Evaluation Benchmarks and Learning Criteria for Discourse-Aware Sentence Representations

Mingda Chen

Joint work with Zewei Chu and Kevin Gimpel

Prior work on evaluation benchmarks

- Focus on capabilities of representations for stand-alone sentences
  - Sentiment analysis
  - Linguistic properties, e.g. verb tense prediction
  - …
- What about the broader context (i.e. discourse) for a sentence?

Our contributions

- An evaluation suite for evaluating discourse knowledge encoded in **sentence representations**.
- Benchmark and compare several pretrained sentence representations.
- Novel learning criteria for capturing discourse structures.

Discourse Evaluation (DiscoEval)

- Focus on evaluating the role of a sentence in its discourse context.
- 7 task groups, covering multiple domains (e.g. Wikipedia, stories, dialogues, and scientific literature).
- Probing tasks. Pretrained embeddings are kept fixed and we only use simple classifiers.

Discourse Evaluation (DiscoEval)

- In general, we follow SentEval and use following input for tasks involving pairs of sentences x_1, x_2

$$[x_1, x_2, x_1 \odot x_2, |x_1 - x_2|]$$

Discourse Evaluation (DiscoEval)

- In general, we follow SentEval and use following input for tasks involving pairs of sentences x_1, x_2

$$[x_1, x_2, x_1 \odot x_2, |x_1 - x_2|]$$

Discourse Evaluation (DiscoEval)

- In general, we follow SentEval and use following input for tasks involving pairs of sentences x_1, x_2

$$[x_1, x_2, x_1 \odot x_2, |x_1 - x_2|]$$

Discourse Evaluation (DiscoEval)

- In general, we follow SentEval and use following input for tasks involving pairs of sentences x_1, x_2

$$[x_1, x_2, x_1 \odot x_2, |x_1 - x_2|]$$

What is a discourse?

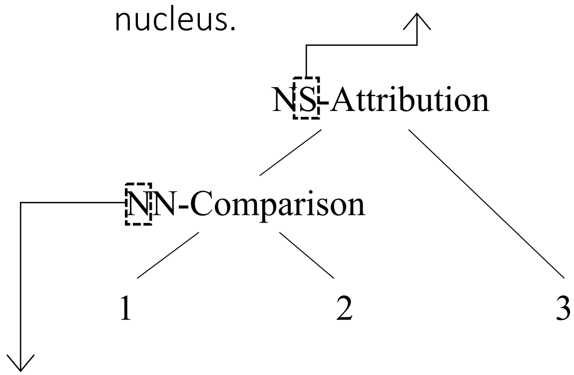
- A discourse is a coherent, structured group of sentences that acts as a fundamental type of structure in natural language.

What is a discourse?

- Linearly-structured, e.g. sentence ordering.
  - The timing of introducing entities.
- Tree-structured, e.g. RST discourse tree.

1. The European Community's consumer price index rose a provisional 0.6% in September from August
2. and was up 5.3% from September 1988,
3. according to Eurostat, the EC's statistical agency.

“S” represents “satellite”, containing additional information about the nucleus.



“N” represents “nucleus”, containing basic information for the relation.

Discourse Relations

- Two human-annotated datasets: Penn Discourse Treebank (PDTB) and RST Discourse Treebank (RST-DT).
- PDTB provides discourse markers for **adjacent sentences**, whereas RST-DT offers **document-level** discourse trees.

Discourse Relations – PDTB

- Use a pair of sentences to predict discourse relations.
- We focus on predicting implicit relations (PDTB-I) and explicit relations (PDTB-E).

PDTB-E

1. In any case, the brokerage firms are clearly moving faster to create new ads than they did in the fall of 1987.
2. ~~But~~ it remains to be seen whether their ads will be any more effective.

Label: Comparison.Contrast

PDTB-I

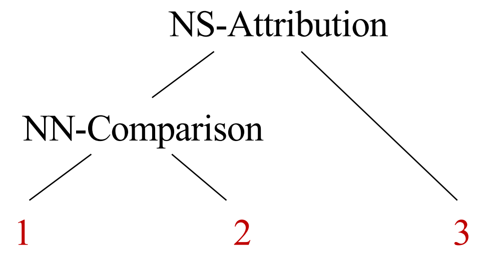
1. "A lot of investor confidence comes from the fact that they can speak to us," he says.
2. [so] "To maintain that dialogue is absolutely crucial."

Label: Contingency.Cause

Discourse Relations – RST-DT

- Text is segmented into basic units, elementary discourse units (EDUs), upon which a discourse tree is built recursively.
- We use 18 fine-grained relations.

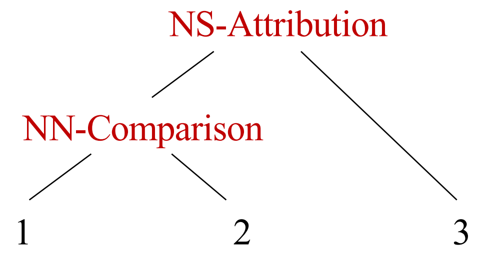
1. The European Community's consumer price index rose a provisional 0.6% in September from August
2. and was up 5.3% from September 1988,
3. according to Eurostat, the EC's statistical agency.



Discourse Relations – RST-DT

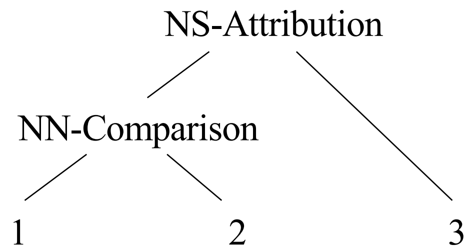
- Text is segmented into basic units, elementary discourse units (EDUs), upon which a discourse tree is built recursively.
- We use 18 fine-grained relations.

1. The European Community's consumer price index rose a provisional 0.6% in September from August
2. and was up 5.3% from September 1988,
3. according to Eurostat, the EC's statistical agency.



