Part V: Advanced Text Mining Applications Empowered by Embeddings

KDD 2021 Tutorial

On the Power of Pre-Trained Text Representations: Models and Applications in Text Mining

Yu Meng, Jiaxin Huang, Yu Zhang, Jiawei Han

Computer Science, University of Illinois at Urbana-Champaign

August 14, 2020

Outline

- Aspect-based Sentiment Analysis
  - Weakly-Supervised Aspect-Based Sentiment Analysis via Joint Aspect-Sentiment Topic Embedding
- Text Summarization
- 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

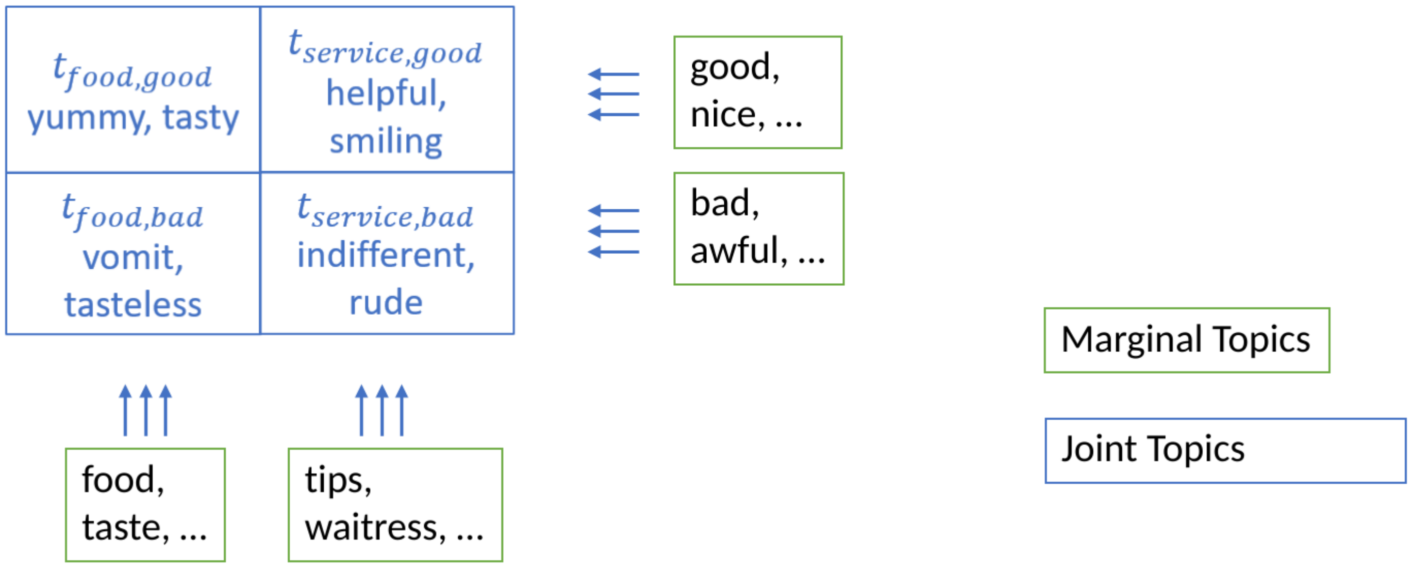
