SemEval-2020 Task 7: Assessing Humor in Edited News Headlines

Nabil Hossain[†], **John Krumm**[‡], **Michael Gamon**[‡] and **Henry Kautz**[†]

[†]Department of Computer Science, University of Rochester
[‡]Microsoft Research AI, Microsoft Corporation, Redmond, WA
{nhossain,kautz}@cs.rochester.edu, {jckrumm,mgamon}@microsoft.com

Task: https://competitions.codalab.org/competitions/20970

Nabil Hossain
Intel Corporation
nabil.hossain@intel.com

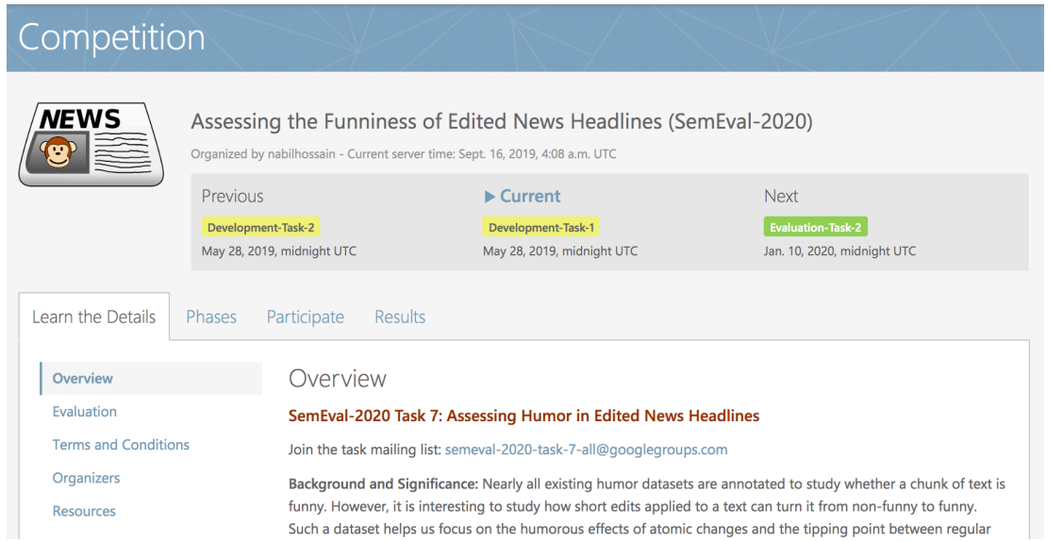
From left: John, Nabil, Henry. Learning to surf.

Introduction

- Computational Humor has been slow
- Need unified, shared tasks and datasets

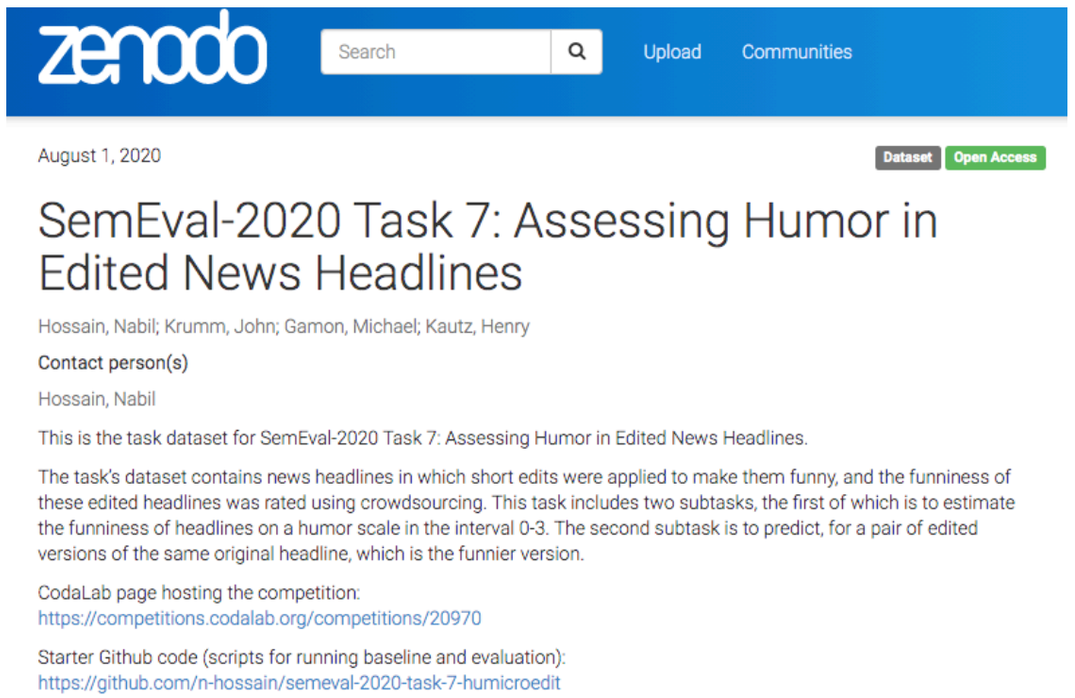
Introduction

- Computational Humor has been slow
- Need unified, shared tasks and datasets
- Our task:
  - Assessing humor in edited news headlines



