

# Mining Discourse Markers for Unsupervised Sentence Representation Learning

Damien Sileo<sup>3,1</sup>, Tim Van De Cruys<sup>2</sup>, Camille Pradel<sup>1</sup>, and Philippe Muller<sup>3</sup>

<sup>1</sup>Synapse Développement, Toulouse, France

<sup>2</sup>IRIT, CNRS, France

<sup>3</sup>IRIT, University of Toulouse, France

[damien.sileo@synapse-fr.com](mailto:damien.sileo@synapse-fr.com)

## Abstract

Current state of the art systems in NLP heavily rely on manually annotated datasets, which are expensive to construct. Very little work adequately exploits unannotated data – such as discourse markers between sentences – mainly because of data sparseness and ineffective extraction methods. In the present work, we propose a method to automatically discover sentence pairs with relevant discourse markers, and apply it to massive amounts of data. Our resulting dataset contains 174 discourse markers with at least 10K examples each, even for rare markers such as *coincidentally* or *amazingly*. We use the resulting data as supervision for learning transferable sentence embeddings. In addition, we show that even though sentence representation learning through prediction of discourse markers yields state of the art results across different transfer tasks, it is not clear that our models made use of the semantic relation between sentences, thus leaving room for further improvements. Our datasets are publicly available <sup>1</sup>

## 1 Introduction

An important challenge within the domain of natural language processing is the construction of adequate semantic representations for textual units – from words over sentences to whole documents. Recently, numerous approaches have been proposed for the construction of vector-based representations for larger textual units, especially sentences. One of the most popular frameworks aims to induce sentence embeddings as an intermediate representation for predicting relations between sentence pairs. For instance, similarity judgments (paraphrases) or inference relations have been used as prediction tasks, and the resulting embeddings perform well in practice, even when

the representations are transferred to other semantic tasks (Conneau et al., 2017). However, the kind of annotated data that is needed for such supervised approaches is costly to obtain, prone to bias, and arguably fairly limited with regard to the kind of semantic information captured, as they single out a narrow aspect of the entire semantic content.

Unsupervised approaches have also been proposed, based on sentence distributions in large corpora in relation to their discourse context. For instance, Kiros et al. (2015) construct sentence representations by trying to reconstruct neighbouring sentences, which allows them to take into account different contextual aspects of sentence meaning. In the same vein, Logeswaran et al. (2016) propose to predict if two sentences are consecutive, even though such local coherence can be straightforwardly predicted with relatively shallow features (Barzilay and Lapata, 2008). A more elaborate setting is the prediction of the semantic or rhetorical relation between two sentences, as is the goal of discourse parsing. A number of annotated corpora exist, such as RST-DT (Carlson et al., 2001) and PDTB (Prasad et al., 2008), but in general the available data is fairly limited, and the task of discourse relation prediction is rather difficult. The problem, however, is much easier when there is a marker that makes the semantic link explicit (Pitler et al., 2008), and this observation has often been used in a semi-supervised setting to predict discourse relations in general (Rutherford and Xue, 2015). Building on this observation, one approach to learn sentence representations is to predict such markers or clusters of markers explicitly (Jernite et al., 2017; Malmi et al., 2018; Nie et al., 2017). Consider the following sentence pair:

*I live in Paris. But I'm often abroad.*

The discourse marker *but* highlights an opposition between the first sentence (the speaker

<sup>1</sup><https://github.com/synapse-developpement/Discovery>

s1	Paul Prudhomme’s Louisiana Kitchen created a sensation when it was published in 1984.
c	happily,
s2’	This family collective cookbook is just as good

Table 1: Sample from our *Discovery* dataset

lives in Paris) and the second sentence (the speaker is often abroad). The marker can thus be straightforwardly used as a label between sentence pairs. In this case, the task is to predict  $c = \textit{but}$  (among other markers) for the pair (*I live in Paris, I’m often abroad*). Note that discourse markers can be considered as noisy labels for various semantic tasks, such as entailment ( $c = \textit{therefore}$ ), subjectivity analysis ( $c = \textit{personally}$ ) or sentiment analysis ( $c = \textit{sadly}$ ). More generally, discourse markers indicate how a sentence contributes to the meaning of a text, and they provide an appealing supervision signal for sentence representation learning based on language use.