Motivation

- 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

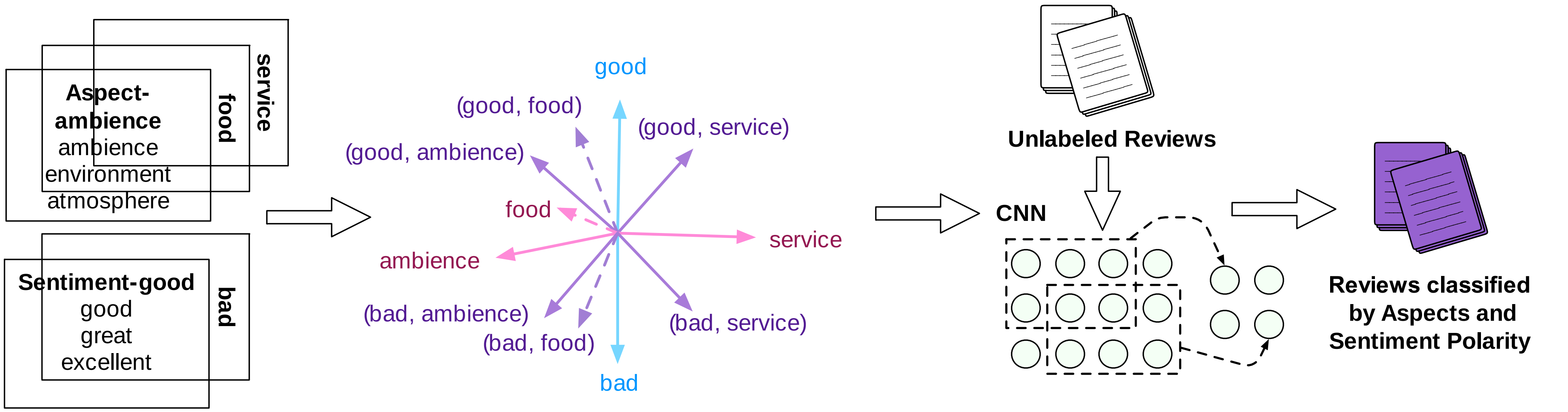
Joint “Sentiment-Aspect” topic



- 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

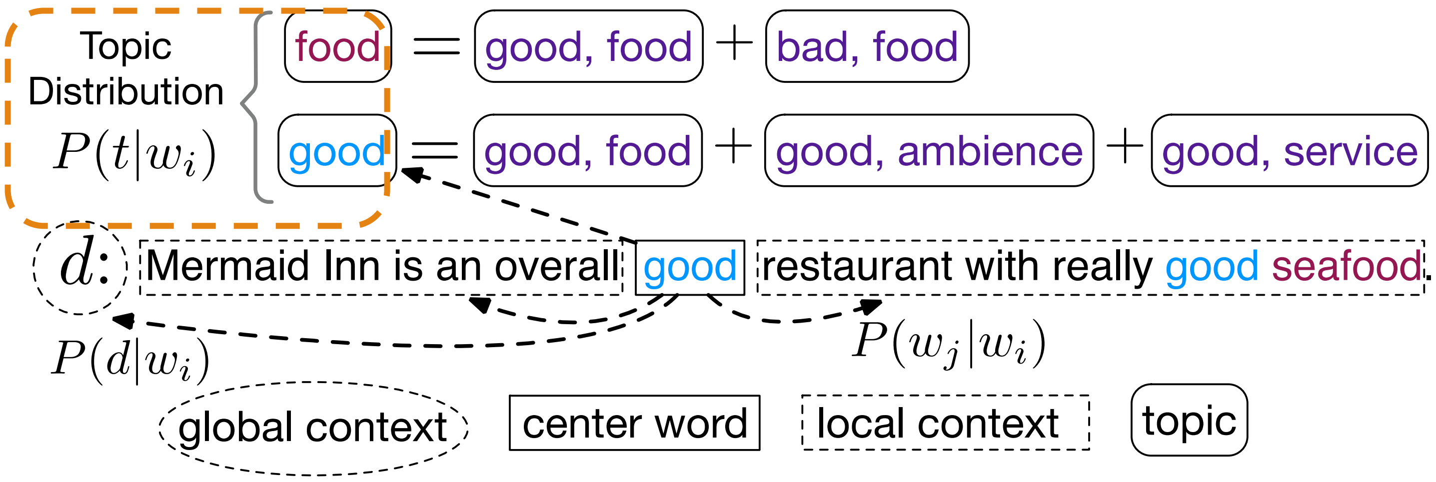
Our Framework

- 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

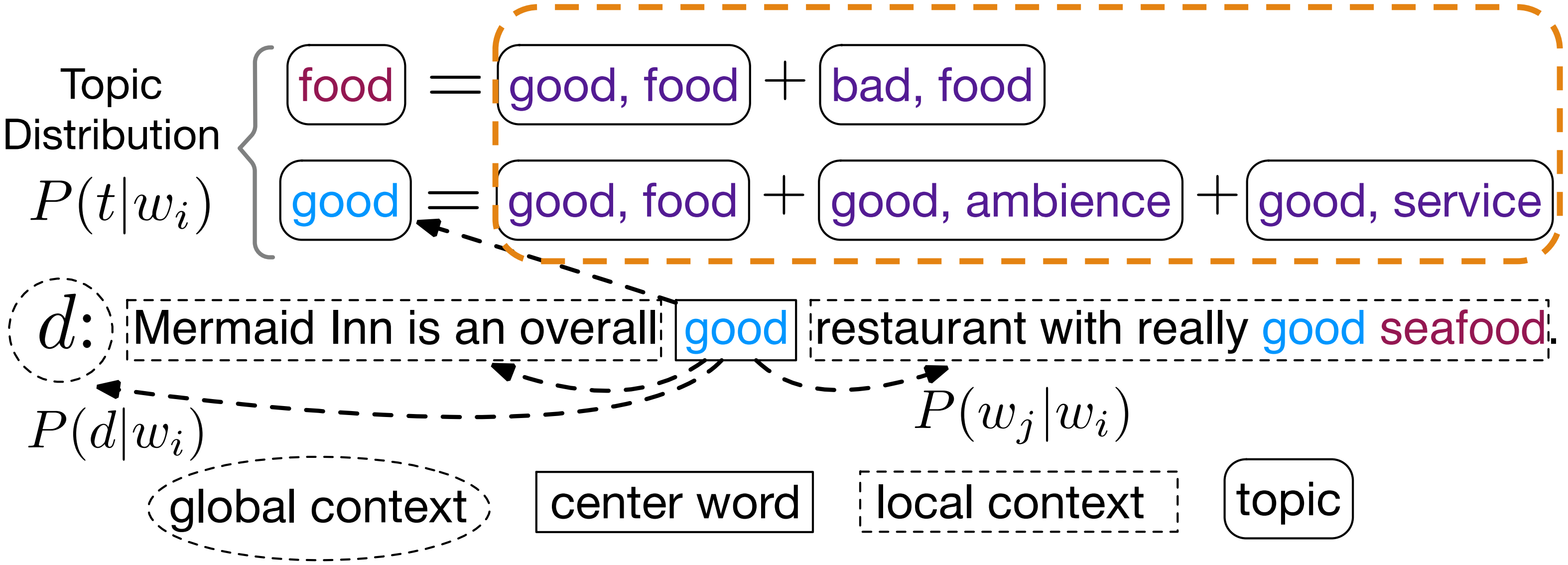


- Regularizing Pure Aspect/Sentiment Topics. We regularize the aspect topic embeddings t_a and sentiment topic embeddings t_s so that different topics are pushed apart
  - Marginal topic regularization:

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

  - 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



- 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

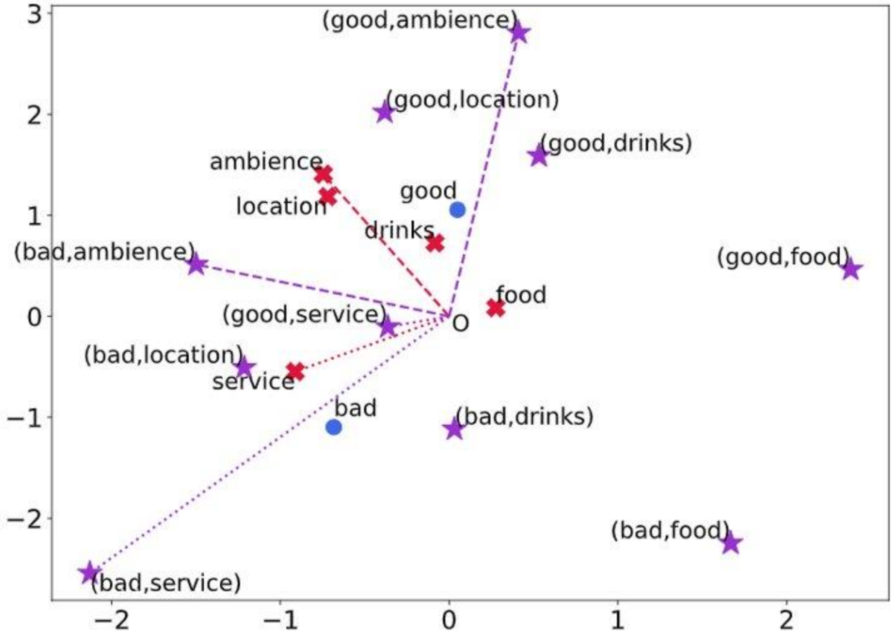
Methods	Restaurant				Laptop			
	Accuracy	Precision	Recall	macro-F1	Accuracy	Precision	Recall	macro-F1
CosSim	61.43	50.12	50.26	42.31	53.84	58.79	54.64	52.18
ABAE(He et al., 2017)	67.34	46.63	50.79	45.31	59.84	59.96	59.60	56.21
CAt(Tulkens and van Cranenburgh, 2020)	66.30	49.20	50.61	46.18	57.95	65.23	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