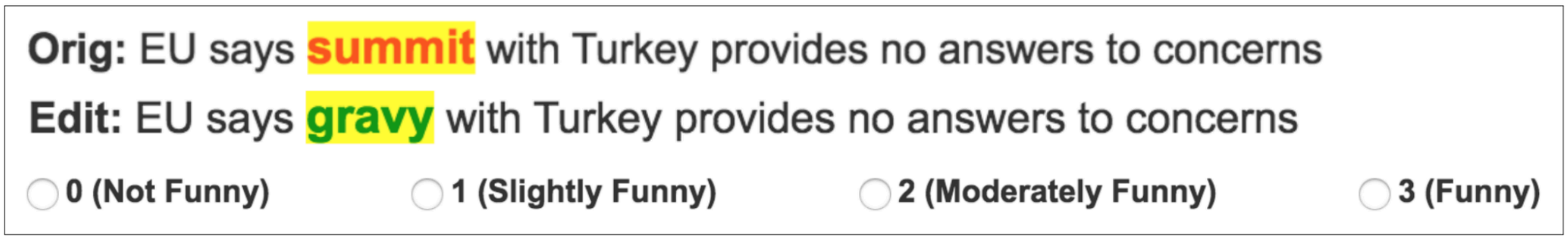
EU says ~~summit~~ gravy with Turkey provides no answers to concerns

- Address “continuous” humor
- Two humor subtasks
- Attracted record participants

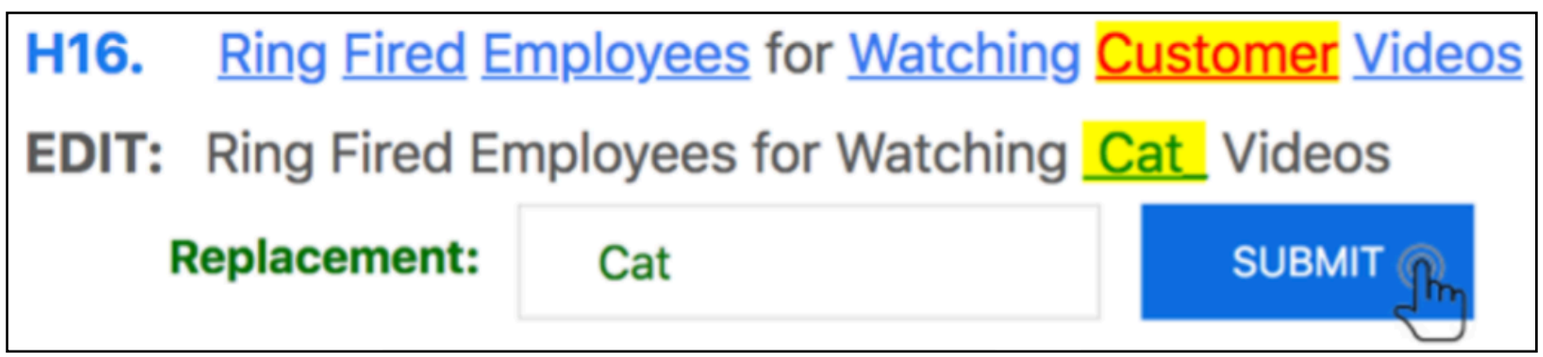


Datasets

Rating Annotation



Editing Annotation



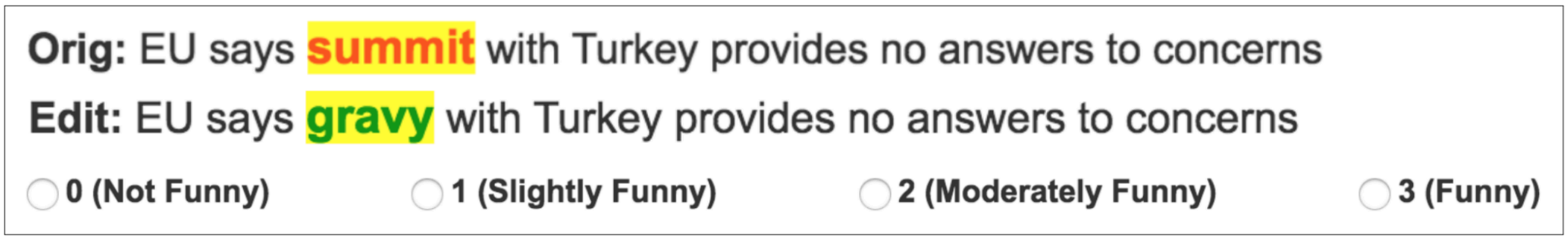
- **Humicroedit**: 15k headlines edited and rated for humor

  [Hossain et al. NAACL 2019. "President vows to cut taxes hair": Dataset and Analysis of Creative Text Editing for Humorous Headlines]

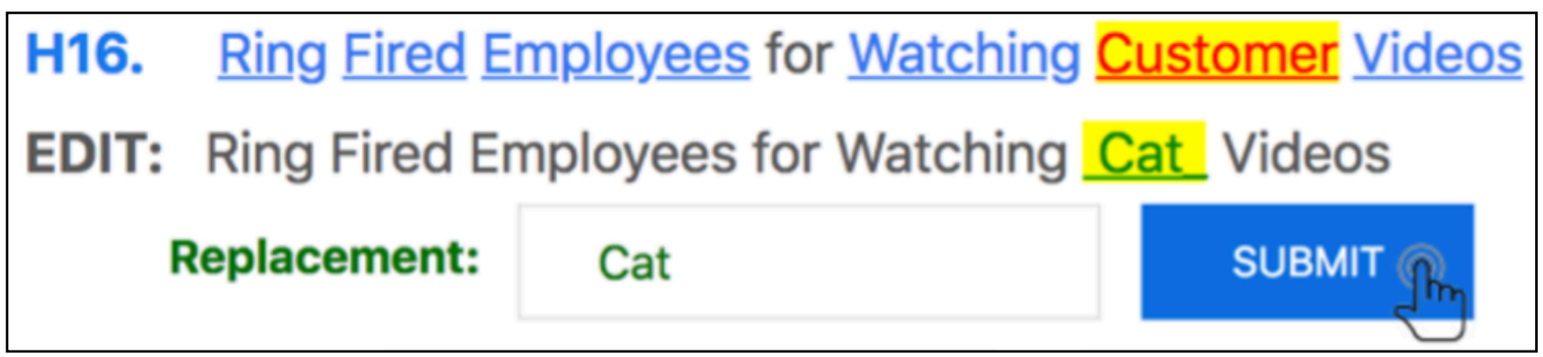
Datasets: https://zenodo.org/record/3969509#.XyWh6fhKh24