Discourse Relations – RST-DT

1. The European Community's consumer price index rose a provisional 0.6% in September from August
2. and was up 5.3% from September 1988,
3. according to Eurostat, the EC's statistical agency.

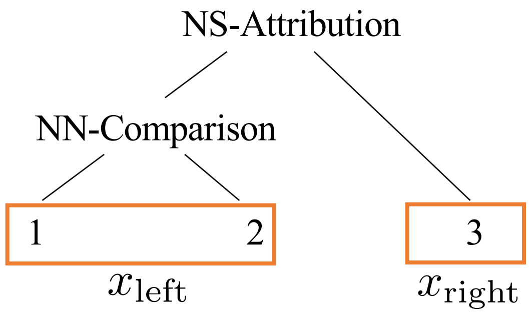


- We first encode EDUs into vectors, then use averaged vectors of EDUs of subtrees as the representation of the subtrees.
- The target prediction is the label of nodes in discourse trees.
- We use a linear classifier and the input is

$$[x_{\text{left}}, x_{\text{right}}, x_{\text{left}} \odot x_{\text{right}}, |x_{\text{left}} - x_{\text{right}}|]$$

Discourse Relations – RST-DT

1. The European Community's consumer price index rose a provisional 0.6% in September from August
2. and was up 5.3% from September 1988,
3. according to Eurostat, the EC's statistical agency.



- We first encode EDUs into vectors, then use averaged vectors of EDUs of subtrees as the representation of the subtrees.
- The target prediction is the label of nodes in discourse trees.
- We use a linear classifier and the input is

$$\left[x_{\text{left}}, x_{\text{right}}, x_{\text{left}} \odot x_{\text{right}}, \left|x_{\text{left}} - x_{\text{right}}\right|\right]$$

Sentence Position (SP)

- Probe the knowledge of a linearly-structured discourse.
- Data source: Wikipedia article, ROC Stories corpus, and arXiv papers.
- We take **five consecutive sentences** from a corpus, randomly move one of these five sentences to the first position, and ask models to **predict the true position of the first sentence** in the modified sequence.

- She was excited thinking she must have lost weight.
- Bonnie hated trying on clothes.
- She picked up a pair of size 12 jeans from the display.
- When she tried them on they were too big!
- Then she realized they actually size 14s, and 12s.

Sentence Position (SP)

- Probe the knowledge of a linearly-structured discourse.
- Data source: Wikipedia article, ROC Stories corpus, and arXiv papers.
- We take **five consecutive sentences** from a corpus, randomly move one of these five sentences to the first position, and ask models to **predict the true position of the first sentence** in the modified sequence.

True position

- She was excited thinking she must have lost weight.
- Bonnie hated trying on clothes.
- She picked up a pair of size 12 jeans from the display.
- When she tried them on they were too big!
- Then she realized they actually size 14s, and 12s.

Discourse Coherence (DC)

- Binary prediction: determine whether a sequence of 6 sentences forms a coherent paragraph.
- Data source: Ubuntu IRC Channel and Wikipedia.
- We start with a coherent sequence of six sentences, then randomly replace one of the sentences (chosen uniformly among positions 2-5) with a sentence from another discourse.

Discourse Coherence (DC)

- An example from the Wikipedia domain

1. The Broadway production took place on May 1, 1947, at the Ethel Barrymore Theatre.
2. The Metropolitan Opera presented it once, on July 31, 1965.
3. After years on the job, Ramsay has found himself one of the division's few real experts .
4. Despite his attempts to get her attention for sufficient time to ask his question, Lucy is occupied with interminable conversations on the telephone.
5. Between her calls, when Lucy leaves the room, Ben even takes the risk of trying to cut the telephone cord, though his attempt is unsuccessful.
6. Not wanting to miss his train, Ben leaves without asking Lucy for her hand in marriage.

Discourse Coherence (DC)

- An example from the Wikipedia domain

1. The Broadway production took place on May 1, 1947, at the Ethel Barrymore Theatre.
2. The Metropolitan Opera presented it once, on July 31, 1965.
3. After years on the job, Ramsay has found himself one of the division's few real experts .
4. Despite his attempts to get her attention for sufficient time to ask his question, Lucy is occupied with interminable conversations on the telephone.
5. Between her calls, when Lucy leaves the room, Ben even takes the risk of trying to cut the telephone cord, though his attempt is unsuccessful.
6. Not wanting to miss his train, Ben leaves without asking Lucy for her hand in marriage.

Discourse Coherence (DC)

