Evaluation Benchmarks and Learning Criteria for Discourse-Aware Sentence Representations

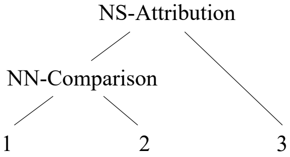
Mingda Chen[2*] **Zewei Chu**[1*] **Kevin Gimpel**[2]
[1]University of Chicago, IL, USA
[2]Toyota Technological Institute at Chicago, IL, USA
{mchen,kgimpel}@ttic.edu, zeweichu@uchicago.edu

Abstract

Prior work on pretrained sentence embeddings and benchmarks focuses on the capabilities of representations for stand-alone sentences. We propose DiscoEval, a test suite of tasks to evaluate whether sentence representations include information about the role of a sentence in its discourse context. We also propose a variety of training objectives that make use of natural annotations from Wikipedia to build sentence encoders capable of modeling discourse information. We benchmark sentence encoders trained with our proposed objectives, as well as other popular pretrained sentence encoders, on DiscoEval and other sentence evaluation tasks. Empirically, we show that these training objectives help to encode different aspects of information from the surrounding document structure. Moreover, BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018a) demonstrate strong performance across DiscoEval tasks with individual hidden layers showing different characteristics.[1]

1 Introduction

Pretrained sentence representations have been found useful in various downstream tasks such as visual question answering (Tapaswi et al., 2016), script inference (Pichotta and Mooney, 2016), and information retrieval (Le and Mikolov, 2014; Palangi et al., 2016). Benchmark datasets (Adi et al., 2017; Conneau and Kiela, 2018; Wang et al., 2018a, 2019) have been proposed to evaluate the encoded knowledge, where the focus has been primarily on natural language understanding capabilities of the representation of a stand-alone sentence, such as its semantic roles, rather than the broader context in which it is situated.



[The European Community's consumer price index rose a provisional 0.6% in September from August]$_1$ [and was up 5.3% from September 1988,]$_2$ [according to Eurostat, the EC's statistical agency.]$_3$

Figure 1: An RST discourse tree from the RST Discourse Treebank. "N" represents "nucleus", containing basic information for the relation. "S" represents "satellite", containing additional information about the nucleus.

In this paper, we seek to incorporate and evaluate discourse knowledge in general purpose sentence representations. A discourse is a coherent, structured group of sentences that acts as a fundamental type of structure in natural language (Jurafsky and Martin, 2009). A discourse structure is often characterized by the arrangement of semantic elements across multiple sentences, such as entities and pronouns. The simplest such arrangement (i.e., linearly-structured) can be understood as sentence ordering, where the structure is manifested in the timing of introducing entities. Deeper discourse structures use more complex relations among sentences (e.g., tree-structured; see Figure 1).

Theoretically, discourse structures have been approached through Centering Theory (Grosz et al., 1995) for studying distributions of entities across text and Rhetorical Structure Theory (RST; Mann and Thompson, 1988) for modelling the logical structure of natural language via discourse trees. Researchers have found modelling discourse useful in a range of tasks (Guzmán et al., 2014; Narasimhan and Barzilay, 2015; Liu and Lapata, 2018; Pan et al., 2018), including summarization (Gerani et al., 2014), text classification (Ji

[*] Equal contribution. Listed in alphabetical order.
[1] Data processing and evaluation scripts are available at https://github.com/ZeweiChu/DiscoEval.

and Smith, 2017), and text generation (Bosselut et al., 2018).

In this paper, we propose DiscoEval, a task suite designed to evaluate discourse-related knowledge in pretrained sentence representations. DiscoEval comprises 7 task groups covering multiple domains, including Wikipedia, stories, dialogues, and scientific literature. The tasks are probing tasks (Shi et al., 2016; Adi et al., 2017; Belinkov et al., 2017; Peters et al., 2018b; Conneau et al., 2018; Poliak et al., 2018; Tenney et al., 2019; Liu et al., 2019a; Ettinger, 2019; Chen et al., 2019, *inter alia*) based on sentence ordering, annotated discourse relations, and discourse coherence. The data is either generated semi-automatically or based on human annotations (Carlson et al., 2001; Prasad et al., 2008; Lin et al., 2009; Kummerfeld et al., 2019).

We also propose a set of novel multi-task learning objectives building upon standard pretrained sentence encoders, which rely on the assumption of distributional semantics of text. These objectives depend only on the natural structure in structured document collections like Wikipedia.

Empirically, we benchmark our models and several popular sentence encoders on DiscoEval and SentEval (Conneau and Kiela, 2018). We find that our proposed training objectives help the models capture different characteristics in the sentence representations. Additionally, we find that ELMo shows strong performance on SentEval, whereas BERT performs the best among the pretrained embeddings on DiscoEval. Both BERT and Skip-thought vectors (Kiros et al., 2015), which have training losses explicitly related to surrounding sentences, perform much stronger compared to their respective prior work, demonstrating the effectiveness of incorporating losses that make use of broader context. Through per-layer analysis, we also find that for both BERT and ELMo, deep layers consistently outperform shallower ones on DiscoEval, showing different trends from SentEval where the shallow layers have the best performance.

