Part V: Advanced Text Mining Applications Empowered by Pretrained Embeddings

KDD 2022 Tutorial

Adapting Pretrained Representations for Text Mining

Yu Meng, Jiaxin Huang, Yu Zhang, Jiawei Han

Computer Science, University of Illinois at Urbana-Champaign

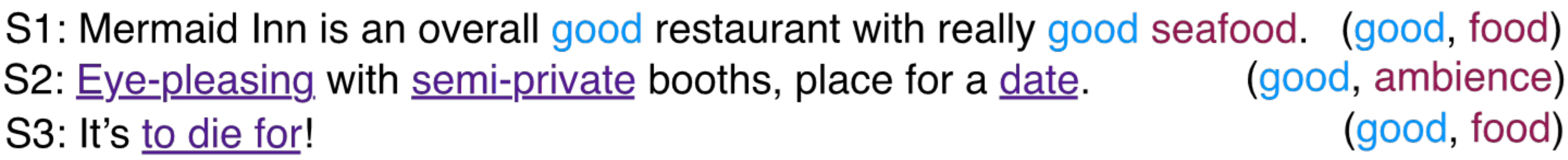
Aug 14, 2022

Outline

- Aspect-based Sentiment Analysis
  - Weakly-Supervised Aspect-Based Sentiment Analysis via Joint Aspect-Sentiment Topic Embedding [EMNLP’20]
- Text Summarization
- Online Story Discovery from News Streams
- Summary & Future Directions

Aspect-based Sentiment Analysis

- Task definition
  - Given an opinionated document about a target entity (e.g., a laptop, a restaurant or a hotel), the goal is to identify the opinion tuple of <aspect, sentiment> of the document



S1: Mermaid Inn is an overall good restaurant with really good seafood. (good, food)

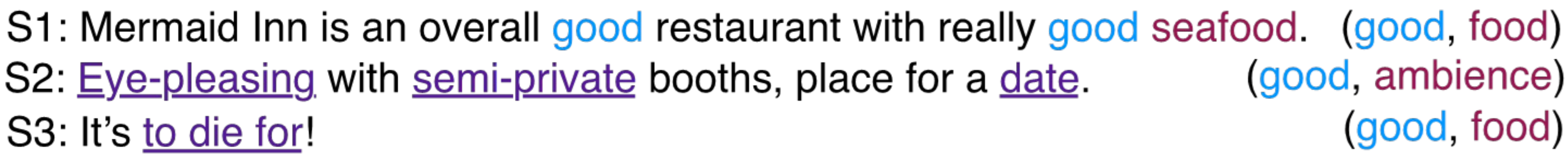
S2: Eye-pleasing with semi-private booths, place for a date. (good, ambience)

S3: It's to die for! (good, food)

- Most previous studies deal with the tasks of aspect extraction and sentiment polarity classification individually or sequentially
- Other methods jointly solve these two sub-tasks by first separating target words from opinion words and then learning joint topic distributions over words

Motivation

- Sample Reviews

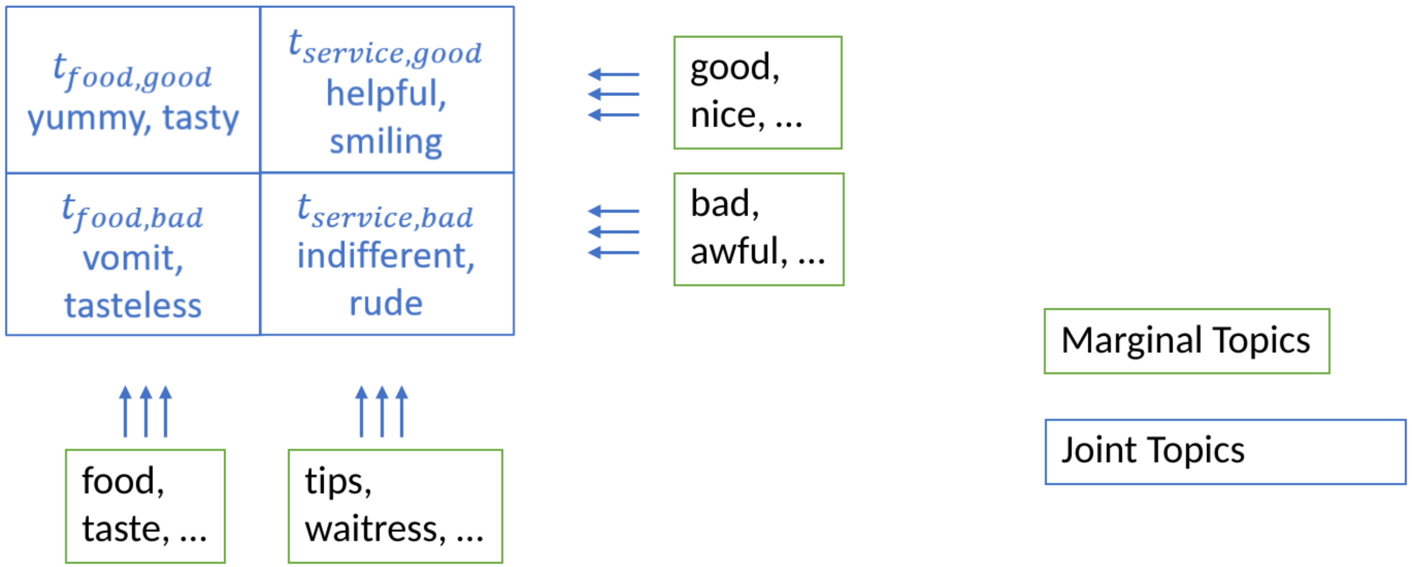


S1: Mermaid Inn is an overall good restaurant with really good seafood. (good, food)
S2: Eye-pleasing with semi-private booths, place for a date. (good, ambience)
S3: It's to die for! (good, food)

- Pure aspect words are in red, and general opinion words are in blue
- Words implying both aspects and opinions (which we define as **joint topics**) are underlined and in purple

- S1: general aspect, opinion words
- S2 and S3: Target is not explicitly addressed. Fine-grained words are used to imply both aspect and polarity

Joint “Sentiment-Aspect” topic



$t_{food,good}$ yummy, tasty

$t_{service,good}$ helpful, smiling

$t_{food,bad}$ vomit, tasteless

$t_{service,bad}$ indifferent, rude

good, nice, …

bad, awful, …

food, taste, …

tips, waitress, …

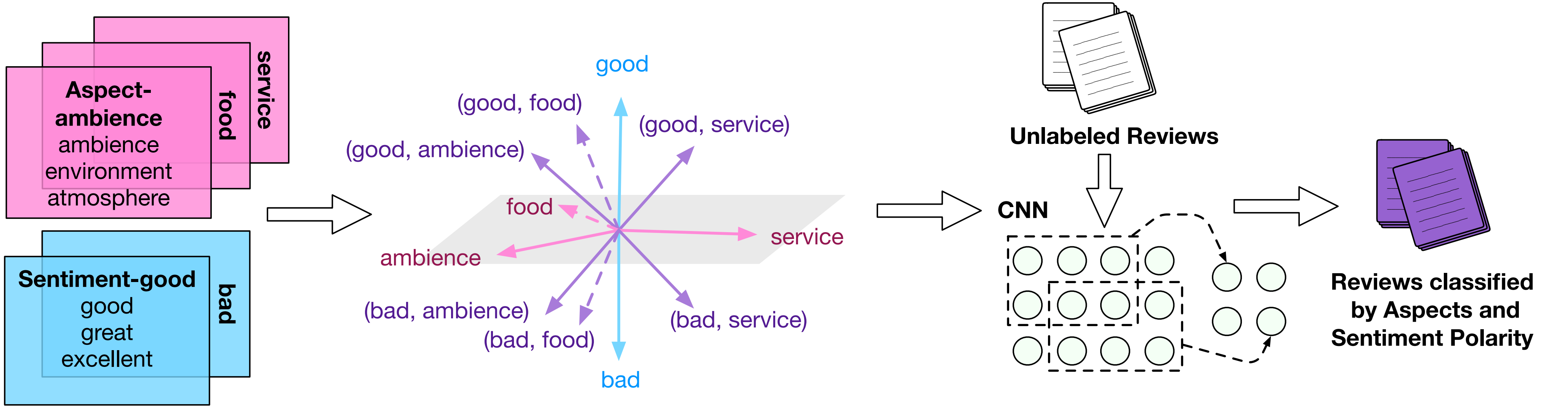
Marginal Topics

Joint Topics

- ❑ If the semantics of each joint topic of <sentiment, aspect> can be automatically captured, machines will be able to identify representative terms of the joint topics such as “semi-private” for <good, ambience>
- ❑ Thus, it will benefit both aspect extraction and sentiment classification
- ❑ Our general idea is to learn and regularize the joint topics in the embedding space to enhance both tasks

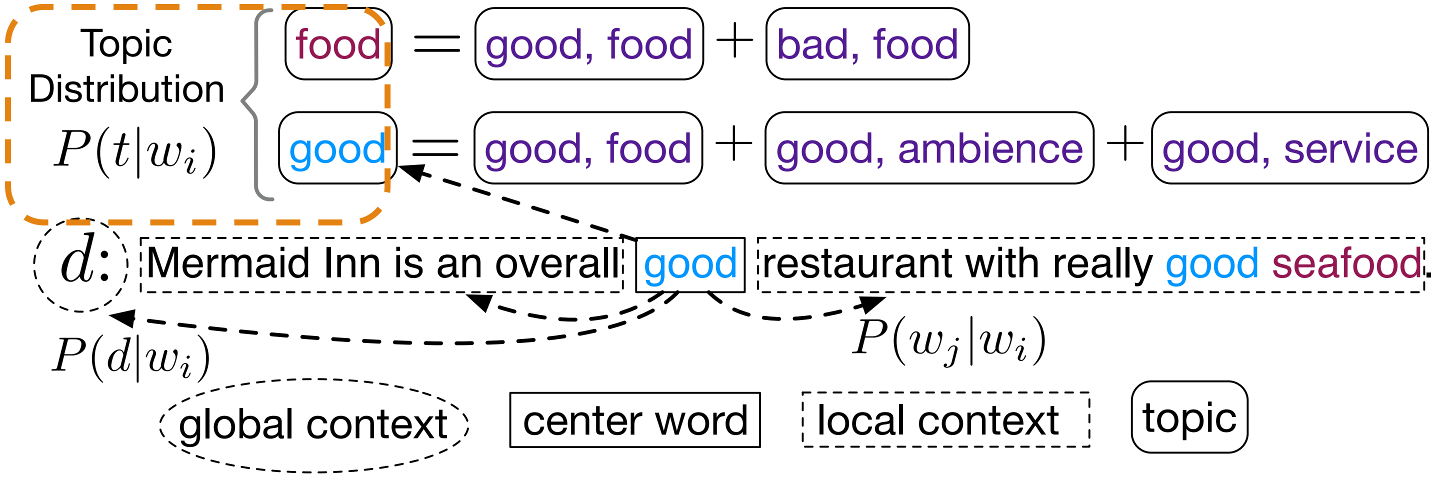
Our Framework

- ❑ Weakly-Supervised Aspect-Based Sentiment Analysis via Joint Aspect-Sentiment Topic Embedding [EMNLP'20]



- ❑ Step 1: Leverage the in-domain training corpus and user-given keywords to learn joint topic representation in the word embedding space
- ❑ Step 2: Embedding-based prediction on unlabeled data are then leveraged by neural models for pre-training and self-training

Joint-Topic Representation Learning

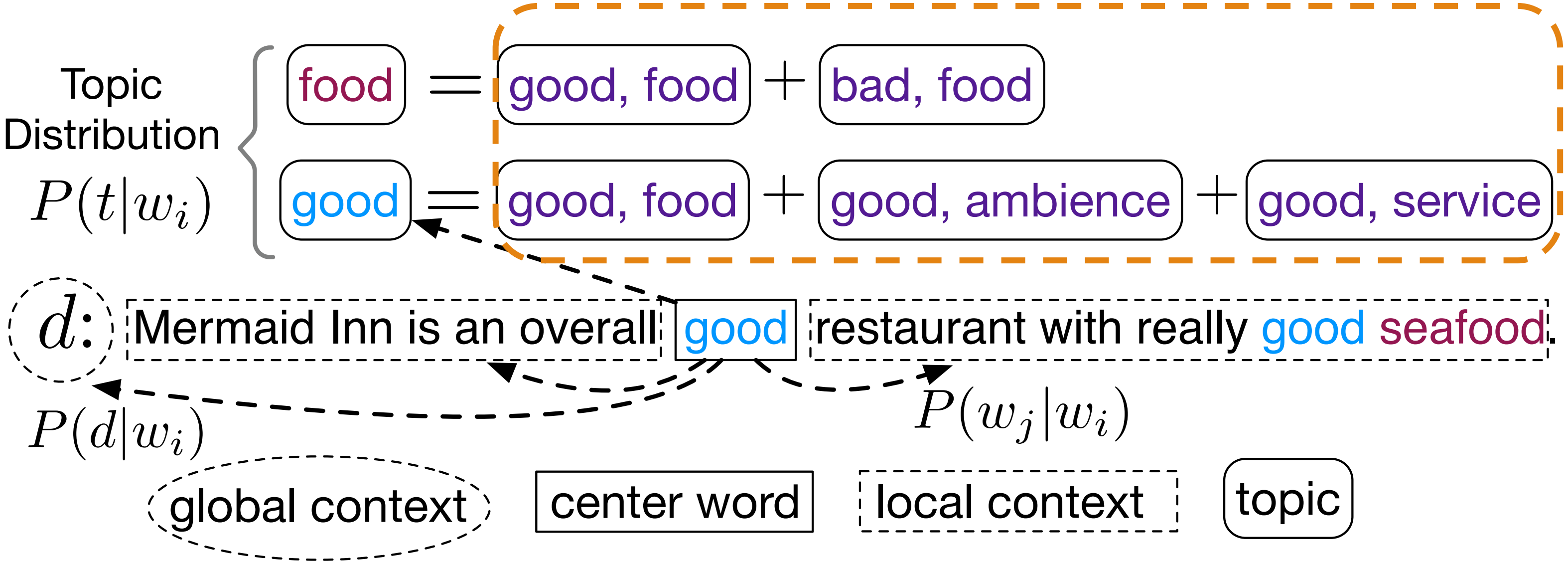

Topic
Distribution
$P(t|w_i)$
food = good, food + bad, food
good = good, food + good, ambience + good, service
d: Mermaid Inn is an overall good restaurant with really good seafood.
$P(d|w_i)$
$P(w_j|w_i)$
global context
center word
local context
topic

