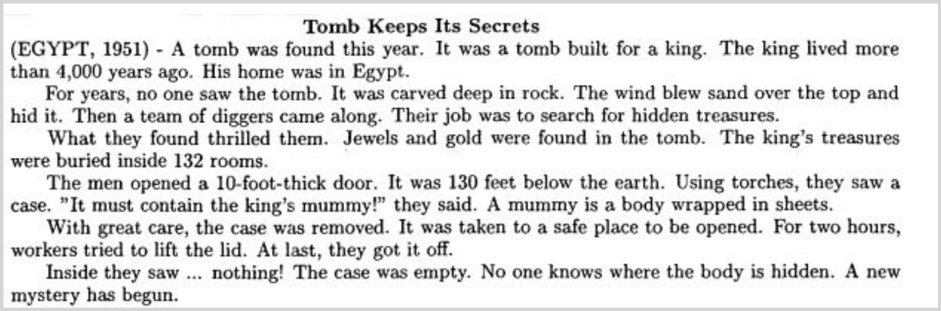
Figure 3: WHAT Rules

Source: https://aclanthology.org/W00-0603.pdf

First Idea: Rule Based Approach

Problem:

- Built for very simple texts



Tomb Keeps Its Secrets

(EGYPT, 1951) - A tomb was found this year. It was a tomb built for a king. The king lived more than 4,000 years ago. His home was in Egypt.

For years, no one saw the tomb. It was carved deep in rock. The wind blew sand over the top and hid it. Then a team of diggers came along. Their job was to search for hidden treasures.

What they found thrilled them. Jewels and gold were found in the tomb. The king's treasures were buried inside 132 rooms.

The men opened a 10-foot-thick door. It was 130 feet below the earth. Using torches, they saw a case. "It must contain the king's mummy!" they said. A mummy is a body wrapped in sheets.

With great care, the case was removed. It was taken to a safe place to be opened. For two hours, workers tried to lift the lid. At last, they got it off.

Inside they saw ... nothing! The case was empty. No one knows where the body is hidden. A new mystery has begun.

Source: https://aclanthology.org/W00-0603.pdf

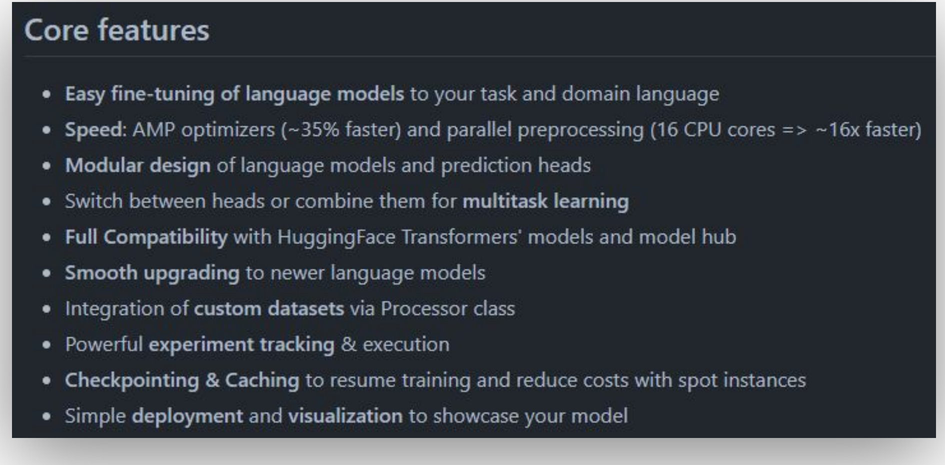
- Identifying **Wh-Questions** for actual questions

→ Already low accuracy: 40 %

Transformer Model:

Usage of **FARM** library (*deepset-ai*)