2 Related Work

Discourse modelling and discourse parsing have a rich history (Marcu, 2000; Barzilay and Lapata, 2008; Zhou et al., 2010; Kalchbrenner and Blunsom, 2013; Ji and Eisenstein, 2015; Li and Jurafsky, 2017; Wang et al., 2018c; Liu et al., 2018; Lin et al., 2019, *inter alia*), much of it based on recovering linguistic annotations of discourse structure.

Several researchers have defined tasks related to discourse structure, including sentence ordering (Chen et al., 2016; Logeswaran et al., 2016; Cui et al., 2018), sentence clustering (Wang et al., 2018b), and disentangling textual threads (Elsner and Charniak, 2008, 2010; Lowe et al., 2015; Mehri and Carenini, 2017; Jiang et al., 2018; Kummerfeld et al., 2019).

There is a great deal of prior work on pretrained representations (Le and Mikolov, 2014; Kiros et al., 2015; Hill et al., 2016; Wieting et al., 2016; McCann et al., 2017; Gan et al., 2017; Peters et al., 2018a; Logeswaran and Lee, 2018; Devlin et al., 2019; Tang and de Sa, 2019; Yang et al., 2019; Liu et al., 2019b, *inter alia*). Skip-thought vectors form an effective architecture for general-purpose sentence embeddings. The model encodes a sentence to a vector representation, and then predicts the previous and next sentences in the discourse context. Since Skip-thought performs well in downstream evaluation tasks, we use this neighboring-sentence objective as a starting point for our models.

There is also work on incorporating discourse related objectives into the training of sentence representations. Jernite et al. (2017) propose binary sentence ordering, conjunction prediction (requiring manually-defined conjunction groups), and next sentence prediction. Similarly, Sileo et al. (2019) and Nie et al. (2019) create training datasets automatically based on discourse relations provided in the Penn Discourse Treebank (PDTB; Lin et al., 2009).

Our work differs from prior work in that we propose a general-purpose pretrained sentence embedding evaluation suite that covers multiple aspects of discourse knowledge and we propose novel training signals based on document structure, including sentence position and section titles, without requiring additional human annotation.

3 Discourse Evaluation

We propose DiscoEval, a test suite of 7 tasks to evaluate whether sentence representations include semantic information relevant to discourse processing. Below we describe the tasks and datasets, as well as the evaluation framework. We closely follow the SentEval sentence embedding evaluation suite, in particular its supervised sentence and sentence pair classification tasks, which use predefined neural architectures with slots for fixed-dimensional sentence embeddings. All DiscoEval tasks are modelled by logistic regression unless otherwise stated in later sections.

We also experimented with adding hidden layers to the DiscoEval classification models. However, we find simpler linear classifiers to provide a clearer comparison among sentence embedding methods. More complex classification models lead to noisier results, as more of the modelling burden is shifted to the optimization of the classifiers. Hence we decide to evaluate the sentence embeddings with simple classification models.

In the rest of this section, we will use $[\cdot, \cdot, \cdots]$ to denote concatenation of vectors, $\odot$ for element-wise multiplication, and $|\cdot|$ for element-wise absolute value.

3.1 Discourse Relations

As the most direct way to probe discourse knowledge, we consider the task of predicting annotated discourse relations among sentences. We use two human-annotated datasets: the RST Discourse Treebank (RST-DT; Carlson et al., 2001) and the Penn Discourse Treebank (PDTB; Prasad et al., 2008). They have different labeling schemes. PDTB provides discourse markers for adjacent sentences, whereas RST-DT offers document-level discourse trees, which recently was used to evaluate discourse knowledge encoded in document-level models (Ferracane et al., 2019). The difference allows us to see if the pretrained representations capture local or global information about discourse structure.

More specifically, as shown in Figure 1, in RST-DT, text is segmented into basic units, elementary discourse units (EDUs), upon which a discourse tree is built recursively. Although a relation can take multiple units, we follow prior work (Ji and Eisenstein, 2014) to use right-branching trees for non-binary relations to binarize the tree structure and use the 18 coarse-grained relations defined by Carlson et al. (2001).

When evaluating pretrained sentence encoders on RST-DT, 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 and the input to the classifier is $[x_{\text{left}}, x_{\text{right}}, x_{\text{left}} \odot x_{\text{right}}, |x_{\text{left}} - x_{\text{right}}|]$, where x_{left} and x_{right} are vector representations of the left and right subtrees respectively. For example, the input for target "NN-Attribution" in Figure 1 would be $x_{\text{left}} = \frac{x_1+x_2}{2}$, $x_{\text{right}} = x_3$, where x_i is the encoded representation for the ith EDU in the text. We use the standard data splits, where there are 347 documents for training and 38 documents for testing. We choose 35 documents from the training set to serve as a validation set.

