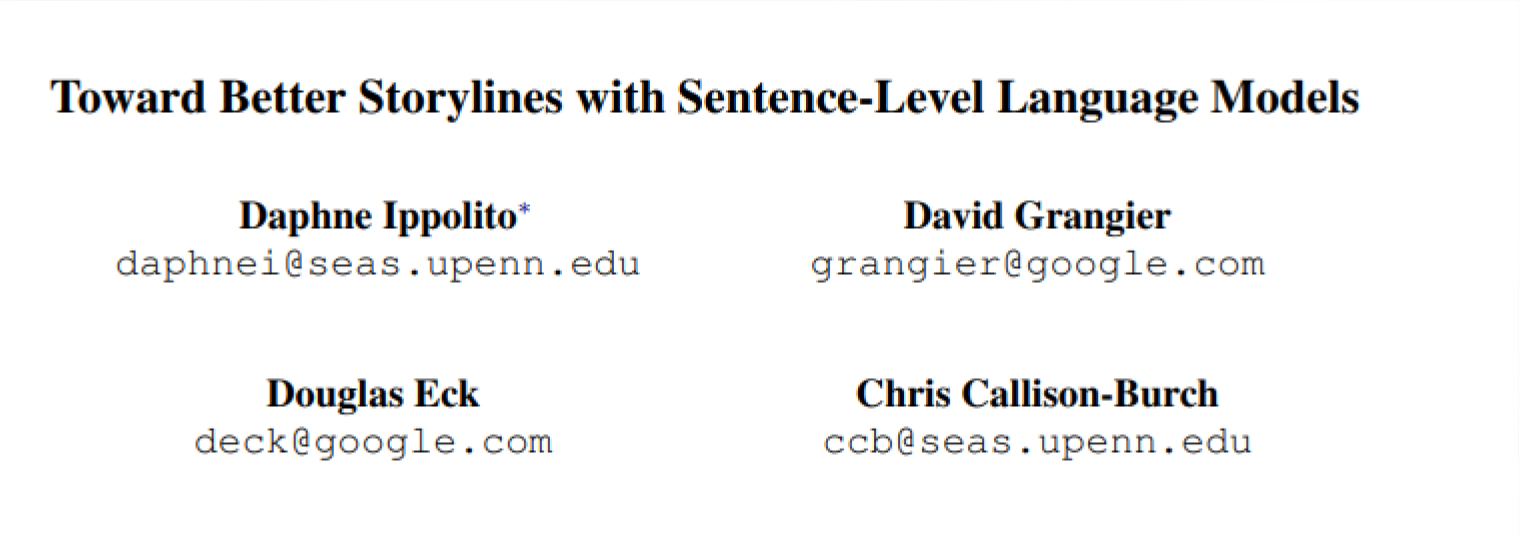
Sentence-level Coherence Modeling

Xiangru Tang

Outline

1. [ACL2020 Google] Toward Better Storylines with Sentence-Level Language Models
2. [ACL2020 Google] Pretraining with Contrastive Sentence Objectives Improves Discourse Performance of Language Models
3. [ACL2020 Stanford] Enabling Language Models to Fill in the Blanks



Toward Better Storylines with Sentence-Level Language Models

Daphne Ippolito[*]
daphnei@seas.upenn.edu

David Grangier
grangier@google.com

Douglas Eck
deck@google.com

Chris Callison-Burch
ccb@seas.upenn.edu

What is a sentence-level language model

- Standard word-level language models predict the next word given previous words.
- Standard word-level language models assume a finite number of words in the vocabulary, and so the probability is always over the same set of words.
- A sentence-level language model predicts the next sentence given previous sentences.
- There is a nearly unlimited number of valid English sentences, so at each step of training/ inference a different set of sentences might be scored.
- At train time, the candidate set contains N - I distractors (which may differ between train iterations) in addition to the true next sentence.

Advantages of a sentence-level language model

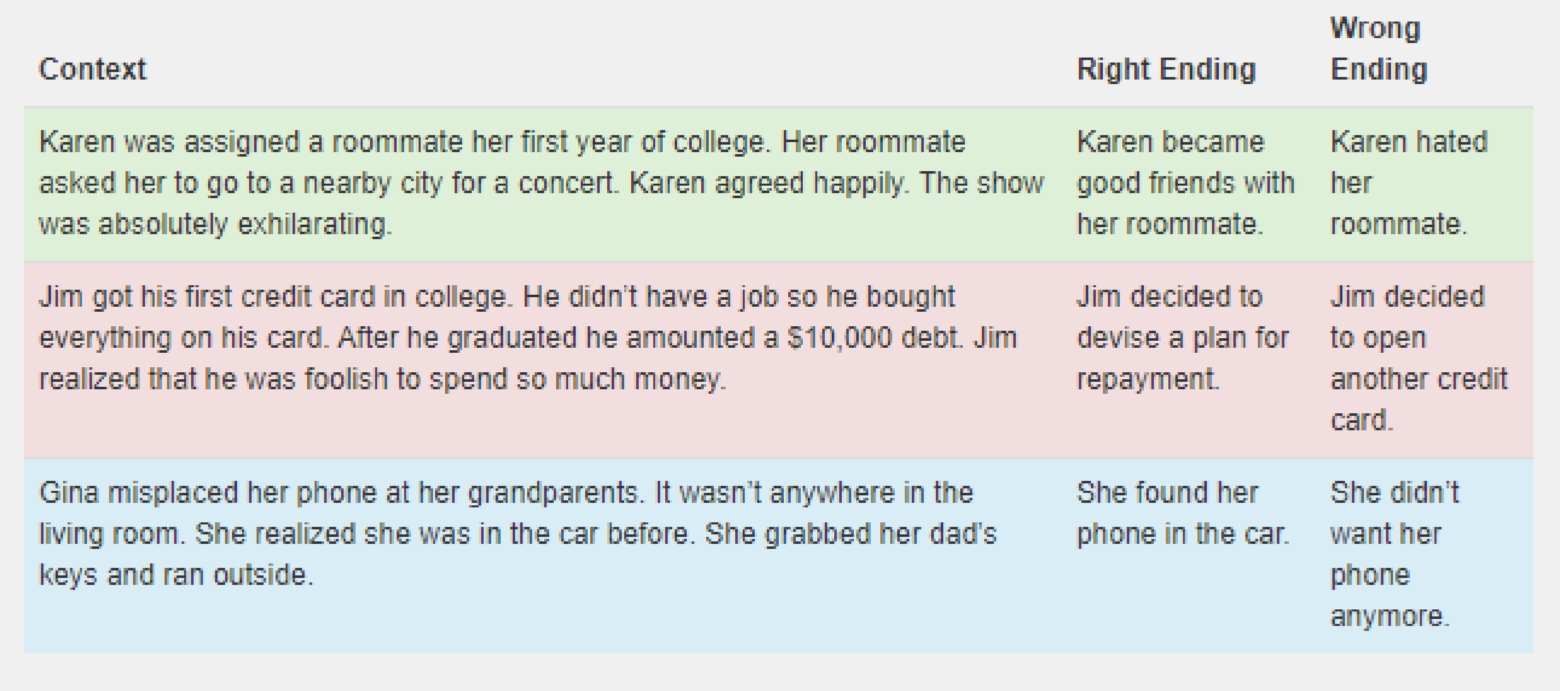
- The task of modeling long-range dependencies is isolated from the task of individual word prediction.
- By taking advantage of strong semantic representations from a pre-trained BERT the sentence LM can be compact and fast to train.
- More text can be seen at each train step than with a word-level LM:

  Sequence length is the number of sentences, allowing a context of hundreds of words to be reduced to only a handful of sentence embeddings.

  100s to 1000s of candidate next sentences can be considered in the train loss.

Applying the sentence LM to the Story Cloze Task

Both candidate endings are fluent and on-topic, but only one is coherent.