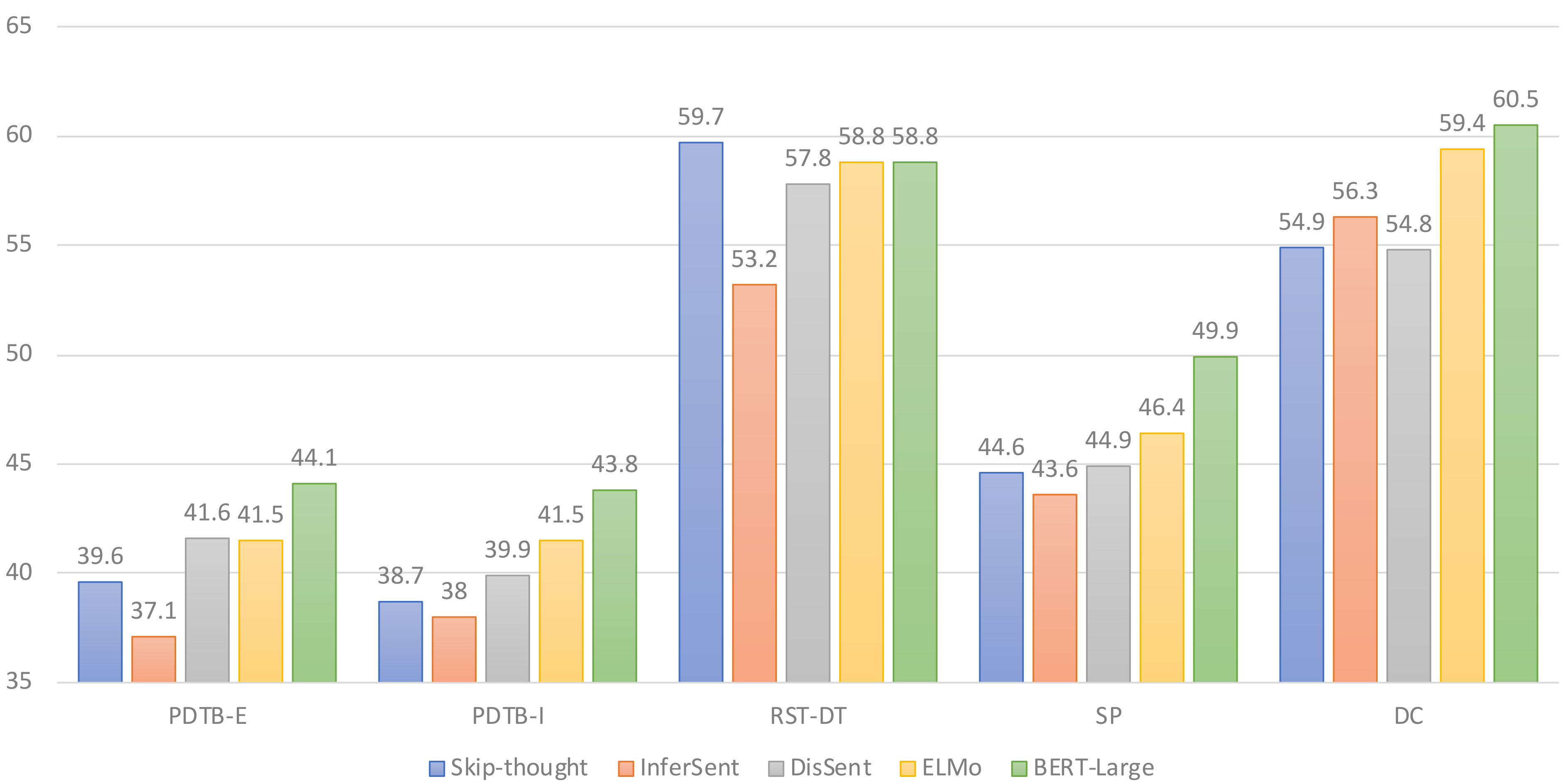
- Solving this task is non-trivial as it may require the ability to perform inference across multiple sentences.

Experiments

indicates models that are trained to encode neighboring sentence information.

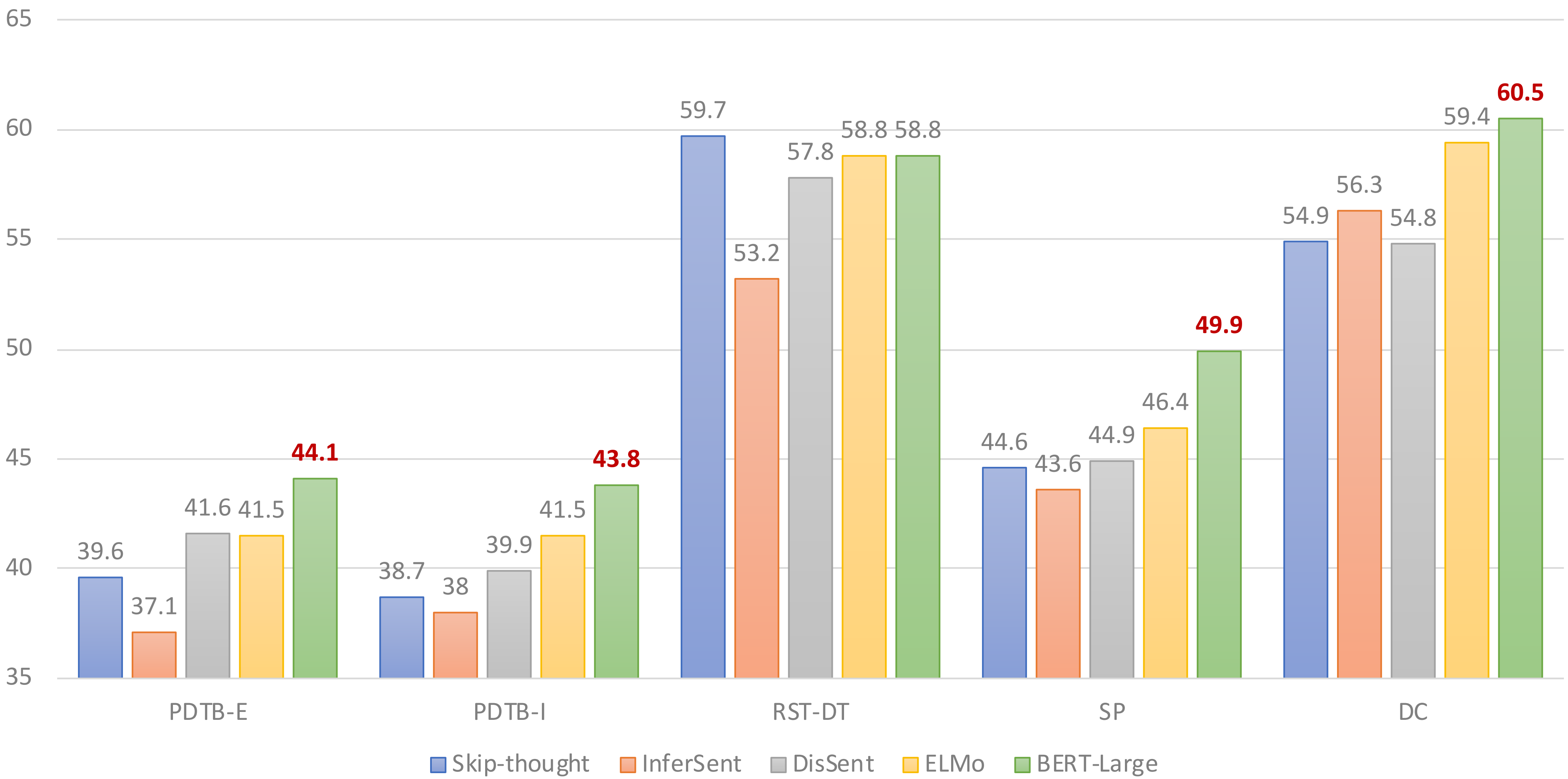
- We benchmark following pretrained models on DiscoEval:
  - Skip-thought
  - DisSent
  - BERT
  - InferSent
  - ELMo

