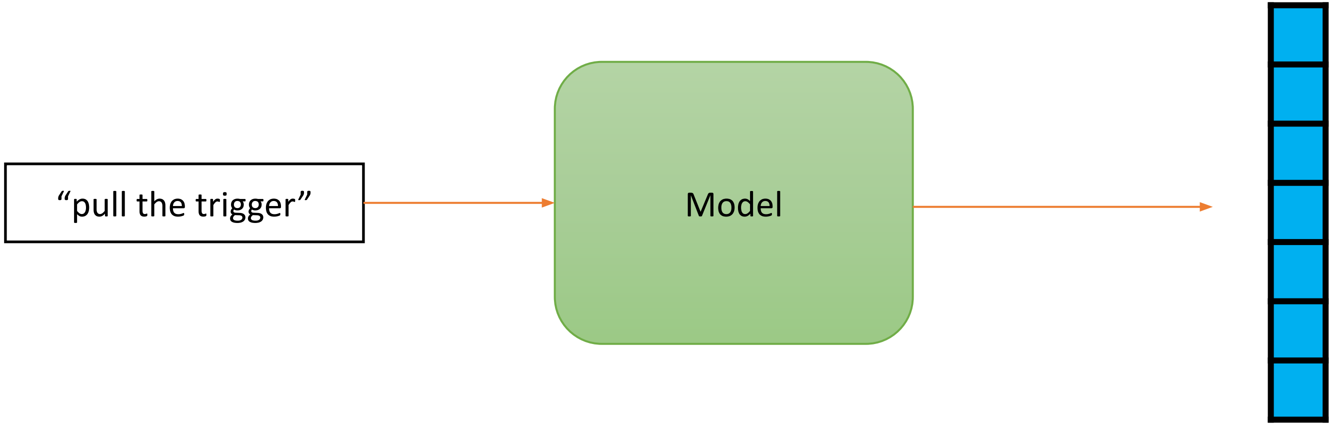
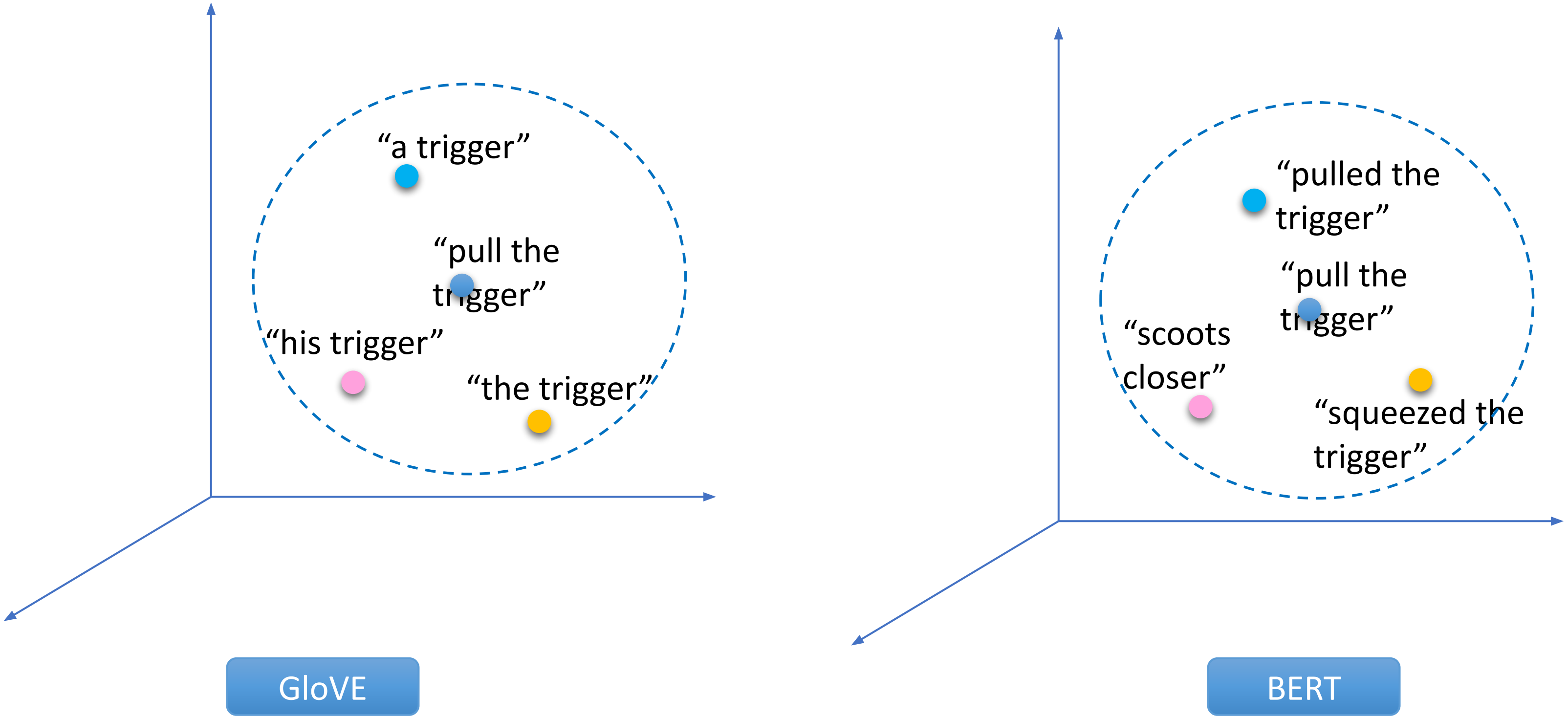
Phrase-BERT:
Improved Phrase Embeddings from BERT with an Application to Corpus Exploration