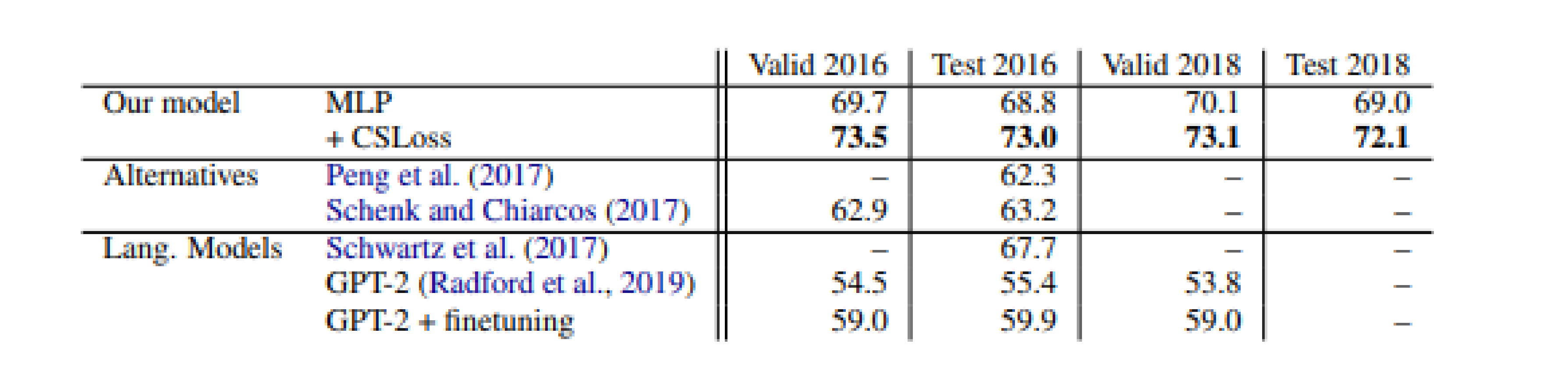


Context	Right Ending	Wrong Ending
Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating.	Karen became good friends with her roommate.	Karen hated her roommate.
Jim got his first credit card in college. He didn't have a job so he bought everything on his card. After he graduated he amounted a $10,000 debt. Jim realized that he was foolish to spend so much money.	Jim decided to devise a plan for repayment.	Jim decided to open another credit card.
Gina misplaced her phone at her grandparents. It wasn't anywhere in the living room. She realized she was in the car before. She grabbed her dad's keys and ran outside.	She found her phone in the car.	She didn't want her phone anymore.

Applying the sentence LM to the Story Cloze Task

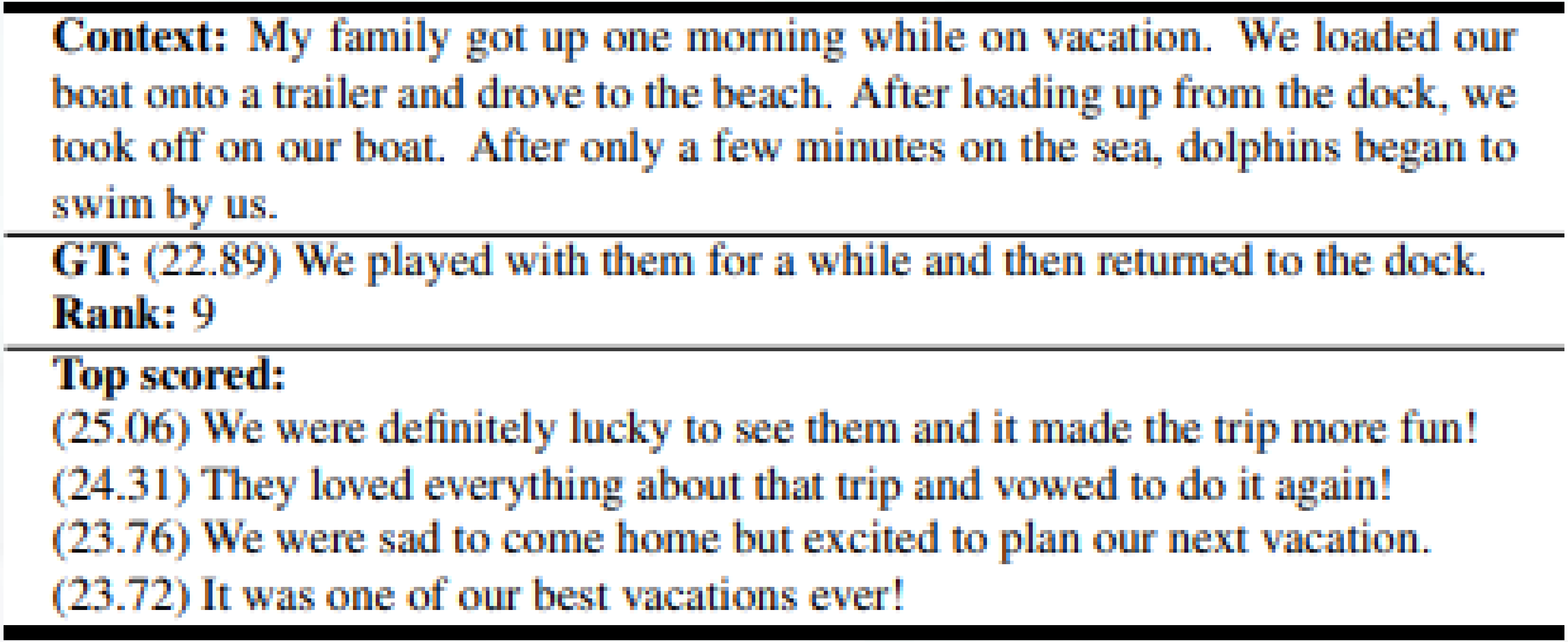
- They train their sentence LM to predict the 5th sentence given the previous 4 sentences. All 98k 5th sentences from the train set are used as the candidate set
- Model architecture is a simple multi-layer perceptron
- They further improve performance by incorporating a secondary loss (CSLoss) that penalizes the model for assigning a high score to any of the 4 context sentences

Applying the sentence LM to the Story Cloze Task



		Valid 2016	Test 2016	Valid 2018	Test 2018
Our model	MLP	69.7	68.8	70.1	69.0
	+ CSLoss	**73.5**	**73.0**	**73.1**	**72.1**
Alternatives	Peng et al. (2017)	–	62.3	–	–
	Schenk and Chiarcos (2017)	62.9	63.2	–	–
Lang. Models	Schwartz et al. (2017)	–	67.7	–	–
	GPT-2 (Radford et al., 2019)	54.5	55.4	53.8	–
	GPT-2 + finetuning	59.0	59.9	59.0	–