- Regularizing Pure Aspect/Sentiment Topics. We regularize the aspect topic embeddings t_a and sentiment topic embeddings t_s so that different topics are pushed apart
  - Marginal topic regularization:

$$\mathcal{L}_{reg}^{A} = -\sum_{a \in A} \sum_{w_i \in l_a} \log P(t_a|w_i) \qquad \mathcal{L}_{reg}^{S} = -\sum_{s \in S} \sum_{w_i \in l_s} \log P(t_s|w_i). \qquad P(t|w_i) \propto \exp(\boldsymbol{u}_i^\top \boldsymbol{t})$$

  - Words can be "classified" into topics based on embedding similarity
  - User-provided keywords are used for initialization, and more keywords are expanded based on cosine similarity in each embedding training epoch

Joint-Topic Representation Learning


Topic
Distribution
$P(t|w_i)$
food = good, food + bad, food
good = good, food + good, ambience + good, service
d: Mermaid Inn is an overall good restaurant with really good seafood.
$P(d|w_i)$
$P(w_j|w_i)$
global context
center word
local context
topic

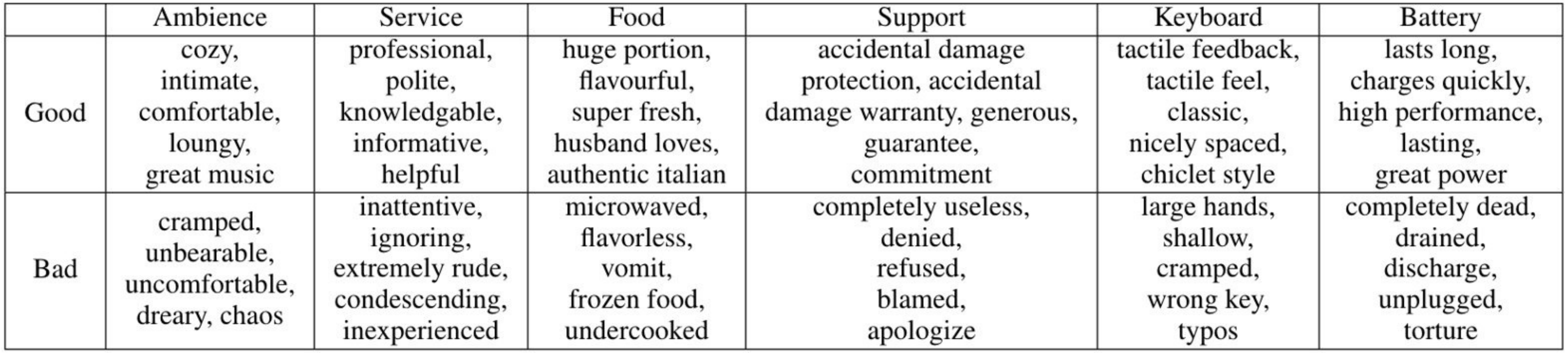
- ❑ Regularizing Joint <Sentiment, Aspect> Topics
- ❑ We connect the learning of joint topic embeddings with pure aspect/sentiment topics by exploring the relationship between marginal distribution and joint distribution

$$P(t_a|w_i) = \sum_{s \in S} P\left(t_{\langle s,a \rangle} \middle| w_i\right) \qquad P(t_s|w_i) = \sum_{a \in A} P\left(t_{\langle s,a \rangle} \middle| w_i\right)$$

  - ❑ To form the joint topic regularization objective, we can replace the probability term in the pure aspect/sentiment regularization objective with the sum of joint probability

Representative Terms for Joint Topics

- To evaluate the quality of the joint topic representation, we retrieve their representative terms by ranking the embedding cosine similarity between words and each joint topic vector

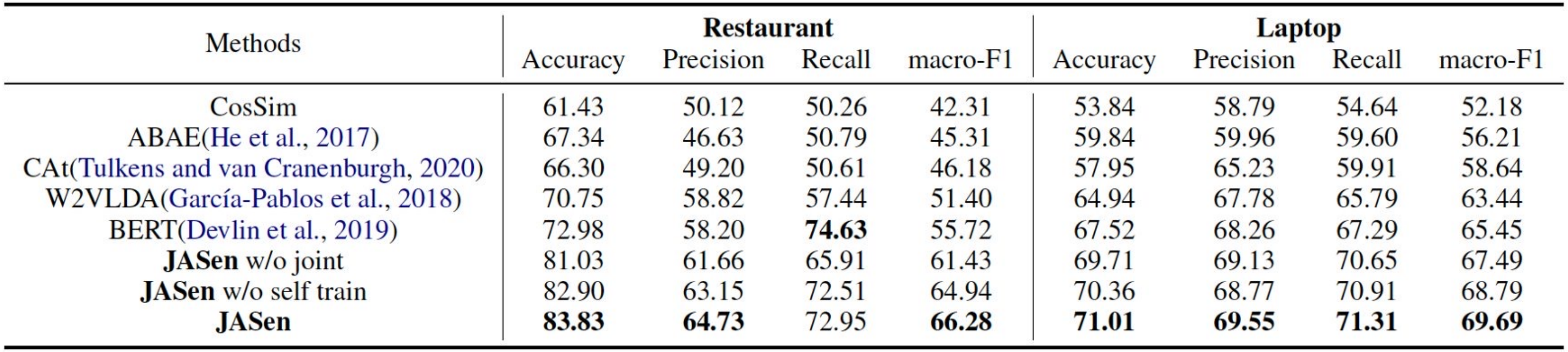


	Ambience	Service	Food	Support	Keyboard	Battery
Good	cozy, intimate, comfortable, loungy, great music	professional, polite, knowledgable, informative, helpful	huge portion, flavourful, super fresh, husband loves, authentic italian	accidental damage protection, accidental damage warranty, generous, guarantee, commitment	tactile feedback, tactile feel, classic, nicely spaced, chiclet style	lasts long, charges quickly, high performance, lasting, great power
Bad	cramped, unbearable, uncomfortable, dreary, chaos	inattentive, ignoring, extremely rude, condescending, inexperienced	microwaved, flavorless, vomit, frozen food, undercooked	completely useless, denied, refused, blamed, apologize	large hands, shallow, cramped, wrong key, typos	completely dead, drained, discharge, unplugged, torture

- Representative terms are not restricted to be adjectives, such as “vomit” in (bad, food)and “commitment” in (good, support)
- “Cramped” appears in both (bad, ambience) in restaurant domain and (bad, keyboard) in laptop domain

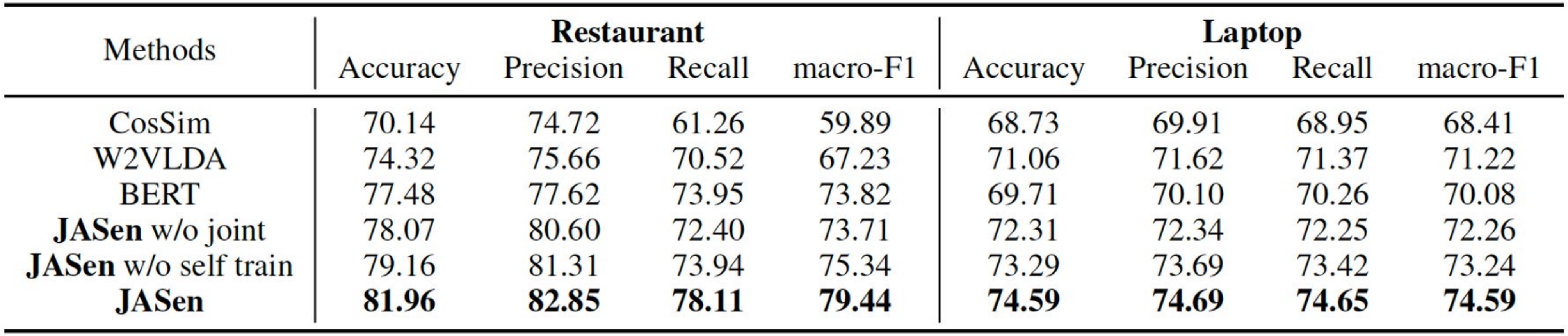
Quantitative Evaluation

- Aspect Extraction



Methods	Restaurant				Laptop			
	Accuracy	Precision	Recall	macro-F1	Accuracy	Precision	Recall	macro-F1
CosSim	61.43	50.12	50.26	42.31	53.84	58.79	54.64	52.18
ABAE(He et al., 2017)	67.34	46.63	50.79	45.31	59.84	59.96	59.60	56.21
CAt(Tulkens and van Cranenburgh, 2020)	66.30	49.20	50.61	46.18	57.95	65.23	59.91	58.64
W2VLDA(García-Pablos et al., 2018)	70.75	58.82	57.44	51.40	64.94	67.78	65.79	63.44
BERT(Devlin et al., 2019)	72.98	58.20	**74.63**	55.72	67.52	68.26	67.29	65.45
JASen w/o joint	81.03	61.66	65.91	61.43	69.71	69.13	70.65	67.49
JASen w/o self train	82.90	63.15	72.51	64.94	70.36	68.77	70.91	68.79
JASen	**83.83**	**64.73**	72.95	**66.28**	**71.01**	**69.55**	**71.31**	**69.69**

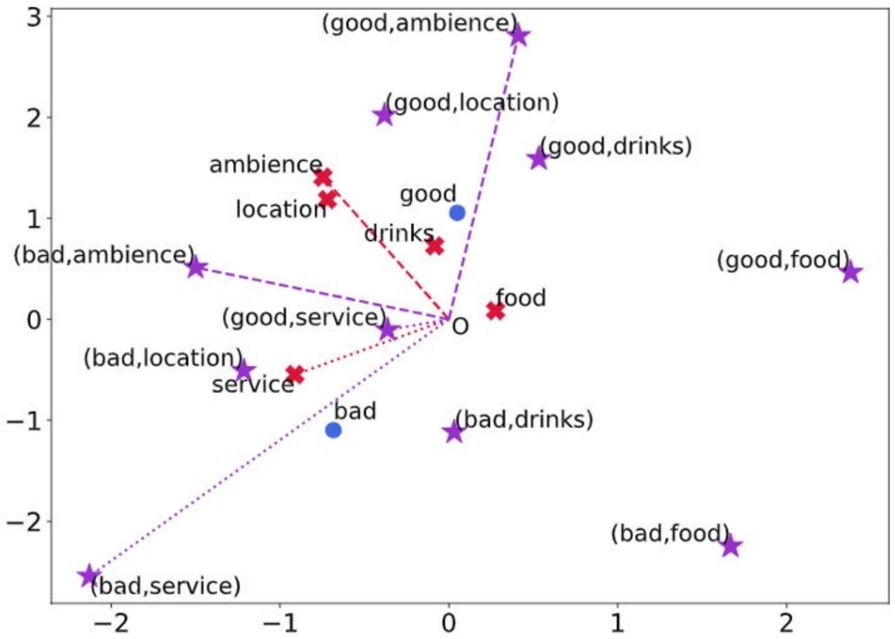
- Sentiment Polarity Classification



Methods	Restaurant				Laptop			
	Accuracy	Precision	Recall	macro-F1	Accuracy	Precision	Recall	macro-F1
CosSim	70.14	74.72	61.26	59.89	68.73	69.91	68.95	68.41
W2VLDA	74.32	75.66	70.52	67.23	71.06	71.62	71.37	71.22
BERT	77.48	77.62	73.95	73.82	69.71	70.10	70.26	70.08
JASen w/o joint	78.07	80.60	72.40	73.71	72.31	72.34	72.25	72.26
JASen w/o self train	79.16	81.31	73.94	75.34	73.29	73.69	73.42	73.24
JASen	**81.96**	**82.85**	**78.11**	**79.44**	**74.59**	**74.69**	**74.65**	**74.59**

Joint Topic Representation Visualization

- Visualization of joint topics (purple stars), aspect topics (red crosses) and sentiment topics (blue dots) in the embedding space



- An interesting observation is that some aspect topics (e.g., ambience) lie approximately in the middle of their joint topics ("good, ambience" and "bad, ambience"), showing that our embedding learning objective understands the joint topics as decomposition of their "marginal" topics