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

Figure 2: Example in the PDTB explicit relation task. The words in [] are taken out from input sentence 2.

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

Figure 3: Example in the PDTB implicit relation task.

For PDTB, we use a pair of sentences to predict discourse relations. Following Lin et al. (2009), we focus on two kinds of relations from PDTB: explicit (PDTB-E) and implicit (PDTB-I). The sentence pairs with explicit relations are two consecutive sentences with a particular connective word in between. Figure 2 is an example of an explicit relation.

In the PDTB, annotators insert an implicit connective between adjacent sentences to reflect their relations, if such an implicit relation exists. Figure 3 shows an example of an implicit relation. The PDTB provides a three-level hierarchy of relation tags. In DiscoEval, we use the second level of types (Lin et al., 2009), as they provide finer semantic distinctions compared to the first level. To ensure there is a reasonable amount of evaluation data, we use sections 2-14 as training set, 15-18 as development set, and 19-23 as test set. In addition, we filter out categories that have less than 10 instances. This leaves us 12 categories for explicit relations and 11 for implicit ones. Category names are listed in the supplementary material.

We use the sentence embeddings to infer sentence relations with supervised training. As input to the classifier, we encode both sentences to vector representations x_1 and x_2, concatenated with their element-wise product and absolute difference: $[x_1, x_2, x_1 \odot x_2, |x_1 - x_2|]$.

- 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.

Figure 4: Example from the ROC Stories domain of the Sentence Position task. The first sentence should be in the fourth position.

3.2 Sentence Position (SP)

We create a task that we call Sentence Position. It can be seen as way to probe the knowledge of linearly-structured discourse, where the ordering corresponds to the timings of events. When constructing this dataset, 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.

We create three versions of this task, one for each of the following three domains: the first five sentences of the introduction section of a Wikipedia article (Wiki), the ROC Stories corpus (ROC; Mostafazadeh et al., 2016), and the first 5 sentences in the abstracts of arXiv papers (arXiv; Chen et al., 2016). Figure 4 shows an example of this task for the ROC Stories domain. The first sentence should be in the fourth position among these sentences. To make correct predictions, the model needs to be aware of both typical orderings of events as well as how events are described in language. In the example shown, Bonnie's excitement comes from her imagination so it must happen after she picked up the jeans and tried them on but right before she realized the actual size.

To train classifiers for these tasks, we do the following. We first encode the five sentences to vector representations x_i. As input to the classifier, we include x_1 and the concatenation of $x_1 - x_i$ for all i: $[x_1, x_1 - x_2, x_1 - x_3, x_1 - x_4, x_1 - x_5]$.

3.3 Binary Sentence Ordering (BSO)

Similar to sentence position prediction, Binary Sentence Ordering (BSO) is a binary classification task to determine the order of two sentences. The fact that BSO only has a pair of sentences as input makes it different from Sentence Position, where there is more context, and we hope that BSO can evaluate the ability of capturing local discourse coherence in the given sentence representations. The data comes from the same three domains as Sentence Position, and each instance is a pair of consecutive sentences.

1. These functions include fast and synchronized response to environmental change, or long-term memory about the transcriptional status.
2. Focusing on the collective behaviors on a population level, we explore potential regulatory functions this model can offer.

Figure 5: Example from the arXiv domain of the Binary Sentence Ordering task (incorrect ordering shown).

Figure 5 shows an example from the arXiv domain of the Binary Sentence Ordering task. The order of the sentences in this instance is incorrect, as the "functions" are referenced before they are introduced. To detect the incorrect ordering in this example, the encoded representations need to be able to provide information about new and old information in each sentence.

To form the input when training classifiers, we concatenate the embeddings of both sentences with their element-wise difference: $[x_1, x_2, x_1 - x_2]$.

3.4 Discourse Coherence (DC)

Inspired by prior work on chat disentanglement (Elsner and Charniak, 2008, 2010) and sentence clustering (Wang et al., 2018b), we propose a sentence disentanglement task. The task is to determine whether a sequence of six sentences forms a coherent paragraph. 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. This task, which we call Discourse Coherence (DC), is a binary classification task and the datasets are balanced between positive and negative instances.

We use data from two domains for this task: Wikipedia and the Ubuntu IRC channel.[2] For Wikipedia, we begin by choosing a sequence of six sentences from a Wikipedia article. For purposes of choosing difficult distractor sentences, we use the Wikipedia categories of each document as an indication of its topic. To create a negative instance, we randomly sample a sentence from another document with a similar set of categories (measured by the percentage of overlapping categories). This sampled sentence replaces one of the six consecutive sentences in the original sequence. When splitting the train, development,

[2] irclogs.ubuntu.com/

1. It is possible he was the youngest of the family as the name “Sextus” translates to sixth in English implying he was the sixth of two living and three stillborn brothers.
2. According to Roman tradition, his rape of Lucretia was the precipitating event in the overthrow of the monarchy and the establishment of the Roman Republic.
3. Tarquinius Superbus was besieging Ardea, a city of the Rutulians.
4. The place could not be taken by force, and the Roman army lay encamped beneath the walls.
5. **He was soon elected to the Academy’s membership (although he had to wait until 1903 to be elected to the Society of American Artists), and in 1883 he opened a New York studio, dividing his time for several years between Manhattan and Boston.**
6. As nothing was happening in the field, they mounted their horses to pay a surprise visit to their homes.