Applying the sentence LM to the Story Cloze Task

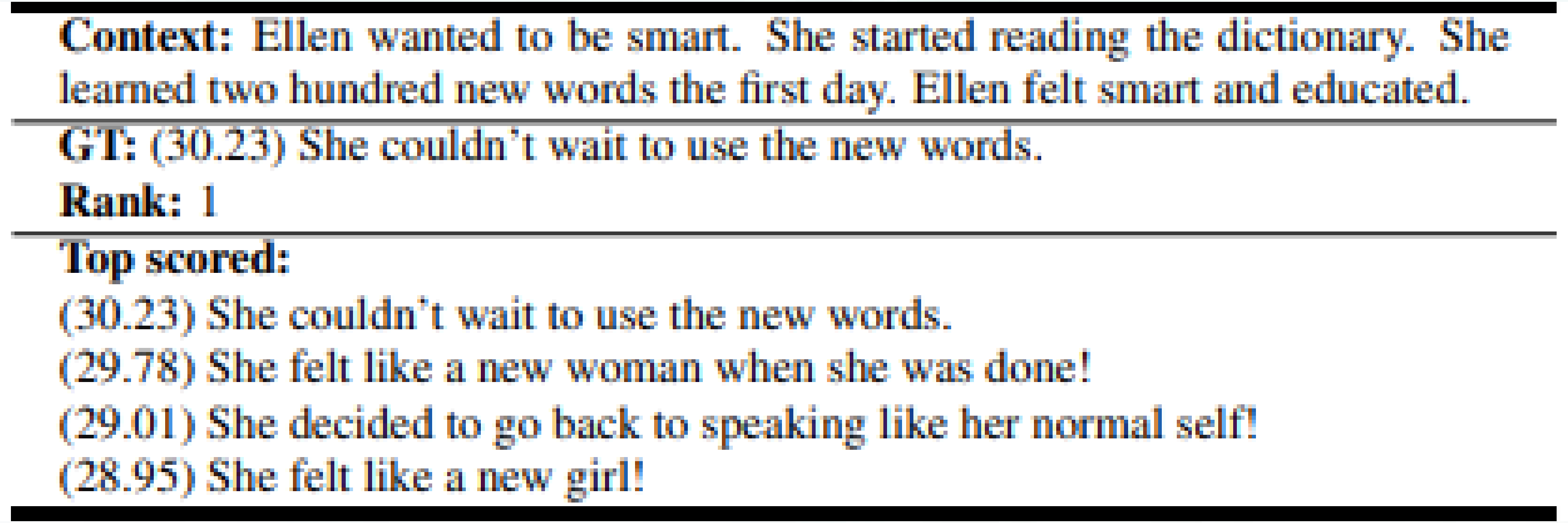


Context: My family got up one morning while on vacation. We loaded our boat onto a trailer and drove to the beach. After loading up from the dock, we took off on our boat. After only a few minutes on the sea, dolphins began to swim by us.

GT: (22.89) We played with them for a while and then returned to the dock.
Rank: 9

Top scored:
(25.06) We were definitely lucky to see them and it made the trip more fun!
(24.31) They loved everything about that trip and vowed to do it again!
(23.76) We were sad to come home but excited to plan our next vacation.
(23.72) It was one of our best vacations ever!

Applying the sentence LM to the Story Cloze Task



Context: Ellen wanted to be smart. She started reading the dictionary. She learned two hundred new words the first day. Ellen felt smart and educated.

GT: (30.23) She couldn't wait to use the new words.

Rank: 1

Top scored:

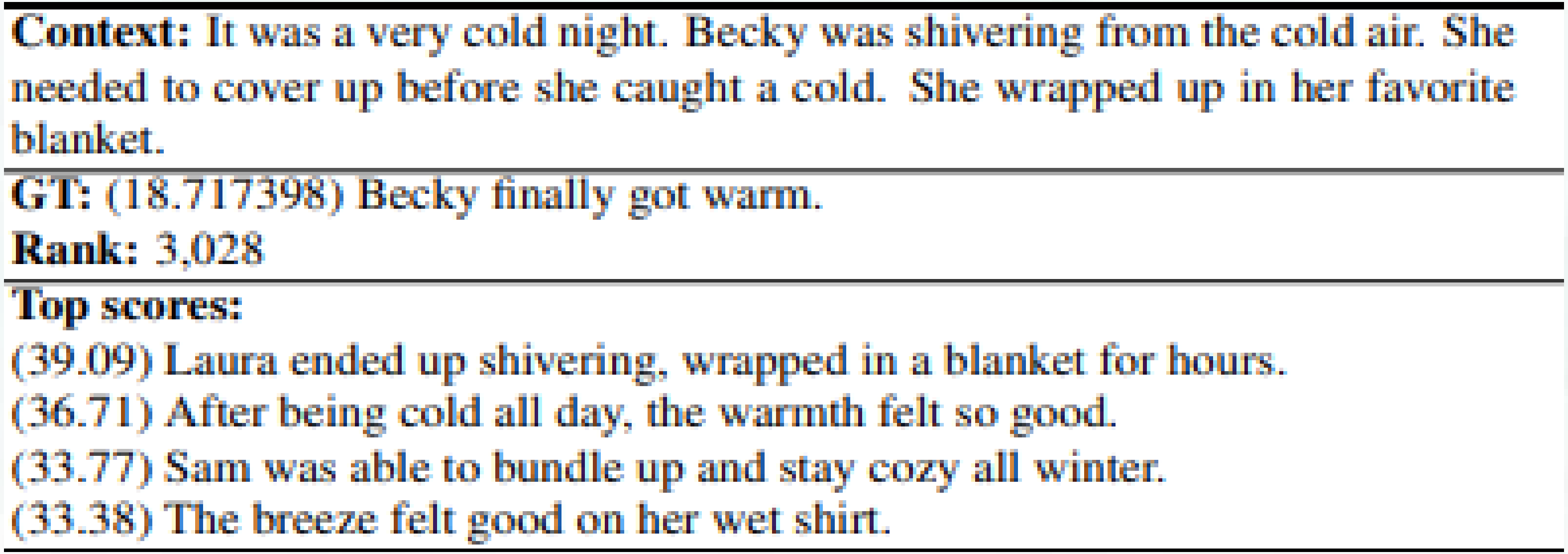
(30.23) She couldn't wait to use the new words.

(29.78) She felt like a new woman when she was done!

(29.01) She decided to go back to speaking like her normal self!

(28.95) She felt like a new girl!

Applying the sentence LM to the Story Cloze Task



Context: It was a very cold night. Becky was shivering from the cold air. She needed to cover up before she caught a cold. She wrapped up in her favorite blanket.

GT: (18.717398) Becky finally got warm.

Rank: 3,028

Top scores:

(39.09) Laura ended up shivering, wrapped in a blanket for hours.

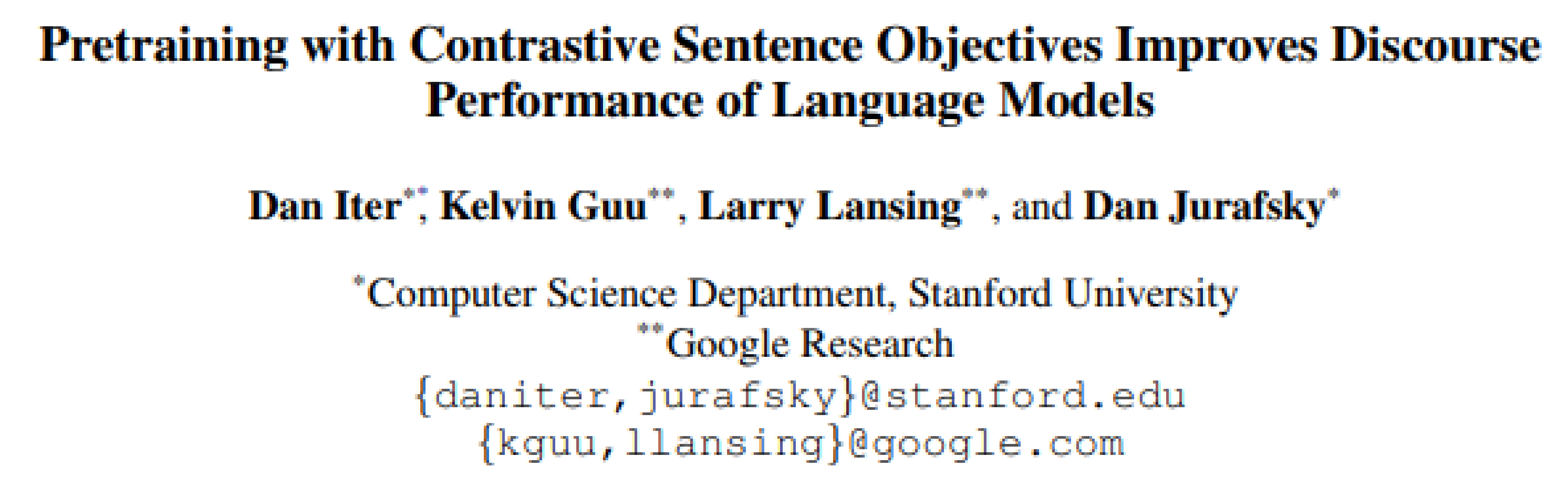
(36.71) After being cold all day, the warmth felt so good.