Shufan Wang Laure Thompson Mohit Iyyer

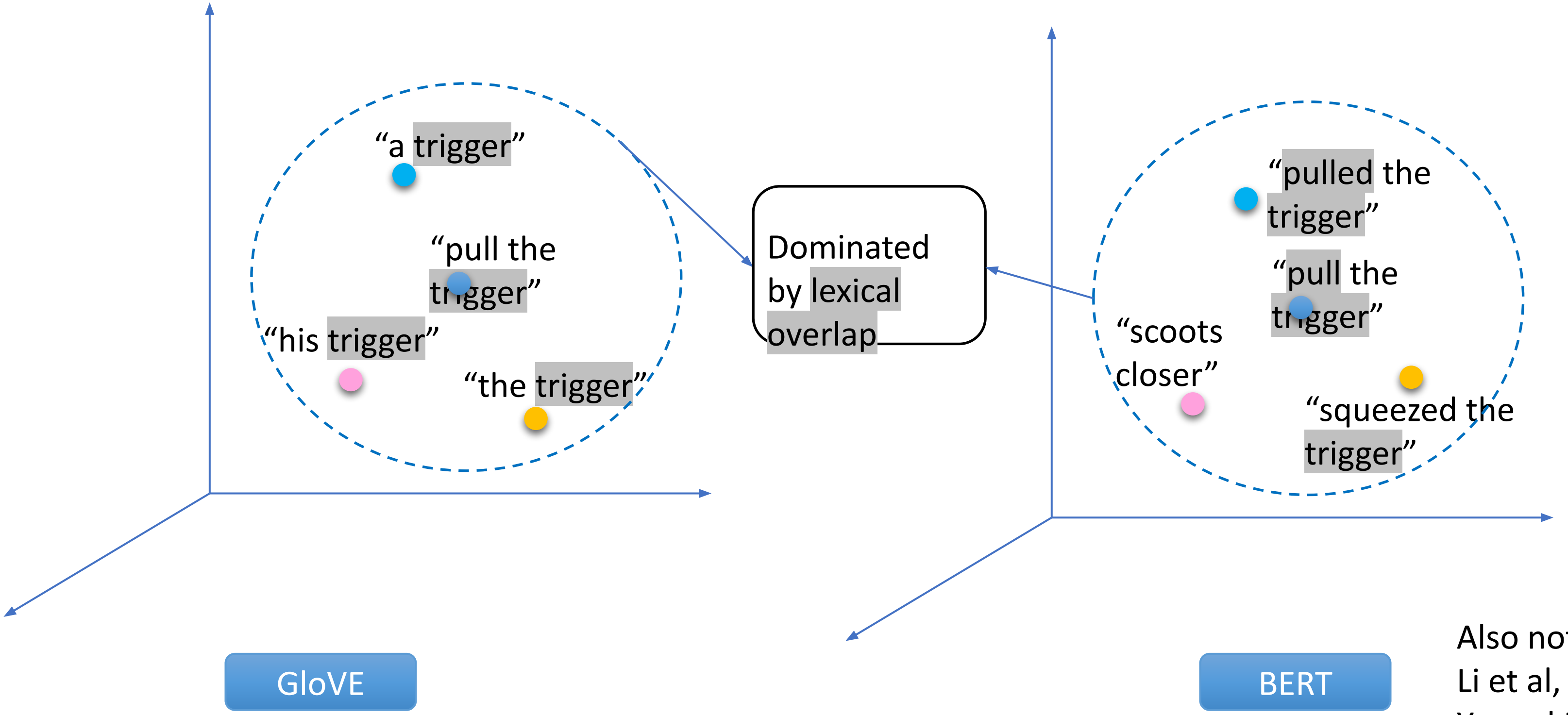
A general-purpose phrase embedding



Phrase Nearest Neighbours

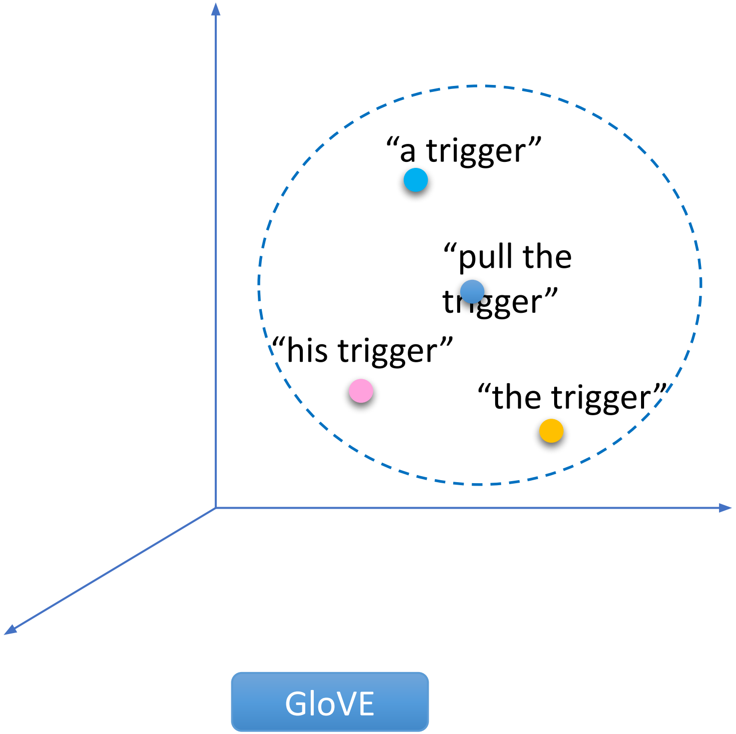


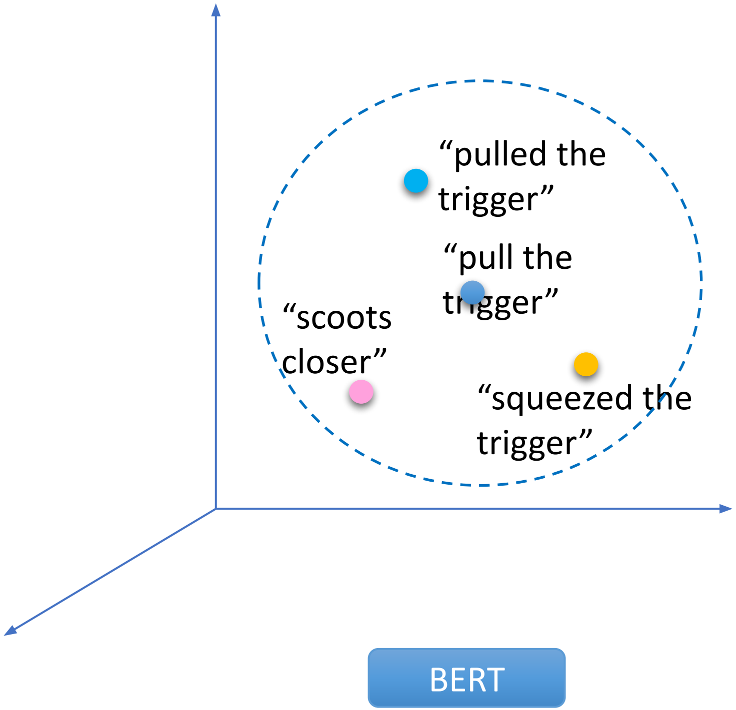
Phrase Nearest Neighbours



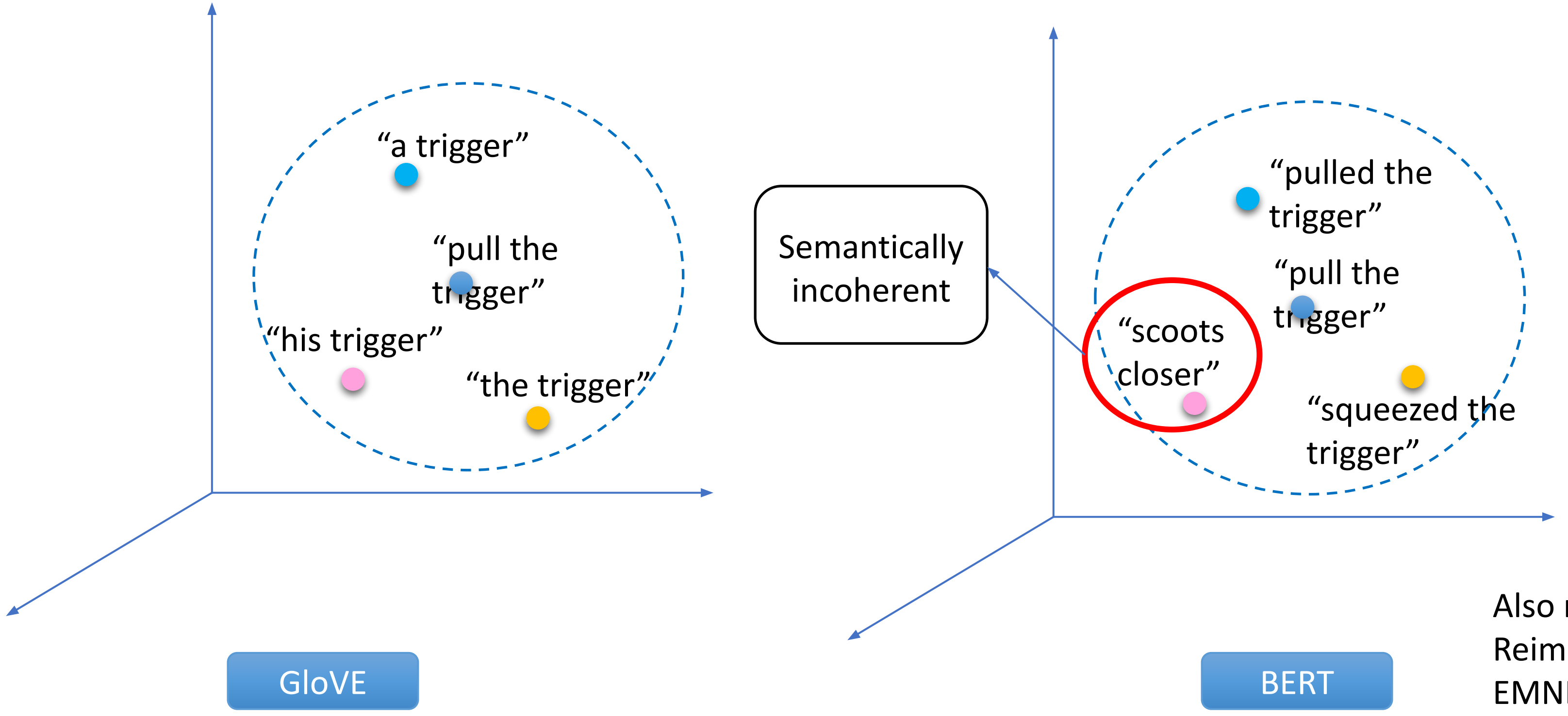
Also noted by:
Li et al, EMNLP 2020;
Yu and Ettinger, ACL 2020