- Based on torch and transformers



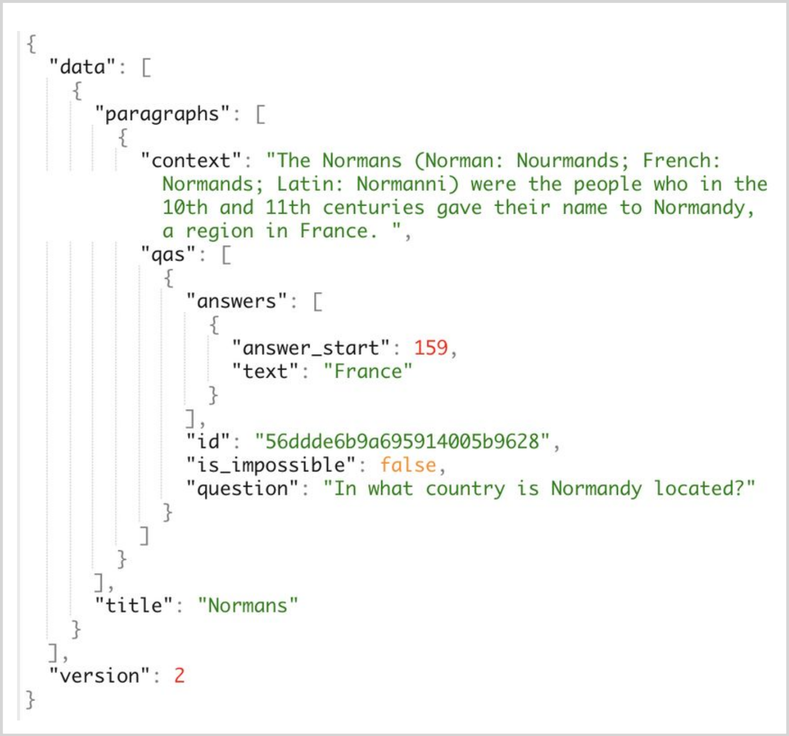
Core features

- **Easy fine-tuning of language models** to your task and domain language
- **Speed**: AMP optimizers (~35% faster) and parallel preprocessing (16 CPU cores => ~16x faster)
- **Modular design** of language models and prediction heads
- Switch between heads or combine them for **multitask learning**
- **Full Compatibility** with HuggingFace Transformers' models and model hub
- **Smooth upgrading** to newer language models
- Integration of **custom datasets** via Processor class
- Powerful **experiment tracking** & execution
- **Checkpointing & Caching** to resume training and reduce costs with spot instances
- Simple **deployment** and **visualization** to showcase your model

Source: https://github.com/deepset-ai/FARM

Transformer Model: Preprocessing

1. **Reformat** files into *Squad2.0 format*
2. **Exclude** abstractive spoilers
3. **Tokenization** through transformer model
4. **Create** case sensitive tokens (no lowercasing)