Figure 6: An example from the Wikipedia domain of the Discourse Coherence task. This sequence is not coherent; the boldface sentence was substituted in for the true fifth sentence from another article.

and test sets, we ensure there are no overlapping documents among them.

Our proposed dataset differs from the sentence clustering task of Wang et al. (2018b) in that it preserves sentence order and does not anonymize or lemmatize words, because they play an important role in conveying information about discourse coherence.

For the Ubuntu domain, we use the human annotations of conversation thread structure from Kummerfeld et al. (2019) to provide us with a coherent sequence of utterances. We filter out sentences by heuristic rules to avoid overly technical and unsolvable cases. The negative sentence is randomly picked from other conversations. Similarly, when splitting the train, development, and test sets, we ensure there are no overlapping conversations among them.

Figure 6 is an instance of the Wikipedia domain of the Discourse Coherence task. This instance is not coherent and the boldfaced text is from a different document. The incoherence can be found either by comparing characteristics of the entity being discussed or by the topic of the sentence group. Solving this task is non-trivial as it may require the ability to perform inference across multiple sentences.

In this task, we encode all sentences to vector representations and concatenate all of them ($[x_1, x_2, x_3, x_4, x_5, x_6]$) as input to the classification model. Note that in this task, we use a hidden layer of 2000 dimensions with sigmoid activation in the classification model, as this is necessary for the classifier to use features based on multiple inputs simultaneously given the simple concatenation as input. We could have developed richer ways to encode the input so that a linear classifier would be feasible (e.g., use the element-wise products of all pairs of sentence embeddings), but we wish to keep the input dimensionality of the classifier small enough that the classifier will be learnable given fixed sentence embeddings and limited training data.

1. The theory behind the SVM and the naive Bayes classifier is explored.
2. This relocation of the active target may be repeated an arbitrary number of times.

Figure 7: Examples from Sentence Section Prediction. The first is from an Abstract while the second is not.

Task	PDTB-E	PDTB-I	Ubuntu	RST-DT	Others
Train	9383	8693	5816	17051	10000
Dev.	3613	2972	1834	2045	4000
Test	3758	3024	2418	2308	4000

Table 1: Size of datasets in DiscoEval.

3.5 Sentence Section Prediction (SSP)

The Sentence Section Prediction (SSP) task is defined as determining the section of a given sentence. The motivation behind this task is that sentences within certain sections typically exhibit similar patterns because of the way people write coherent text. The pattern can be found based on connectives or specificity of a sentence. For example, “Empirically” is usually used in the abstract or introduction sections in scientific writing.

We construct the dataset from PeerRead (Kang et al., 2018), which consists of scientific papers from a variety of fields. The goal is to predict whether or not a sentence belongs to the Abstract section. After eliminating sentences that are too easy for the task (e.g., equations), we randomly sample sentences from the Abstract or from a section in the middle of a paper.[3] Figure 7 shows two sentences from this task, where the first sentence is more general and from an Abstract whereas the second is more specific and is from another section. In this task, the input to the classifier is simply the sentence embedding.

Table 1 shows the number of instances in each DiscoEval task introduced above.

[3] We avoid sentences from the Introduction or Conclusion sections to make the task more solvable.

4 Models and Learning Criteria

Having described DiscoEval, we now discuss methods for incorporating discourse information into sentence embedding training. All models in our experiments are composed of a single encoder and multiple decoders. The encoder, parameterized by a bidirectional Gated Recurrent Unit (BiGRU; Chung et al., 2014), encodes the sentence, either in training or in evaluation of the downstream tasks, to a fixed-length vector representation (i.e., the average of the hidden states across positions).

The decoders take the aforementioned encoded sentence representation, and predict the targets we define in the sections below. We first introduce Neighboring Sentence Prediction, the loss for our baseline model. We then propose additional training losses to encourage our sentence embeddings to capture other context information.

4.1 Neighboring Sentence Prediction (NSP)

Similar to prior work on sentence embeddings (Kiros et al., 2015; Hill et al., 2016), we use an encoded sentence representation to predict its surrounding sentences. In particular, we predict the immediately preceding and succeeding sentences. All of our sentence embedding models use this loss. Formally, the loss is defined as

$$\text{NSP} = -\log p_\theta(s_{t-1}|s_t) - \log p_\phi(s_{t+1}|s_t)$$

where we parameterize p_θ and p_ϕ as separate feedforward neural networks and compute the log-probability of a target sentence using its bag-of-words representation.