Phrase Nearest Neighbours



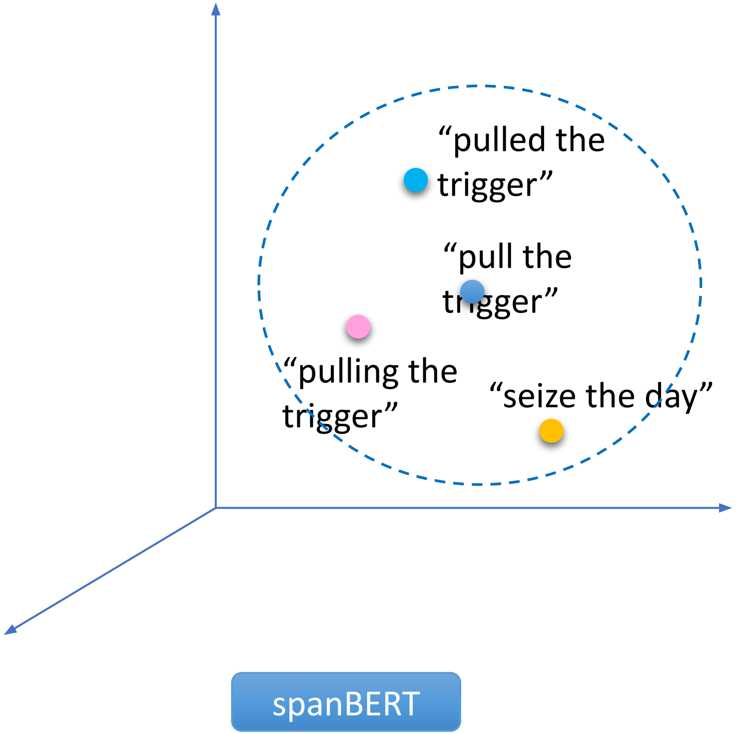


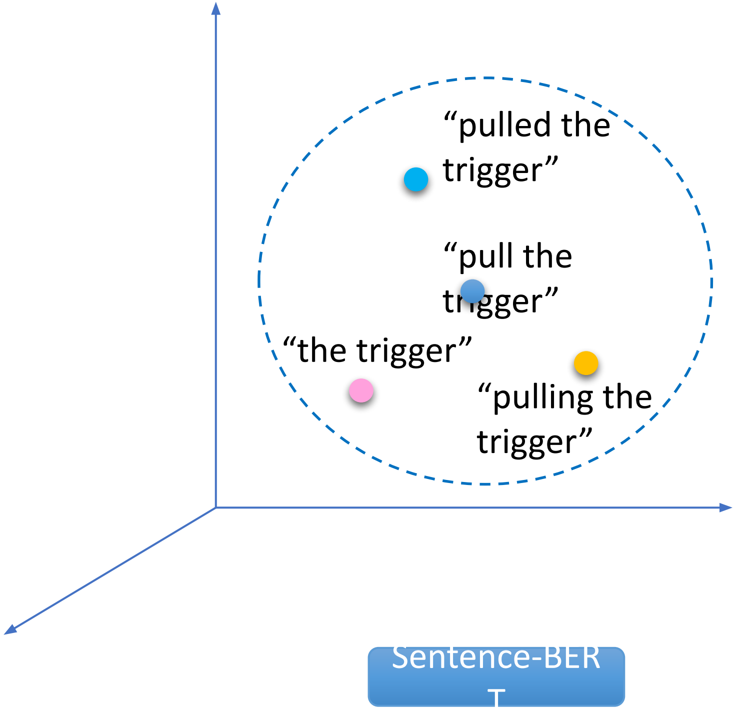
Phrase Nearest Neighbours



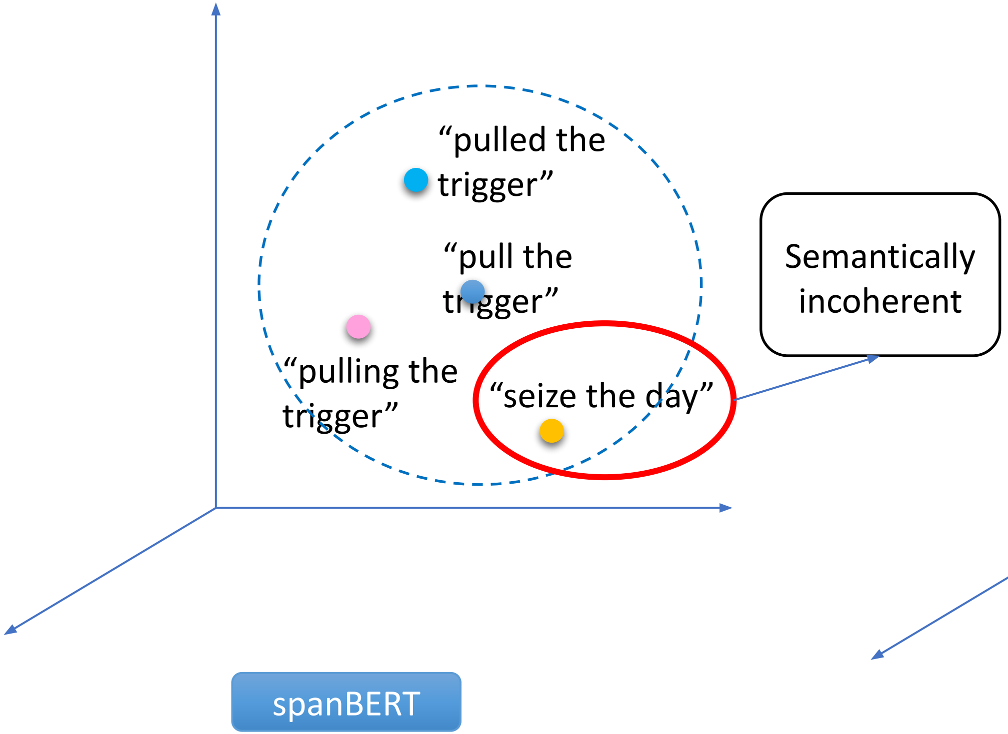
Also noted by:
Reimers and Gurevych, EMNLP 2019;
Yu and Ettinger , ACL 2020