(33.77) Sam was able to bundle up and stay cozy all winter.

(33.38) The breeze felt good on her wet shirt.

Takeaways and Future Work

- Taking advantage of pre-trained BERT sentence embeddings allows us to train a model that avoids having to learn token-level fluency and can instead focus solely on inter-sentence coherence.
- Future work could include building out a two-step generative process, where first a good candidate next sentence is retrieved, and then it is refined to better fit the context.
- However, further work is needed to improve performance on "real-world" story text.



Pretraining with Contrastive Sentence Objectives Improves Discourse Performance of Language Models

Dan Iter[*,**], **Kelvin Guu**[**], **Larry Lansing**[**], and **Dan Jurafsky**[*]

[*]Computer Science Department, Stanford University
[**]Google Research
{daniter,jurafsky}@stanford.edu
{kguu,llansing}@google.com

Discourse Coherence

- The aspect of text quality that measure the connectedness and organization of sentences
- Evaluated on pretrained text embeddings
- Previous work shows pretrained language models perform well

This work

- Show that language models can improve their discourse-level representation by pretraining with an inter-sentence objective.
- Present their model, CONtrastive Position and Ordering with Negatives Objective (CONPONO) .
- Evaluate their model on a discourse representation benchmark, DiscoEval • Model design choices and ablations.
- Applying CONPONO on non-discourse tasks that can benefit from better inter-sentence representations.

Previous

- Next Sentence Prediction
- Sentence Ordering

Conpono: Which sentence is 2 before?

- Popular techniques include the use of word embeddings to capture semantic properties of words.

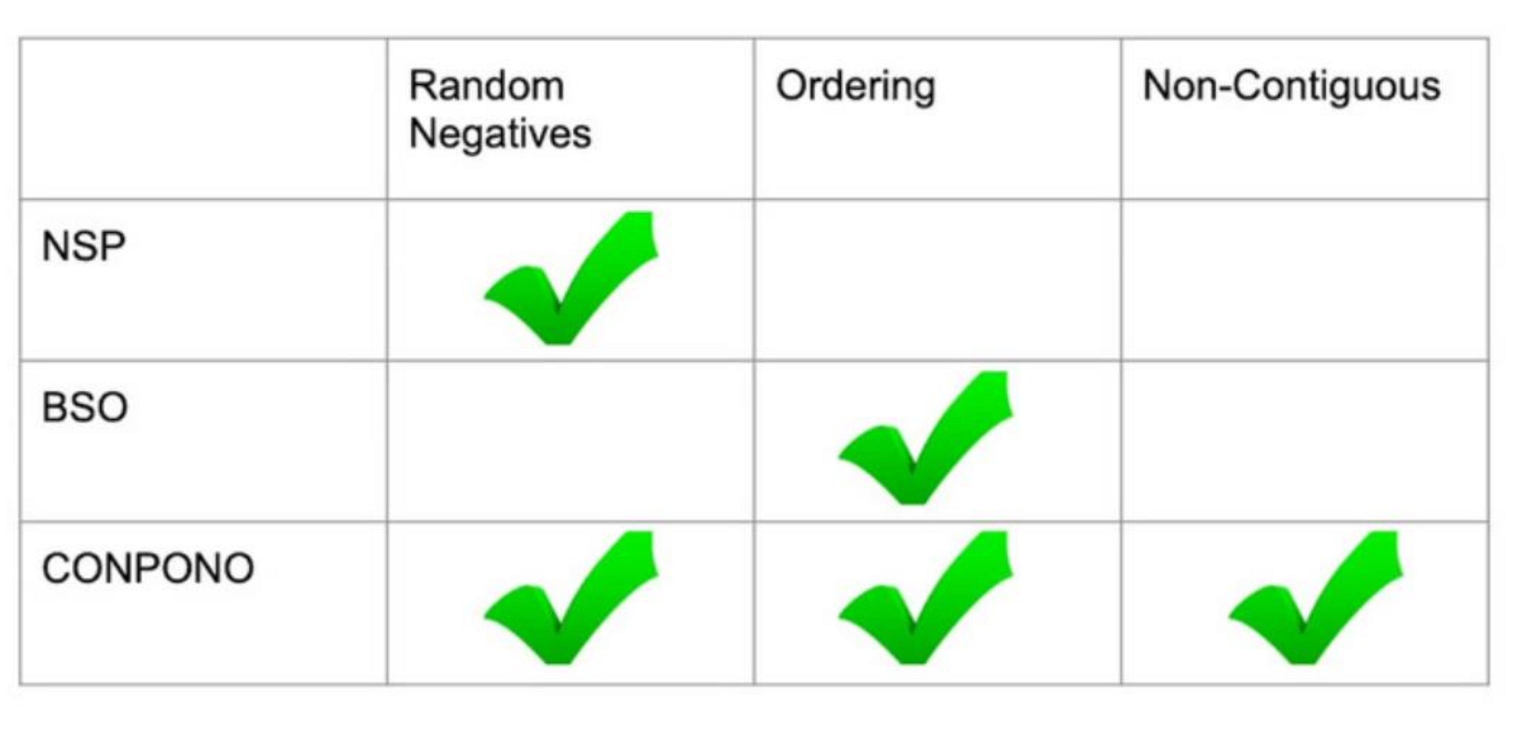
A) In the 2010s, representation learning and deep neural network-style machine learning method became widespread in natural language processing.

B) Many results showed that such techniques can achieve state-of-the-art results in many natural language tasks.

C) In some areas, this shift has entailed substantial changes in how NLP systems are designed

D) Keynesian economics derives from John Maynard Keynes, in particular his book The General Theory of Employment, Interest and Money (1936), which ushered in contemporary macroeconomics as a distinct field.