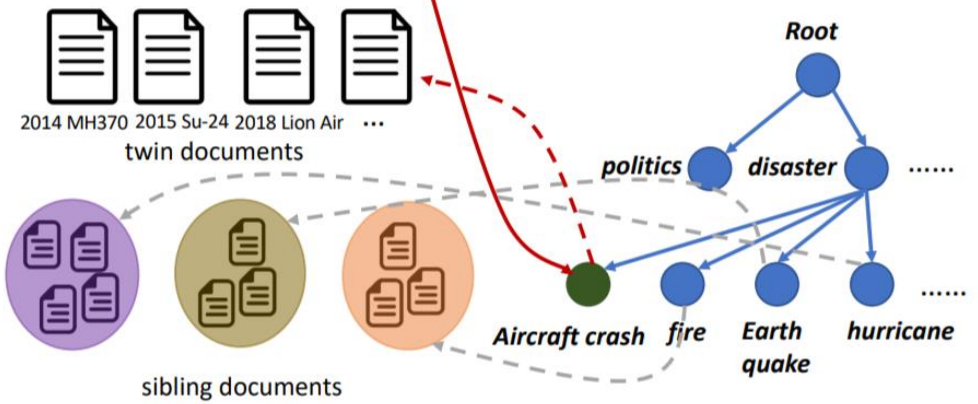
Outline

- Aspect-based Sentiment Analysis
- Text Summarization
  - SUMDocs: Extractive Summarization with Background Corpus
  - Pre-trained Language Models on Summarization
- 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.

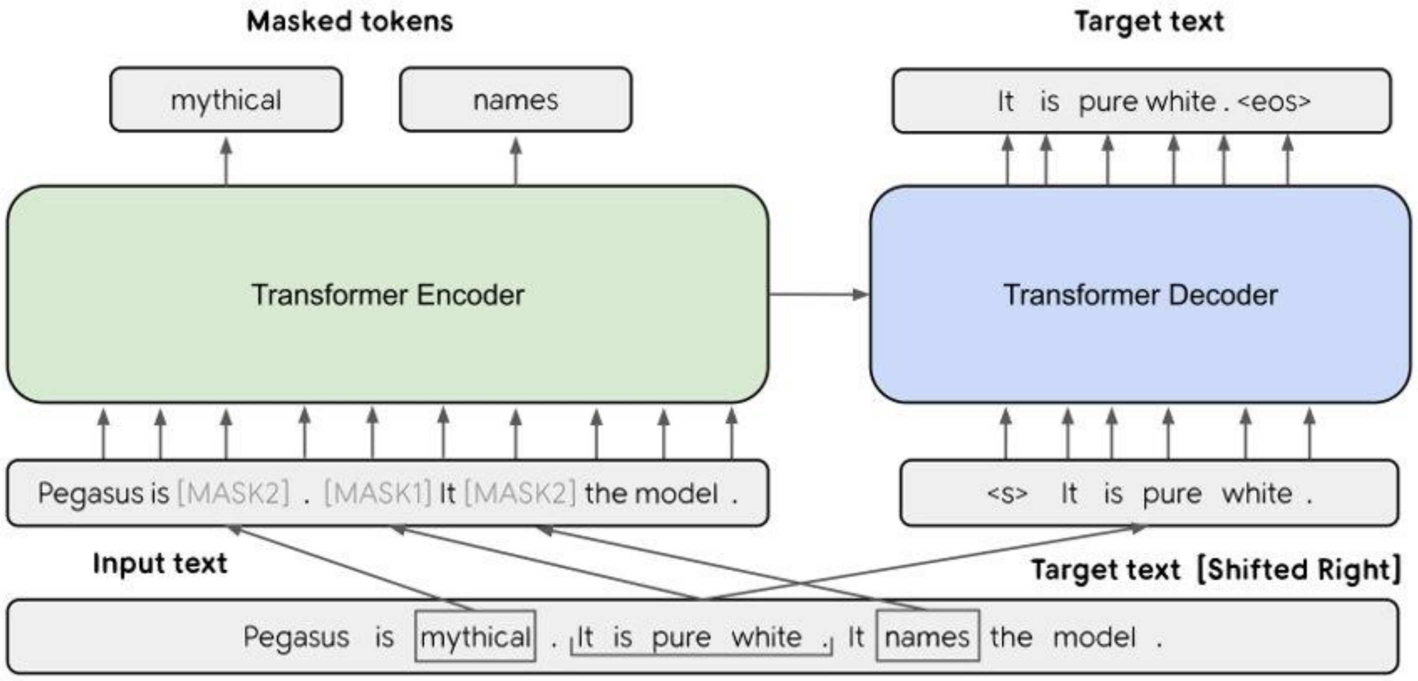
	SUMDocS
keywords	79, abbott, god, february, patriot, statement, 13, appeared, natural, 2016
summary	breaking : u.s. supreme court justice antonin scalia found dead at west texas ranch at 79 cbs news (@cbsnews) february 13, 2016 cbs news reported scalia appeared to die of natural causes, according to 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 on Summarization
- Summary & Future Directions