Experiments – Benchmark pretrained models on DiscoEval



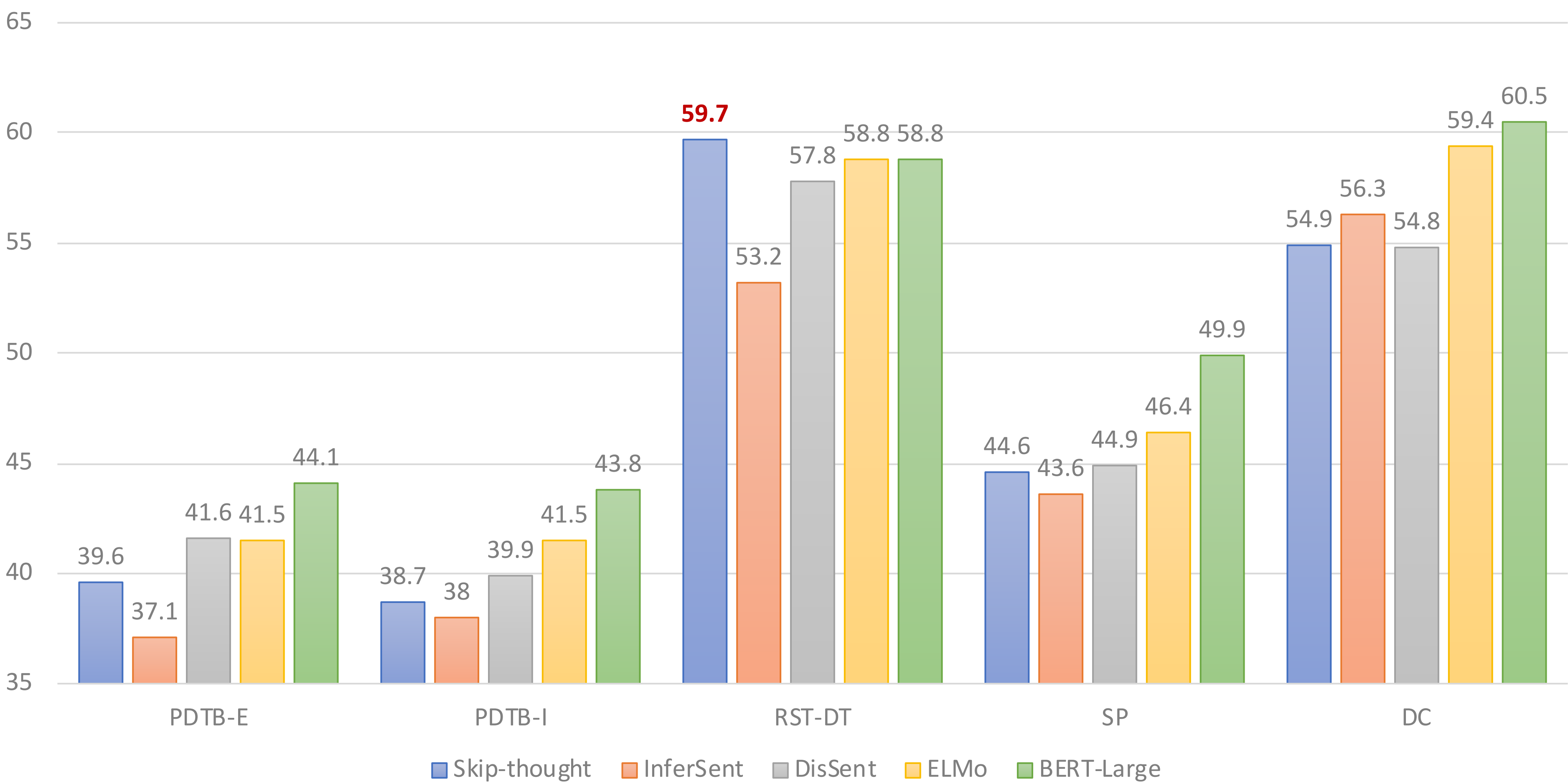
Experiments – Benchmark pretrained models on DiscoEval

- BERT-Large performs best for the most of tasks.