4.2 Nesting Level (NL)

A table of contents serves as a high level description of an article, outlining its organizational structure. Wikipedia articles, for example, contain rich tables of contents with many levels of hierarchical structure. The "nesting level" of a sentence (i.e., how many levels deep it resides) provides information about its role in the overall discourse. To encode this information into our sentence representations, we introduce a discriminative loss to predict a sentence's nesting level in the table of contents:

$$\text{NL} = -\log p_\theta(l_t|s_t)$$

where l_t represents the nesting level of the sentence s_t and p_θ is parameterized by a feedforward neural network. Note that sentences within the same paragraph share the same nesting level. In Wikipedia, there are up to 7 nesting levels.

4.3 Sentence and Paragraph Position (SPP)

Similar to nesting level, we add a loss based on using the sentence representation to predict its position in the paragraph and in the article. The position of the sentence can be a strong indication of the relations between the topics of the current sentence and the topics in the entire article. For example, the first several sentences often cover the general topics to be discussed more thoroughly in the following sentences. To encourage our sentence embeddings to capture such information, we define a position prediction loss

$$\text{SPP} = -\log p_\theta(sp_t|s_t) - \log p_\phi(pp_t|s_t)$$

where sp_t is the sentence position of s_t within the current paragraph and pp_t is the position of the current paragraph in the whole document.

4.4 Section and Document Title (SDT)

Unlike the previous position-based losses, this loss makes use of section and document titles, which gives the model more direct access to the topical information at different positions in the document. The loss is defined as

$$\text{SDT} = -\log p_\theta(st_t|s_t) - \log p_\phi(dt_t|s_t)$$

Where st_t is the section title of sentence s_t, dt_t is the document title of sentence s_t, and p_θ and p_ϕ are two different bag-of-words decoders.

5 Experiments

5.1 Setup

We train our models on Wikipedia as it is a knowledge rich textual resource and has consistent structures over all documents. Details on hyperparameters are in the supplementary material. When evaluating on DiscoEval, we encode sentences with pretrained sentence encoders. Following SentEval, we freeze the sentence encoders and only learn the parameters of the downstream classifier. The "Baseline" row in Table 2 are embeddings trained with only the NSP loss. The subsequent rows are trained with extra losses defined in Section 4 in addition to the NSP loss.

	SentEval				DiscoEval							
	USS	SSS	SC	Probing	SP	BSO	DC	SSP	PDTB-E	PDTB-I	RST-DT	avg.
Skip-thought	41.7	81.2	78.4	70.1	47.5	64.6	55.2	77.5	39.3	40.2	**59.7**	54.8
InferSent	**63.4**	**83.3**	79.7	71.8	45.8	62.9	56.3	62.2	37.3	38.8	52.3	50.8
DisSent	50.0	79.2	80.5	74.0	47.7	64.9	54.8	62.2	42.2	40.7	57.8	52.9
ELMo	60.9	77.6	80.8	74.7	47.8	65.6	**60.7**	79.0	41.3	41.8	57.5	56.2
BERT-Base	30.1	66.3	81.4	73.9	53.1	68.5	58.9	80.3	41.9	42.4	58.8	57.7
BERT-Large	43.6	70.7	**83.4**	**75.0**	**53.8**	**69.3**	59.6	**80.4**	**44.3**	**43.6**	59.1	**58.6**
Baseline (NSP)	57.8	77.1	77.0	70.6	47.3	63.8	61.0	77.8	36.5	39.1	56.7	54.6
+ SDT	59.0	77.3	76.8	69.7	45.8	62.9	60.3	78.0	36.6	39.1	55.7	54.1
+ SPP	56.0	77.5	77.4	70.7	48.4	65.3	60.2	78.4	38.1	39.9	56.4	55.2
+ NL	56.7	78.2	77.2	70.6	46.9	64.0	61.0	78.9	37.6	39.9	56.5	55.0
+ SPP + NL	55.4	76.7	77.0	70.4	48.5	64.7	59.9	78.9	37.8	40.5	56.7	55.3
+ SDT + NL	58.5	76.9	77.2	70.2	46.1	63.0	60.8	78.1	36.7	38.1	56.2	54.1
+ SDT +SPP	58.4	77.4	76.6	70.2	46.5	63.9	60.4	77.6	35.2	38.6	56.3	54.1
ALL	58.8	76.3	77.0	70.2	46.1	63.7	60.0	78.6	36.3	37.6	55.3	53.9

Table 2: Results for SentEval and DiscoEval. The highest number in each column is boldfaced. The highest number for our models in each column is underlined. “All” uses all four losses. “avg.” is the averaged accuracy for all tasks in DiscoEval.

Additionally, we benchmark several popular pretrained sentence encoders on DiscoEval, including Skip-thought,[4] InferSent (Conneau et al., 2017),[5] DisSent (Nie et al., 2019),[6] ELMo,[7] and BERT.[8] For ELMo, we use the averaged vector of all three layers and time steps as the sentence representations. For BERT, we use the averaged vector at the position of the “[CLS]” token across all layers. We also evaluate per-layer performance for both models in Section 6.