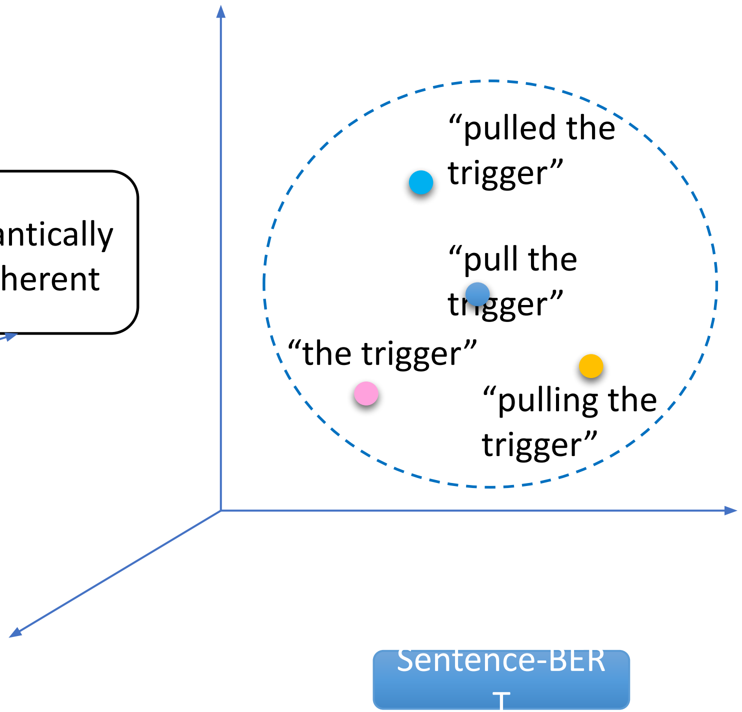
Phrase Nearest Neighbours



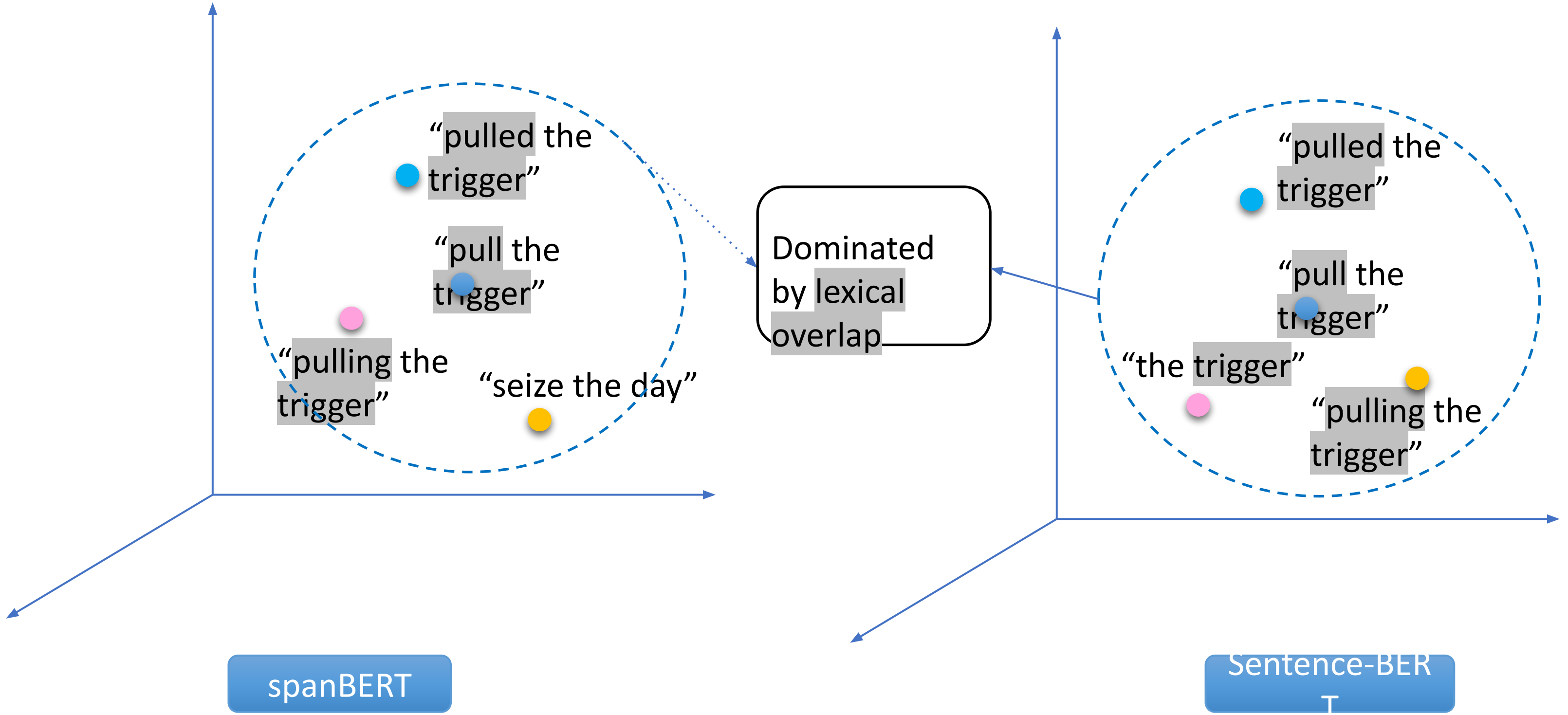


Phrase Nearest Neighbours

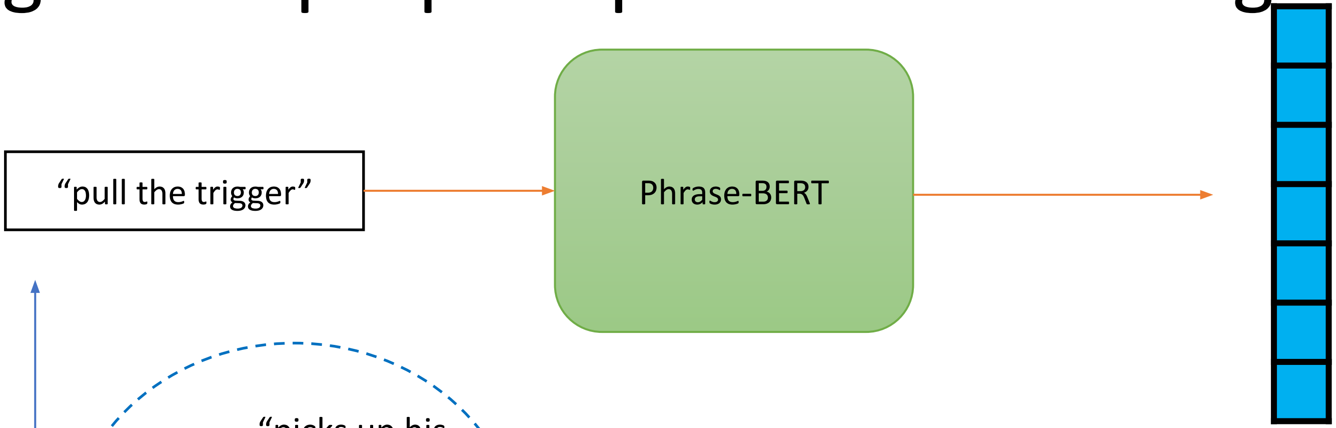


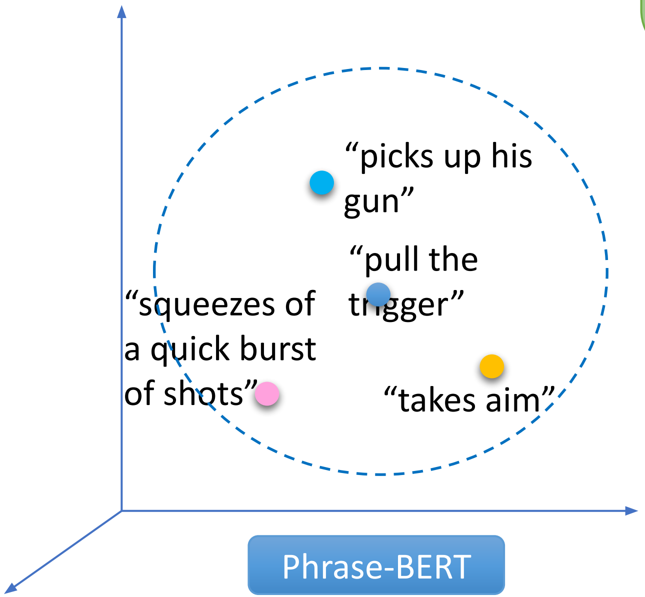


Phrase Nearest Neighbours



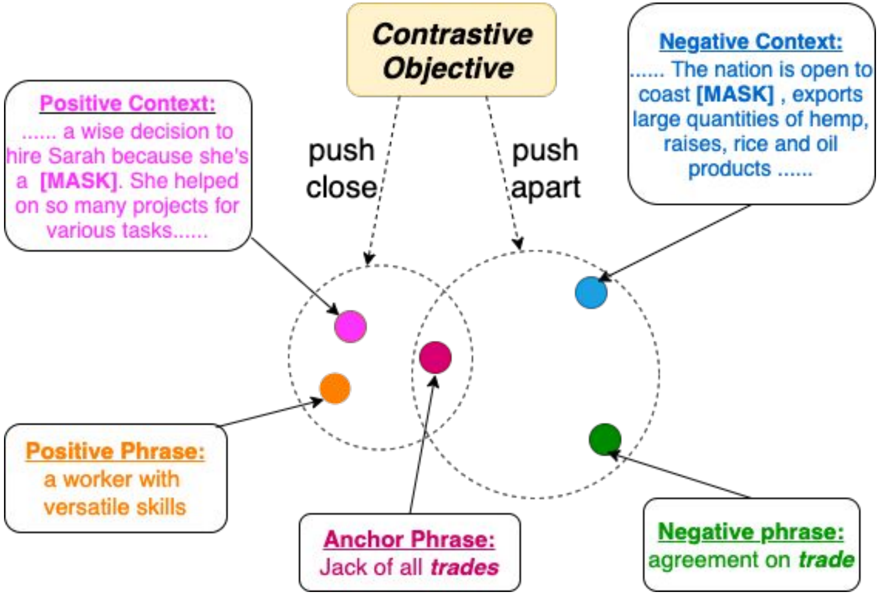
A general-purpose phrase embedding



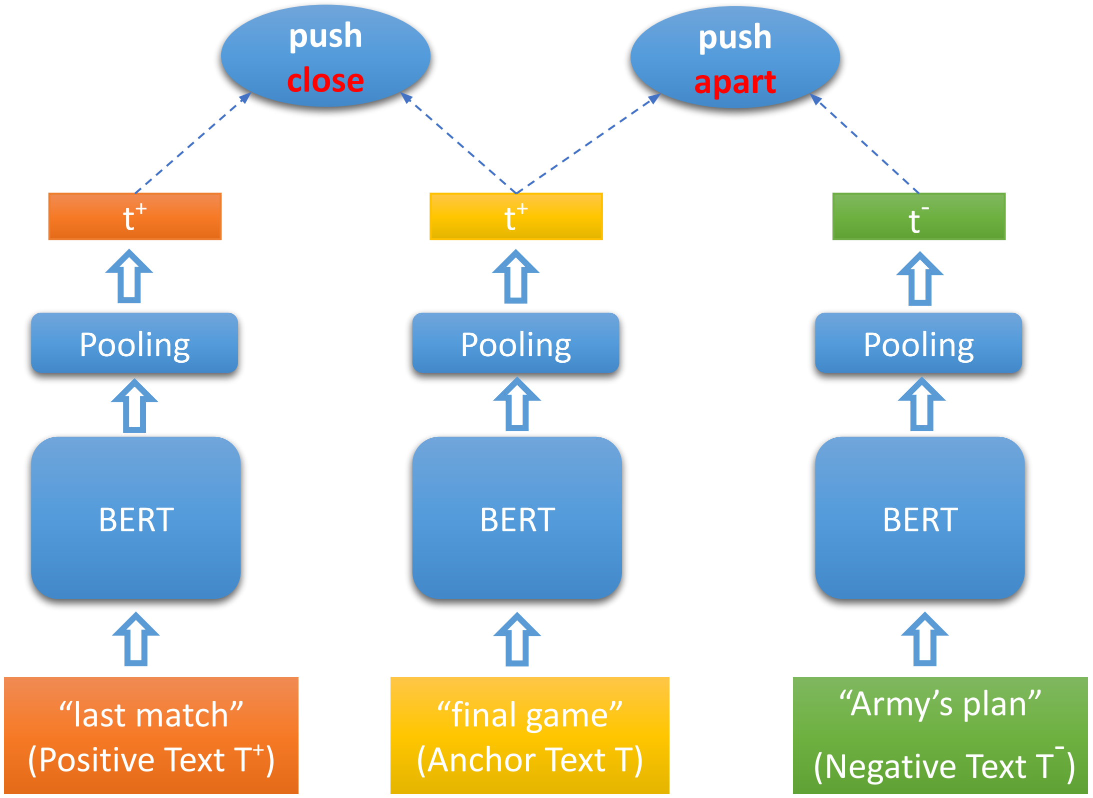


- Meaningful phrase representation
- Lexical diversity in embedding space
- Easily integrated with a phrase-based topic model

A Contrastive Learning approach

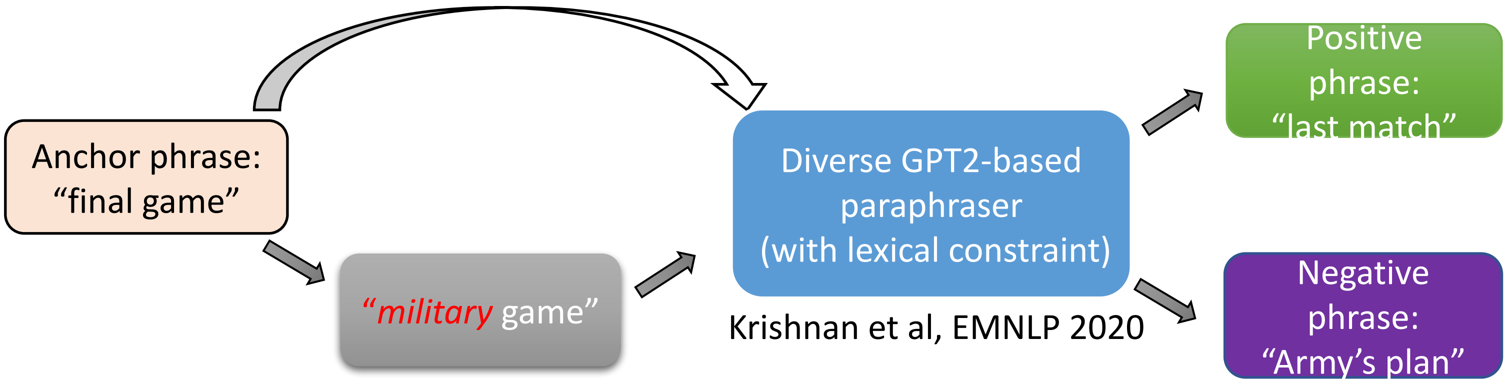


A Triplet Loss Objective



A dataset of diverse paraphrases

- Construct diverse paraphrases to remove lexical cues:



A dataset of phrase-in-context

- Context are from Books 3 Corpus (a large scale 100 GB collection of books from various genres)