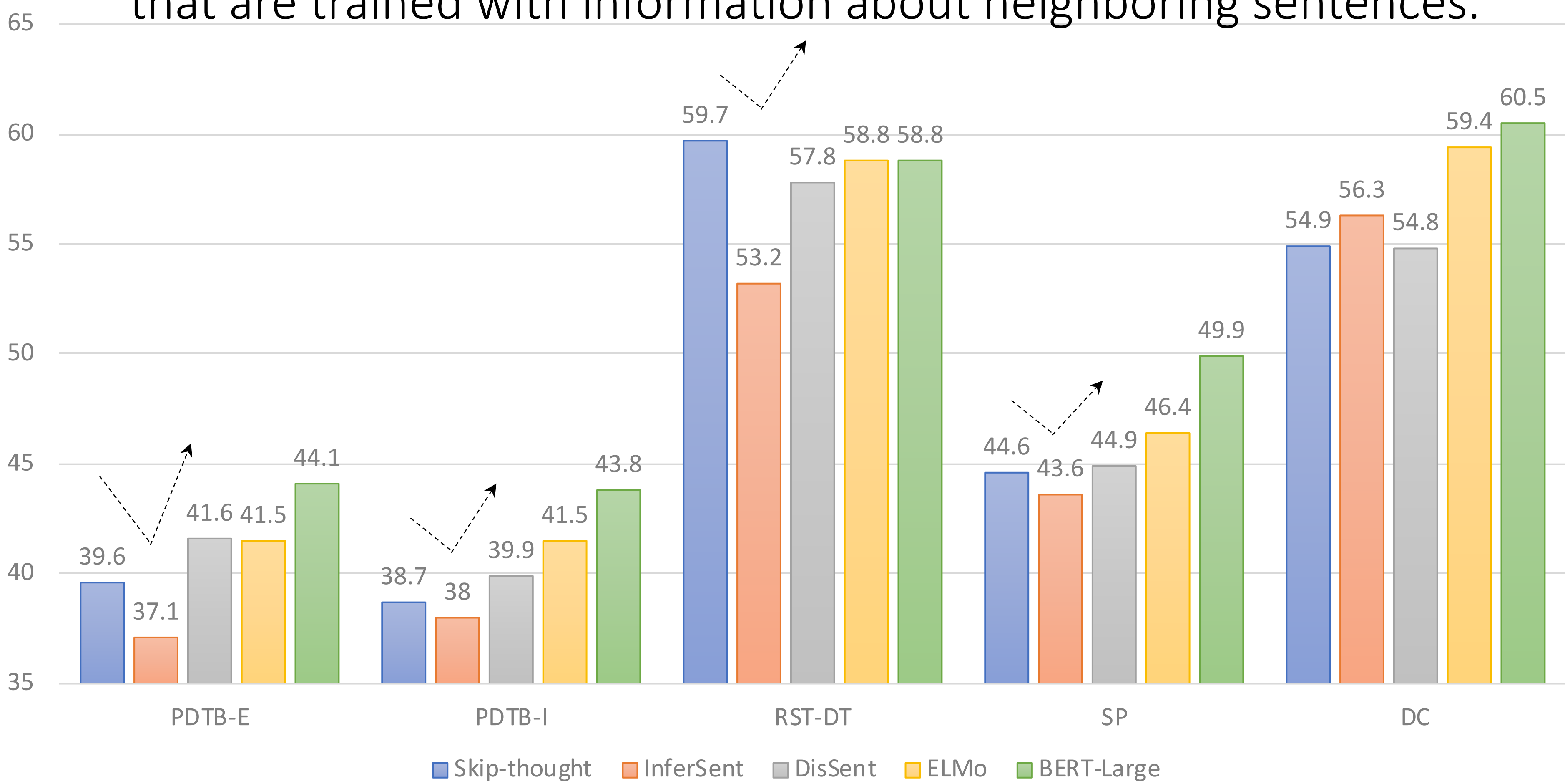
Experiments – Benchmark pretrained models on DiscoEval

- Skip-thought performs best on RST-DT.



Experiments – Benchmark pretrained models on DiscoEval

- InferSent performs much worse than other pretrained embeddings that are trained with information about neighboring sentences.