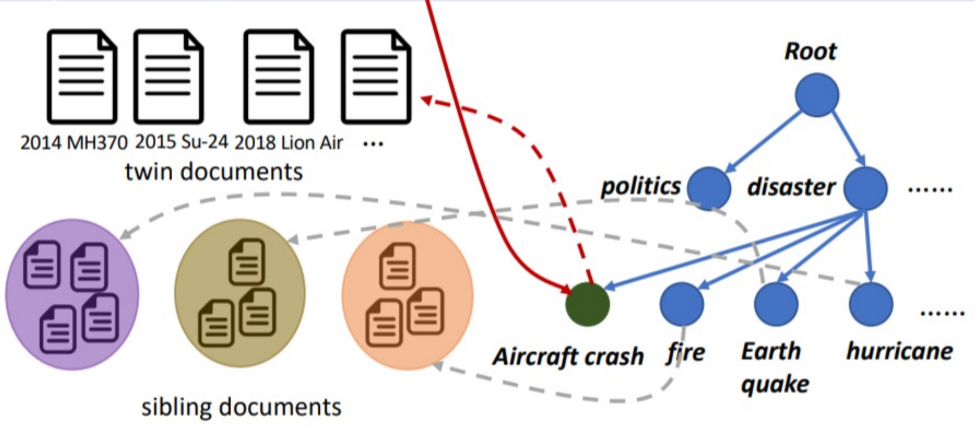
Outline

- ❑ Aspect-based Sentiment Analysis
- ❑ Text Summarization
  - ❑ SUMDocs: Extractive Summarization with Background Corpus [SDM’21]
  - ❑ Pre-trained Language Models for Summarization
    - ❑ Generating Representative Headlines for News Stories [WWW’20]
- ❑ Online Story Discovery from News Streams
- ❑ Summary & Future Directions

SUMDocS

- SUMDocS: Surrounding-aware Unsupervised Multi-Document Summarization (SDM’21)
- Leverage surrounding documents from the background corpus to obtain salient and discriminative extractive summarization

	Ethiopian Airlines Crash
d1	A doomed Ethiopian Airlines jet suffered from faulty readings by a key sensor, and pilots followed Boeing’s recommended procedures when …
d2	Sunday’s devasting plane crash in Ethiopia could renew safety questions about the newest version of Boeing’s popular 737 airliner…
d3	Will the Boeing 737 MAX aircraft be cleared to fly again and, if so, will the traveling public be willing to fly on those airplanes? Obviously…

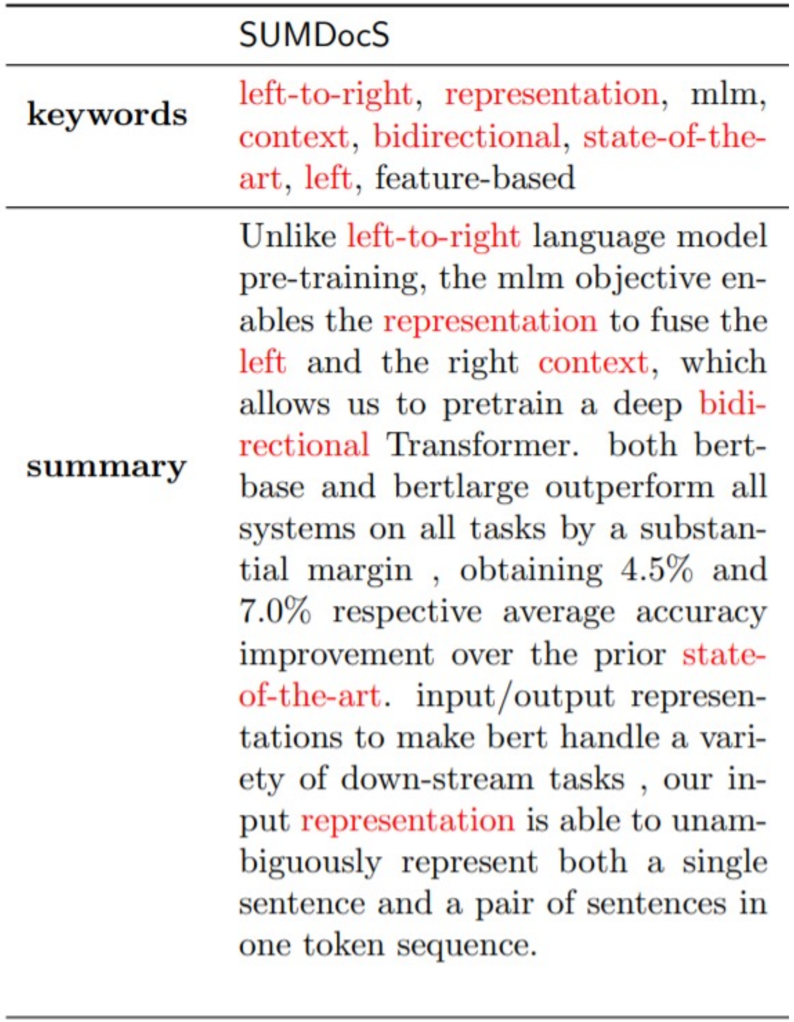


SUMDocS

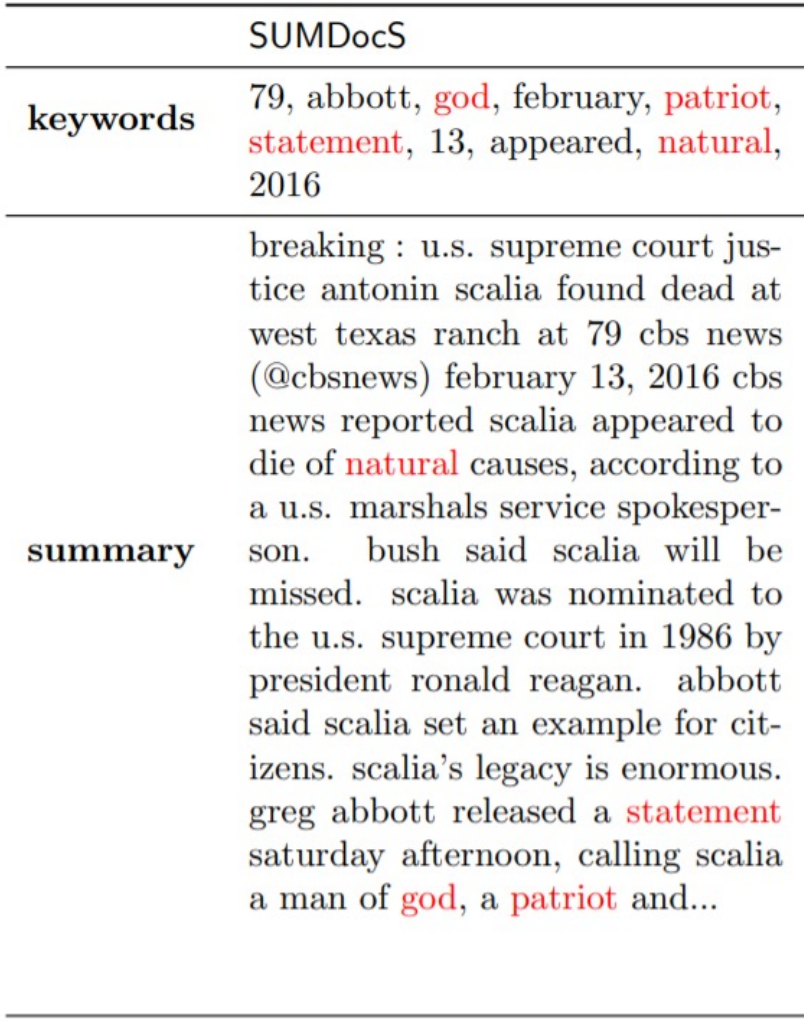
- How to leverage the background corpus?
  - Twin documents: Documents belonging to the same category
  - Sibling documents: Documents belonging to orthogonal categories
- Consider three factors when generating extractive summarizations
  - Global novelty: Category-level frequent and discriminative phrases are likely to be salient phrases
  - Local consistency: Frequently co-occurred phrases should have similar salient score
  - Local saliency: Phrases that are salient in target documents but less salient in twin documents should be promoted

SUMDocS: Results

- Identified keywords and generated summaries on NLP corpus (left) and news corpus (right)



	SUMDocS
keywords	left-to-right, representation, mlm, context, bidirectional, state-of-the-art, left, feature-based
summary	Unlike left-to-right language model pre-training, the mlm objective enables the representation to fuse the left and the right context, which allows us to pretrain a deep bidirectional Transformer. both bert-base and bertlarge outperform all systems on all tasks by a substantial margin , obtaining 4.5% and 7.0% respective average accuracy improvement over the prior state-of-the-art. input/output representations to make bert handle a variety of down-stream tasks , our input representation is able to unambiguously represent both a single sentence and a pair of sentences in one token sequence.



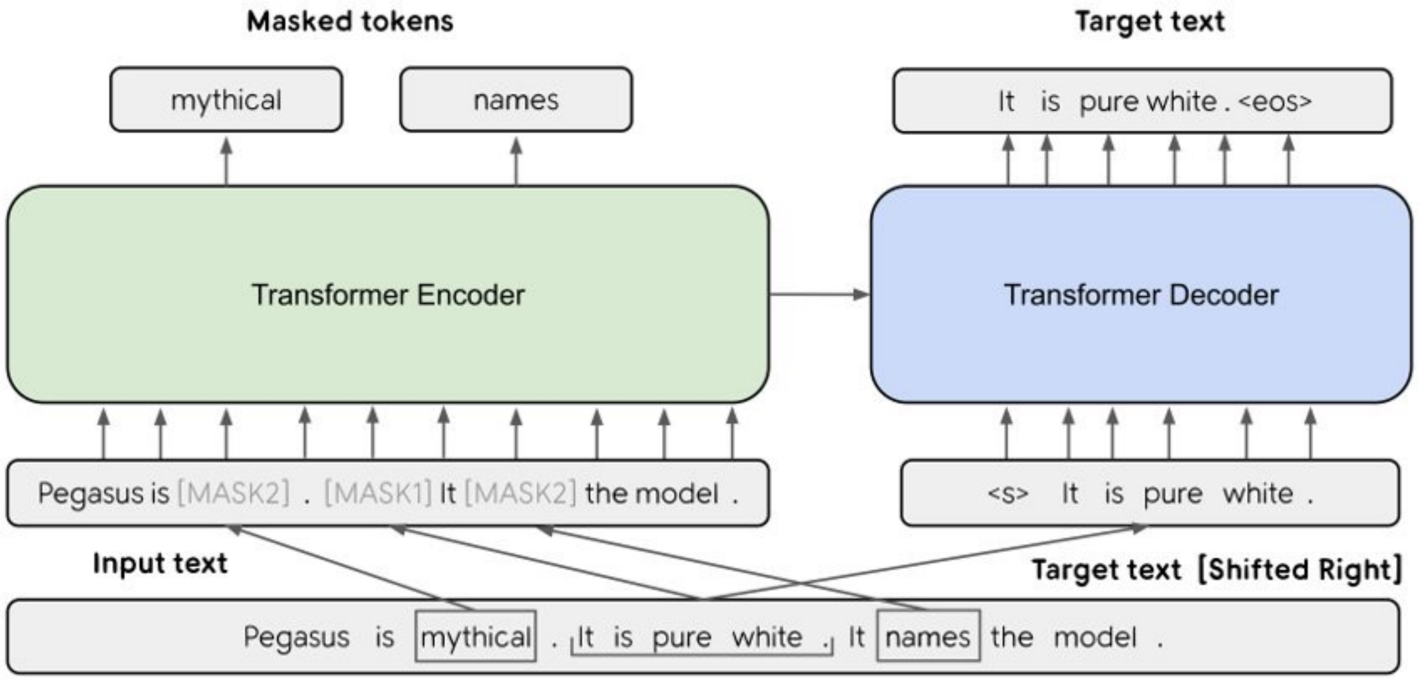
	SUMDocS
keywords	79, abbott, god, february, patriot, statement, 13, appeared, natural, 2016
summary	breaking : u.s. supreme court justice antonin scalia found dead at west texas ranch at 79 cbs news (@cbsnews) february 13, 2016 cbs news reported scalia appeared to die of natural causes, according to a u.s. marshals service spokesperson. bush said scalia will be missed. scalia was nominated to the u.s. supreme court in 1986 by president ronald reagan. abbott said scalia set an example for citizens. scalia's legacy is enormous. greg abbott released a statement saturday afternoon, calling scalia a man of god, a patriot and...