Positive Context:
…… The current **legal education** consists of a 5-year-long course after which the scholar is granted a bachelor's degree ……

Anchor phrase: "Legal education"

- Negative contexts are randomly sampled
- Mask off the actual occurrence of the anchor phrase

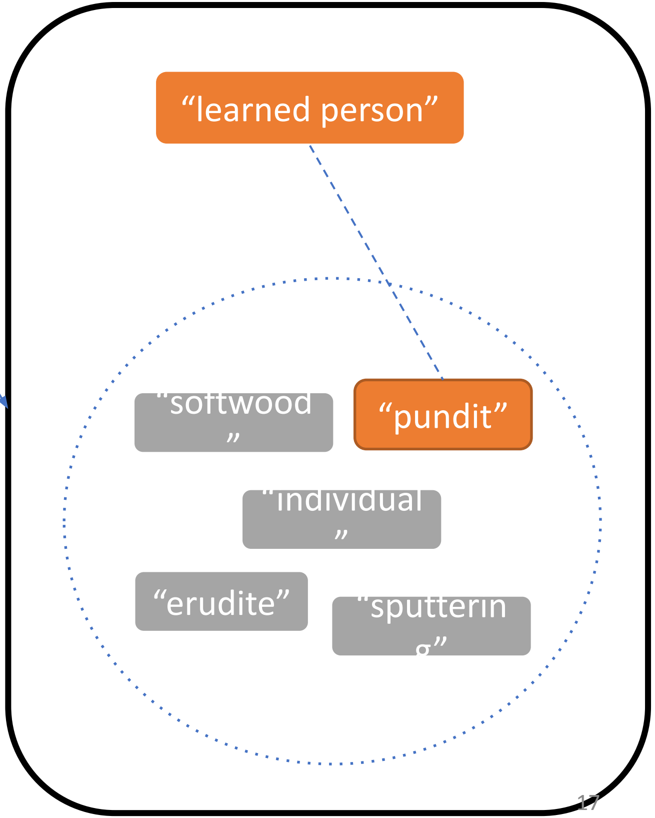
Phrase Semantics Evaluation

A diverse set of phrase-level evaluation tasks

- Turney: unigram-bigram concept match
- BiRD: bi-gram Relatedness
- PPDB: paraphrase detection
- PPDB-filtered: paraphrase detection
- PAWS-short: paraphrase detection

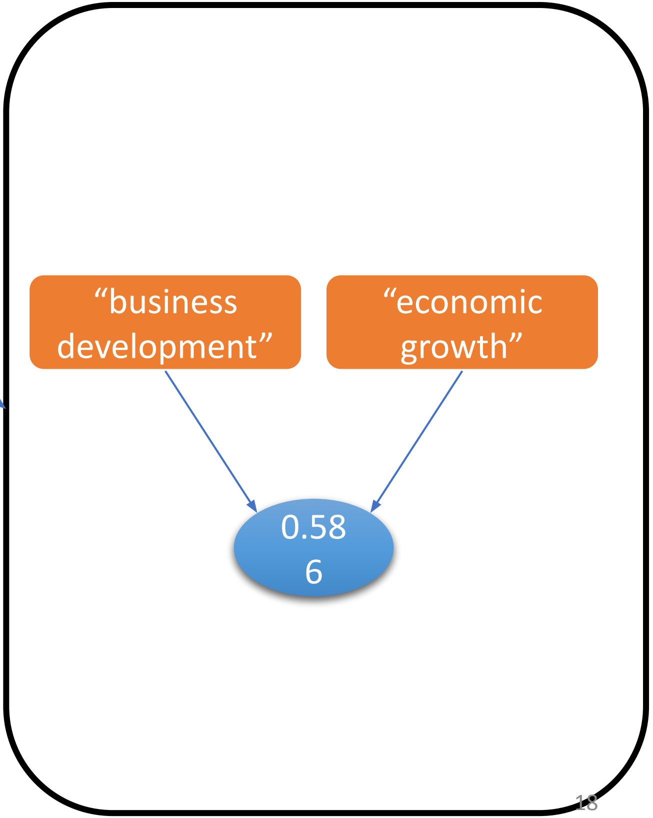
A diverse set of phrase-level evaluation tasks

- Turney: unigram-bigram concept match
- BiRD: bi-gram Relatedness
- PPDB: paraphrase detection
- PPDB-filtered: paraphrase detection
- PAWS-short: paraphrase detection