Inter-sentence representation Task



	Random Negatives	Ordering	Non-Contiguous
NSP	✓		
BSO		✓	
CONPONO	✓	✓	✓

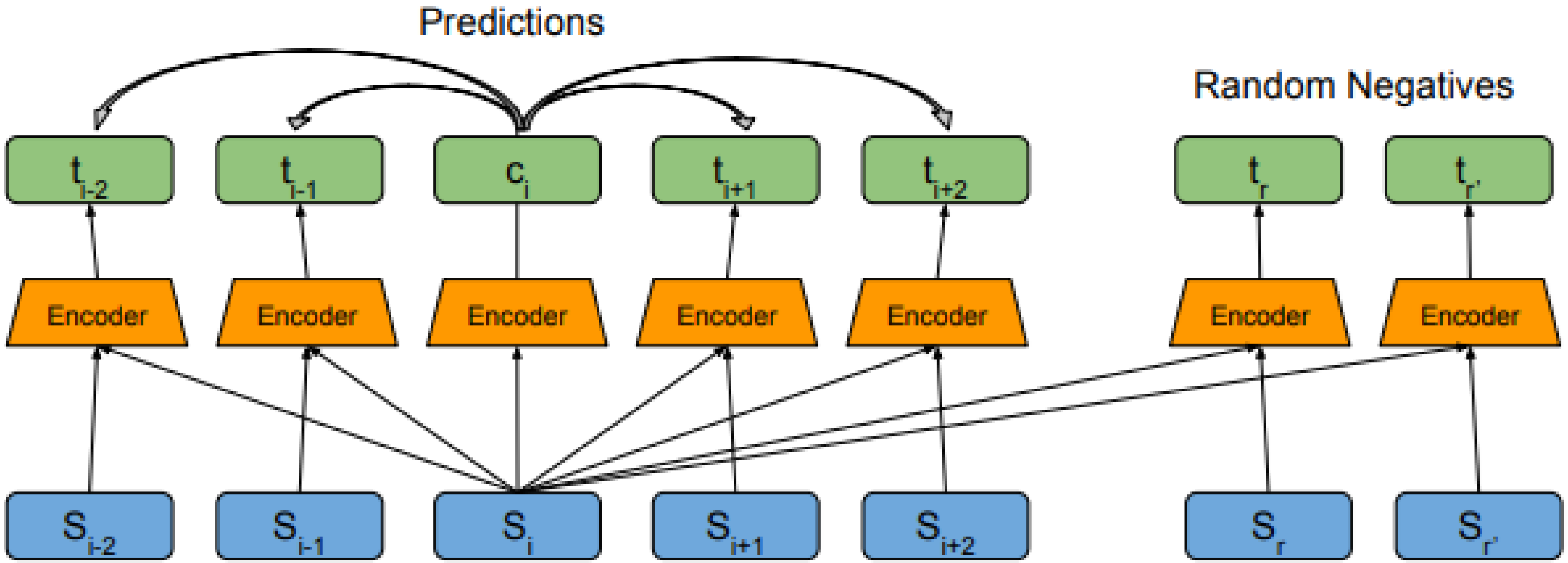


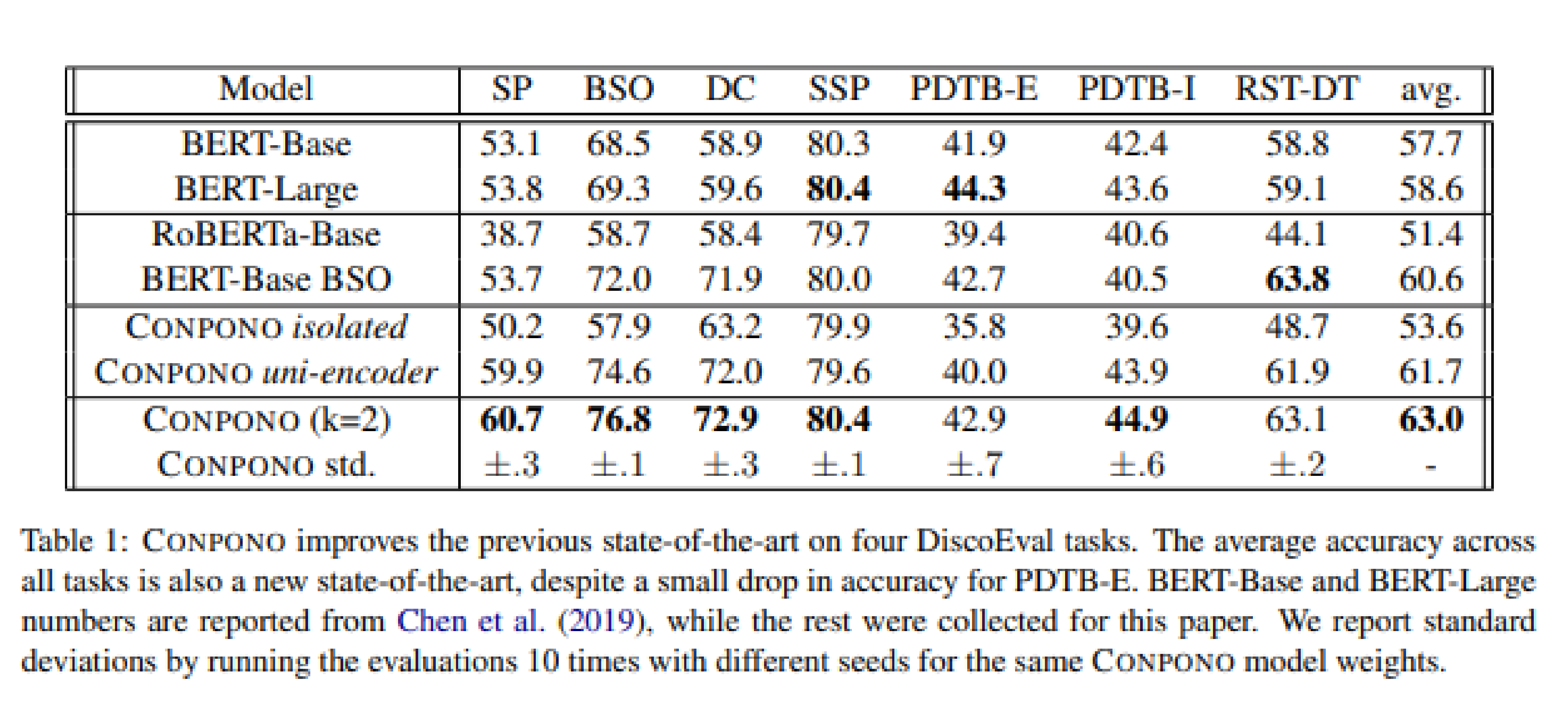
Figure 1: During training, a text segment is selected as the anchor (S_i). The anchor as well as all the targets, $S_{i-2}...S_{i+2}$ plus random samples S_r are encoded with the transformer masked language model. The encoded representation of the anchor is used to predict each target at its target distance. The S_i objects are raw text sentences, the *encoder* is the transformer model, and c_i and t_i are vectors.

Key Contributions

- **Fine-grained sentence ordering** : expand beyond binary sentence ordering and easy negatives
- **Non-contiguous inter-sentence representation** : window size>1
- **Discriminative non-binary objective** : not binary, no generation
- **Cross-attention for discourse coherence** : joint encode segment A and B
- **Model size :**

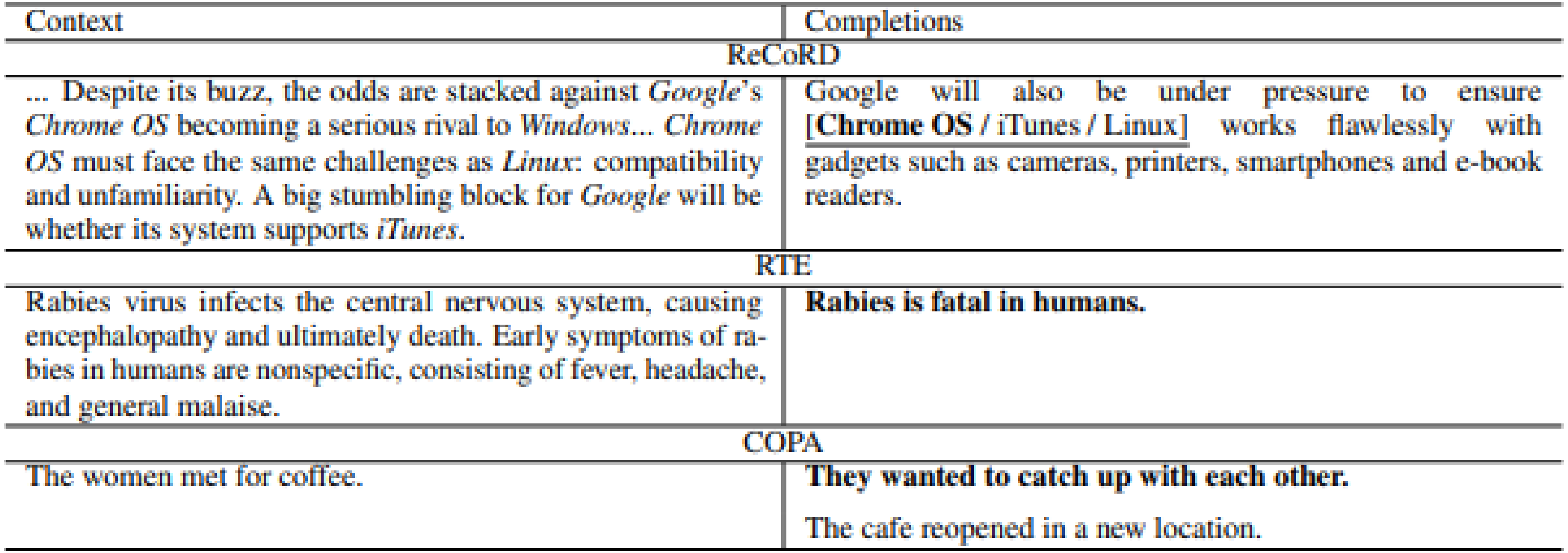
Evaluation - DiscoEval Tasks

- Sentence Position
- Binary Sentence Ordering
- Discourse Coherence
- Sentence Section Prediction
- Penn Discourse Tree Bank
- Rhetorical Structure Theory