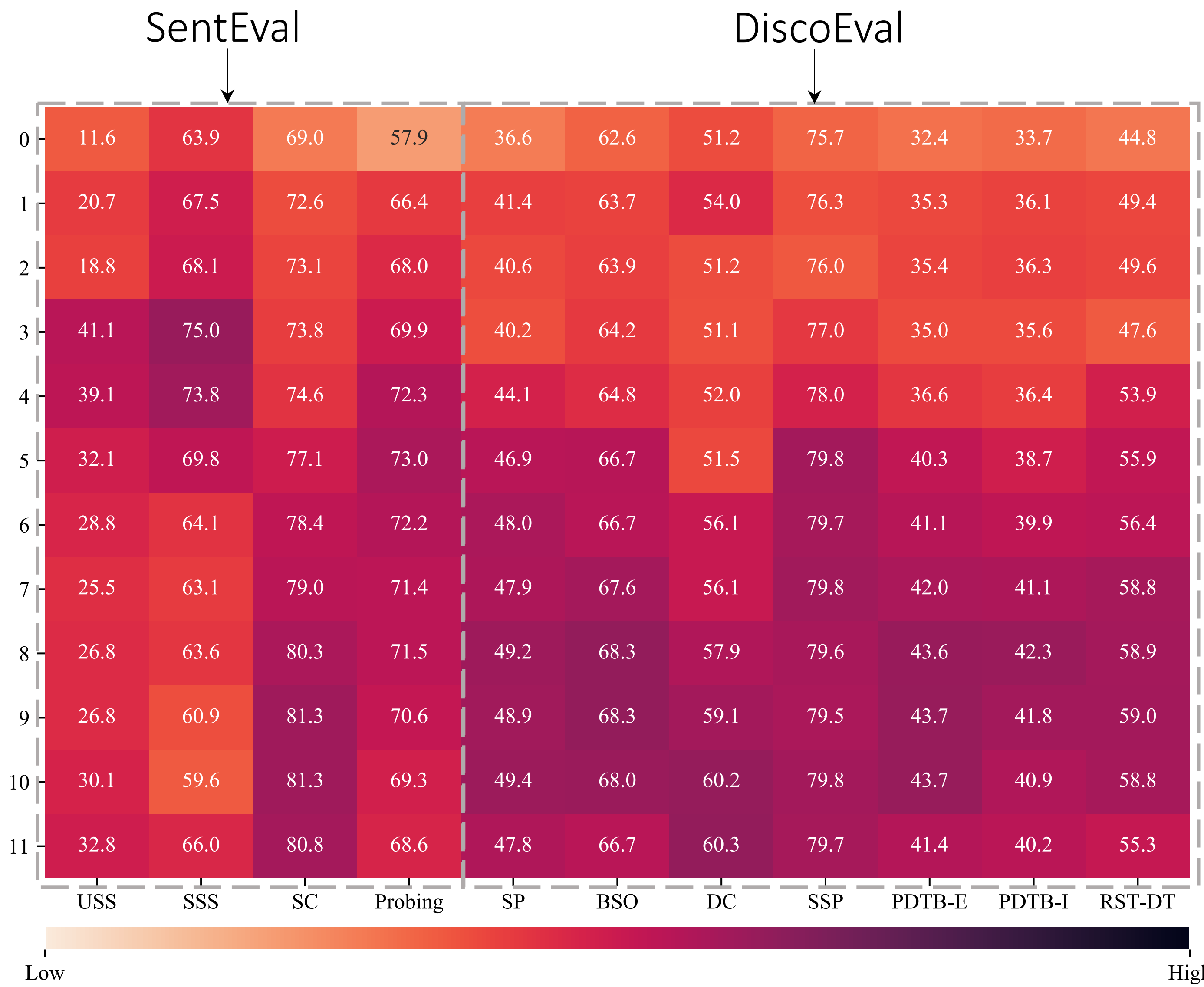


Experiments – Per-Layer analysis based on BERT

	USS	SSS	SC	Probing	SP	BSO	DC	SSP	PDTB-E	PDTB-I	RST-DT
0	11.6	63.9	69.0	57.9	36.6	62.6	51.2	75.7	32.4	33.7	44.8
1	20.7	67.5	72.6	66.4	41.4	63.7	54.0	76.3	35.3	36.1	49.4
2	18.8	68.1	73.1	68.0	40.6	63.9	51.2	76.0	35.4	36.3	49.6
3	41.1	75.0	73.8	69.9	40.2	64.2	51.1	77.0	35.0	35.6	47.6
4	39.1	73.8	74.6	72.3	44.1	64.8	52.0	78.0	36.6	36.4	53.9
5	32.1	69.8	77.1	73.0	46.9	66.7	51.5	79.8	40.3	38.7	55.9
6	28.8	64.1	78.4	72.2	48.0	66.7	56.1	79.7	41.1	39.9	56.4
7	25.5	63.1	79.0	71.4	47.9	67.6	56.1	79.8	42.0	41.1	58.8
8	26.8	63.6	80.3	71.5	49.2	68.3	57.9	79.6	43.6	42.3	58.9
9	26.8	60.9	81.3	70.6	48.9	68.3	59.1	79.5	43.7	41.8	59.0
10	30.1	59.6	81.3	69.3	49.4	68.0	60.2	79.8	43.7	40.9	58.8
11	32.8	66.0	80.8	68.6	47.8	66.7	60.3	79.7	41.4	40.2	55.3

Low High

Experiments – Per-Layer analysis based on BERT



Experiments – Per-Layer analysis

	ELMo	BERT-Base
SentEval	0.8	5.0
DiscoEval	1.3	8.9

Average of the layer number for the best layers in SentEval and DiscoEval.

- Assumption: deeper layers ➔ higher-level structures

🡇

Aligns with the information needed to solve the discourse tasks.

Human Evaluation

	Sentence Position			Discourse Coherence	
Human	77.3			87.0	
BERT-Large	49.9			60.5	
	Wiki	arXiv	ROC	Wiki	Ubuntu
Human	84.0	76.0	94.0	98.0	74.0
BERT-Large	43.0	56.0	50.9	64.9	56.1

- Human still outperforms BERT-Large by a large margin.

Learning Criteria

- General idea: make use of document structures.
- Document structures are related to discourse comprehension, showing how are the information units unfolded.
- Naturally annotated data from structured document collections, e.g. Wikipedia.

Learning Criteria

Section and Document Title (SDT)


Natural language processing

From Wikipedia, the free encyclopedia

Not to be confused with Nonlinear programming.
Not to be confused with Neuro-linguistic programming.
This article is about language processing by computers. For the processing of language by the human brain, see Language processing in the brain.

Natural language processing (**NLP**) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation.

Contents [hide]

Gift shop

Hi. I'm your automated online assistant. How may I help you?

Ask

An automated online assistant providing customer service on a web page, an example of an application where natural language processing is a major component.[1]

History [edit]

The history of natural language processing (NLP) generally started in the 1950s, although work can be found from earlier periods. In 1950, Alan Turing published an article titled "Computing Machinery and Intelligence" which proposed what is now called the Turing test as a criterion of intelligence.

The Georgetown experiment in 1954 involved fully automatic translation of more than sixty Russian sentences into English. The authors claimed that within three or five years, machine translation would be a solved problem.[2] However, real progress was much slower, and after the ALPAC report in 1966, which found that ten-year-long research had failed to fulfill the expectations, funding for machine translation was dramatically reduced. Little further research in machine translation was conducted until the late 1980s, when the first statistical machine translation systems were developed.

Some notably successful natural language processing systems developed in the 1960s were SHRDLU, a natural language system working in restricted "blocks worlds" with restricted vocabularies, and ELIZA, a simulation of a Rogerian psychotherapist, written by Joseph Weizenbaum between 1964 and 1966. Using almost no information about human thought or emotion, ELIZA sometimes provided a startlingly human-like interaction. When the "patient" exceeded the very small knowledge base, ELIZA might provide a generic response, for example, responding to "My head hurts" with "Why do you say your head hurts?".

During the 1970s, many programmers began to write "conceptual ontologies", which structured real-world information into computer-understandable data. Examples are MARGIE (Schank, 1975), SAM (Cullingford, 1978), PAM (Wilensky, 1978), TaleSpin (Meehan, 1976), QUALM (Lehnert, 1977), Politics (Carbonell, 1979), and Plot Units (Lehnert 1981). During this time, many chatterbots were written including PARRY, Racter, and Jabberwacky.