Outline

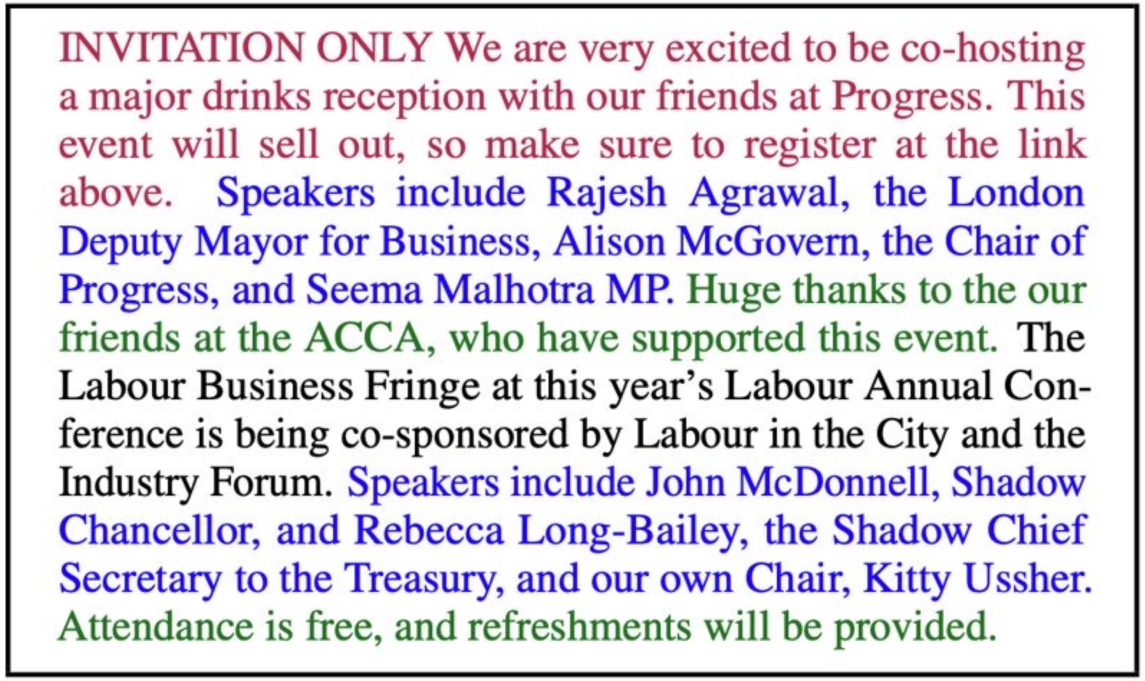
- Aspect-based Sentiment Analysis
- Text Summarization
  - SUMDocs: Extractive Summarization with Background Corpus
  - Pre-trained Language Models for Summarization
    - Generating Representative Headlines for News Stories [WWW’20]
- Online Story Discovery from News Streams
- Summary & Future Directions

Self-supervised Pre-trained Summarization Model

- PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization (ICML’20)
- Transformer based encoder decoder framework
- Two Pre-training objectives:
- **Encoder**: masked language model
- **Decoder**: gap sentence generation
  - Choose important sentence by rouge score with remaining sentences in the document

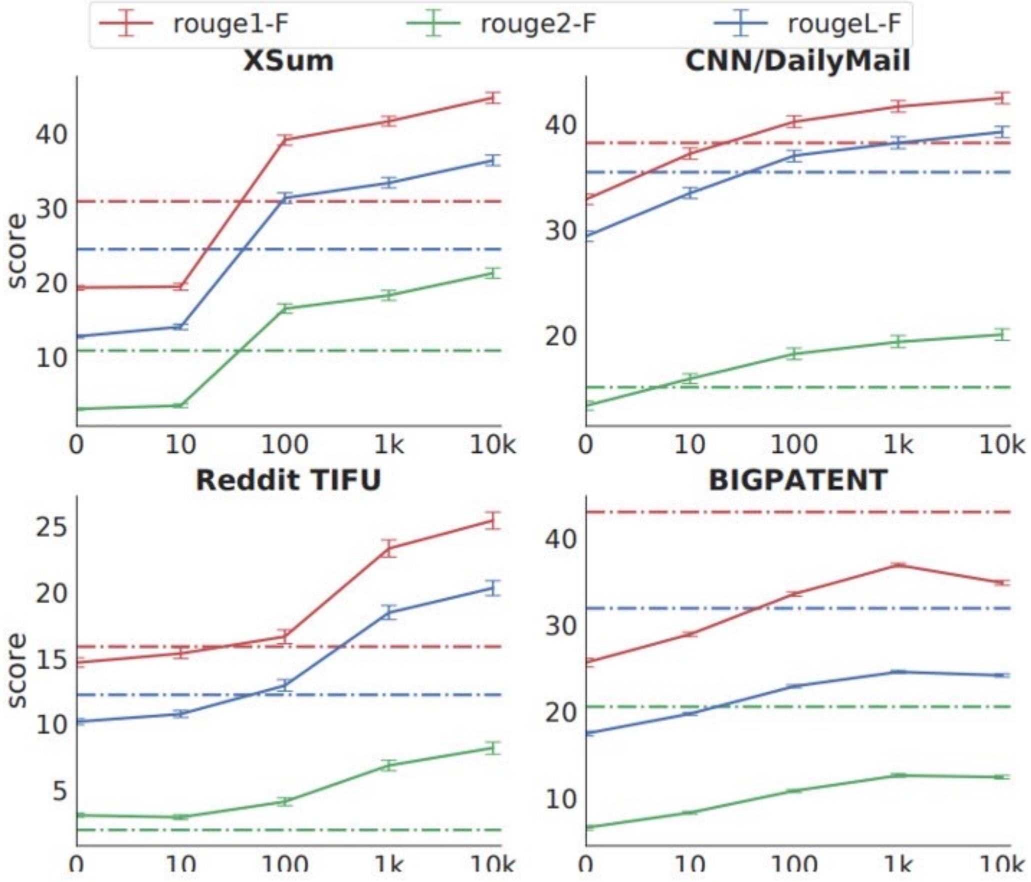

Masked tokens
mythical
names
Target text
It is pure white . <eos>
Transformer Encoder
Transformer Decoder
Pegasus is [MASK2] . [MASK1] It [MASK2] the model .
<s> It is pure white .
Input text
Target text [Shifted Right]
Pegasus is mythical . It is pure white . It names the model .

Selected Sentence for Gap Sentence Generation



INVITATION ONLY We are very excited to be co-hosting a major drinks reception with our friends at Progress. This event will sell out, so make sure to register at the link above. Speakers include Rajesh Agrawal, the London Deputy Mayor for Business, Alison McGovern, the Chair of Progress, and Seema Malhotra MP. Huge thanks to the our friends at the ACCA, who have supported this event. The Labour Business Fringe at this year's Labour Annual Conference is being co-sponsored by Labour in the City and the Industry Forum. Speakers include John McDonnell, Shadow Chancellor, and Rebecca Long-Bailey, the Shadow Chief Secretary to the Treasury, and our own Chair, Kitty Ussher. Attendance is free, and refreshments will be provided.

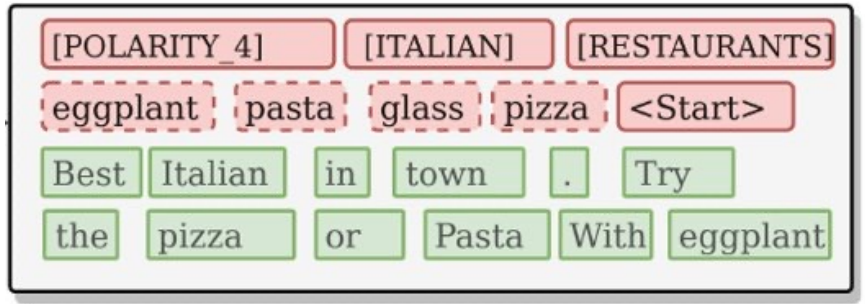
Figure 2: An example of sentences (from the C4 corpus) selected by Random, Lead and Ind-Orig respectively. Best viewed in color.



Fine-tuning with limited supervised samples
Solid: few-shot with pre-trained weights
Dashed: supervised with initial weights

Keyword-Guided Summarization

- Self-Supervised and Controlled Opinion Summarization [EACL'21]
  - Control tokens are used to let the generated summary align with the input documents.
- Inputs to the model:



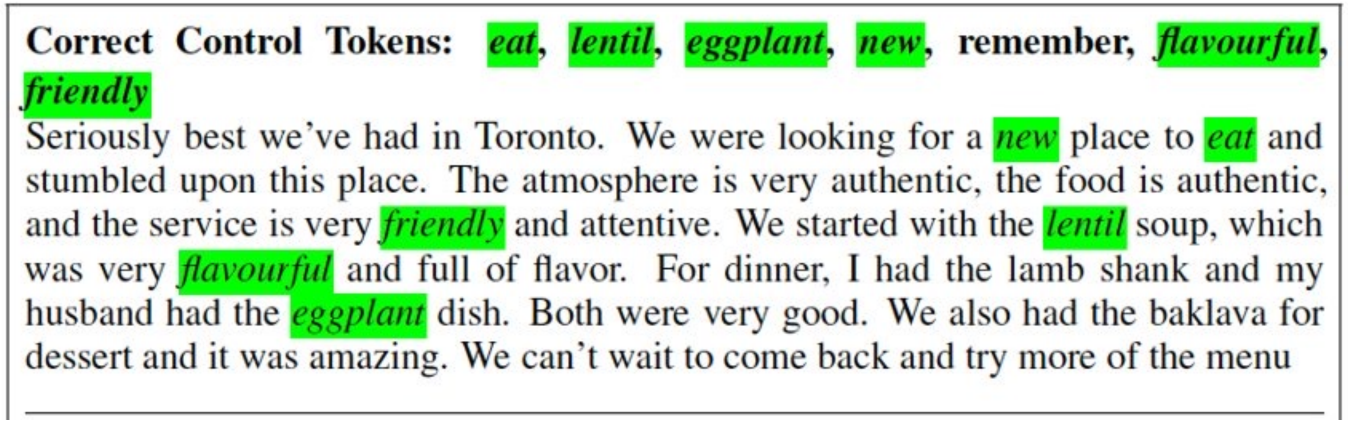
[POLARITY_4] [ITALIAN] [RESTAURANTS]

eggplant pasta glass pizza <Start>

Best Italian in town . Try

the pizza or Pasta With eggplant

- Summary guided by tokens:



Correct Control Tokens: ***eat***, ***lentil***, ***eggplant***, ***new***, **remember**, ***flavourful***, ***friendly***

Seriously best we've had in Toronto. We were looking for a *new* place to *eat* and stumbled upon this place. The atmosphere is very authentic, the food is authentic, and the service is very *friendly* and attentive. We started with the *lentil* soup, which was very *flavourful* and full of flavor. For dinner, I had the lamb shank and my husband had the *eggplant* dish. Both were very good. We also had the baklava for dessert and it was amazing. We can't wait to come back and try more of the menu

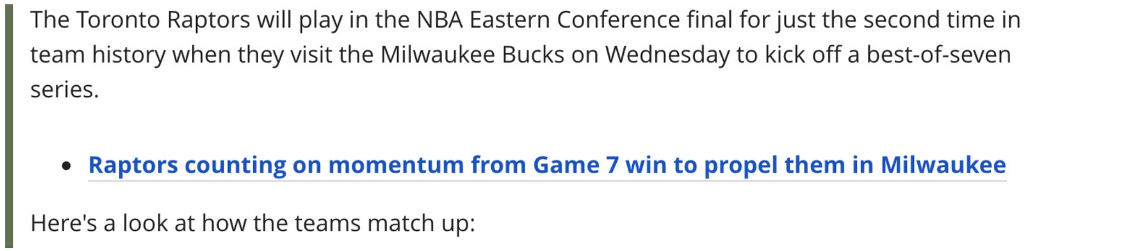
Outline

- Aspect-based Sentiment Analysis
- Text Summarization
  - SUMDocs: Extractive Summarization with Background Corpus
  - Pre-trained Language Models for Summarization
    - Generating Representative Headlines for News Stories [WWW’20]
- Online Story Discovery from News Streams
- Summary & Future Directions

Generating Representative Headlines for News Stories

Raptors vs. Bucks Prediction

Article 1

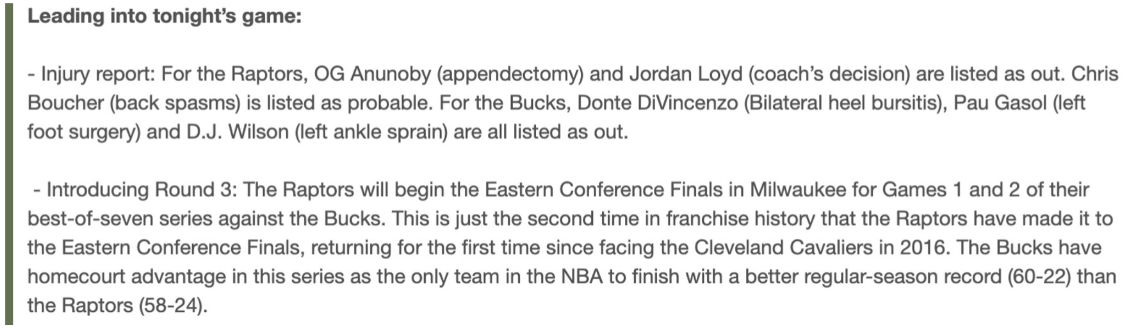


The Toronto Raptors will play in the NBA Eastern Conference final for just the second time in team history when they visit the Milwaukee Bucks on Wednesday to kick off a best-of-seven series.

- **Raptors counting on momentum from Game 7 win to propel them in Milwaukee**

Here's a look at how the teams match up:

Article 2

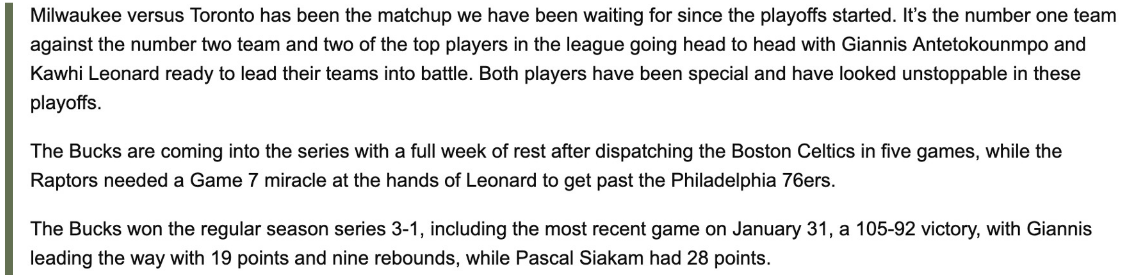


Leading into tonight's game:

- Injury report: For the Raptors, OG Anunoby (appendectomy) and Jordan Loyd (coach's decision) are listed as out. Chris Boucher (back spasms) is listed as probable. For the Bucks, Donte DiVincenzo (Bilateral heel bursitis), Pau Gasol (left foot surgery) and D.J. Wilson (left ankle sprain) are all listed as out.