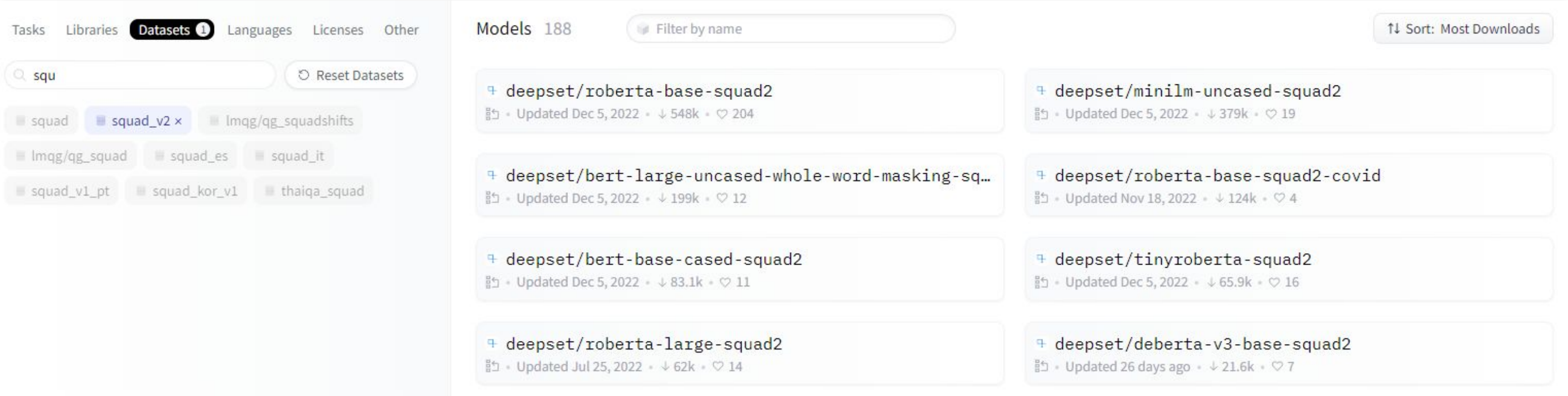


```
{
  "data": [
    {
      "paragraphs": [
        {
          "context": "The Normans (Norman: Nourmands; French:
            Normands; Latin: Normanni) were the people who in the
            10th and 11th centuries gave their name to Normandy,
            a region in France. ",
          "qas": [
            {
              "answers": [
                {
                  "answer_start": 159,
                  "text": "France"
                }
              ],
              "id": "56ddde6b9a695914005b9628",
              "is_impossible": false,
              "question": "In what country is Normandy located?"
            }
          ]
        }
      ],
      "title": "Normans"
    }
  ],
  "version": 2
}
```

Source: https://www.researchgate.net/figure/An-example-of-the-SQuAD-dataset-Each-dataset-contains-an-array-of-instances-each_fig2_341852018

Transformer Model: Hyperparameter

Used language model: *roberta-based-squad2*



deepset/roberta-base-squad2
Updated Dec 5, 2022 • ↓ 548k • ♡ 204

deepset/minilm-uncased-squad2
Updated Dec 5, 2022 • ↓ 379k • ♡ 19

deepset/bert-large-uncased-whole-word-masking-sq…
Updated Dec 5, 2022 • ↓ 199k • ♡ 12

deepset/roberta-base-squad2-covid
Updated Nov 18, 2022 • ↓ 124k • ♡ 4

deepset/bert-base-cased-squad2
Updated Dec 5, 2022 • ↓ 83.1k • ♡ 11

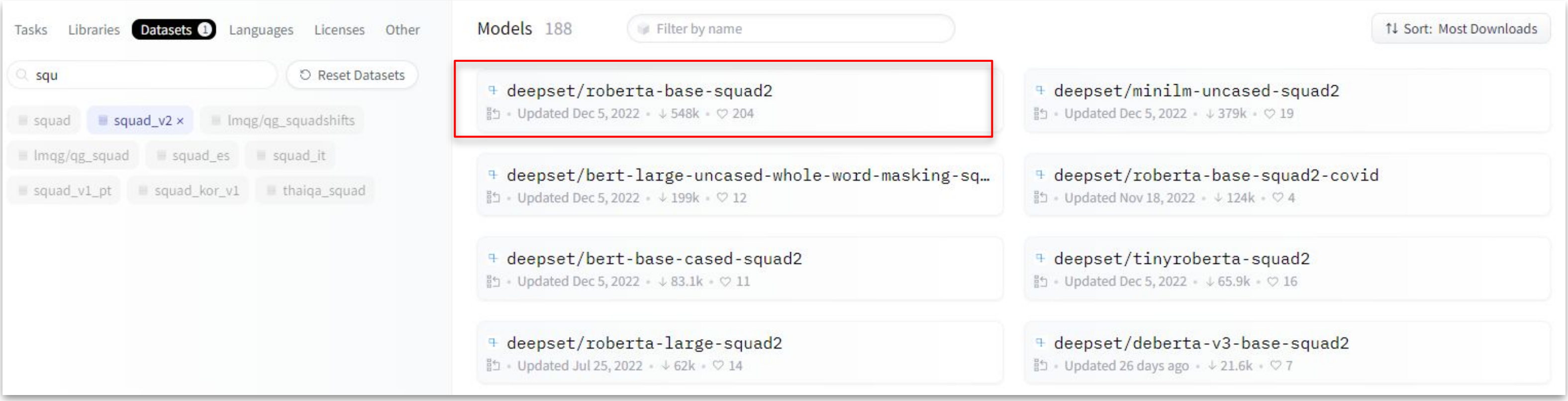
deepset/tinyroberta-squad2
Updated Dec 5, 2022 • ↓ 65.9k • ♡ 16