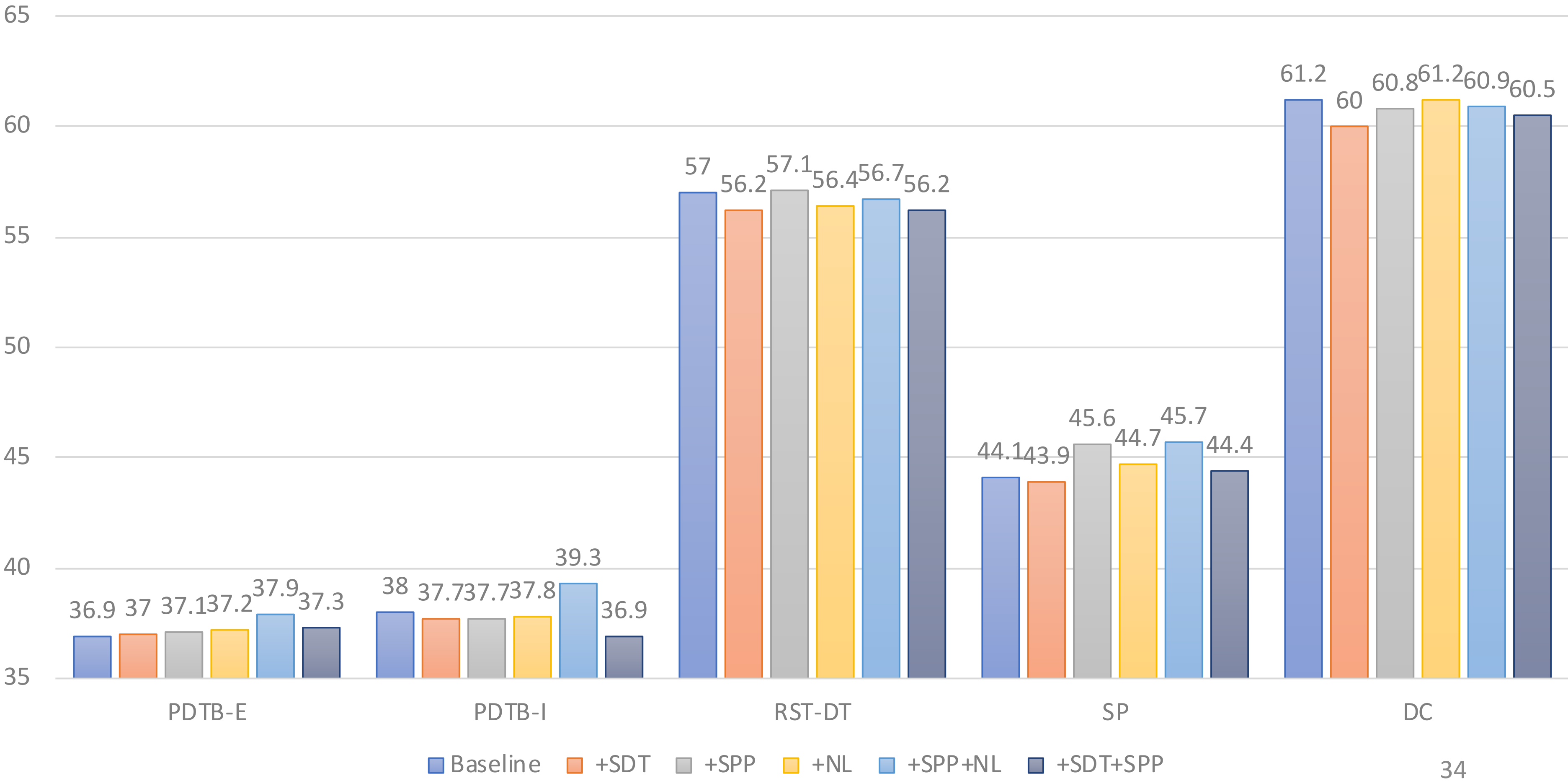
Nesting Level (NL)

Sentence and Paragraph Position (SPP)

Learning Criteria

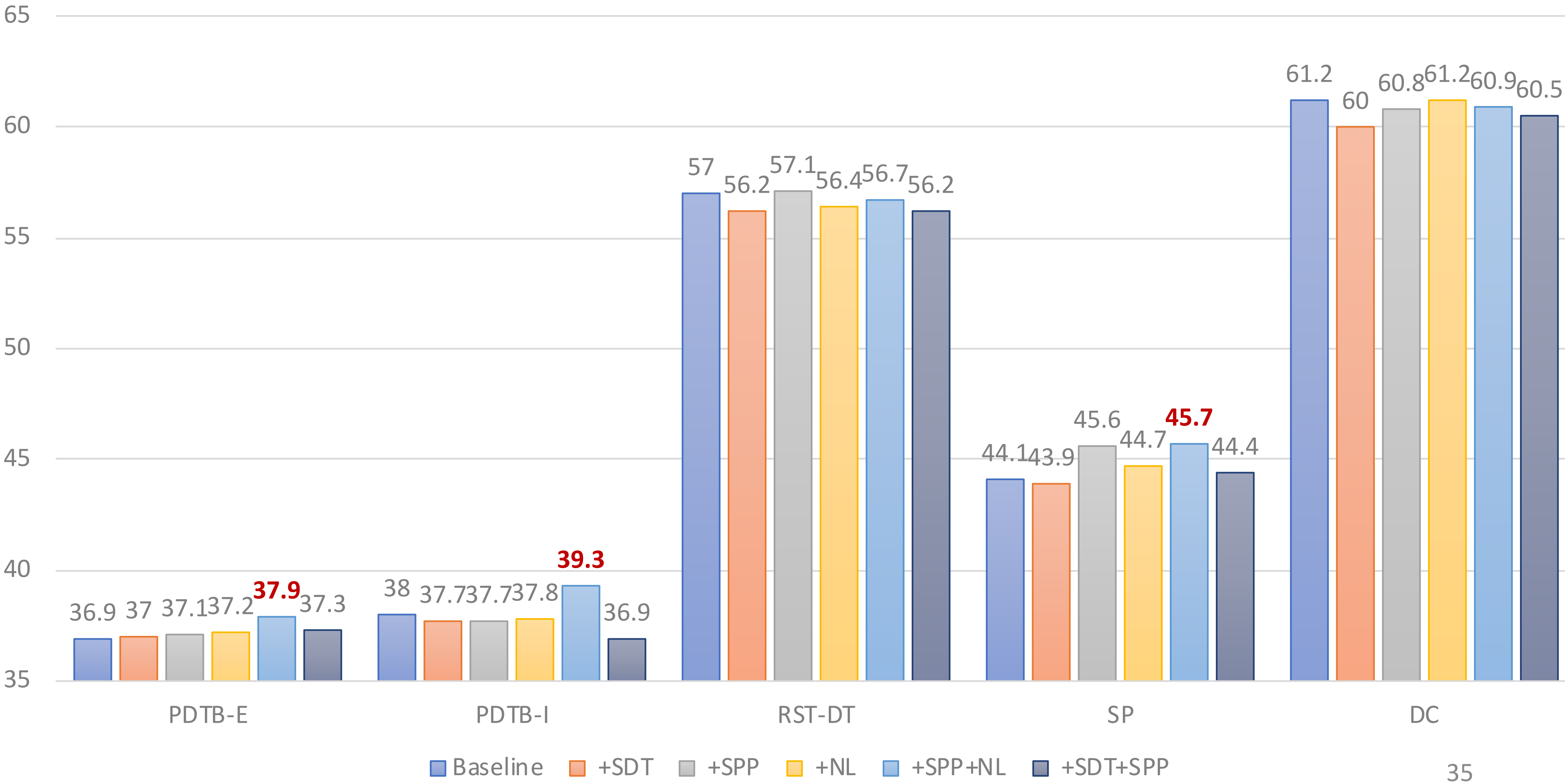
- Our models are built upon Skip-thought. All are trained with Neighboring Sentence Prediction (NSP).
- Models are trained to reconstruct bag-of-words representations of target sequences in NSP and SDT.