- Introducing Round 3: The Raptors will begin the Eastern Conference Finals in Milwaukee for Games 1 and 2 of their best-of-seven series against the Bucks. This is just the second time in franchise history that the Raptors have made it to the Eastern Conference Finals, returning for the first time since facing the Cleveland Cavaliers in 2016. The Bucks have homecourt advantage in this series as the only team in the NBA to finish with a better regular-season record (60-22) than the Raptors (58-24).

Article 3



Milwaukee versus Toronto has been the matchup we have been waiting for since the playoffs started. It's the number one team against the number two team and two of the top players in the league going head to head with Giannis Antetokounmpo and Kawhi Leonard ready to lead their teams into battle. Both players have been special and have looked unstoppable in these playoffs.

The Bucks are coming into the series with a full week of rest after dispatching the Boston Celtics in five games, while the Raptors needed a Game 7 miracle at the hands of Leonard to get past the Philadelphia 76ers.

The Bucks won the regular season series 3-1, including the most recent game on January 31, a 105-92 victory, with Giannis leading the way with 19 points and nine rebounds, while Pascal Siakam had 28 points.

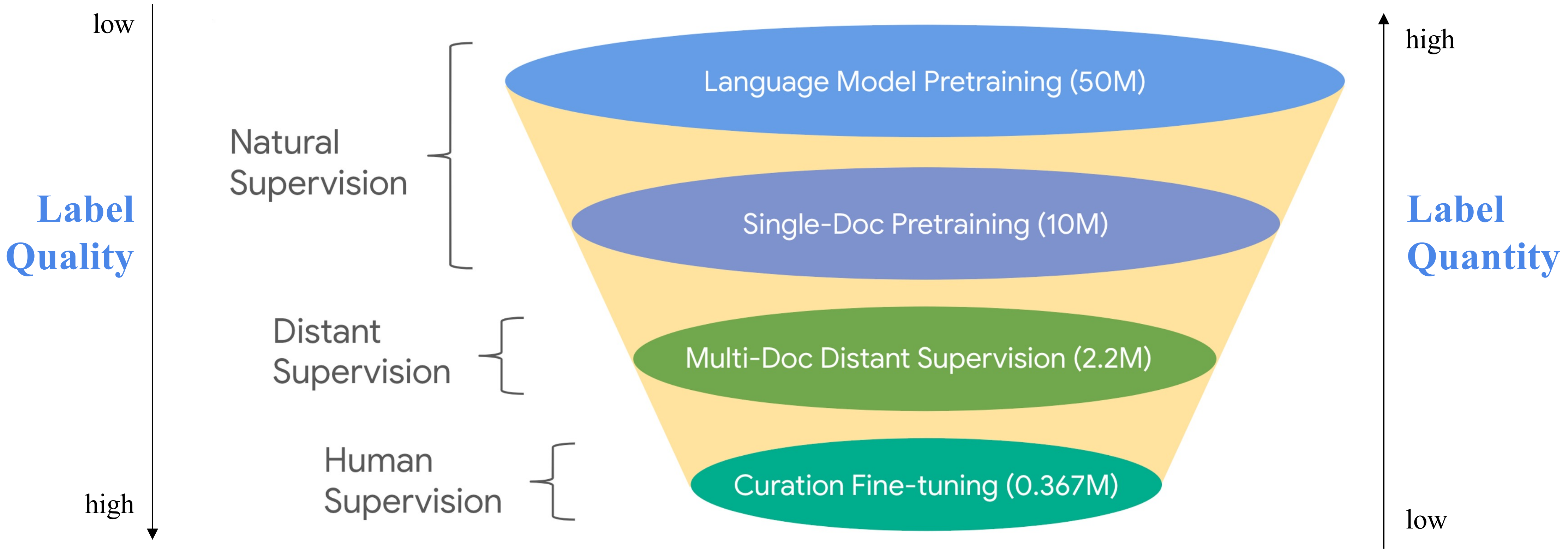
The NewSHead Dataset

- For standard research and evaluation
  - Release the first dataset for **news** story **head**line generation
  - 3-5 news articles per story; Label: human edition + validation
  - 357K stories, >1M articles, 20x larger than the biggest MDS dataset
- Still **a drop** in the **ocean** compared with massive unlabeled news (**50M** articles)!

The NewSHead$_{\text{heuristic}}$ Dataset

- Heuristically generated from unlabeled news articles
  - Cluster news articles into news stories by embedding (same as NewSHead)
    - Label: select an existing article title from the cluster as story headline
    - train a title scorer: given the content of an article and a title, predict whether they match
    - the scorer can be directly trained from existing article/title pairs: no human annotation
    - given all article titles in the cluster, rank them by average matching score with other articles
    - the top article title is representative: can match all articles well
    - prune those under threshold, too long or too short labels
  - Leading to 2.2M (6x larger) news stories with free-to-get labels
  - Question: how far can we go without expensive manual story headline annotations?
- Propose: a three-level pretraining framework

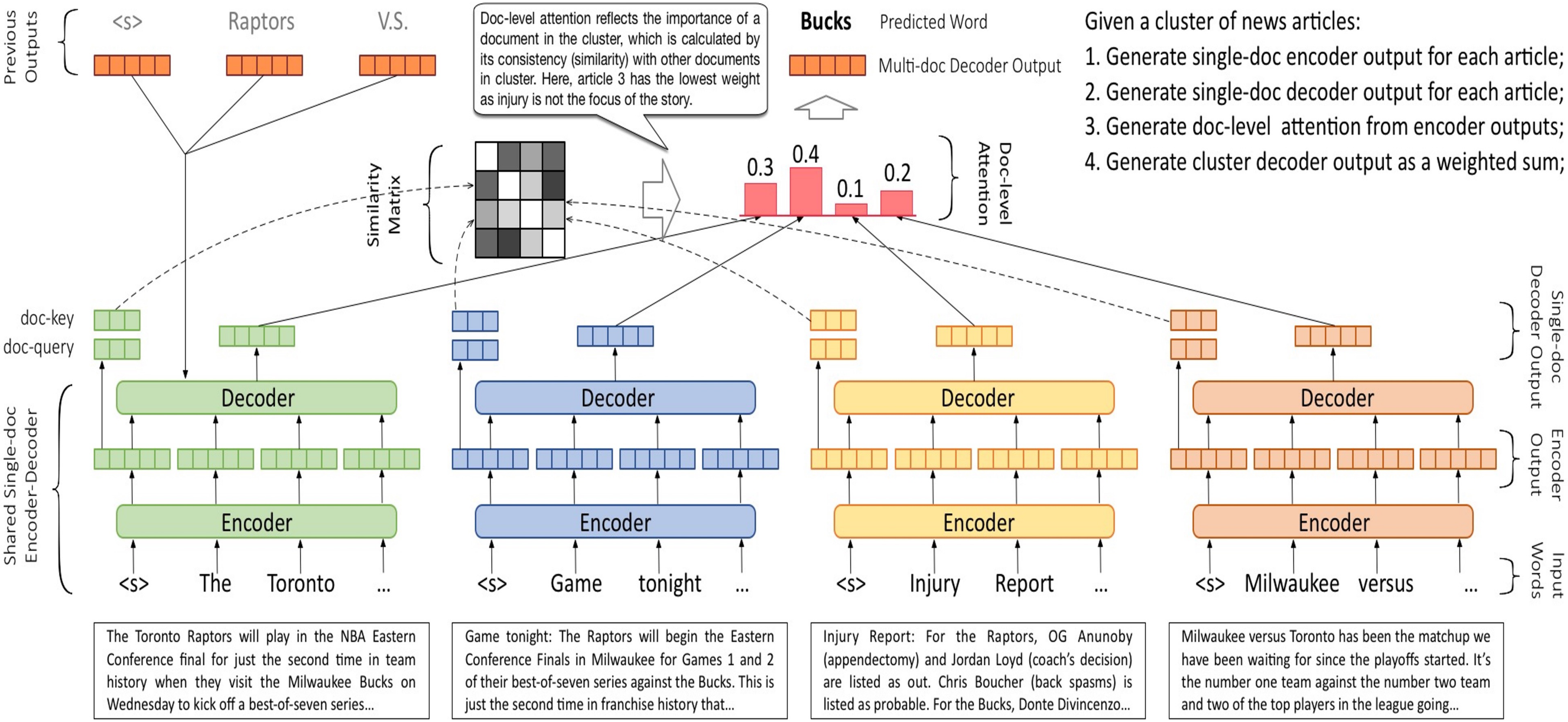
Multi-level Pretraining


low
Label Quality
high
Natural Supervision
Distant Supervision
Human Supervision
Language Model Pretraining (50M)
Single-Doc Pretraining (10M)
Multi-Doc Distant Supervision (2.2M)
Curation Fine-tuning (0.367M)
high
Label Quantity
low

Framework

Given a cluster of news articles:

1. Generate single-doc encoder output for each article;
2. Generate single-doc decoder output for each article;
3. Generate doc-level attention from encoder outputs;
4. Generate cluster decoder output as a weighted sum;



Doc-level attention reflects the importance of a document in the cluster, which is calculated by its consistency (similarity) with other documents in cluster. Here, article 3 has the lowest weight as injury is not the focus of the story.

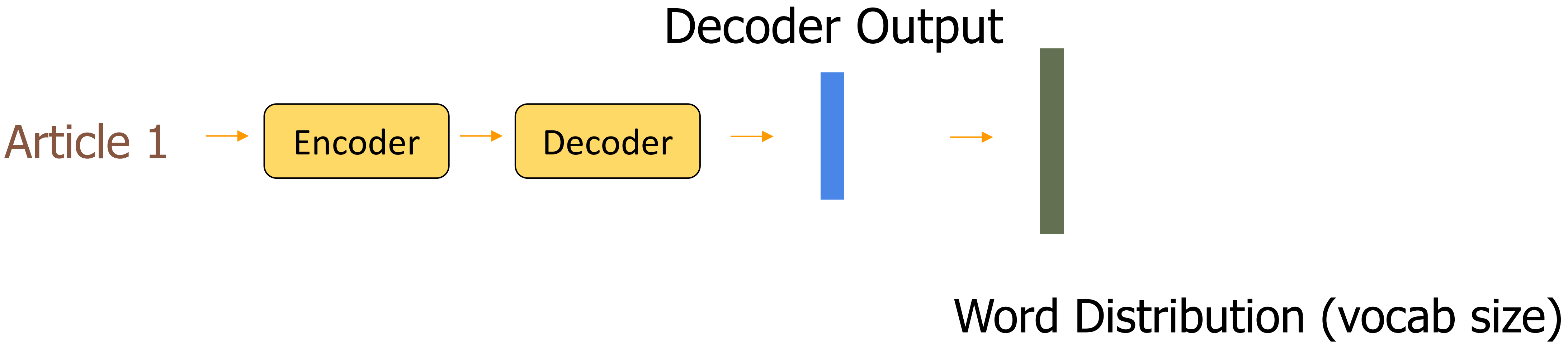
The Toronto Raptors will play in the NBA Eastern Conference final for just the second time in team history when they visit the Milwaukee Bucks on Wednesday to kick off a best-of-seven series...

Game tonight: The Raptors will begin the Eastern Conference Finals in Milwaukee for Games 1 and 2 of their best-of-seven series against the Bucks. This is just the second time in franchise history that...

Injury Report: For the Raptors, OG Anunoby (appendectomy) and Jordan Loyd (coach's decision) are listed as out. Chris Boucher (back spasms) is listed as probable. For the Bucks, Donte Divincenzo...