Datasets

Rating Annotation



Editing Annotation



- **Humicroedit**: 15k headlines edited and rated for humor

  [Hossain et al. NAACL 2019. "President vows to cut taxes hair": Dataset and Analysis of Creative Text Editing for Humorous Headlines]

- **FunLines**: Competitive game to generate humorous headlines

  [Hossain et al. ACL 2020. Stimulating Creativity with FunLines: A Case Study of Humor Generation in Headlines]

Datasets: https://zenodo.org/record/3969509#.XyWh6fhKh24
FunLines Demo: https://youtu.be/5OXJMxDBaLY

Examples

ID	**Original Headline** (replaced word in **bold**)	**Substitute**	**Rating**
R1	CNN 's Jake Tapper to **interview** Paul Ryan following retirement announcement	wrestle	2.8
R2	4 arrested in Sydney raids to stop terrorist **attack**	kangaroo	2.6
R3	Man Sets Off Explosive Device at L.A.-Area Cheesecake Factory, no **Injuries**	complaints	2.4
R4	5 dead, 9 injured in **shooting** at Fort Lauderdale Airport	delay	1.2
R5	Congress Struggles to **Confront** Sexual Harassment as Stories Pile Up	increase	1.2
R6	Congress Achieves the Impossible on **Tax** Reform	toilet	0.8
R7	Overdoses now leading **cause** of death of Americans under 50	sign	0.0
R8	Noor Salman, widow of Orlando massacre **shooter** Omar Mateen, arrested	columnist	0.0

Tasks

H1: EU says ~~summit~~ **gravy** with Turkey provides no answers to concerns

H2: EU says summit with Turkey provides no answers to ~~concerns~~ **questions**

- Subtask 1
  - **Regression**: Rate edited headlines on a 0-3 humor scale [48 teams]
- Subtask 2
  - **Classification**: Predict funnier of the two edits of a headline [31 teams]

Task: https://competitions.codalab.org/competitions/20970

Quickstart: github.com/n-hossain/semeval-2020-task-7-humicroedit

Tasks

H1: EU says ~~summit~~ **gravy** with Turkey provides no answers to concerns

H2: EU says summit with Turkey provides no answers to ~~concerns~~ **questions**

- Subtask 1
  - **Regression**: Rate edited headlines on a 0-3 humor scale [48 teams]
- Subtask 2
  - **Classification**: Predict funnier of the two edits of a headline [31 teams]

Task	Type	Metric	Train	FunLines (Train)	Dev	Test
Subtask 1	Regression	RMSE	9,653	8,248	2,420	3,025
Subtask 2	Classification	Accuracy	9,382	1,959	2,356	2,961

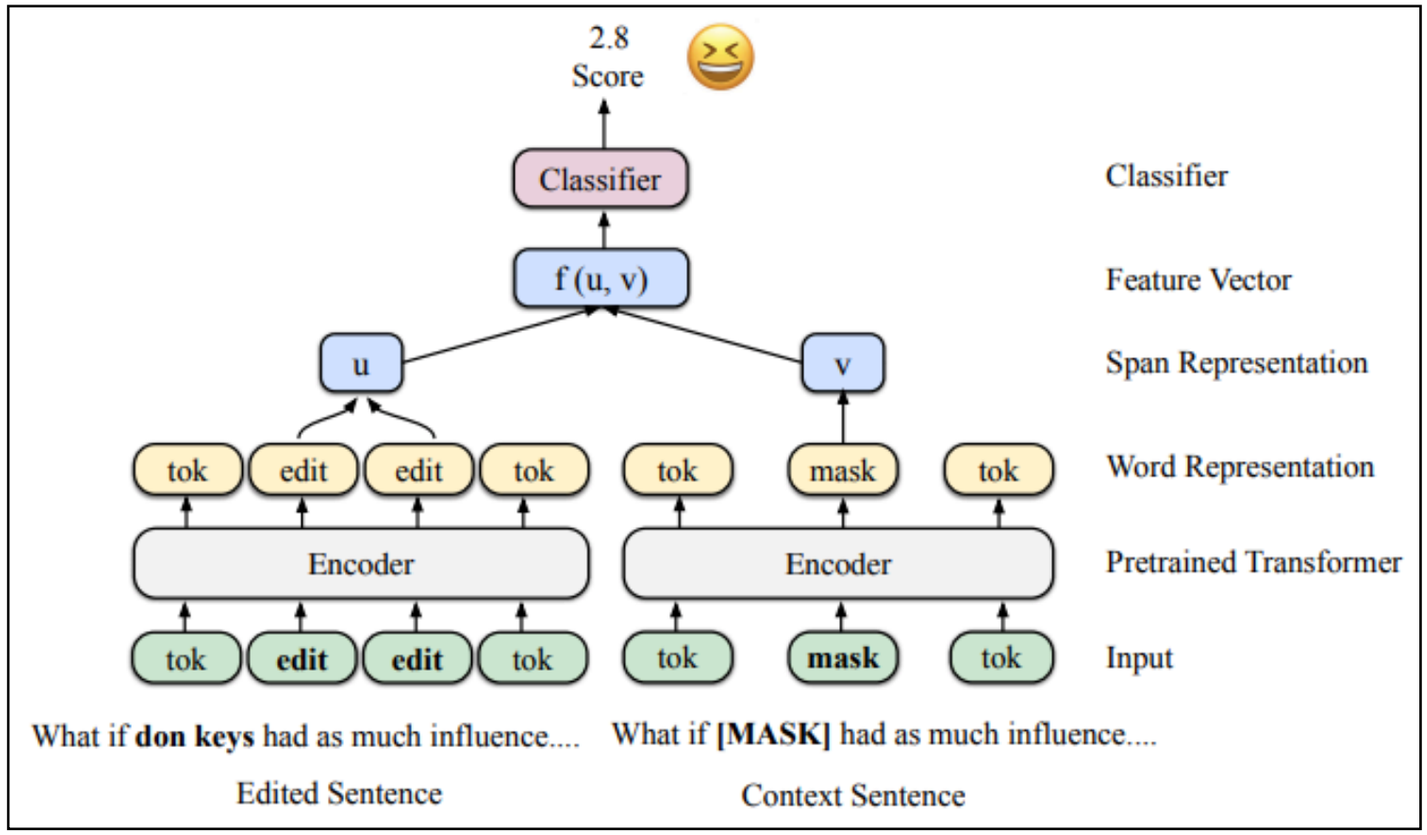
Table 2: Summary of the subtasks and their datasets.