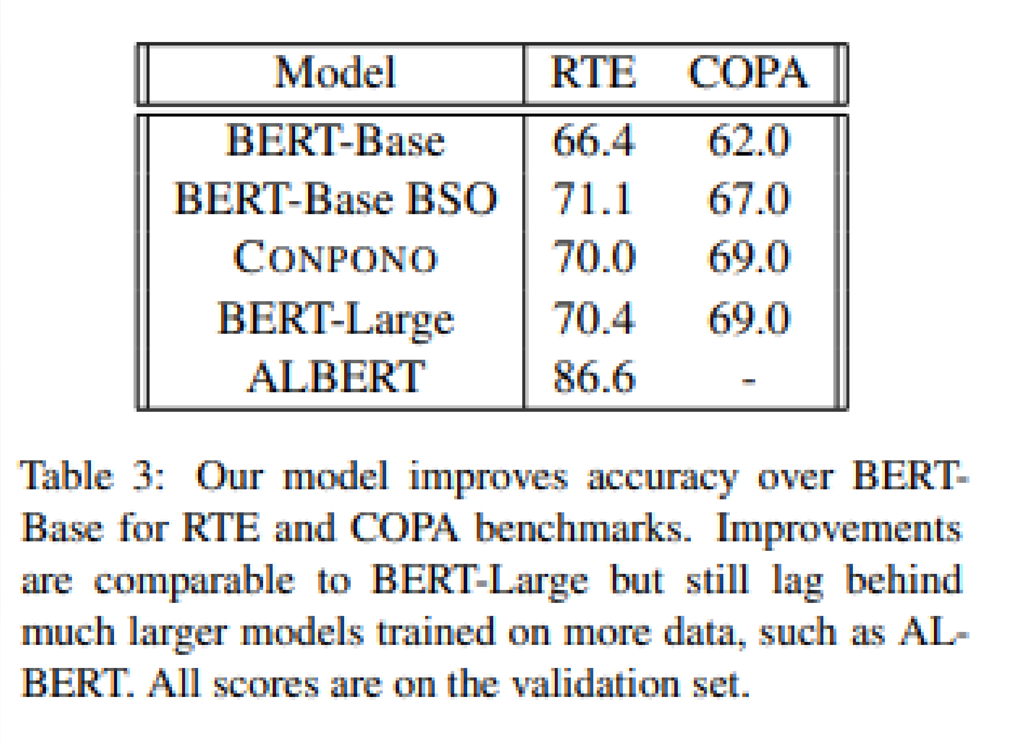
Model	SP	BSO	DC	SSP	PDTB-E	PDTB-I	RST-DT	avg.
BERT-Base	53.1	68.5	58.9	80.3	41.9	42.4	58.8	57.7
BERT-Large	53.8	69.3	59.6	**80.4**	**44.3**	43.6	59.1	58.6
RoBERTa-Base	38.7	58.7	58.4	79.7	39.4	40.6	44.1	51.4
BERT-Base BSO	53.7	72.0	71.9	80.0	42.7	40.5	**63.8**	60.6
CONPONO *isolated*	50.2	57.9	63.2	79.9	35.8	39.6	48.7	53.6
CONPONO *uni-encoder*	59.9	74.6	72.0	79.6	40.0	43.9	61.9	61.7
CONPONO (k=2)	**60.7**	**76.8**	**72.9**	**80.4**	42.9	**44.9**	63.1	**63.0**
CONPONO std.	±.3	±.1	±.3	±.1	±.7	±.6	±.2	-

Table 1: CONPONO improves the previous state-of-the-art on four DiscoEval tasks. The average accuracy across all tasks is also a new state-of-the-art, despite a small drop in accuracy for PDTB-E. BERT-Base and BERT-Large numbers are reported from Chen et al. (2019), while the rest were collected for this paper. We report standard deviations by running the evaluations 10 times with different seeds for the same CONPONO model weights.



Context	Completions
ReCoRD	
... Despite its buzz, the odds are stacked against *Google*'s *Chrome OS* becoming a serious rival to *Windows*... *Chrome OS* must face the same challenges as *Linux*: compatibility and unfamiliarity. A big stumbling block for *Google* will be whether its system supports *iTunes*.	Google will also be under pressure to ensure [**Chrome OS** / iTunes / Linux] works flawlessly with gadgets such as cameras, printers, smartphones and e-book readers.
RTE	
Rabies virus infects the central nervous system, causing encephalopathy and ultimately death. Early symptoms of rabies in humans are nonspecific, consisting of fever, headache, and general malaise.	**Rabies is fatal in humans.**
COPA	
The women met for coffee.	**They wanted to catch up with each other.** The cafe reopened in a new location.

Table 2: These are examples from ReCoRD, RTE, and COPA that exhibit aspects of discourse coherence. For ReCoRD, candidate entities are in italics and replaced terms in the completion are underlined. True completions are bold.



Model	RTE	COPA
BERT-Base	66.4	62.0
BERT-Base BSO	71.1	67.0
CONPONO	70.0	69.0
BERT-Large	70.4	69.0
ALBERT	86.6	-

Table 3: Our model improves accuracy over BERT-Base for RTE and COPA benchmarks. Improvements are comparable to BERT-Large but still lag behind much larger models trained on more data, such as ALBERT. All scores are on the validation set.

Model	SP	BSO	DC	SSP	PDTB-E	PDTB-I	RST-DT	avg.
k=4	59.84	76.05	**73.62**	**80.65**	42.28	44.25	63.00	62.81
k=3	60.47	76.68	72.74	80.30	**43.40**	44.28	62.56	62.92
k=2	**60.67**	**76.75**	72.85	80.38	42.87	**44.87**	**63.13**	**63.07**
k=1	47.56	66.03	72.62	80.15	42.79	43.55	62.31	59.29
- MLM	54.92	75.37	68.35	80.2	41.67	43.88	61.27	60.81
Small	45.41	61.70	67.71	75.58	35.26	36.18	46.58	52.63

Table 5: The ablation analysis shows the effects of different k values (ie. window sizes) in our objective, removing the MLM objective during pretraining and training with a small transformer encoder.