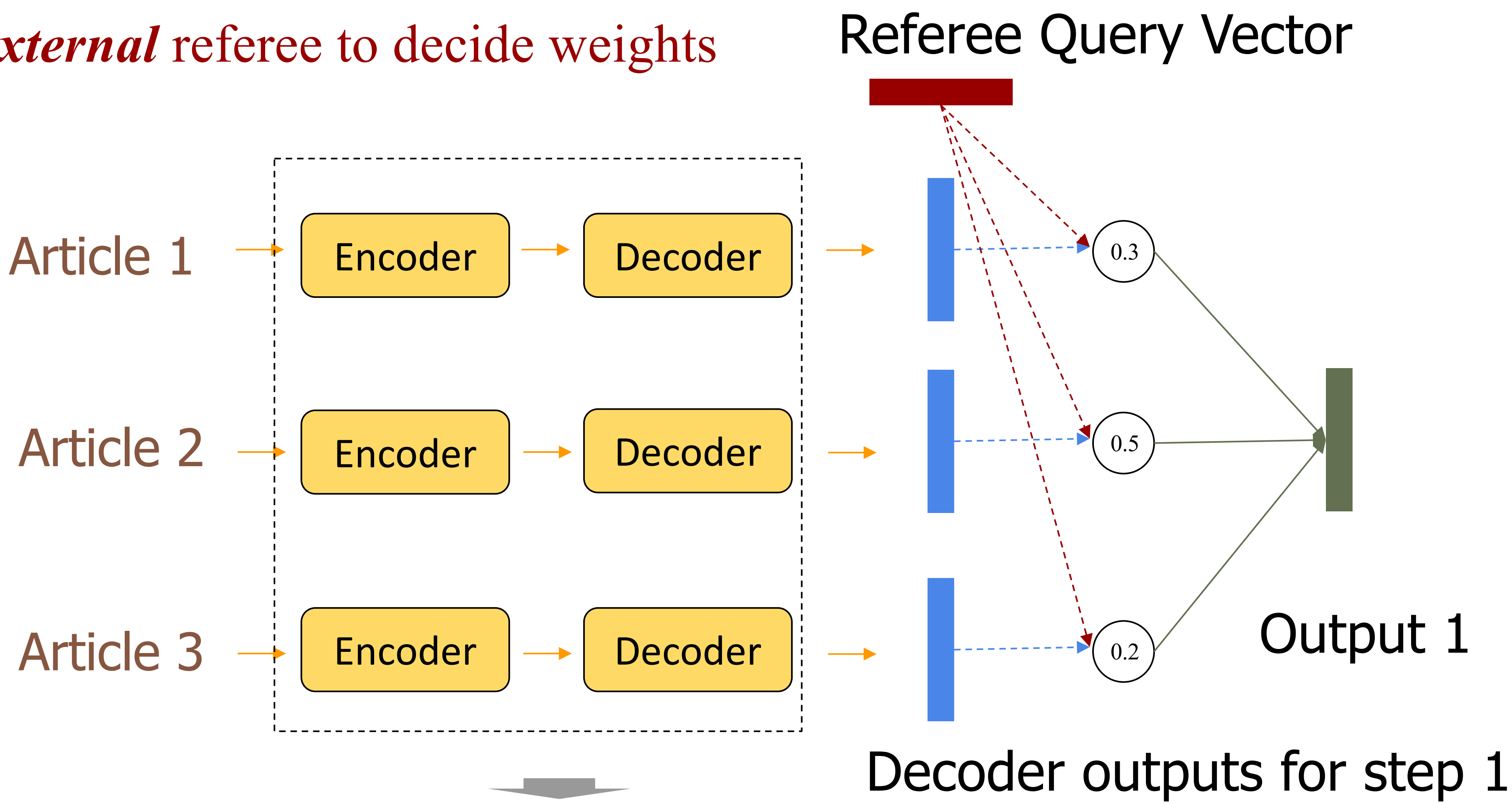
Milwaukee versus Toronto has been the matchup we have been waiting for since the playoffs started. It's the number one team against the number two team and two of the top players in the league going...

Architecture: Single-doc Headline Generation



Doc-level Attention 1: Referee Attention

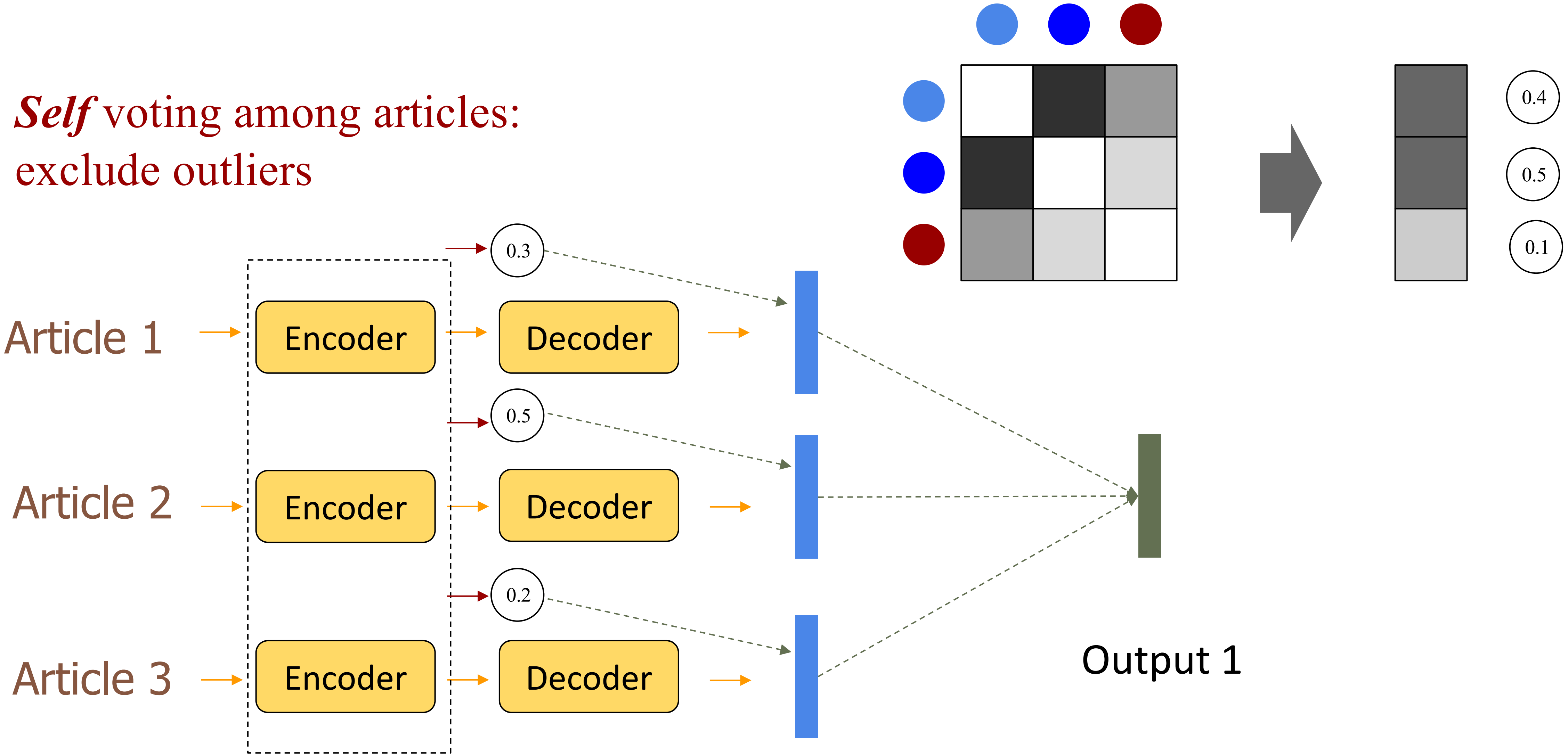
External referee to decide weights



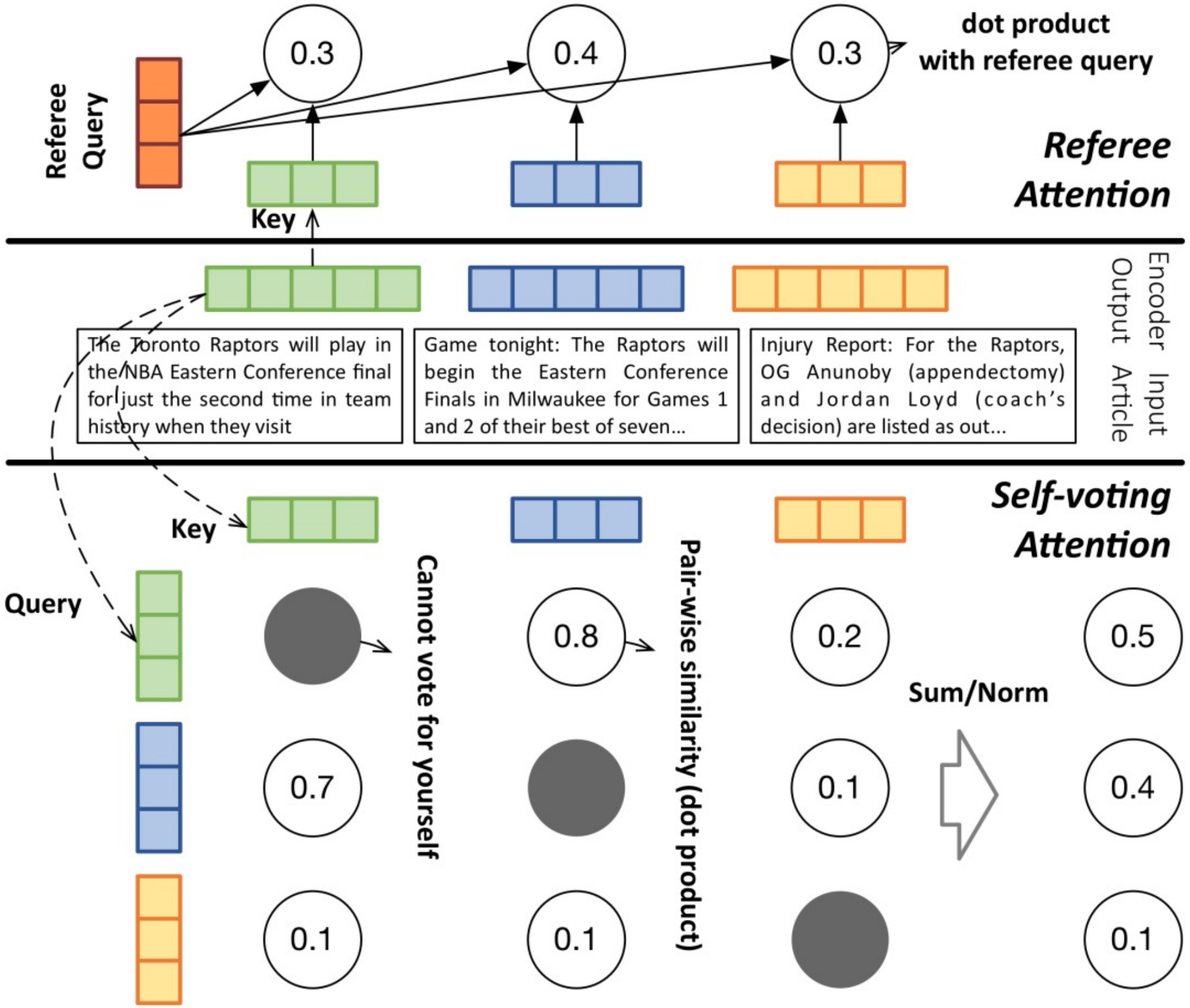
Keep the architecture of single-doc encoder decoder to fully leverage pretraining

Doc-level Attention 2: Self-voted Attention

Self voting among articles:
exclude outliers



Doc-Attention: Comparison


Referee Query
0.3
0.4
0.3
dot product with referee query
Referee Attention
Key
Encoder Output
Input Article
The Toronto Raptors will play in the NBA Eastern Conference final for just the second time in team history when they visit
Game tonight: The Raptors will begin the Eastern Conference Finals in Milwaukee for Games 1 and 2 of their best of seven...
Injury Report: For the Raptors, OG Anunoby (appendectomy) and Jordan Loyd (coach's decision) are listed as out...
Self-voting Attention
Key
Query
Cannot vote for yourself
0.8
Pair-wise similarity (dot product)
0.2
0.7
0.1
0.1
0.1
Sum/Norm
0.5
0.4
0.1