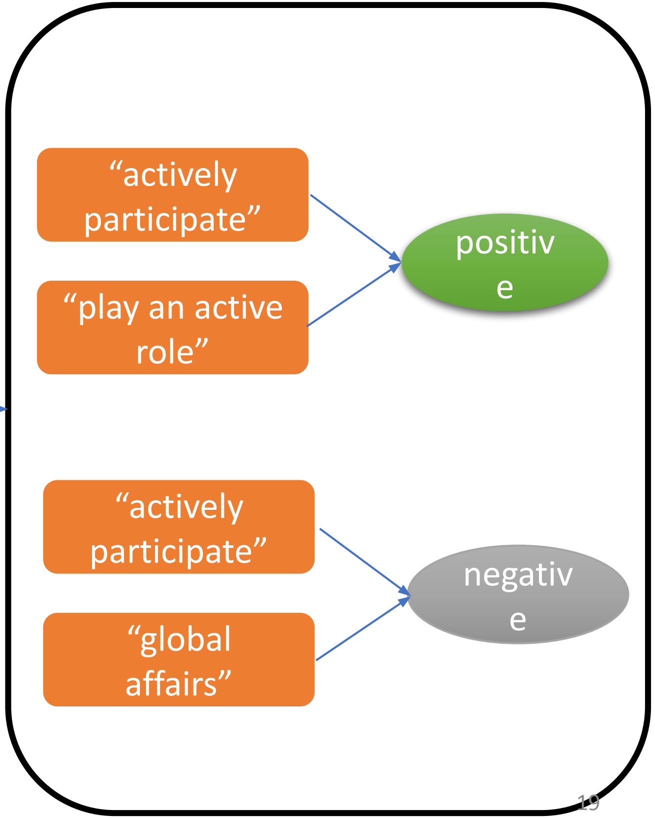
A diverse set of phrase-level evaluation tasks

- Turney: unigram-bigram concept match
- BiRD: bi-gram Relatedness
- PPDB: paraphrase detection
- PPDB-filtered: paraphrase detection
- PAWS-short: paraphrase detection



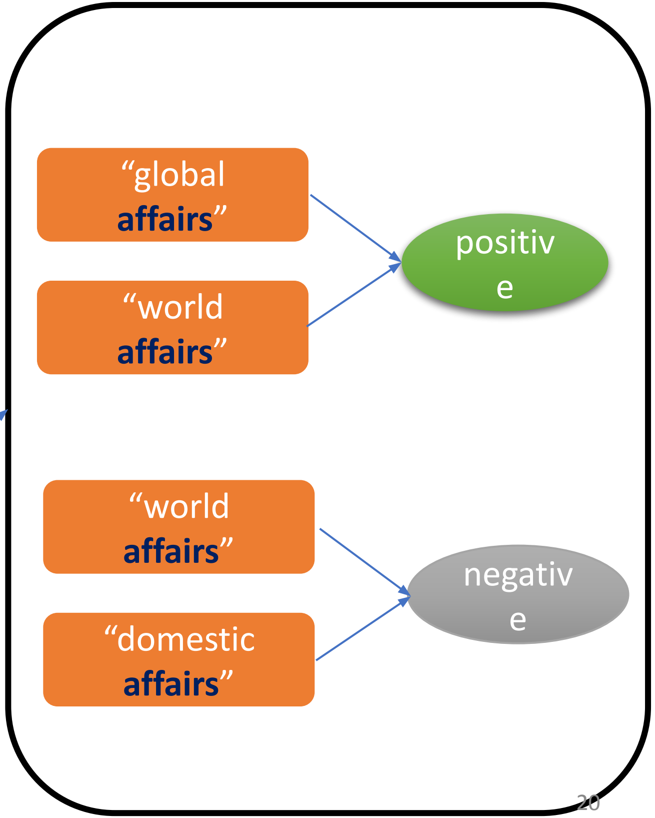
A diverse set of phrase-level evaluation tasks

- Turney: unigram-bigram concept match
- BiRD: bi-gram Relatedness
- PPDB: paraphrase detection
- PPDB-filtered: paraphrase detection
- PAWS-short: paraphrase detection



A diverse set of phrase-level evaluation tasks

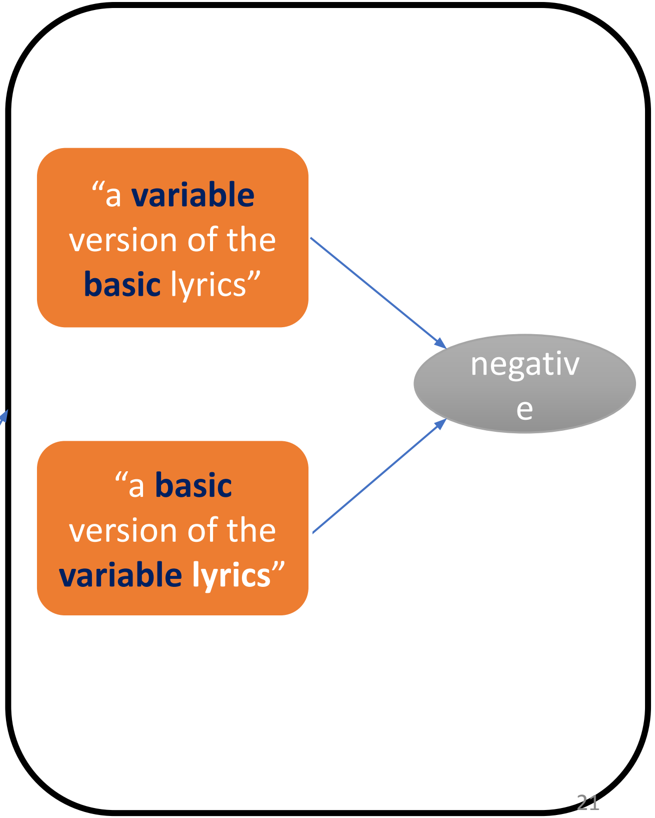
- Turney: unigram-bigram concept match
- BiRD: bi-gram Relatedness
- PPDB: paraphrase detection
- PPDB-filtered*: paraphrase detection
- PAWS-short: paraphrase detection



* Yu and Ettinger, ACL 2020

A diverse set of phrase-level evaluation tasks

- Turney: unigram-bigram concept match
- BiRD: bi-gram Relatedness
- PPDB: paraphrase detection
- PPDB-filtered: paraphrase detection
- PAWS-short: paraphrase detection



Phrases semantics evaluation

Model	Turney	BiRD	PPDB-filtered	PPDB	PAWS-short
Glove	37.8	0.560	44.2	47.2	50.0
BERT	42.6	0.444	60.1	86.2	50.0
SpanBERT	38.7	0.258	57.3	95.1	50.1
Sentence-BERT	51.8	0.687	64.2	95.8	50.0
Phrase-BERT	**57.2**	**0.688**	**68.0**	**97.6**	**58.9**

Phrases semantics evaluation

Model	Turney	BiRD	PPDB-filtered	PPDB	PAWS-short
Glove	37.8	0.560	44.2	47.2	50.0
BERT	42.6	0.444	60.1	86.2	50.0
SpanBERT	38.7	0.258	57.3	95.1	50.1
Sentence-BERT	51.8	0.687	64.2	95.8	50.0
Phrase-BERT	**57.2**	**0.688**	**68.0**	**97.6**	**58.9**

Phrase-BERT consistently outperforms baselines across all tasks

Phrases semantics evaluation

Model	Turney	BiRD	PPDB-filtered	PPDB	PAWS-short
Glove	**37.8**	**0.560**	**44.2**	**47.2**	**50.0**
BERT	**42.6**	**0.444**	**60.1**	**86.2**	**50.0**
SpanBERT	38.7	0.258	57.3	95.1	50.1
Sentence-BERT	51.8	0.687	64.2	95.8	50.0
Phrase-BERT	57.2	0.688	68.0	97.6	58.9

BERT sometimes underperforms un-contextualized baselines

Phrases semantics evaluation

Model	Turney	BiRD	PPDB-filtered	PPDB	PAWS-short
Glove	37.8	0.560	44.2	47.2	50.0
BERT	42.6	0.444	60.1	86.2	50.0
SpanBERT	**38.7**	**0.258**	**57.3**	**95.1**	**50.1**
Sentence-BERT	51.8	0.687	64.2	95.8	50.0
Phrase-BERT	57.2	0.688	68.0	97.6	58.9

SpanBERT is not the solution for phrases representation

Phrases semantics evaluation

Model	Turney	BiRD	PPDB-filtered	PPDB	PAWS-short
Glove	37.8	0.560	44.2	47.2	50.0
BERT	42.6	0.444	60.1	86.2	50.0
SpanBERT	38.7	0.258	57.3	95.1	50.1
Sentence-BERT	**51.8**	**0.687**	**64.2**	**95.8**	**50.0**
Phrase-BERT	57.2	0.688	68.0	97.6	58.9

Sentence-BERT (contrastively fine-tuned) offers closer performance to Phrase-BERT