Task: https://competitions.codalab.org/competitions/20970

Quickstart: github.com/n-hossain/semeval-2020-task-7-humicroedit

Benchmarks

- BASELINE:
  - mean rating (Subtask 1)
  - majority label (Subtask 2)
- CBOW — GloVe word vectors
- BERT & RoBERTa encodings



Model	Subtask 1 RMSE	Subtask 2 Acc.
BASELINE	0.575	0.490
CBOW		
with CONTEXT+FREEZE	**0.542**	0.599
+ORIG	0.559	0.599
+FUNLINES	0.544	0.605
+ORIG+FUNLINES	0.558	0.601
+FT	0.544	0.604
+FT+ORIG	0.561	0.592
+FT+FUNLINES	0.548	**0.606**
+FT+ORIG+FUNLINES	0.563	0.589
BERT		
with CONTEXT+FREEZE	0.531	0.616
+ORIG	0.534	0.603
+FUNLINES	**0.530**	0.615
+ORIG+FUNLINES	0.541	0.615
+FT	0.536	**0.635**
+FT+ORIG	0.536	0.628
+FT+FUNLINES	0.541	0.630
+FT+ORIG+FUNLINES	0.533	0.629
RoBERTa		
with CONTEXT+FREEZE	0.528	0.635
+ORIG	0.536	0.625
+FUNLINES	0.528	0.640
+ORIG+FUNLINES	0.533	0.618
+FT	0.534	0.649
+FT+ORIG	0.527	**0.650**
+FT+FUNLINES	0.526	0.638
+FT+ORIG+FUNLINES	**0.522**	0.626

Benchmarks: Jin et al. 2020. SemEval-2020. Duluth at SemEval-2020 task 7: Using surprise as a key to unlock humorous headlines

Subtask 1 Results

Rank	Team	RMSE
1	Hitachi	0.49725
2	Amobee	0.50726
3	YNU-HPCC	0.51737
4	MLEngineer	0.51966
5	LMML	0.52027
6	ECNU	0.52187
bench.	**RoBERTa**	**0.52207**
7	LT3	0.52532
8	WMD	0.52603
9	Ferryman	0.52776
10	zxchen	0.52886
bench.	**BERT**	**0.53036**
11	Duluth	0.53108
12	will_go	0.53228
13	XSYSIGMA	0.53308
14	LRG	0.53318
15	MeisterMorxrc	0.53383
16	JUST_Farah	0.53396
17	Lunex	0.53518
18	UniTuebingenCL	0.53954
bench.	**CBOW**	**0.54242**
19	IRLab_DAIICT	0.54670
20	O698	0.54754
21	UPB	0.54803
22	Buhscitu	0.55115
23	Fermi	0.55226
24	INGEOTEC	0.55391

Rank	Team	RMSE
25	JokeMeter	0.55791
26	testing	0.55838
27	HumorAAC	0.56454
28	ELMo-NB	0.56829
29	prateekgupta2533	0.56983
30	funny3	0.57237
31	WUY	0.57369
32	XTHL	0.57470
bench.	**BASELINE**	**0.57471**
33	HWMT_Squad	0.57471
34	moonalasad	0.57479
35	dianehu	0.57488
36	Warren	0.57527
37	tangmen	0.57768
38	Lijunyi	0.57946
39	Titowak	0.58157
40	xenia	0.58286
41	Smash	0.59202
42	KdeHumor	0.61643
43	uir	0.62401
44	SO	0.65099
45	heidy	0.68338
46	Hasyarasa	0.70333
47	frietz58	0.72252
48	SSN_NLP	0.84476