When reporting results for SentEval, we compute the averaged Pearson correlations for Semantic Textual Similarity tasks from 2012 to 2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016). We refer to the average as unsupervised semantic similarity (USS) since those tasks do not require training data. We compute the averaged results for the STS Benchmark (Cer et al., 2017), textual entailment, and semantic relatedness (Marelli et al., 2014) and refer to the average as supervised semantic similarity (SSS). We compute the average accuracy for movie review (Pang and Lee, 2005); customer review (Hu and Liu, 2004); opinion polarity (Wiebe et al., 2005); subjectivity classification (Pang and Lee, 2004); Stanford sentiment treebank (Socher et al., 2013); question classification (Li and Roth, 2002); and paraphrase detection (Dolan et al., 2004), and refer to it as sentence classification (SC). For the rest of the linguistic probing tasks (Conneau et al., 2018), we report the average accuracy and report it as “Probing”.

5.2 Results

Table 2 shows the experiment results over all SentEval and DiscoEval tasks. Different models and training signals have complex effects when performing various downstream tasks. We summarize our findings below:

- On DiscoEval, Skip-thought performs best on RST-DT. DisSent performs strongly for PDTB tasks but it requires discourse markers from PDTB for generating training data. BERT has the highest average by a large margin, but ELMo has competitive performance on multiple tasks.
- The NL or SPP loss alone has complex effects across tasks in DiscoEval, but when they are combined, the model achieves the best performance, outperforming our baseline by 0.7% on average. In particular, it yields 40.5% accuracy on PDTB-I, outperforming Skip-thought by 0.3%. This is presumably caused by the differing, yet complementary, effects of these two losses (NL and SPP).
- The SDT loss generally hurts performance on DiscoEval, especially on the position-related tasks (SP, BSO). This can be explained by the notion that consecutive sentences in the same section are encouraged to have the same sentence representations when using the SDT loss. However, the SP and BSO tasks involve differentiating neighboring sentences in terms of their position and ordering information.

[4] github.com/ryankiros/skip-thoughts

[5] github.com/facebookresearch/InferSent

[6] github.com/windweller/DisExtract

[7] github.com/allenai/allennlp

[8] github.com/huggingface/pytorch-pretrained-BERT

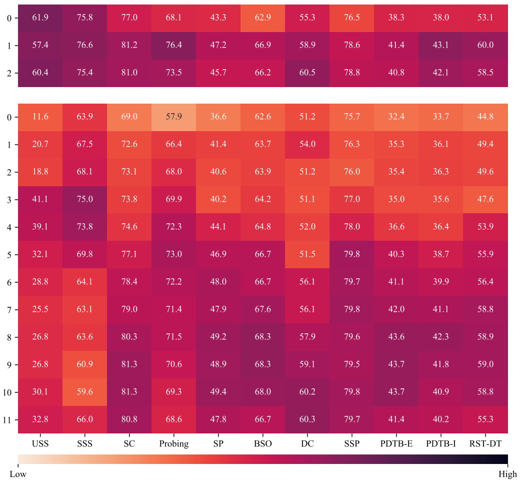


Figure 8: Heatmap for individual hidden layers of BERT-Base (lower part) and ELMo (upper part).

	ELMo	BERT-Base
SentEval	0.8	5.0
DiscoEval	1.3	8.9

Table 3: Average of the layer number for the best layers in SentEval and DiscoEval.

- On SentEval, SDT is most helpful for the USS tasks, presumably because it provides the most direct information about the topic of each sentence, which is a component of semantic similarity. SDT helps slightly on the SSS tasks. NL gives the biggest improvement in SSS.
- In comparing BERT to ELMo and Skip-thought to InferSent on DiscoEval, we can see the benefit of adding information about neighboring sentences. Our proposed training objectives show complementary improvements over NSP, which suggests that they can potentially benefit these pretrained representations.

6 Analysis

Per-Layer analysis. To investigate the performance of individual hidden layers, we evaluate ELMo and BERT on both SentEval and DiscoEval using each hidden layer. For ELMo, we use the averaged vector from the targeted layer. For BERT-Base, we use the vector from the position of the "[CLS]" token. Figure 8 shows the heatmap of performance for individual hidden layers. We note that for better visualization, colors in each column are standardized. On SentEval, BERT-Base performs better with shallow layers on USS, SSS, and Probing (though not on SC), but on DiscoEval, the results using BERT-Base gradually increase with deeper layers. To evaluate this phenomenon quantitatively, we compute the average of the layer number for the best layers for both ELMo and BERT-Base and show it in Table 3. From the table, we can see that DiscoEval requires deeper layers to achieve better performance. We assume this is because deeper layers can capture higher-level structure, which aligns with the information needed to solve the discourse tasks.

Baseline w/o hidden layer	52.0
Baseline w/ hidden layer	61.0

Table 4: Accuracies with baseline encoder on Discourse Coherence task, with or without a hidden layer in the classifier.