Phrases semantics evaluation

Model	Turney	BiRD	PPDB-filtered	PPDB	PAWS-short
Glove	37.8	0.560	44.2	47.2	50.0
BERT	42.6	0.444	60.1	**86.2**	50.0
SpanBERT	38.7	0.258	57.3	**95.1**	50.1
Sentence-BERT	51.8	0.687	64.2	**95.8**	50.0
Phrase-BERT	57.2	0.688	68.0	**97.6**	58.9

BERT-based baselines are great at taking lexical cues

Phrases semantics evaluation

Model	Turney	BiRD	PPDB-filtered	PPDB	PAWS-short
Glove	37.8	0.560	44.2	47.2	50.0
BERT	42.6	0.444	**60.1**	**86.2**	50.0
SpanBERT	38.7	0.258	**57.3**	**95.1**	50.1
Sentence-BERT	51.8	0.687	**64.2**	**95.8**	50.0
Phrase-BERT	57.2	0.688	**68.0**	**97.6**	58.9

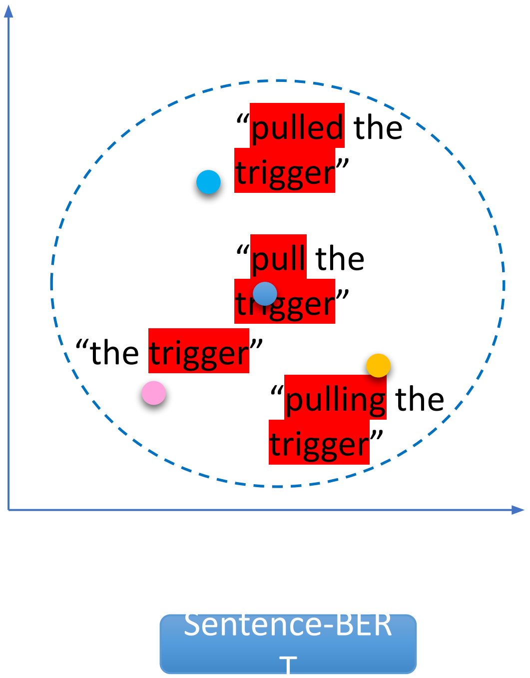
But over-reliance on lexical overlaps hurts phrase semantics understanding

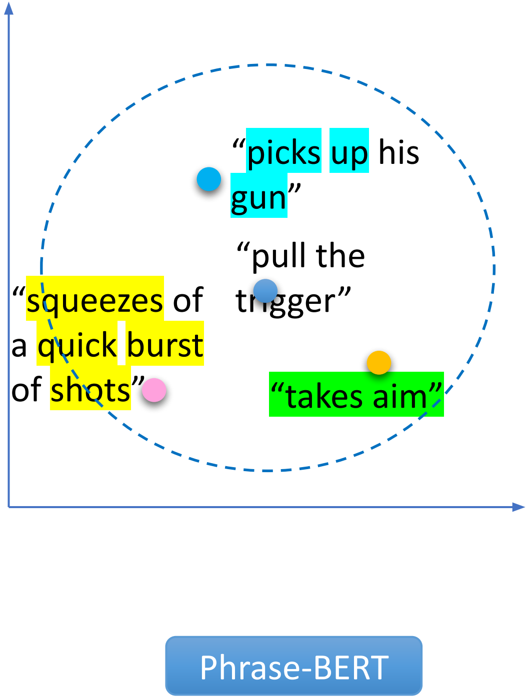
Phrases semantics evaluation

Model	Turney	BiRD	PPDB-filtered	PPDB	PAWS-short
Glove	37.8	0.560	44.2	47.2	**50.0**
BERT	42.6	0.444	60.1	86.2	**50.0**
SpanBERT	38.7	0.258	57.3	95.1	**50.1**
Sentence-BERT	51.8	0.687	64.2	95.8	**50.0**
Phrase-BERT	57.2	0.688	68.0	97.6	**58.9**

Compositionality is hard to capture for all models

Lexical diversity





Lexical diversity

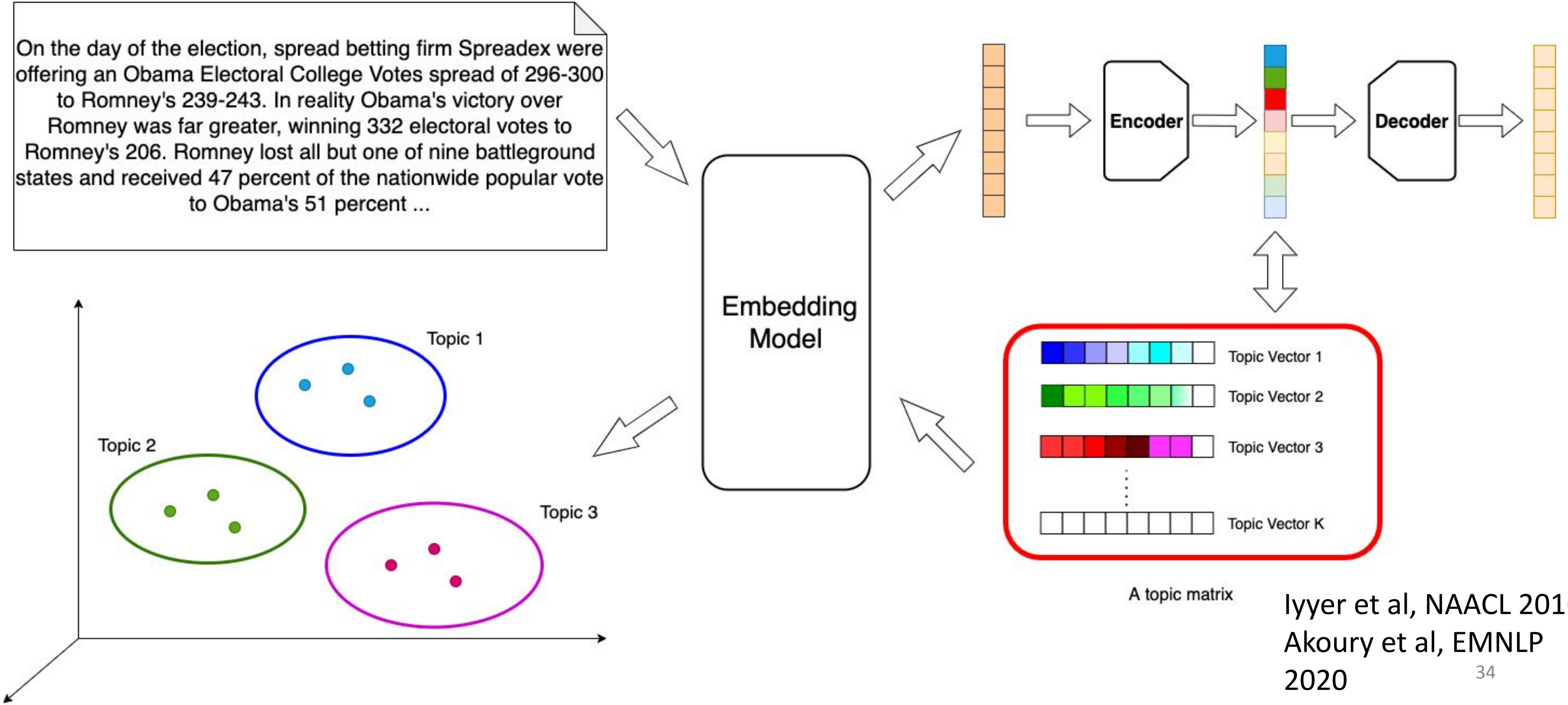
Model	percentage of new tokens (↑= more diverse)	LCS-Precision (↓= more diverse)	Levenshtein distance (↑= more diverse)
Sentence-BERT	5.0	51.1	8.5
Phrase-BERT	**5.3**	**47.6**	**8.7**