Subtask 2 Results

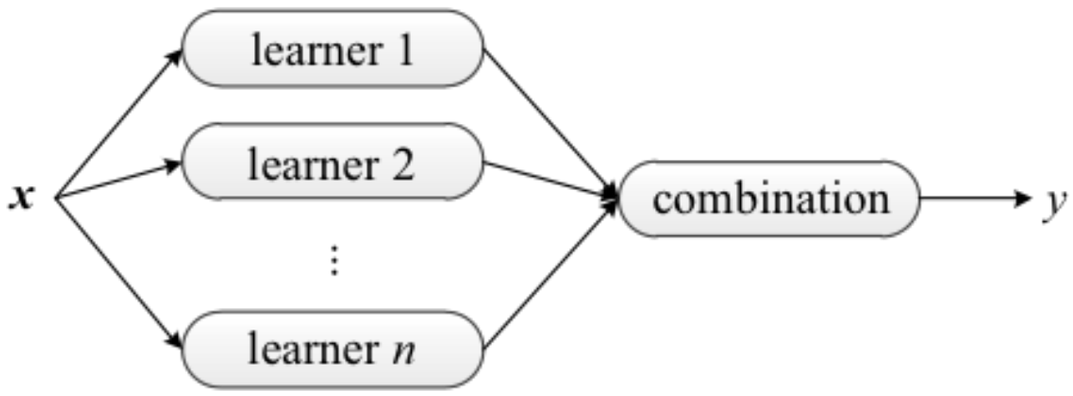
Rank	Team	Accuracy
1	Hitachi	0.6743
2	Amobee	0.6606
3	YNU-HPCC	0.6591
bench.	**RoBERTa**	**0.6495**
4	LMML	0.6469
5	XSYSIGMA	0.6446
6	ECNU	0.6438
7	Fermi	0.6393
bench.	**BERT**	**0.6355**
8	zxchen	0.6347
9	Duluth	0.6320
10	WMD	0.6294
11	Buhscitu	0.6271
12	MLEngineer	0.6229
13	LRG	0.6218
14	UniTuebingenCL	0.6183
15	O698	0.6134
16	JUST_Farah	0.6088
bench.	**CBOW**	**0.6057**
17	INGEOTEC	0.6050
18	Ferryman	0.6027
19	UPB	0.6001
20	Hasyarasa	0.5970
21	JokeMeter	0.5776
22	UTFPR	0.5696
23	Smash	0.5426
24	SSN_NLP	0.5377
25	WUY	0.5320
26	uir	0.5213
27	KdeHumor	0.5190
28	Titowak	0.5038
bench.	**BASELINE**	**0.4950**
29	heidy	0.4197
30	SO	0.3291
31	HumorAAC	0.3204

Popular Approaches

- PLMs: BERT, RoBERTa, ELMo, GPT-2, XLNet, Transformer-XL

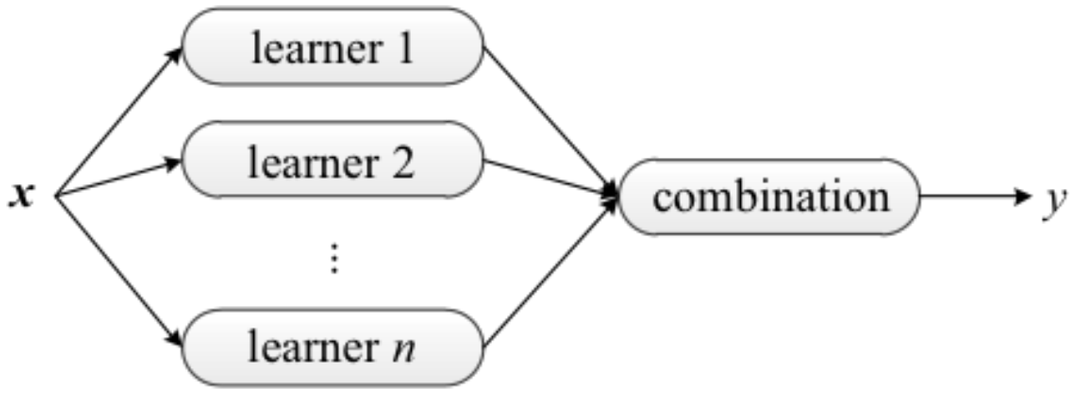
- Context independent: Word2Vec, FastText, GloVe

Popular Approaches



- PLMs: BERT, RoBERTa, ELMo, GPT-2, XLNet, Transformer-XL
- Context independent: Word2Vec, FastText, GloVe
- Ensembling: ridge regression, XGBoost, Naïve Bayes

Popular Approaches



- PLMs: BERT, RoBERTa, ELMo, GPT-2, XLNet, Transformer-XL
- Context independent: Word2Vec, FastText, GloVe
- Ensembling: ridge regression, XGBoost, Naïve Bayes
- Sub-task 2: Use model for Sub-task 1 to find the funnier headline

Top 3 Systems

- **Hitachi**: Ensemble of 700 fine-tuned PLM instances
  - RoBERTa > GPT-2 > BERT > XLM > XLNet > Transformer-XL
  - combined using Ridge regression

Top 3 Systems

- **Hitachi:** Ensemble of 700 fine-tuned PLM instances
  - RoBERTa > GPT-2 > BERT > XLM > XLNet > Transformer-XL
  - combined using Ridge regression

- **Amobee:** Ensemble of 90 BERT, RoBERTa, XLNet
  - Weights: 0.5 RoBERTa, 0.3 XLNet, 0.2 BERT