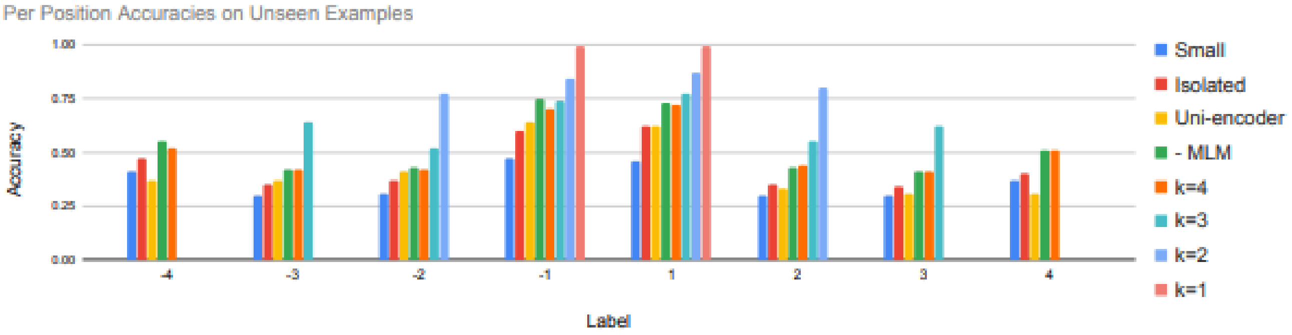
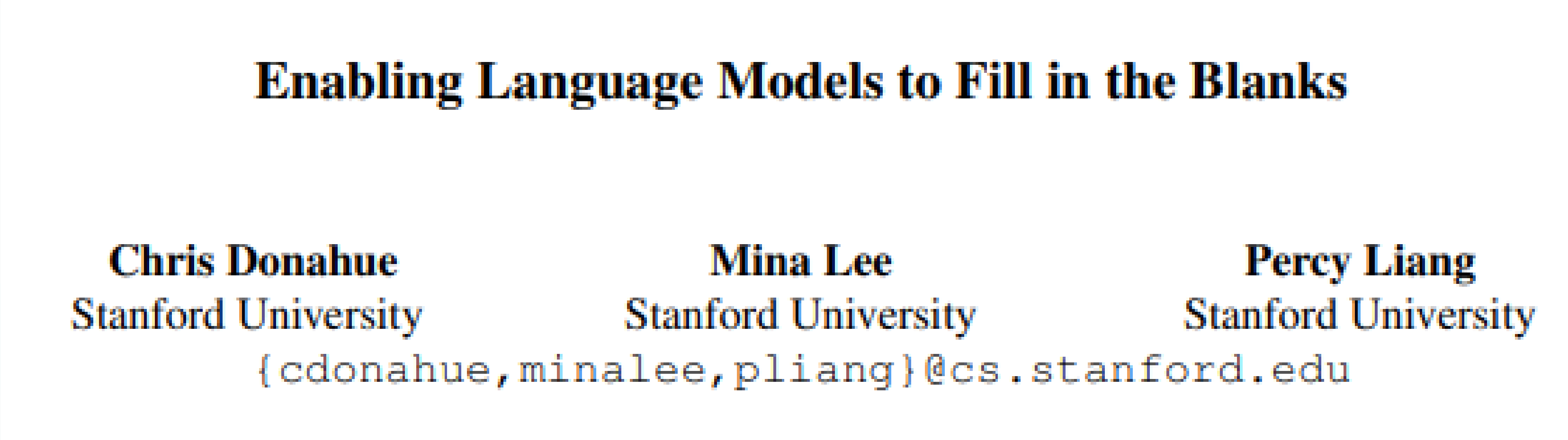


Figure 2: We can evaluate the accuracy on the CONPONO objective for each label (ie. distance between anchor and target sentence) on a set of 5,000 examples held-out from training. We observe that higher accuracy does not necessarily correlate with better downstream performance on DiscoEval.



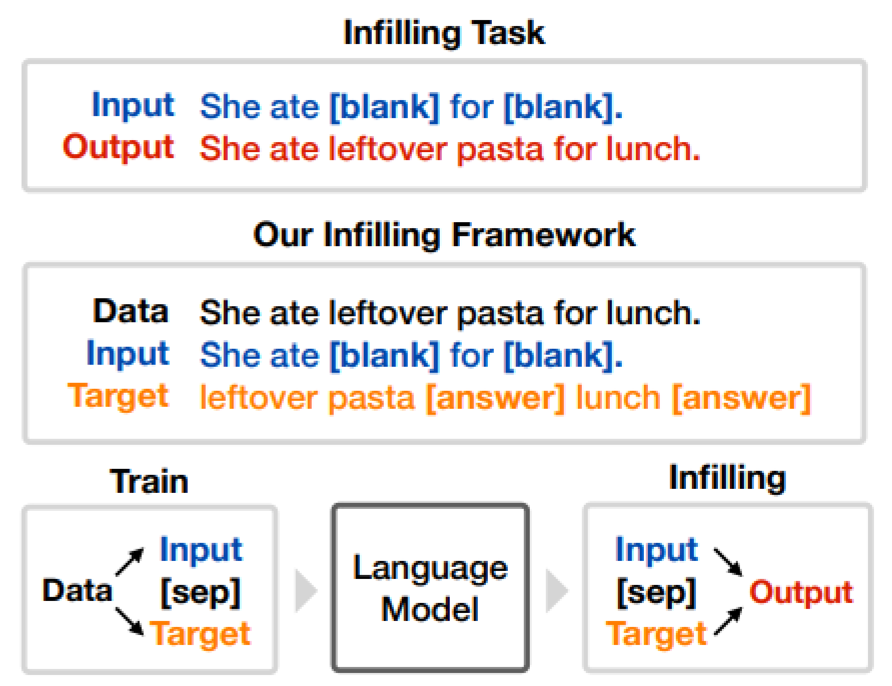
Enabling Language Models to Fill in the Blanks

Chris Donahue
Stanford University

Mina Lee
Stanford University

Percy Liang
Stanford University

{cdonahue,minalee,pliang}@cs.stanford.edu



Infilling Task

Input She ate [blank] for [blank].

Output She ate leftover pasta for lunch.

Our Infilling Framework

Data She ate leftover pasta for lunch.

Input She ate [blank] for [blank].

Target leftover pasta [answer] lunch [answer]

Train

Data → Input [sep] Target

Language Model

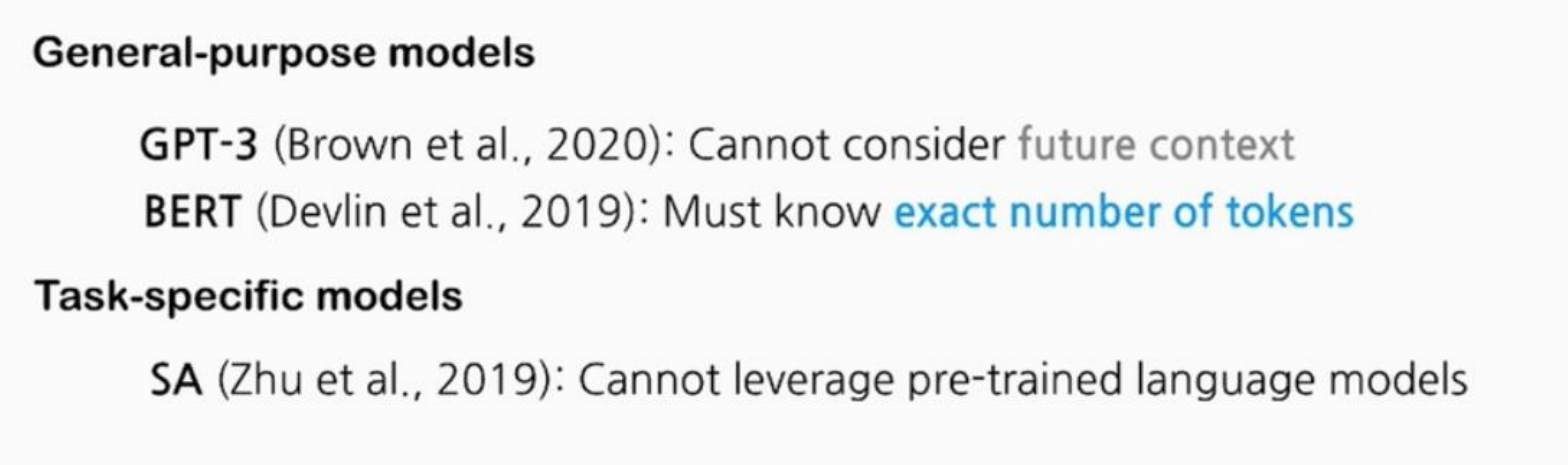
Infilling

Input [sep] Target → Output

This task takes incomplete text as input and outputs completed text.