DiscoEval architectures. In all DiscoEval tasks except DC, we use no hidden layer in the neural architectures, following the example of SentEval. However, some tasks are unsolvable with this simple architecture. In particular, the DC tasks have low accuracies with all models unless a hidden layer is used. As shown in Table 4, when adding a hidden layer of 2000 to this task, the performance on DC improves dramatically. This shows that DC requires more complex comparison and inference among input sentences. Our human evaluation below on DC also shows that human accuracies exceed those of the classifier based on sentence embeddings by a large margin.

Human Evaluation. We conduct a human evaluation on the Sentence Position, Binary Sentence Ordering, and Discourse Coherence datasets. A native English speaker was provided with 50 examples per domain for these tasks. While the results in Table 5 show that the overall human accuracies exceed those of the classifier based on BERT-Large by a large margin, we observe that within some specific domains, for example Wiki in BSO, BERT-Large demonstrates very strong performance.

Does context matter in Sentence Position? In the SP task, the inputs are the target sentence together with 4 surrounding sentences. We study the effect of removing the surrounding 4 sentences, i.e., only using the target sentence to predict its position from the start of the paragraph.

	Sentence Position			Binary Sentence Ordering			Discourse Coherence	
Human	77.3			84.7			87.0	
BERT-Large	53.8			69.3			59.6	
	Wiki	arXiv	ROC	Wiki	arXiv	ROC	Wiki	Ubuntu
Human	84.0	76.0	94.0	64.0	72.0	96.0	98.0	74.0
BERT-Large	50.7	47.3	63.4	70.4	66.8	70.8	65.1	54.2

Table 5: Accuracies (%) for a human annotator and BERT-Large on Sentence Position, Binary Sentence Ordering, and Discourse Coherence tasks.

Random	20
Baseline w/o context	43.2
Baseline w/ context	47.3

Table 6: Accuracies (%) for baseline encoder on Sentence Position task when using downstream classifier with or without context.

Table 6 shows the comparison of the baseline model performance on Sentence Position with or without the surrounding sentences and a random baseline. Since our baseline model is already trained with NSP, it is expected to see improvements over a random baseline. The further improvement from using surrounding sentences demonstrates that the context information is helpful in determining the sentence position.

7 Conclusion

We proposed DiscoEval, a test suite of tasks to evaluate discourse-related knowledge encoded in pretrained sentence representations. We also proposed a variety of training objectives to strengthen encoders' ability to incorporate discourse information. We benchmarked several pretrained sentence encoders and demonstrated the effects of the proposed training objectives on different tasks. While our learning criteria showed benefit on certain classes of tasks, our hope is that the DiscoEval evaluation suite can inspire additional research in capturing broad discourse context in fixed-dimensional sentence embeddings.

Acknowledgments

We thank Jonathan Kummerfeld for helpful discussions about the IRC Disentanglement dataset, Davis Yoshida for discussions about BERT, and the anonymous reviewers for their feedback that improved this paper. This research was supported in part by a Bloomberg data science research grant to K. Gimpel.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR*.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *sem 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43. Association for Computational Linguistics.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 173–184, New Orleans, Louisiana. Association for Computational Linguistics.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14. Association for Computational Linguistics.

Mingda Chen, Zewei Chu, Karl Stratos, and Kevin Gimpel. 2019. Enteval: A holistic evaluation benchmark for entity representations. In *Proc. of EMNLP*.

Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2016. Neural sentence ordering. *arXiv preprint arXiv:1607.06952*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Baiyun Cui, Yingming Li, Ming Chen, and Zhongfei Zhang. 2018. Deep attentive sentence ordering network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4340–4349. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 350–356, Geneva, Switzerland.

Micha Elsner and Eugene Charniak. 2008. You talking to me? a corpus and algorithm for conversation disentanglement. In *Proceedings of ACL-08: HLT*, pages 834–842, Columbus, Ohio. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2010. Disentangling chat. *Computational Linguistics*, 36(3):389–409.

Allyson Ettinger. 2019. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *arXiv preprint arXiv:1907.13528*.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2019. Evaluating discourse in structured text representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 646–653, Florence, Italy. Association for Computational Linguistics.

Zhe Gan, Yunchen Pu, Ricardo Henao, Chunyuan Li, Xiaodong He, and Lawrence Carin. 2017. Learning generic sentence representations using convolutional neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2390–2400, Copenhagen, Denmark. Association for Computational Linguistics.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613, Doha, Qatar. Association for Computational Linguistics.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2).

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–698, Baltimore, Maryland. Association for Computational Linguistics.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Yacine Jernite, Samuel R Bowman, and David Sontag. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv preprint arXiv:1705.00557*.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland. Association for Computational Linguistics.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

Yangfeng Ji and Noah A. Smith. 2017. Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1005, Vancouver, Canada. Association for Computational Linguistics.