Top 3 Systems

- **Hitachi:** Ensemble of 700 fine-tuned PLM instances
  - RoBERTa > GPT-2 > BERT > XLM > XLNet > Transformer-XL
  - combined using Ridge regression

- **Amobee:** Ensemble of 90 BERT, RoBERTa, XLNet
  - Weights: 0.5 RoBERTa, 0.3 XLNet, 0.2 BERT
- **YNU-HPCC:** Ensemble of 11 edited headline encodings
  - FastText, Word2Vec, ELMo, BERT encoders
  - combined using XGBoost regressor

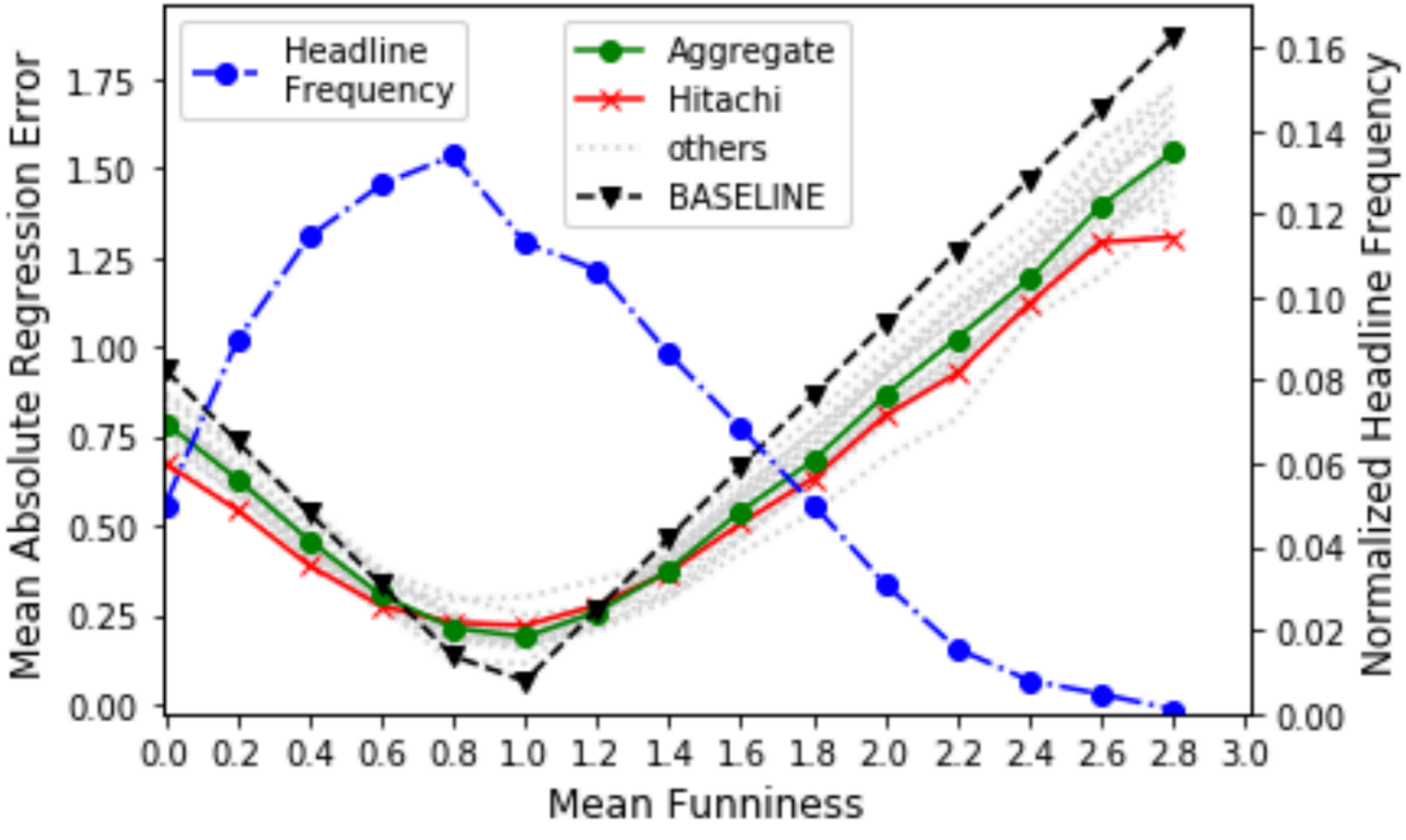
Notable Approaches

System	Approach
ECNU	sentiment & humor lexicons for feature extraction
LT3 (baseline)	lexical, entity, readability, length, positional, word embedding similarity, perplexity, string similarity features
IRLab_DAIICT	5 BERT classifiers, one per headline rating
Buhscitu	knowledge bases, language model, hand-crafted features
Hasyarasa	word embedding + knowledge graph to find contextual absurdity
UTFPR	unsupervised, using word co-occurrence for unexpectedness

Analysis
(Top 20 systems)

Error bins (Subtask 1)

- Top 20 systems for Sub task 1
- Min error around 1.0 funniness
- Difficult to estimate more extreme humor



Systematic Estimation Errors (Subtask 1)