Phrase-based topic model: a case study

Baseline model: LDA

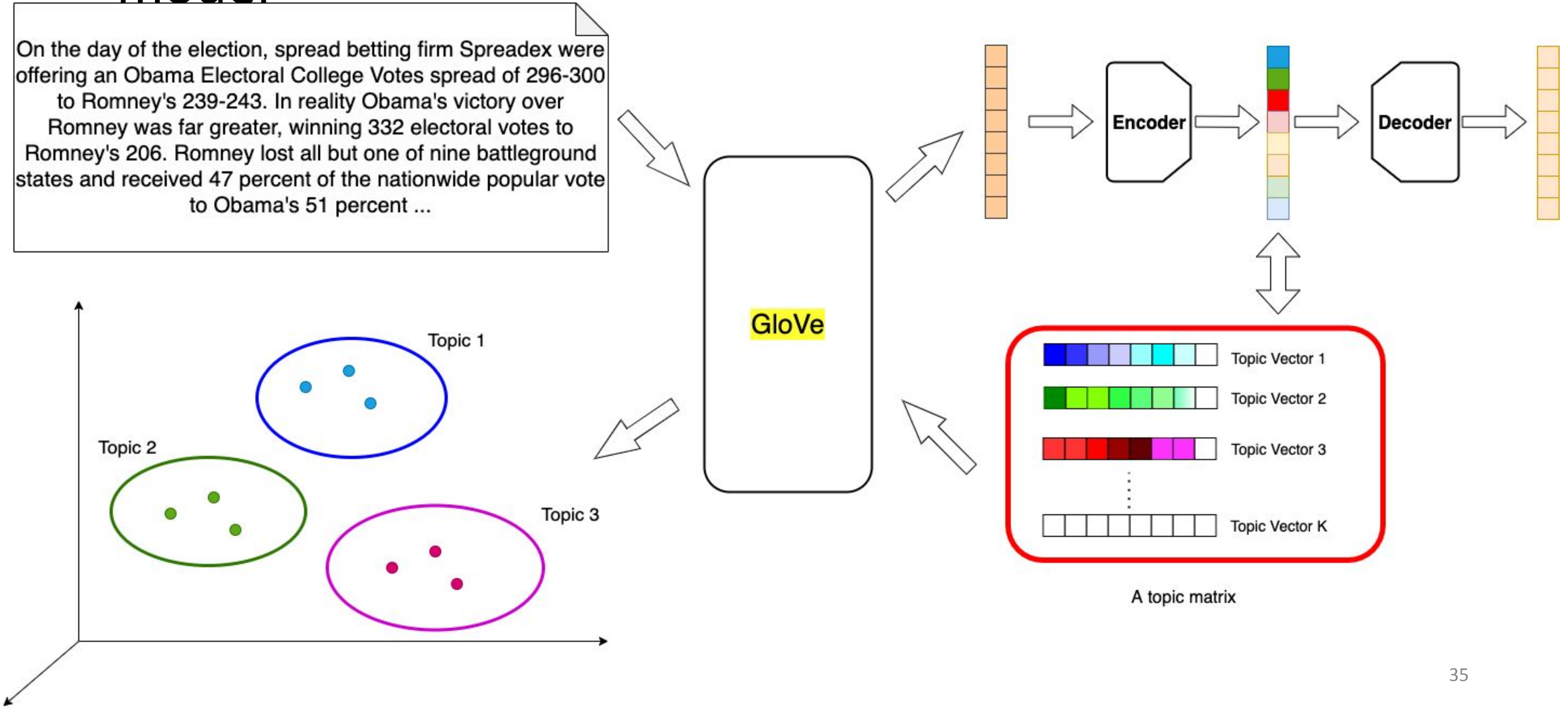
- LDA = latent dirichlet allocation
- The dominant approach for corpus exploration
- Each topic is represented by a list of unigrams.

Autoencoder + phrase-BERT = Phrase-based topic model

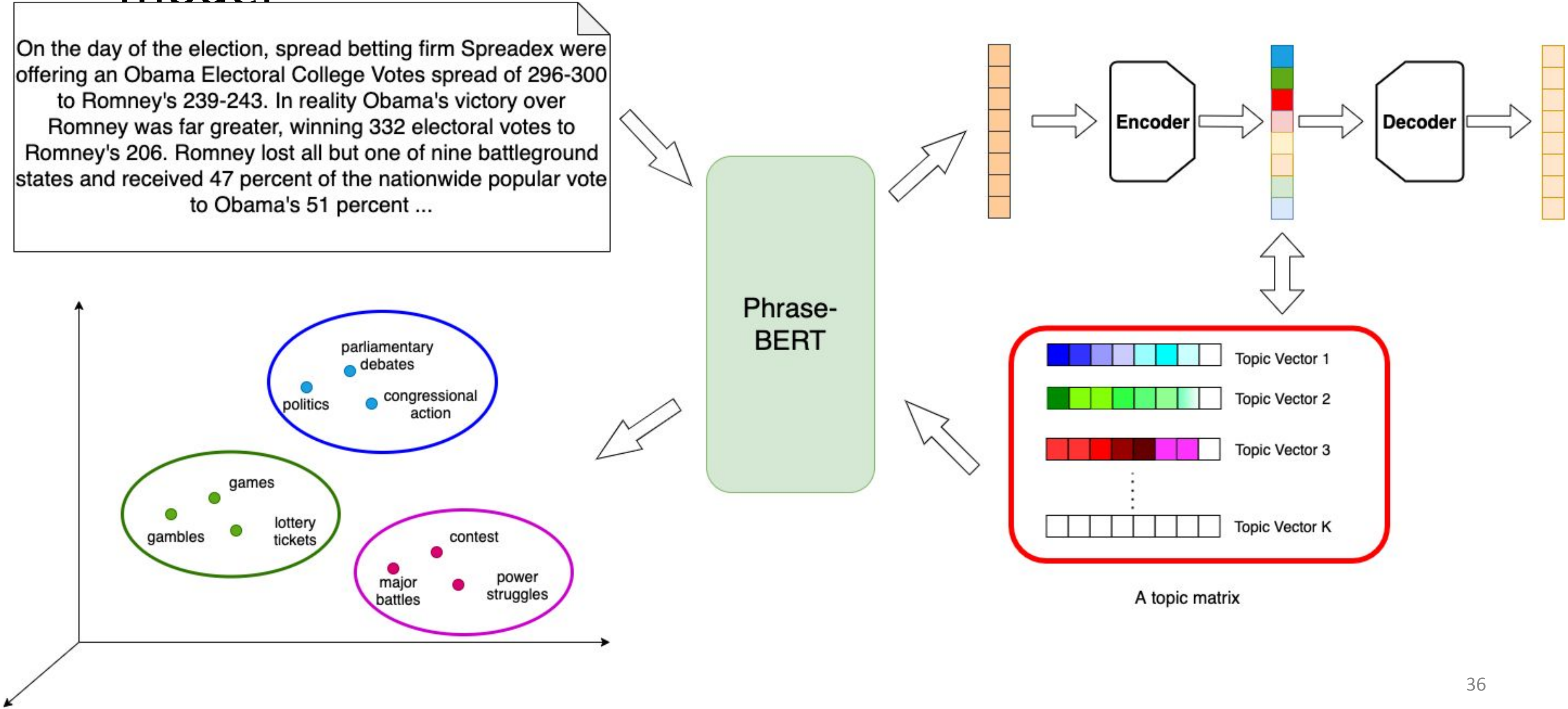


Iyyer et al, NAACL 2016
Akoury et al, EMNLP 2020

Autoencoder + phrase-BERT = Phrase-based topic model



Autoencoder + phrase-BERT = Phrase-based topic model



Human Evaluation on Topic Coherence

Model	Wiki	Story	Reviews	Average
Phrase-BERT based topic model (**PNTM**)	**83.3**	**76.7**	**77.3**	**79.1**
Phrase-LDA (**pLDA**)	55.3	48.7	50.7	51.6
Topical N-Gram (**TNG**)	37.3	58.7	60.0	52.1
Unigram (GloVe) based topic model (**UNTM**)	76.9	70.0	62.0	69.6
Latent Dirichlet Allocation (**LDA**)	48.7	48.0	52.7	49.8

Word Intrusion Test for Topic Coherence
(↑= more coherent)

Human Evaluation on Topic-to-document relatedness

Model	**Accuracy**
Phrase-BERT based topic model (**PNTM**)	**89.3**
Phrase-LDA (**pLDA**)	89.3
Topical N-Gram (**TNG**)	78.7
Unigram (GloVe) based topic model (**UNTM**)	80.7
Latent Dirichlet Allocation (**LDA**)	90.0

Topic Intrusion Test for Topic Assignment (↑= more agreement with humans in assigning topics to documents)

Phrase-BERT enables meaningful topic by reducing lexical overlap

Topic on "sports"

winning, semifinalist, finisher, a race, raceme, race, the race, the race's, side rowing competition, formula one

the semifinal, Olympic, marathon, raceme, bicyclist, semifinalist, side rowing race, race, place finish, side rowing competition

Topic on "music"

the Beatles, his album, the album, discography, the album, an album, Beatles, this album, the Beatles', their album

musician, his music, musical, concerto, chorale, liver performance, a concert, accompaniment, pianistic, antiphonal

BERT

Phrase-BERT

Conclusion and Future work

- Our contributions:
  - Phrase-BERT, a more powerful phrase embedding models
  - ↑ capture phrase semantics, ↓ lexical overlap
  - Easily integrated into a phrase-based topic model

- Future directions:
  - Towards a benchmark of holistic phrase-evaluation
  - Multi-lingual phrase embeddings,
  - Other downstream tasks involving phrases

Thank you!