Self-supervised Pre-trained Summarization Model

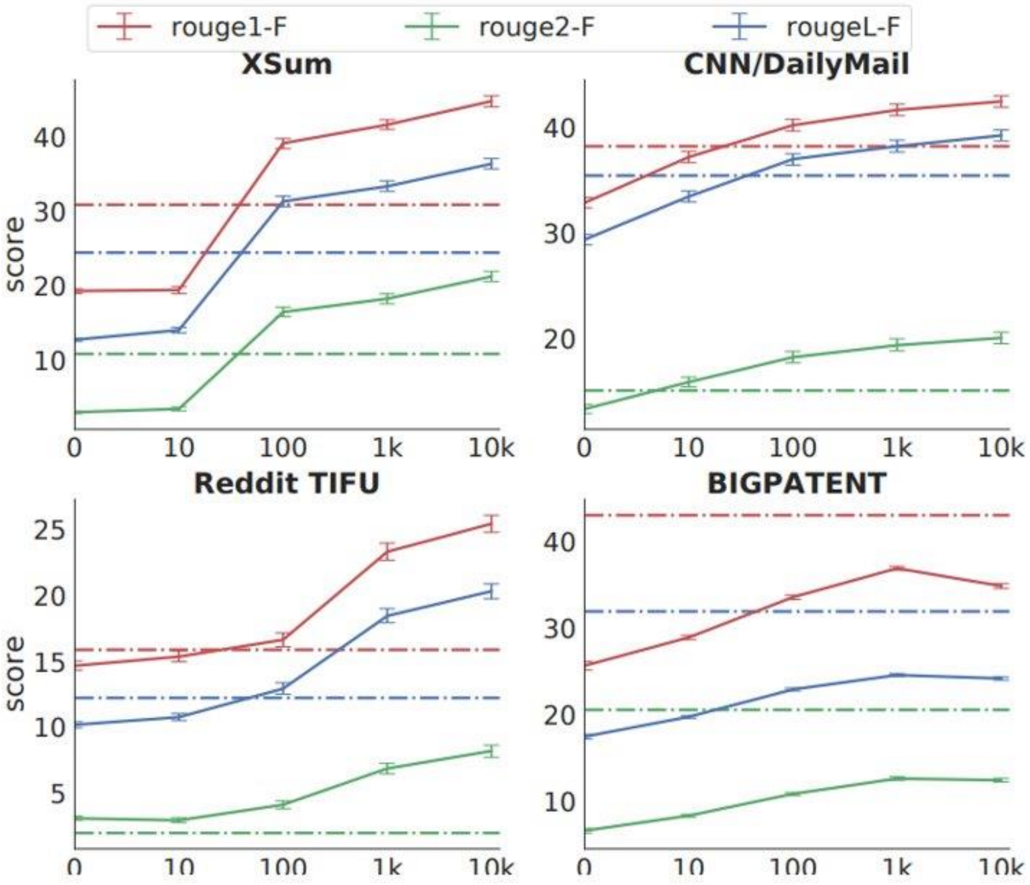
- PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization (ICML'20)
- Transformer based encoder decoder framework
- Two Pre-training objectives:
- **Encoder**: masked language model
- **Decoder**: gap sentence generation
  - Choose important sentence by rouge score with remaining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



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.