IMPORTANT: infill spans corrupted by arbitrary mask functions (words, n-grams, sentences, paragraphs, and documents)

Previous work on Text Infilling



General-purpose models

GPT-3 (Brown et al., 2020): Cannot consider future context

BERT (Devlin et al., 2019): Must know exact number of tokens

Task-specific models

SA (Zhu et al., 2019): Cannot leverage pre-trained language models

Their work: ILM Infilling by Language Modeling

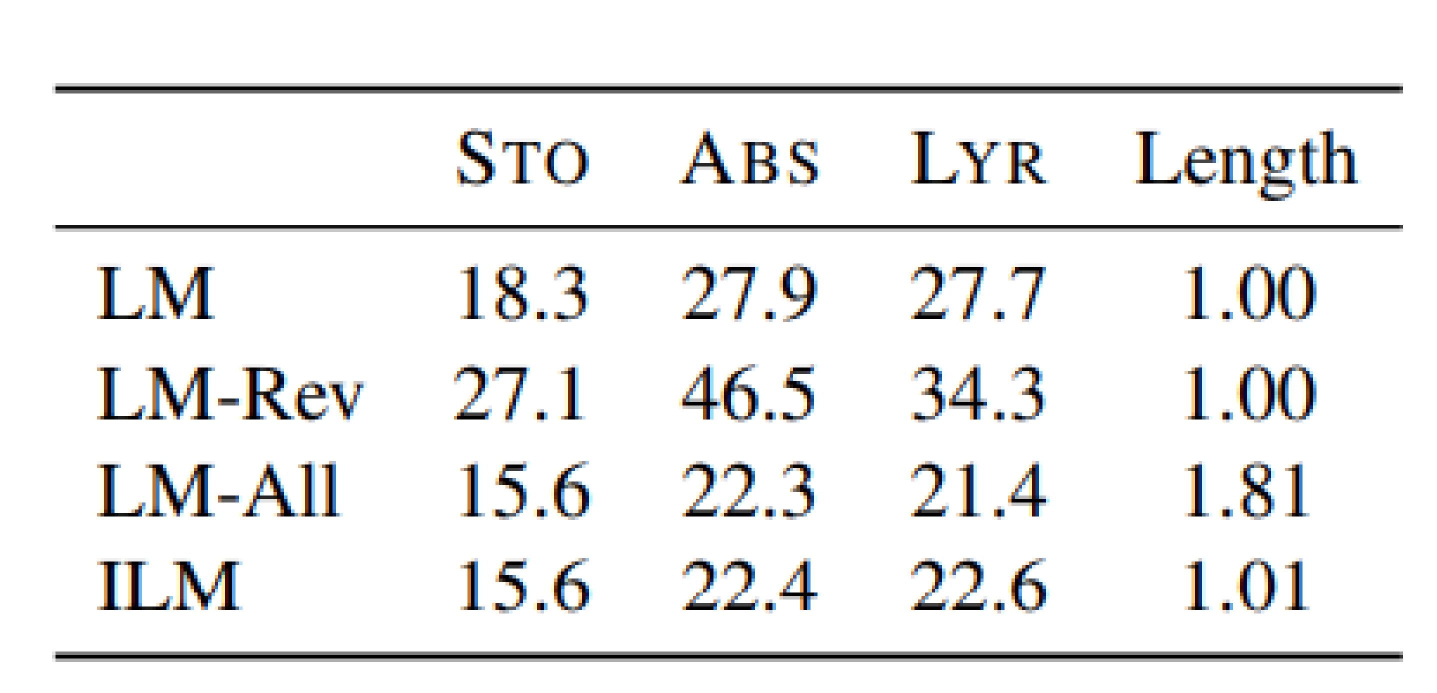
1. Download any language model

2. Finetune the model on infilling examples

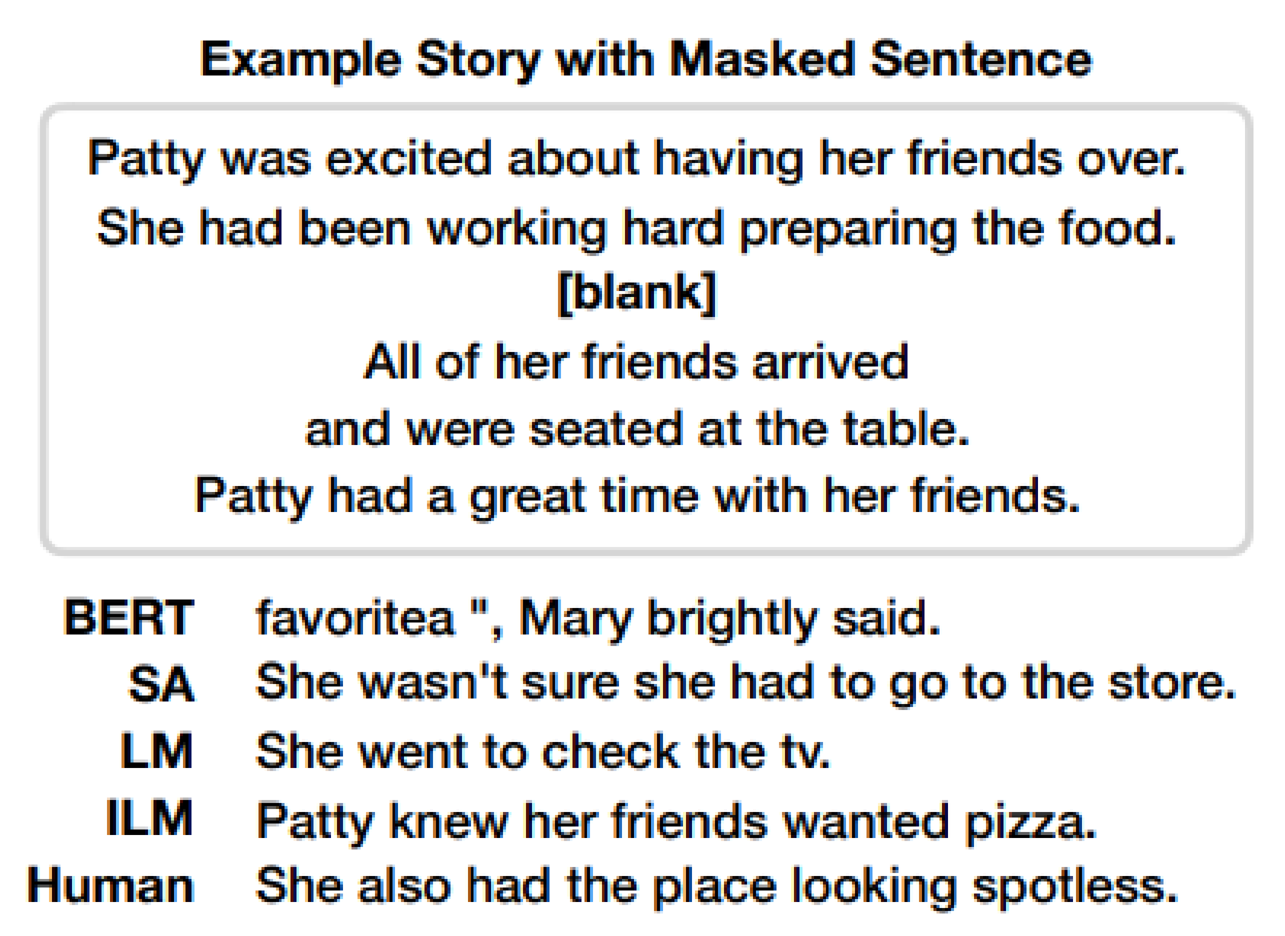
Experimental setup

- Data: Stories (Mostafazadeh et al., 2016), Abstracts, Lyrics
- Metric: Score, Perplexity
- Model: BERT, SA (Zhu et al., 2019), LM, ILM (ours)

- 1. Human evaluation
- 2. Quantitative evaluation



	Sto	Abs	Lyr	Length
LM	18.3	27.9	27.7	1.00
LM-Rev	27.1	46.5	34.3	1.00
LM-All	15.6	22.3	21.4	1.81
ILM	15.6	22.4	22.6	1.01



Example Story with Masked Sentence

Patty was excited about having her friends over.
She had been working hard preparing the food.
[blank]
All of her friends arrived
and were seated at the table.
Patty had a great time with her friends.

BERT	favoritea ", Mary brightly said.
SA	She wasn't sure she had to go to the store.
LM	She went to check the tv.
ILM	Patty knew her friends wanted pizza.
Human	She also had the place looking spotless.

Takeaways

- Conceptual simplicity: minimal change to LM
- Model-agnostic framework： Leverage massively pretrianed LM

Thank [blank] for [blank]! → Thank you for listening!