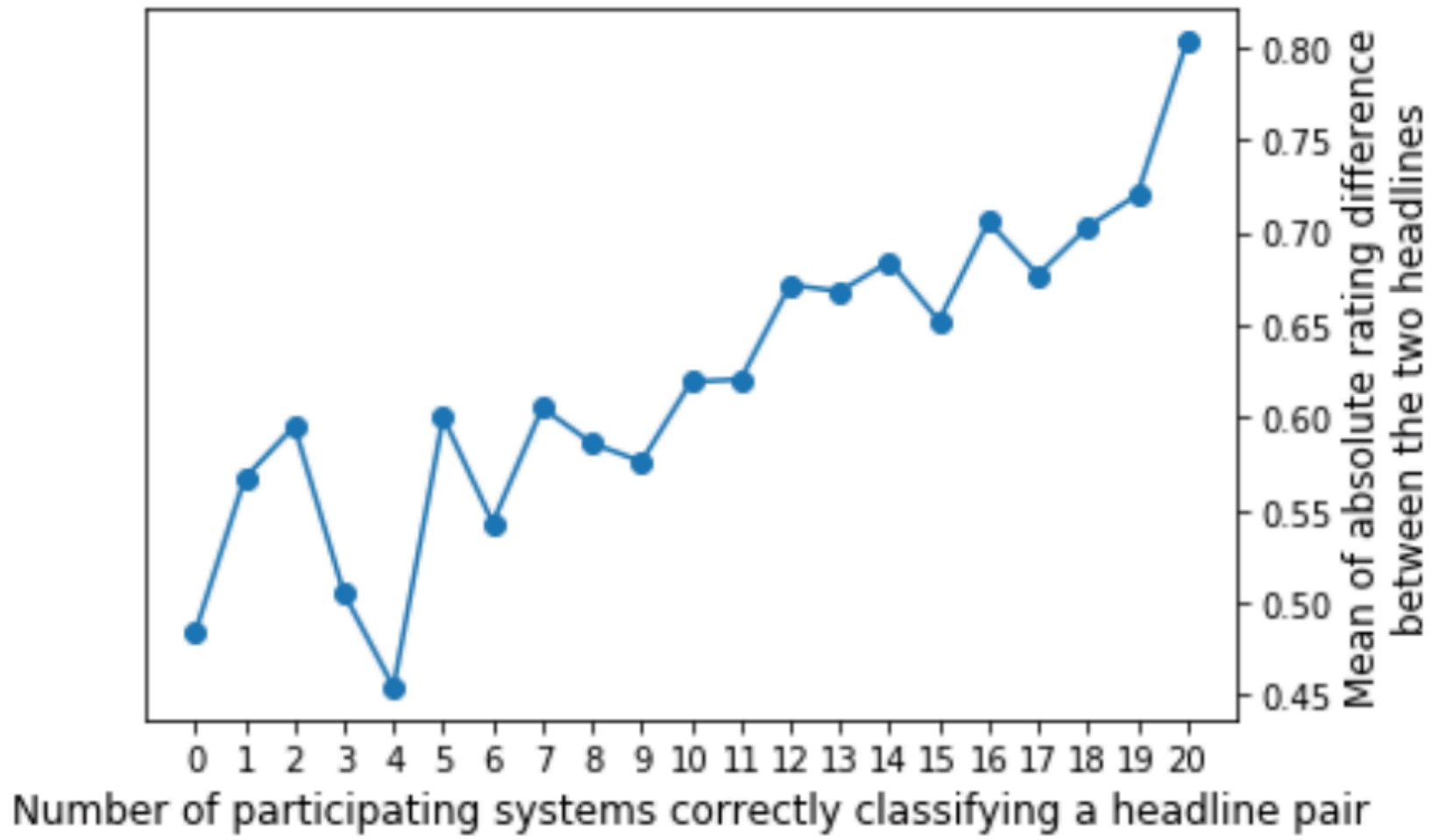
ID	**Original Headline** (replaced word in **bold**)	Substitute	Rating	Est.	Err.
R1	CNN 's Jake Tapper to **interview** Paul Ryan following retirement announcement	wrestle	2.8	1.17	-1.63
R2	4 arrested in Sydney raids to stop terrorist **attack**	kangaroo	2.6	1.06	-1.54
R3	Man Sets Off Explosive Device at L.A.-Area Cheesecake Factory, no **Injuries**	complaints	2.4	0.80	-1.60
R4	5 dead, 9 injured in **shooting** at Fort Lauderdale Airport	delay	1.2	0.49	-0.71
R5	Congress Struggles to **Confront** Sexual Harassment as Stories Pile Up	increase	1.2	0.66	-0.54
R6	Congress Achieves the Impossible on **Tax** Reform	toilet	0.8	1.35	+0.55
R7	Overdoses now leading **cause** of death of Americans under 50	sign	0.0	0.52	+0.52
R8	Noor Salman, widow of Orlando massacre **shooter** Omar Mateen, arrested	columnist	0.0	0.43	+0.43

- Examples (over/under)-estimated collectively by top 20 systems
- Challenges
  - world knowledge (R1)
  - cultural reference (R2)
  - sarcasm (R3, R4, R5)
  - negative sentiment (R7, R8)

Funniness Gaps (Subtask 2)