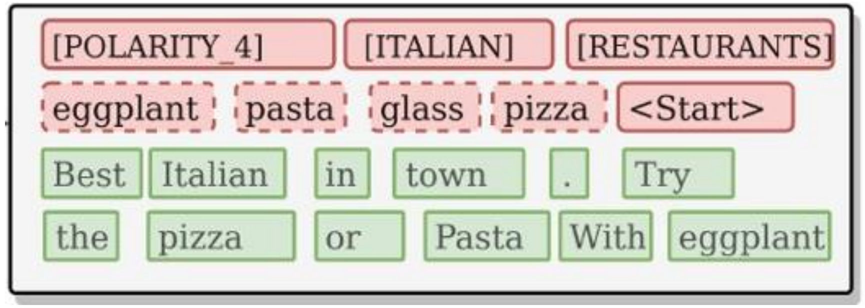
Figure 2: An example of sentences (from the C4 corpus) selected by Random, Lead and Ind-Orig respectively. Best viewed in color.



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:



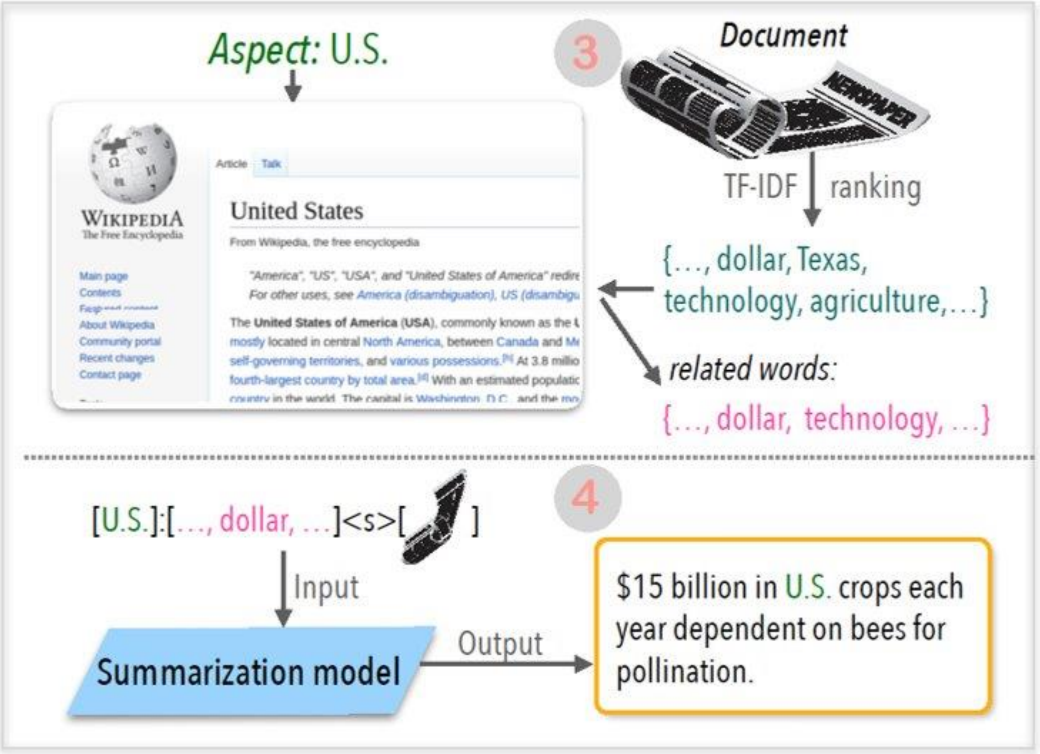
- Summary guided by tokens:

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

Aspect-based Summarization

- Summarizing Text on Any Aspects: A Knowledge-Informed Weakly-Supervised Approach [EMNLP'20]
- uses external knowledge base such as Concept-Net and Wikipedia to construct weak supervision and an aspect modeling scheme.



(1) With generic summary provided, the authors synthesize aspect-based summary by extracting aspect-related words in the generic summary.
(2) To extract aspect-related words in the document, the authors use the words in the wikipedia page of an entity to intersect with highly ranked TF-IDF words in the document.
(3) They fine-tune the pre-trained BART model with
    input: entity and related words in the document.
    output: synthesized summary.