Performance Comparison

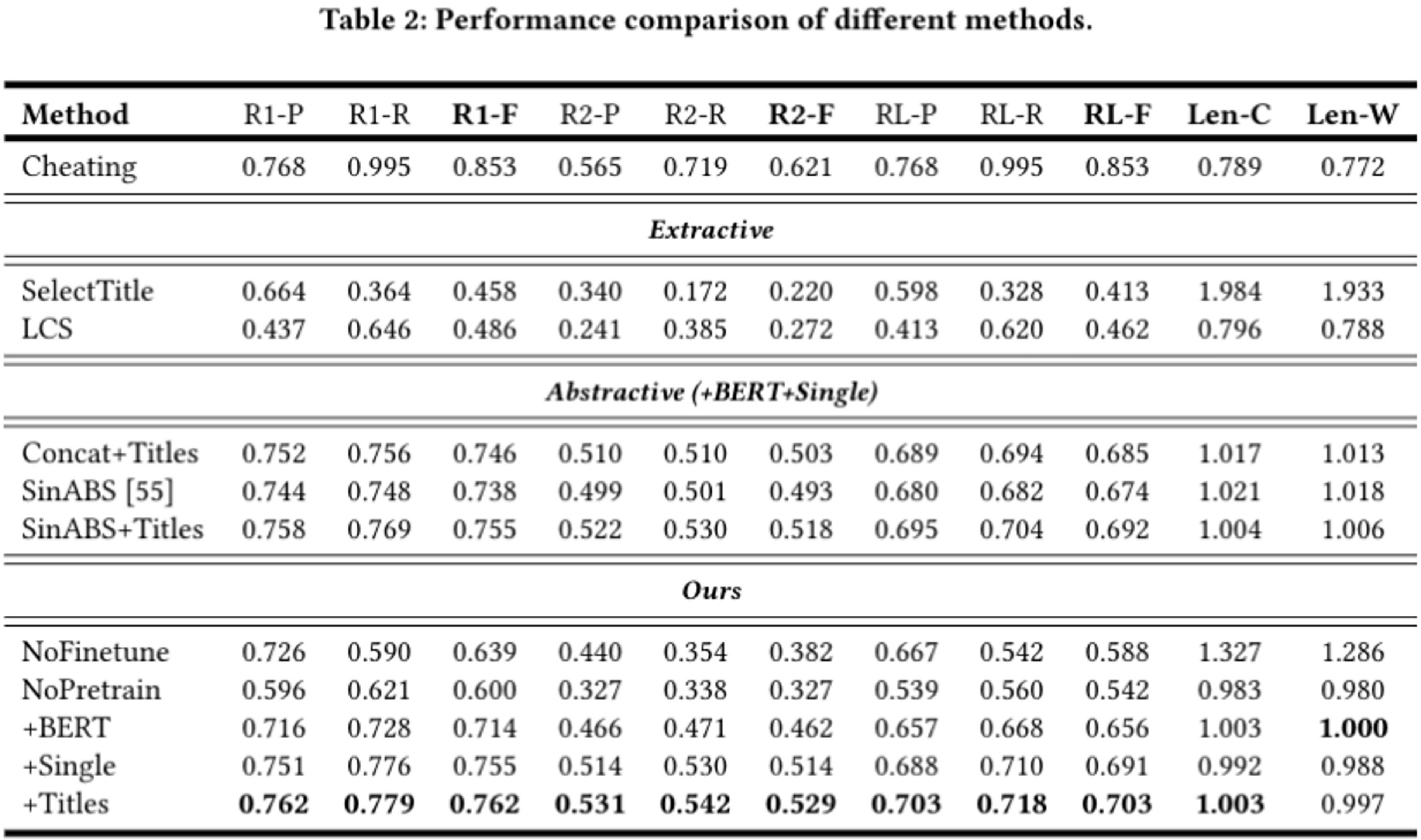
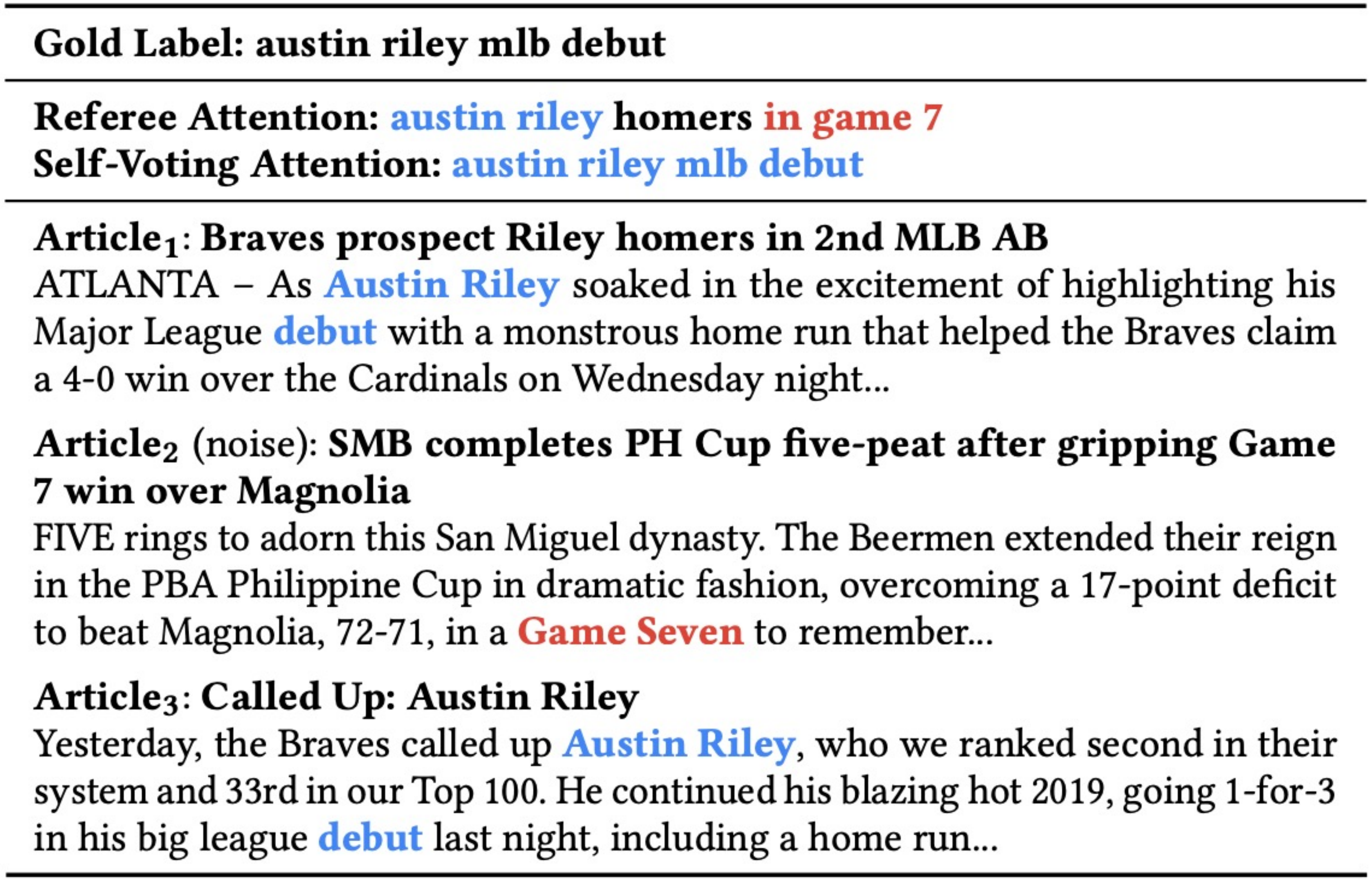


Table 2: Performance comparison of different methods.

Method	R1-P	R1-R	**R1-F**	R2-P	R2-R	**R2-F**	RL-P	RL-R	**RL-F**	**Len-C**	**Len-W**
Cheating	0.768	0.995	0.853	0.565	0.719	0.621	0.768	0.995	0.853	0.789	0.772
Extractive											
SelectTitle	0.664	0.364	0.458	0.340	0.172	0.220	0.598	0.328	0.413	1.984	1.933
LCS	0.437	0.646	0.486	0.241	0.385	0.272	0.413	0.620	0.462	0.796	0.788
Abstractive (+BERT+Single)											
Concat+Titles	0.752	0.756	0.746	0.510	0.510	0.503	0.689	0.694	0.685	1.017	1.013
SinABS [55]	0.744	0.748	0.738	0.499	0.501	0.493	0.680	0.682	0.674	1.021	1.018
SinABS+Titles	0.758	0.769	0.755	0.522	0.530	0.518	0.695	0.704	0.692	1.004	1.006
Ours											
NoFinetune	0.726	0.590	0.639	0.440	0.354	0.382	0.667	0.542	0.588	1.327	1.286
NoPretrain	0.596	0.621	0.600	0.327	0.338	0.327	0.539	0.560	0.542	0.983	0.980
+BERT	0.716	0.728	0.714	0.466	0.471	0.462	0.657	0.668	0.656	1.003	**1.000**
+Single	0.751	0.776	0.755	0.514	0.530	0.514	0.688	0.710	0.691	0.992	0.988
+Titles	**0.762**	**0.779**	**0.762**	**0.531**	**0.542**	**0.529**	**0.703**	**0.718**	**0.703**	**1.003**	0.997

A Case Study on Noisy Dataset



Gold Label: austin riley mlb debut

Referee Attention: austin riley homers in game 7
Self-Voting Attention: austin riley mlb debut

Article$_1$: Braves prospect Riley homers in 2nd MLB AB
ATLANTA – As **Austin Riley** soaked in the excitement of highlighting his Major League **debut** with a monstrous home run that helped the Braves claim a 4-0 win over the Cardinals on Wednesday night...

Article$_2$ (noise): **SMB completes PH Cup five-peat after gripping Game 7 win over Magnolia**
FIVE rings to adorn this San Miguel dynasty. The Beermen extended their reign in the PBA Philippine Cup in dramatic fashion, overcoming a 17-point deficit to beat Magnolia, 72-71, in a **Game Seven** to remember...

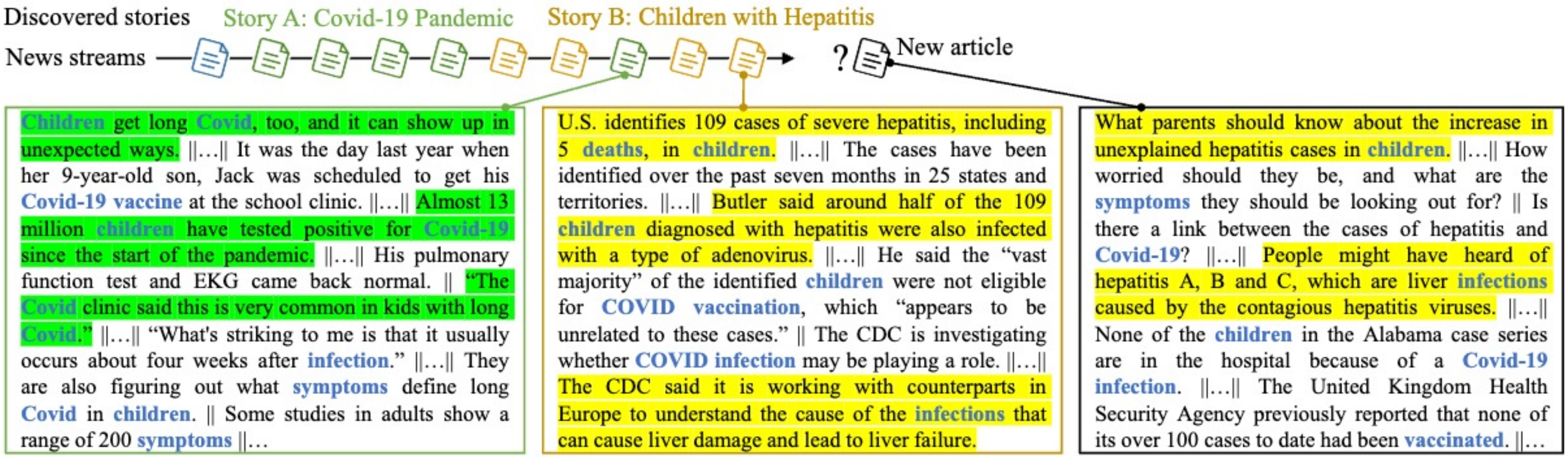
Article$_3$: Called Up: Austin Riley
Yesterday, the Braves called up **Austin Riley**, who we ranked second in their system and 33rd in our Top 100. He continued his blazing hot 2019, going 1-for-3 in his big league **debut** last night, including a home run...

Outline

- ❑ Aspect-based Sentiment Analysis
- ❑ Text Summarization
- ❑ Online Story Discovery from News Streams
  - ❑ SCStory: Self-supervised and Continual Online Story Discovery
- ❑ Summary & Future Directions

Online News Discovery

- Challenges:
  - Diversified local context: Story-indicative semantics need to be identified from the **unstructured** and **diverse** content
  - Evolving global context: Distinctive themes shared among news articles **change over time**
  - No supervision: **Impossible to obtain human annotations** for rapidly and massively published news
  - Efficient adaptation: The model should be **scalable and adaptable** to massive news article streams



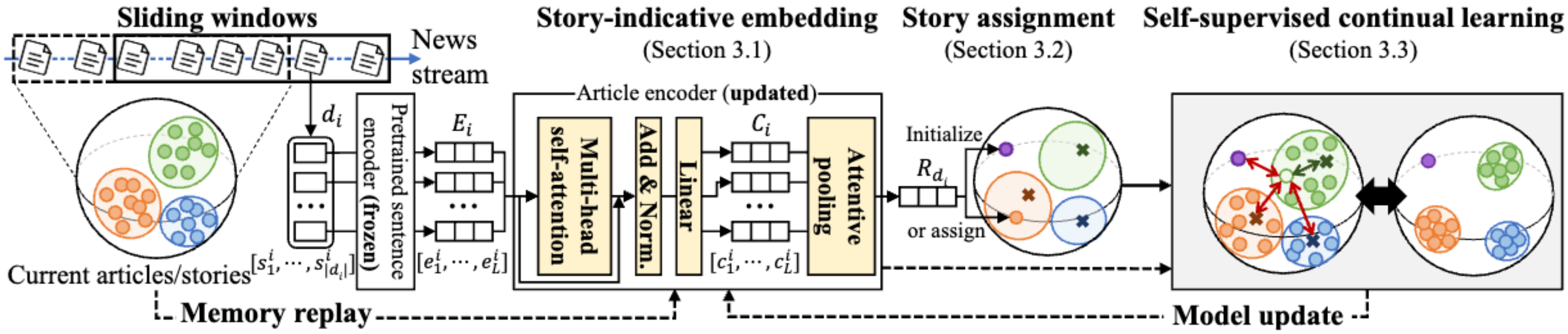
Children get long Covid, too, and it can show up in unexpected ways. ||…|| It was the day last year when her 9-year-old son, Jack was scheduled to get his Covid-19 vaccine at the school clinic. ||…|| Almost 13 million children have tested positive for Covid-19 since the start of the pandemic. ||…|| His pulmonary function test and EKG came back normal. || “The Covid clinic said this is very common in kids with long Covid.” ||…|| “What's striking to me is that it usually occurs about four weeks after infection.” ||…|| They are also figuring out what symptoms define long Covid in children. || Some studies in adults show a range of 200 symptoms ||…

U.S. identifies 109 cases of severe hepatitis, including 5 deaths, in children. ||…|| The cases have been identified over the past seven months in 25 states and territories. ||…|| Butler said around half of the 109 children diagnosed with hepatitis were also infected with a type of adenovirus. ||…|| He said the “vast majority” of the identified children were not eligible for COVID vaccination, which “appears to be unrelated to these cases.” || The CDC is investigating whether COVID infection may be playing a role. ||…|| The CDC said it is working with counterparts in Europe to understand the cause of the infections that can cause liver damage and lead to liver failure.