H1: EU says ~~summit~~ **gravy** with Turkey provides no answers to concerns

H2: EU says summit with Turkey provides no answers to ~~concerns~~ **questions**

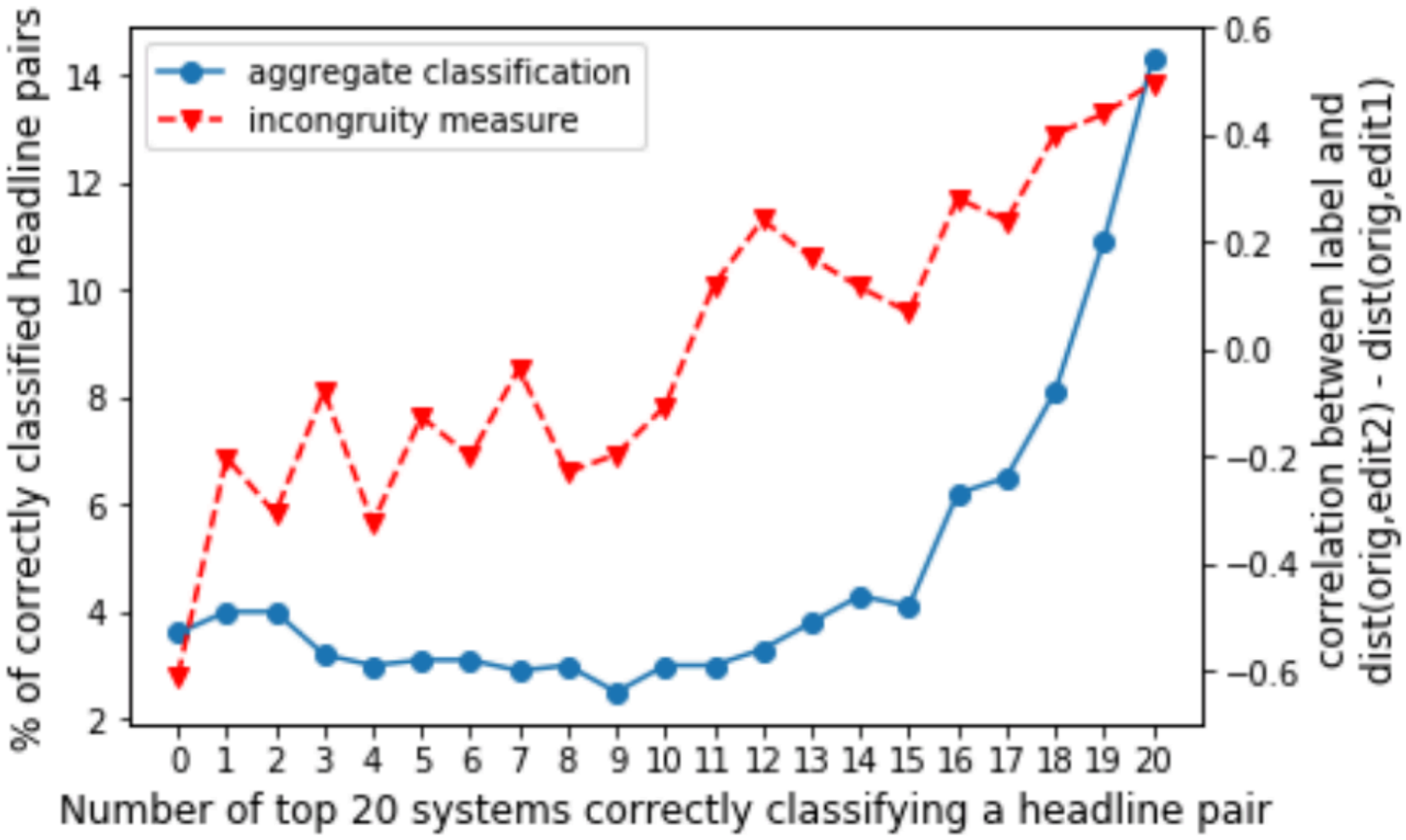


- Headline pairs with larger funniness gaps are easier to classify

Incongruity (Subtask 2)

H1: EU says ~~summit~~ **gravy** with Turkey provides no answers to concerns

H2: EU says summit with Turkey provides no answers to ~~concerns~~ **questions**



- Systems have better classification if headline incongruity correlates with funniness

Extreme Examples (Subtask 2)

ID	**Original Headline** (replaced word in **bold**)	**Substitute**	**Rating**	**Dist.**
C1	Secret **Service** likely wouldn't have intervened in Trump Jr.-Russia meeting	police	0.0	0.72
✔	Secret **Service** likely wouldn't have intervened in Trump Jr.-Russia meeting	Santa	2.6	0.85
C2	Amazon, Facebook and Google could save **billions** thanks to the GOP tax bill	puppies	1.0	0.89
✘	Amazon, Facebook and Google could save **billions** thanks to the GOP tax bill	pennies	2.2	0.54
C3	LA Times editorial board condemns Donald Trump **presidency** as 'trainwreck'	diet	1.2	0.96
✔	LA Times editorial board **condemns** Donald Trump presidency as 'trainwreck'	celebrates	1.0	0.69
C4	US officials drop **mining** cleanup rule after industry objects	floor	1.4	0.86
✘	US officials drop **mining** cleanup rule after industry objects	Bedroom	1.2	1.01

- ✔: all 20 systems correctly classified
- ✘: all 20 systems failed to classify
- Dist.: GloVe word vector distance between replaced-replacement words (measure of incongruity)
- C2: where incongruity does not lead to humor

Conclusion

- Assessing humor in edited news headlines
- Record participants for a humor task
- Winning systems use PLMs in ensemble
- Analysis shows systems mainly capture incongruity
- Future work — funny headlines and Mad Libs generation

[Hossain et al. EMNLP 2017. Filling the Blanks for Mad Libs Humor]

Thanks

- Thanks to all the participants

Subtask	Metric	Baseline	Best (Team)
Regression	RMSE	0.575	0.497 (Hitachi)
Classification	Accuracy	49.5%	67.4% (Hitachi)

- Questions: nabil.hossain@intel.com