Example Results

Document In an exclusive interview with Breitbart News, Republican presidential nominee Donald Trump blasted Bill **Clinton**'s suggestion that the United States use **Syrian refugees** to rebuild Detroit. The populist **billionaire** denounced **Clinton**'s suggested proposal as "**crazy**" and "unfair" to **American** workers who are already living there and are in need of **jobs**. "It's very unfair to the **people** that are living there. I think it's crazy," Trump told Breitbart on Thursday. "I mean, these people ... ... "There are plenty of people in Detroit who you could almost look at as refugees," Carson said. "I mean, we need to take care of our own people. We need to create jobs for them. " **Clinton**'s suggestion that the U. S. ought to give Detroit jobs to foreign **refugees** came during a February discussion at the Clinton Global Initiative with Chobani **billionaire** and mass migration enthusiast, Hamdi Ulukaya. "The truth is that the big loser in this over the long run is ... ... a pretty good deal. " During the discussion, **Clinton** praised Ulukaya for his efforts to fill his **yogurt** plants with **imported** foreign **refugees**. Ulukaya suggested that the U. S. ought to be taking in more **refugees** and said that he was "proud" of Turkey's decision to accept 2 million **Syrian refugees**. Ulukaya told **Clinton** that **Syrian refugees** "bring **flavors** to the community just like in … Twin Falls, [Idaho]" where Ulukaya's **yogurt** factory is based. Clinton's controversial suggestion that ... ... millions of more illegal immigrants, thousands of more violent crimes, and total chaos and lawlessness. **According** to Pew **polling** data, Hillary Clinton's plan to expand immigration is opposed by at least 83 **percent** of the **American electorate** — **voters** whom Clinton has **suggested** are racist for opposing immigration. **According** to a September 2015 Rasmussen survey, 85 **percent black voters** oppose Clinton's refugee agenda to admit more than 100, 000 Middle Eastern refugees — with less than one **percent** of **black voters** (. 56 **percent**) in **favor** of her refugee plan.

Aspect: vote

Summary: Polls show that at least 83 percent of the U.S. electorate is opposed to expanding immigration and that 85 percent of black voters oppose the plan to admit more than 100,000 middle eastern refugees to the country.

Summary & Future Directions

KDD 2021 Tutorial

On the Power of Pre-Trained Text Representations: Models and Applications in Text Mining

Yu Meng, Jiaxin Huang, Yu Zhang, Jiawei Han

Computer Science, University of Illinois at Urbana-Champaign

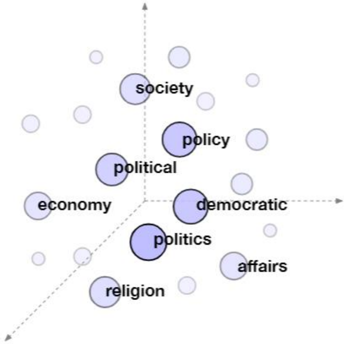
August 14, 2021

Our Roadmap of This Tutorial

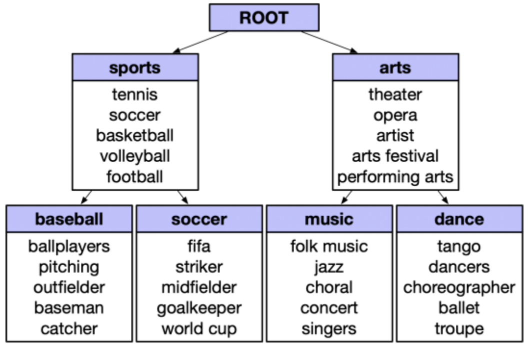
Text Corpus

Existing KB

Part I: Text Representation

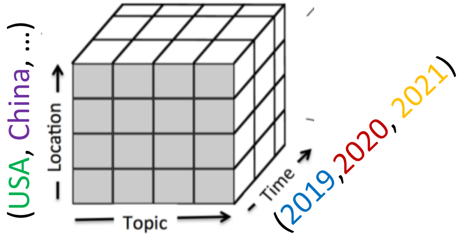


Part II: Text Mining Basics
(Phrase/Entity Mining, Taxonomy)

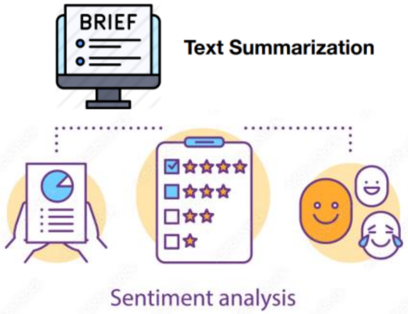


Part III: Topic Discovery

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

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

Yu, Han ... Link...

... Information Networks, 2012

Y. Sun: SIGKDD'13 Dissertation Award

...g 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