deepset/roberta-large-squad2
Updated Jul 25, 2022 • ↓ 62k • ♡ 14

deepset/deberta-v3-base-squad2
Updated 26 days ago • ↓ 21.6k • ♡ 7

Transformer Model: Hyperparameter

Used language model: *roberta-based-squad2*



squ

squad · squad_v2 × · lmqg/qg_squadshifts · lmqg/qg_squad · squad_es · squad_it · squad_v1_pt · squad_kor_v1 · thaiqa_squad

Models 188

deepset/roberta-base-squad2
Updated Dec 5, 2022 • 548k • 204

deepset/minilm-uncased-squad2
Updated Dec 5, 2022 • 379k • 19

deepset/bert-large-uncased-whole-word-masking-sq…
Updated Dec 5, 2022 • 199k • 12

deepset/roberta-base-squad2-covid
Updated Nov 18, 2022 • 124k • 4

deepset/bert-base-cased-squad2
Updated Dec 5, 2022 • 83.1k • 11

deepset/tinyroberta-squad2
Updated Dec 5, 2022 • 65.9k • 16

deepset/roberta-large-squad2
Updated Jul 25, 2022 • 62k • 14

deepset/deberta-v3-base-squad2
Updated 26 days ago • 21.6k • 7

Transformer Model: Hyperparameter

Hyperparameter	Value
language model	model: roberta-based-squad2
Batch Size	24
N-epochs	5
Max_seq_len	384
Doc_stride	192
Embeds_dropout_prob	0.1
Learning_rate	3e-5
Schedule_opts	{LinearWarmup, 0.2}

Transformer Model: Preprocessing

Biggest threat: Multi Spoiler

→ Model mostly suggests just one of many

What kind of multi spoilers are there?

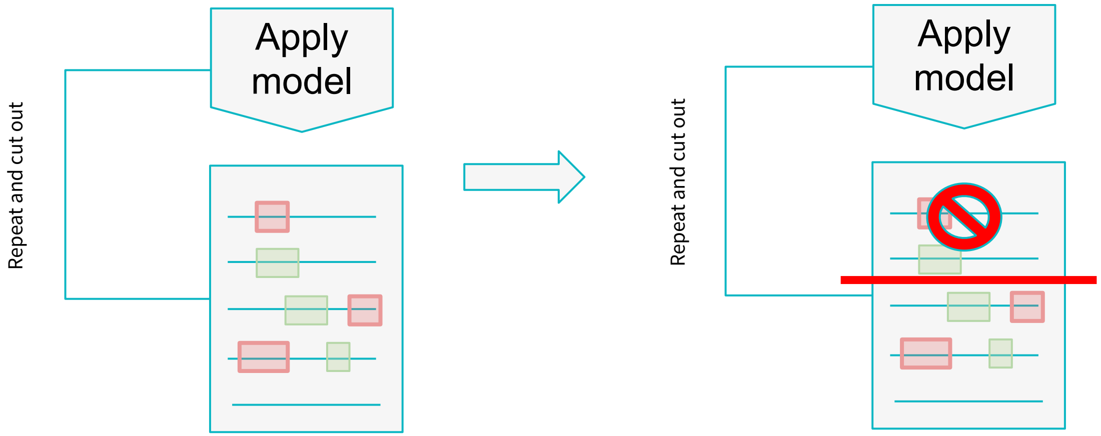
→ Enumerations (*with listing*)