Jyun-Yu Jiang, Francine Chen, Yan-Ying Chen, and Wei Wang. 2018. Learning to disentangle interleaved conversational threads with a siamese hierarchical network and similarity ranking. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1812–1822. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 119–126. Association for Computational Linguistics.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 3294–3302, Cambridge, MA, USA. MIT Press.

Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph J. Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros C Polymenakos, and Walter Lasecki. 2019. A large-scale corpus for conversation disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3846–3856, Florence, Italy. Association for Computational Linguistics.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Jiwei Li and Dan Jurafsky. 2017. Neural net models of open-domain discourse coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198–209. Association for Computational Linguistics.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Xiang Lin, Shafiq Joty, Prathyusha Jwalapuram, and M Saiful Bari. 2019. A unified linear-time framework for sentence-level discourse parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4200, Florence, Italy. Association for Computational Linguistics.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351. Association for Computational Linguistics.

Jiangming Liu, Shay B. Cohen, and Mirella Lapata. 2018. Discourse representation structure parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 429–439, Melbourne, Australia. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2018. Learning structured text representations. *Transactions of the Association for Computational Linguistics*, 6:63–75.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. 2016. Sentence ordering and coherence modeling using recurrent neural networks.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA, USA.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6294–6305. Curran Associates, Inc.

Shikib Mehri and Giuseppe Carenini. 2017. Chat disentanglement: Identifying semantic reply relationships with random forests and recurrent neural networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 615–623, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Karthik Narasimhan and Regina Barzilay. 2015. Machine comprehension with discourse relations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1253–1262, Beijing, China. Association for Computational Linguistics.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics.

Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 24(4):694–707.

Boyuan Pan, Yazheng Yang, Zhou Zhao, Yueting Zhuang, Deng Cai, and Xiaofei He. 2018. Discourse marker augmented network with reinforcement learning for natural language inference. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 989–999, Melbourne, Australia. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Karl Pichotta and Raymond J. Mooney. 2016. Using sentence-level LSTM language models for script inference. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–289, Berlin, Germany. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *In Proceedings of LREC*.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3477–3486, Minneapolis, Minnesota. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Shuai Tang and Virginia R. de Sa. 2019. Exploiting invertible decoders for unsupervised sentence representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4050–4060, Florence, Italy. Association for Computational Linguistics.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018a. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355. Association for Computational Linguistics.

Su Wang, Eric Holgate, Greg Durrett, and Katrin Erk. 2018b. Picking apart story salads. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1455–1465. Association for Computational Linguistics.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018c. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967, Brussels, Belgium. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In *Proceedings of International Conference on Learning Representations*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *Coling 2010: Posters*, pages 1507–1514. Coling 2010 Organizing Committee.

Evaluation Benchmarks and Learning Criteria for Discourse-Aware Sentence Representations

Mingda Chen[2*] **Zewei Chu**[1*] **Kevin Gimpel**[2]
[1]University of Chicago, IL, USA
[2]Toyota Technological Institute at Chicago, IL, USA
{mchen,kgimpel}@ttic.edu,zeweichu@uchicago.edu

RST-DT
Attribution
Background
Cause
Comparison
Condition
Contrast
Elaboration
Enablement
Evaluation
Explanation
Joint
Manner-Means
Same-unit
Summary
Temporal
Textual-organization
Topic-Change
Topic-Comment

Table 1: 18 coarse-grained relations in RST-DT

PDTB-E	PDTB-I
Comparison.Concession	Comparison.Concession
Comparison.Contrast	Comparison.Contrast
Contingency.Cause	Contingency.Cause
Contingency.Condition	Contingency.Prag cause
Contingency.Prag condition	Expansion.Alternative
Expansion.Alternative	Expansion.Conjunction
Expansion.Conjunction	Expansion.Instantiation
Expansion.Instantiation	Expansion.List
Expansion.List	Expansion.Restatement
Expansion.Restatement	Temporal.Asynchronous
Temporal.Asynchronous	Temporal.Synchrony
Temporal.Synchrony	

Table 2: The PDTB relation categories

A Hyperparameters

Our models use 1200 dimensional BiGRUs, resulting in 2400 dimensional sentence representations. The feedforward neural networks used in the decoders are parameterized using two hidden layers and use ReLU activation functions. We intialize our models with 300 dimensional GloVe embeddings (**?**). We use Adam (**?**) as optimizer and train our models for one epoch on Wikipedia without employing early stopping.

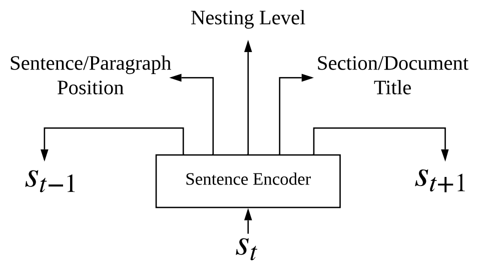


Figure 1: Schematic showing multitask training for our sentence embedding model.

[*] Equal contribution. Listed in alphabetical order.