Experiments – Benchmark proposed learning objectives on DiscoEval



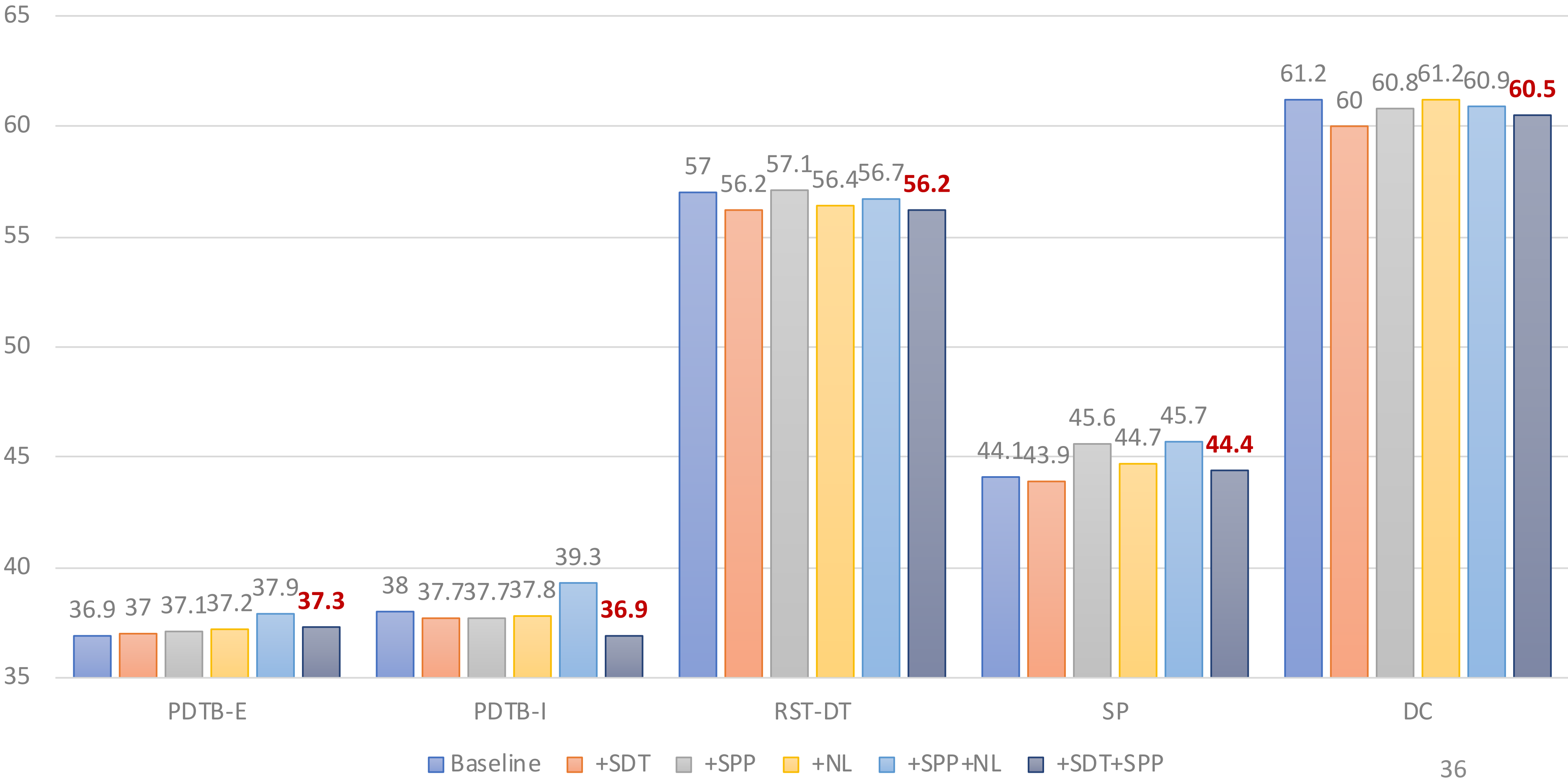
Experiments – Benchmark proposed learning objectives on DiscoEval

- SPP+NL gives the strongest performance compared to other combinations.



Experiments – Benchmark proposed learning objectives on DiscoEval

- Simply adding all the losses is not optimal as some of them could be contradictory.



Conclusion

- We introduce DiscoEval for evaluating discourse knowledge encoded in pretrained sentence representations, which is comprised of 7 task groups and covers multiple domains.
- We also introduce a set of multi-task losses that make use of document structures for learning discourse-aware sentence representations.
- Human evaluations show that humans still outperform BERT-Large by a large margin.

Thanks!

DiscoEval is available at
https://github.com/ZeweiChu/DiscoEval