What parents should know about the increase in unexplained hepatitis cases in children. ||…|| How worried should they be, and what are the symptoms they should be looking out for? || Is there a link between the cases of hepatitis and Covid-19? ||…|| People might have heard of hepatitis A, B and C, which are liver infections caused by the contagious hepatitis viruses. ||…|| None of the children in the Alabama case series are in the hospital because of a Covid-19 infection. ||…|| The United Kingdom Health Security Agency previously reported that none of its over 100 cases to date had been vaccinated. ||…

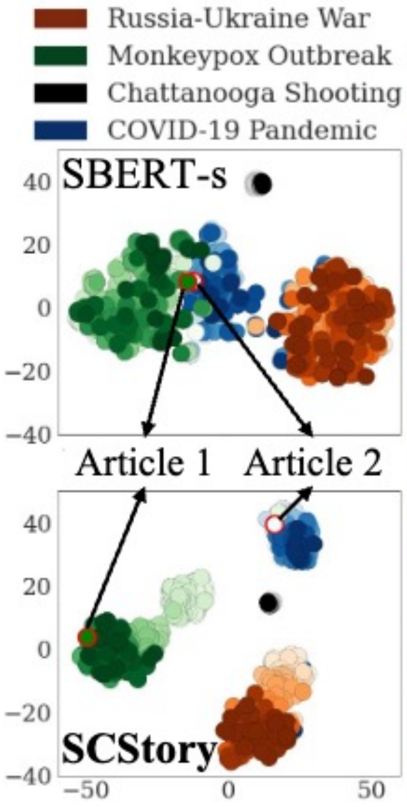
SCStory: The Model

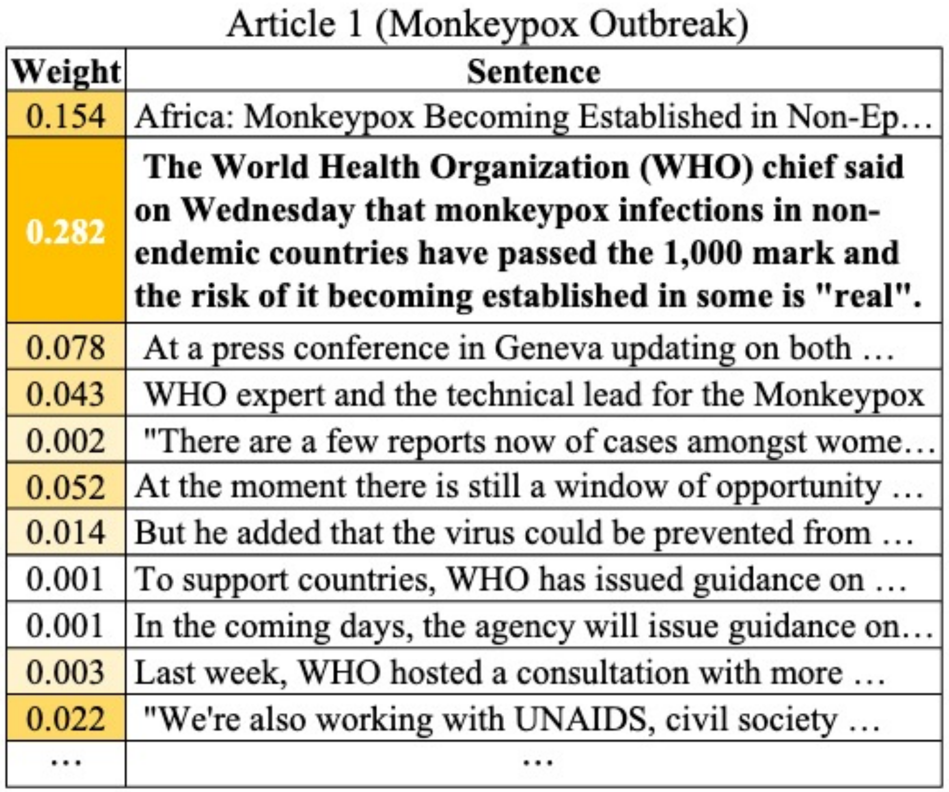
- ❑ Learn story-indicative article representations by modeling the correlations between sentence representations
- ❑ Assign each new article to the most confident current story or form a new story
- ❑ Update the story-indicative embedding module based on the contrastive loss using articles from the current window



SCStory: Case Study

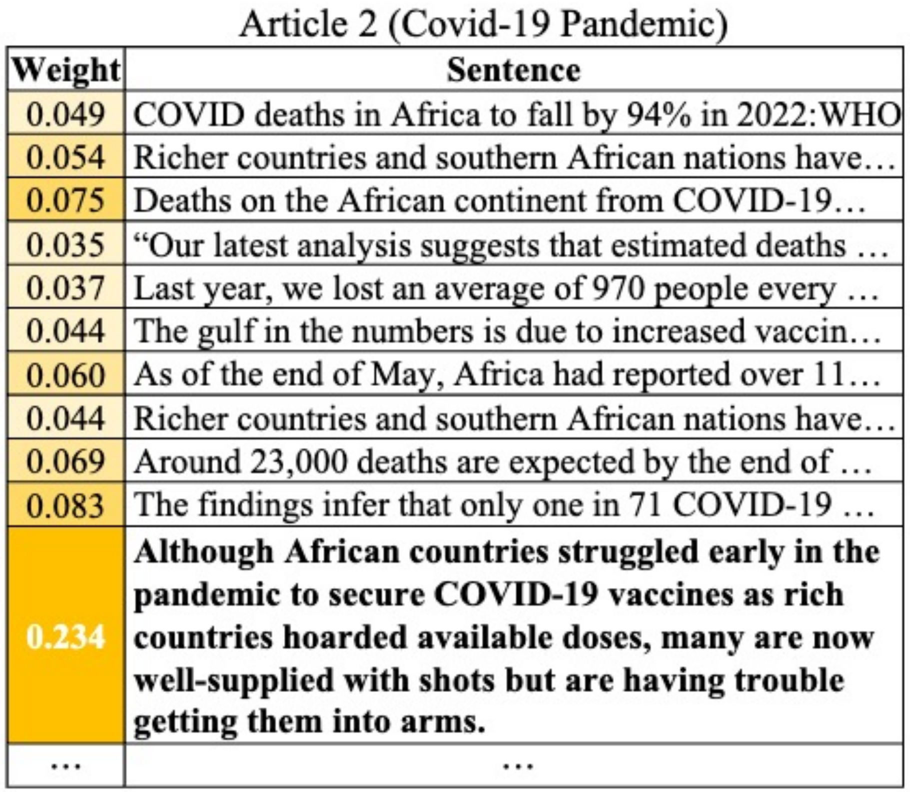
- ❑ Article representations have better separations across different stories
- ❑ Attention weights reflect the discriminativeness of sentences





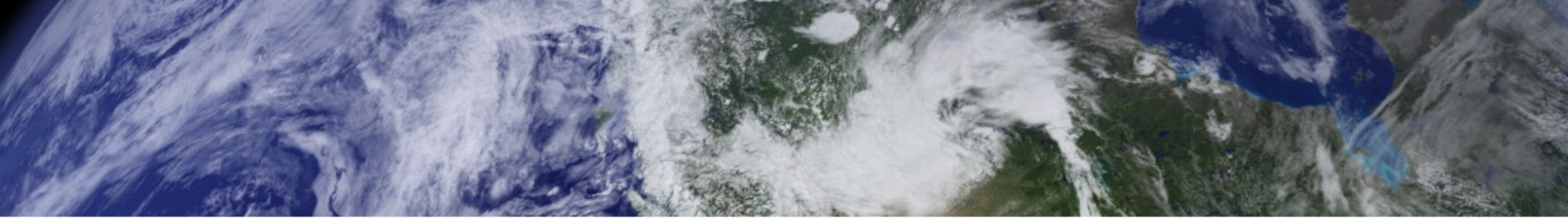
Article 1 (Monkeypox Outbreak)

Weight	Sentence
0.154	Africa: Monkeypox Becoming Established in Non-Ep…
0.282	**The World Health Organization (WHO) chief said on Wednesday that monkeypox infections in non-endemic countries have passed the 1,000 mark and the risk of it becoming established in some is "real".**
0.078	At a press conference in Geneva updating on both …
0.043	WHO expert and the technical lead for the Monkeypox
0.002	"There are a few reports now of cases amongst wome…
0.052	At the moment there is still a window of opportunity …
0.014	But he added that the virus could be prevented from …
0.001	To support countries, WHO has issued guidance on …
0.001	In the coming days, the agency will issue guidance on…
0.003	Last week, WHO hosted a consultation with more …
0.022	"We're also working with UNAIDS, civil society …
…	…



Article 2 (Covid-19 Pandemic)

Weight	Sentence
0.049	COVID deaths in Africa to fall by 94% in 2022:WHO
0.054	Richer countries and southern African nations have…
0.075	Deaths on the African continent from COVID-19…
0.035	"Our latest analysis suggests that estimated deaths …
0.037	Last year, we lost an average of 970 people every …
0.044	The gulf in the numbers is due to increased vaccin…
0.060	As of the end of May, Africa had reported over 11…
0.044	Richer countries and southern African nations have…
0.069	Around 23,000 deaths are expected by the end of …
0.083	The findings infer that only one in 71 COVID-19 …
0.234	**Although African countries struggled early in the pandemic to secure COVID-19 vaccines as rich countries hoarded available doses, many are now well-supplied with shots but are having trouble getting them into arms.**
…	…



Summary & Future Directions

KDD 2022 Tutorial

Adapting Pretrained Representations for Text Mining

Yu Meng, Jiaxin Huang, Yu Zhang, Jiawei Han

Computer Science, University of Illinois at Urbana-Champaign

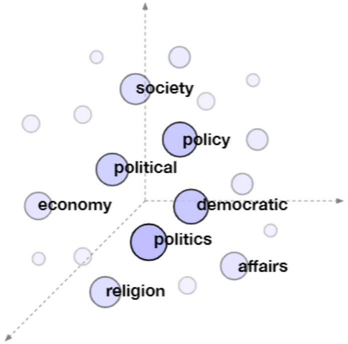
Aug 14, 2022

Our Roadmap of This Tutorial

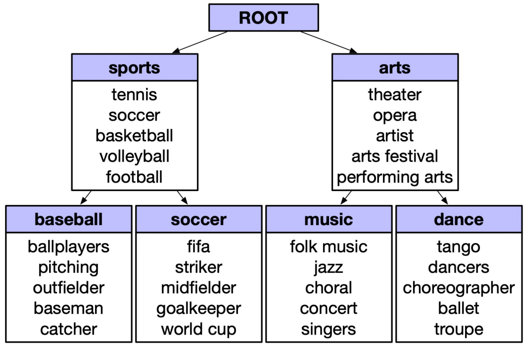
Text Corpus

Existing KB

Part I: Pretrained Language Model



Part II: Text Mining Basics (Phrase/Entity Mining, Taxonomy)



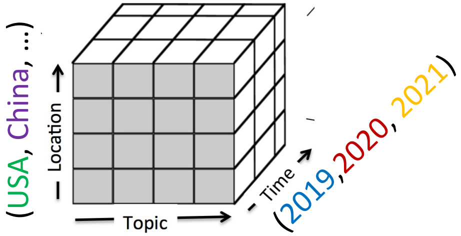
ROOT

sports	arts
tennis soccer basketball volleyball football	theater opera artist arts festival performing arts

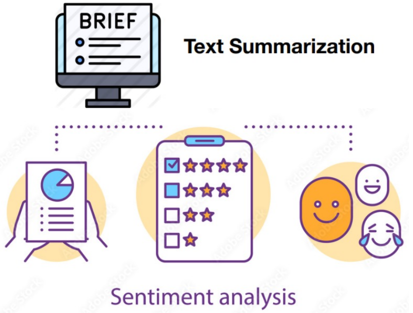
baseball	soccer	music	dance
ballplayers pitching outfielder baseman catcher	fifa striker midfielder goalkeeper world cup	folk music jazz choral concert singers	tango dancers choreographer ballet troupe

Part III: Topic Discovery

Part IV: Weakly-Supervised Text Classification



(Politics, Science, Economy)



Part V: Advanced Text Mining Applications (Sentiment Analysis, Summarization)

Summary: from Unstructured Text to Knowledge

- ❑ Leverage the Power of Text Embedding and Language Models to Transform Unstructured Text into Structured Knowledge
- ❑ Mining Structures from Massive Unstructured Text (Texts → Structures)
  - ❑ Automated Text Representation Learning
  - ❑ Automated Multi-Faceted Taxonomy Construction
  - ❑ Automated Topic Mining
  - ❑ Automated Text Classification for Document Assignment
  - ❑ Automated Comparative Summarization in Multidimensional Text Cube
- ❑ Still a lot of work to do from unstructured text to structured knowledge

Our Journey: From Big Data to Big Structures & Knowledge

Thir

DATA

Concepts

MK

Jiawei Han |

MORGAN & CLAYPOOL PUB

Mining Structu
Factual Know
from Text
An Effort-Light

Xiang Ren
Jiawei Han

SYNTHE
DATA M

Jiawei Ha

MORGAN & CLAYPOOL PUBLISHERS

Multidimensional Mining of Massive Text Data

Chao Zhang
Jiawei Han

Synthesis Lectures on Data Mining and Knowledge Discovery

Jiawei Han, Lise Getoor, Wei Wang, Johannes Gehrke, Robert Grossman, *Series Editors*

MORGAN & CLAYPOOL PUBLISHERS

hrase Mining from
ssive Text and
plications

OVERY

Series Editors

RGAN & CLAYPOOL PUBLISHERS

nt Entity

URES ON
ND KNOWLEDGE DISCOVERY

, Wei Wang, Johannes Gehrke, Robert Grossman, *Series Editors*

Han, Kamber and Pei, Data Mining, 3rd ed. 2011

Yu, Han
Link

Information Networks, 2012
Y. Sun: SIGKDD'13 Dissertation Award

ng and Han, Mining Latent Entity Structures, 2015
. Wang: SIGKDD'15 Dissertation Award

Acknowledgement

- Thanks for the research support from: ARL/NSCTA, NIH, NSF, DHS, ARO, DTRA

Thank you !

Q&A