→ Enumerations in the text (*e.g. multiple tips*)

→ Multiple related informations

Reiterating and reducing Context

Reiterating for Multi Part Spoiler

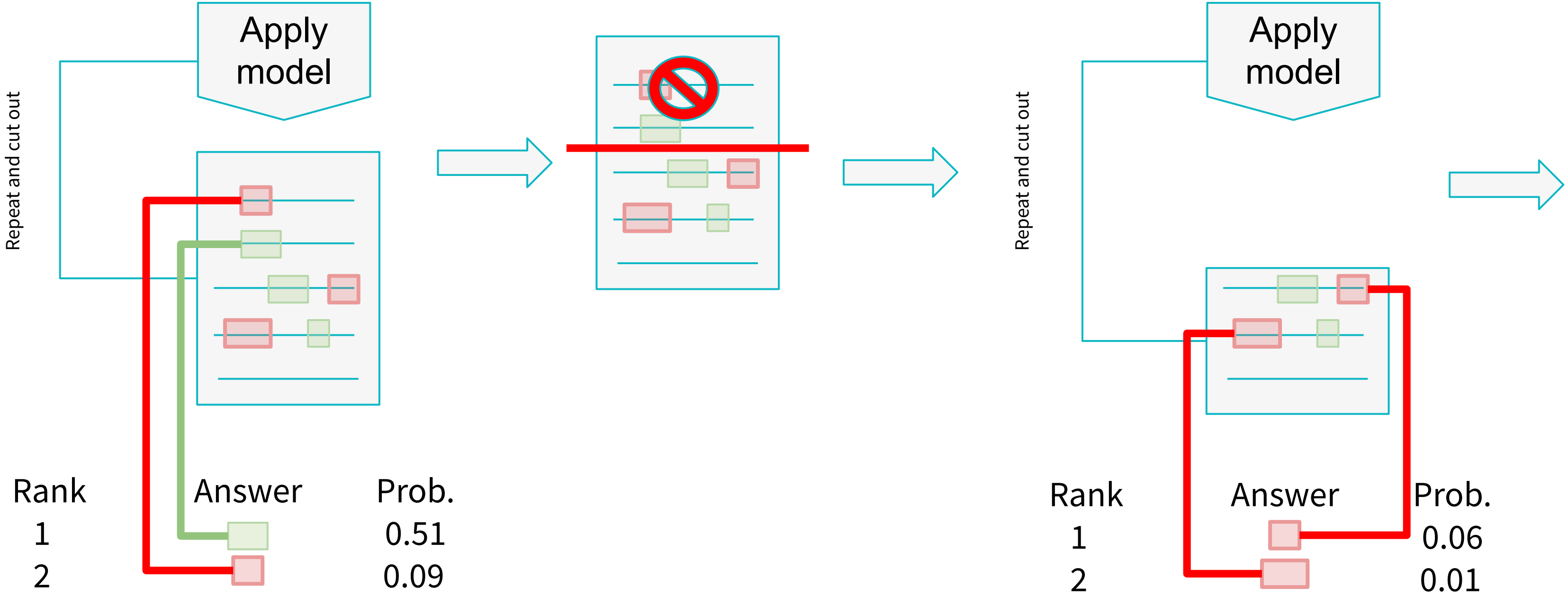


Rank	Answer	Prob.
1	🟩	0.51
2	🟥	0.09

Rank	Answer	Prob.
1	🟥	0.06
2	🟥	0.01

Reiterating and reducing Context

Reiterating for Multi Part Spoiler



Rank	Answer	Prob.
1	🟩	0.51
2	🟥	0.09

Rank	Answer	Prob.
1	🟥	0.06
2	🟥	0.01

Transformer Model: Additional Rules

IDEA

Identifying multi spoiler with **model and rule set**

→ Extract enumerations via regex

Transformer Model: Additional Rules

Biggest threat: Multi Spoiler

- Model mostly suggests just one of many

How to identify multi spoiler?

- Manually analyse questions and context

Pattern	Matches	Correct matches
r”^\d”	141	119
How to …	30	12
There are …	8	8
r"[\.\?\!\s\d\s]"	145	121
r".*these \d"	10	10
"need to know"	8	7

Transformer Model: Additional Rules

Implementation

Apply and select patterns on postText

Catches: 11 Simple Weight Loss Strategies For Fruitful Results

Apply additional pattern on context

➔ Pattern: ".*\d+\s*[\.\)].+\d+?\s*[\.\)].+?\d+\s*[\.\)]"

Catches: 1. Cut Out Fizzy Drinks [...] , and it’s much harder to find yourself snacking guiltily on them! 2. Have 5 Small Meals a Day [...] filling up on snacks! 3. Eat Breakfast [...]

Transformer Model: Additional Rules

Extract targetParagraphs fitting pattern

- [1-9]\d{0,1}\s*[\.\)]\s.+

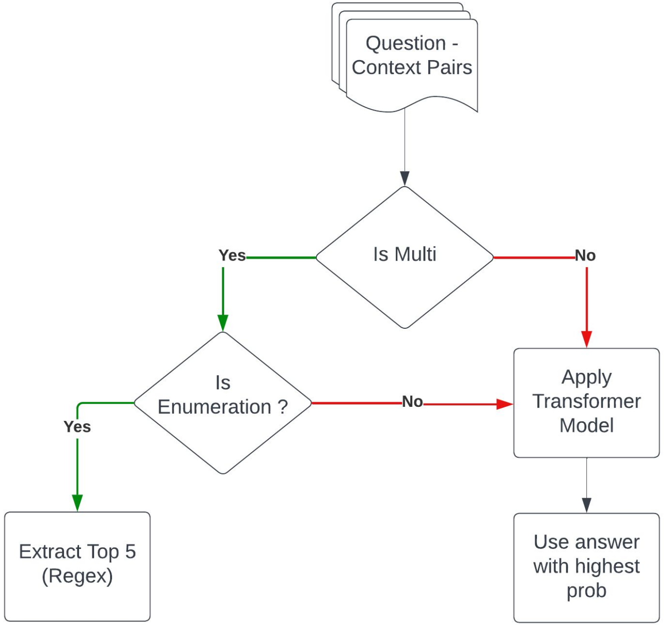
When first string in result list starts with 1

- Consider first 5 values of list as spoiler

Otherwise reverse list

- Consider first 5 values of list as spoiler

Second Idea: Transformer Model + Additional Rules



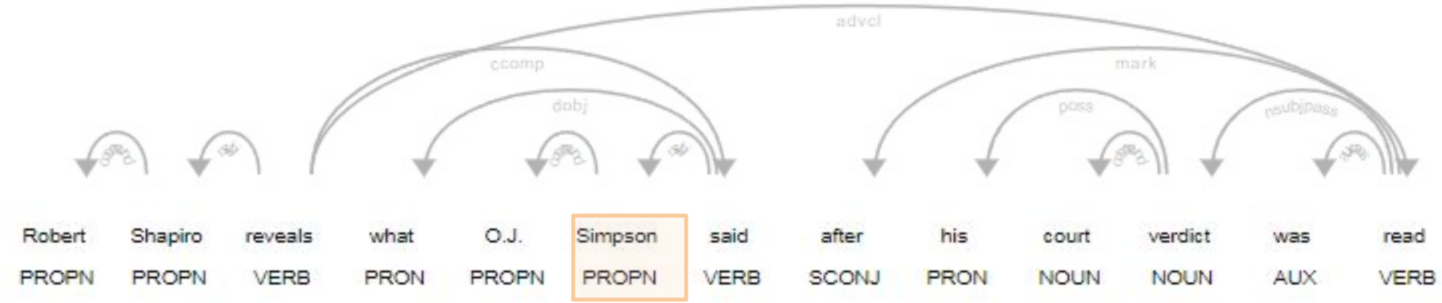
Other (withdrawn) Ideas

Make use of semantic and syntactic patterns with

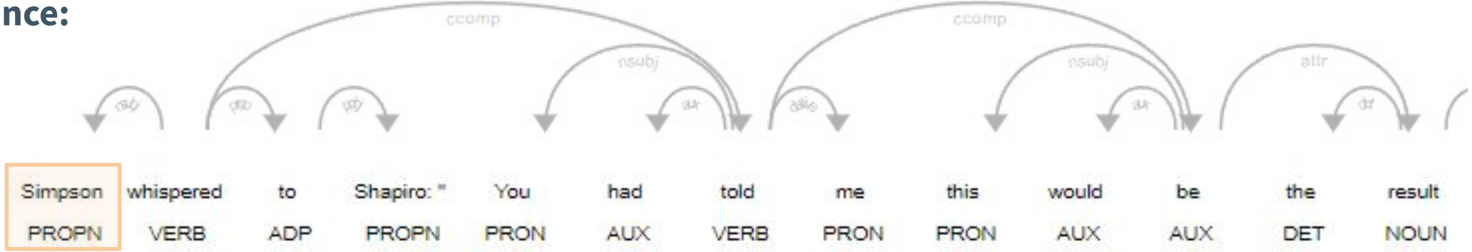
- Spacy Entity Recognition
- Spacy Part of Speech Tagging

Other (withdrawn) Ideas

Searching for last proper noun/entity in postText:



Searching for THIS proper noun/entity in the beginning of a sentence:



Performance on the Validation Set

Submission to tira.io

Against the official **validation data set**

Model	BLEU Score (in %)
Naive (Baseline)	0.021
Transformer (Baseline)	0.382
roBERTa-sQuad v1	0.3171
roBERTa-sQuad v2	0.3258

Performance on the Test Set

Submission to tira.io

Against the official **test data set**

Model	BLEU Score
roBERTa-sQuad v1	0.307
roBERTa-sQuad v2	0.322

6. Conclusion

Conclusion

For Task 1

→ We created a roBERTa Model that is embellished by extra features such as **NER entities**, **spoiler-title ratio** and **natural language** as input

For Task 2

→ Explored **rule-based approaches** for Question Answering

→ Created a **roBERTa-SQuAD2.0** model embellished by regex

Contributions

Anh Huy Tran:

Research, Preprocessing, Task 1: Spoiler Classification, Dockerisation, Testing & Docker Image Submission on Tira

Andreas Kruff:

Research, Preprocessing, SpaCy NER & Enumeration recognition, Task 2: Spoiler Generation,

References

[1] *Clickbait Challenge at SemEval 2023 - Clickbait Spoiling*. (2022). Webis. Retrieved January 4, 2023, from https://pan.webis.de/semeval23/pan23-web/clickbait-challenge.html

[2] *EntityRecognizer · spaCy API Documentation*. (2023). EntityRecognizer. Retrieved January 4, 2023, from https://spacy.io/api/entityrecognizer

[3] *FARM*. (2022, January). deepset.ai. Retrieved January 6, 2023, from https://farm.deepset.ai

[4] Liu, Y. (2019, 26. Juli). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv.org. https://arxiv.org/abs/1907.11692

[4] Riloff, E. (2000, Mai). *A rule-based question answering system for reading comprehension tests*.

The Stanford Question Answering Dataset. (2023). https://rajpurkar.github.io/SQuAD-explorer/

[5] *simpletransformers*. (2023, January). simpletransformers.ai. Retrieved January 6, 2023, from https://simpletransformers.ai/

THANKS FOR LISTENING!

Any questions?