A wide variety of discourse usages would be desirable in order to learn general sentence representations. Extensive research in linguistics has resulted in elaborate discourse marker inventories for many languages.<sup>2</sup> These inventories were created by manual corpus exploration or annotation of small-scale corpora: the largest annotated corpus, the English PDTB consists of a few tens of thousand examples, and provides a list of about 100 discourse markers, organized in a number of categories.

Previous work on sentence representation learning with discourse markers makes use of even more restricted sets of discourse markers, as shown in table 2. Jernite et al. (2017) use 9 categories as labels, accounting for 40 discourse markers in total. It should be noted that the aggregate labels do not allow for any fine-grained distinctions; for instance, the *TIME* label includes both *now* and *next*, which is likely to impair the supervision. Moreover, discourse markers may be ambiguous; for example *now* can be used to express contrast. On the other hand, Nie et al. (2017) make use of 15 discourse markers, 5 of which are accounting for more than 80% of their training data. In order to ensure the quality of their examples, they only select pairs matching a dependency pattern manually specified for each marker. As such,

<sup>2</sup>See for instance a sample of language on the Textlink project website: <http://www.textlink.ii.metu.edu.tr/dsd-view>

both of these studies use a restricted or impoverished set of discourse markers; they also both use the BookCorpus dataset, whose size (4.7M sentences that contain a discourse marker, according to Nie et al., 2017) is prohibitively small for the prediction of rare discourse markers.

In this work we use web-scale data in order to explore the prediction of a wide range of discourse markers, with more balanced frequency distributions, along with application to sentence representation learning. We use English data for the experiments, but the same method could be applied to any language that bears a typological resemblance with regard to discourse usage, and has sufficient amounts of textual data available (e.g. German or French). Inspired by recent work (Dasgupta et al., 2018; Poliak et al., 2018; Levy et al., 2018; Glockner et al., 2018) on the unexpected properties of recent manually labelled datasets (e.g. SNLI), we will also analyze our dataset to check whether labels are easy to guess, and whether the proposed model architectures make use of high-level reasoning for their predictions. Our contributions are as follows:

- we propose a simple and efficient method to discover new discourse markers, and present a curated list of 174 markers for English;
- we provide evidence that many connectives can be predicted with only simple lexical features;
- we investigate whether relation prediction actually makes use of the relation between sentences;
- we carry out extensive experiments based on the Inferred/SentEval framework.

## 2 Discovering discourse markers

### 2.1 Rationale

Our goal is thus to capture semantic aspects of sentences by means of distributional observations. For our training signal, we aim at something more evolved than just plain contextual co-occurrence,

author	discourse markers / classes	classes	markers
Jernite et al. (2017)	ADDITION, CONTRAST, TIME, RESULT, SPECIFIC, COMPARE, STRENGTH, RETURN, RECOGNIZE	9	40
Nie et al. (2017)	<i>and, but, because, if, when, before, though, so, as, while, after, still, also, then, although</i>	15	15
current work	<i>later, often, understandably, gradually, or, ironically, namely, . . .</i>	174	174

Table 2: Discourse markers or classes used by previous work on unsupervised representation learning

but simpler than a full-fledged encoder-decoder *à la* Skip-Thought. In that respect, discourse relations are an interesting compromise, if we can reliably extract them in large quantities. This objective is shared with semi-supervised approaches to discourse relation prediction, where automatically extracted explicit instances feed a model targeting implicit instances (Marcu and Echihiabi, 2002; Sporleder and Lascarides, 2008; Pitler and Nenkova, 2009; Rutherford and Xue, 2015). In this perspective, it is important to collect unambiguous instances of potential discourse markers. To do so, previous work used heuristics based on specific constructs, especially syntactic patterns for intra-sentential relations, based on a fixed list of manually collected discourse markers. Since we focus on sentence representations, we limit ourselves to discourse arguments that are well-formed sentences, thus also avoiding clause segmentation issues.

Following a heuristic from Rutherford and Xue (2015), also considered by Malmi et al. (2018) and Jernite et al. (2017), we collect pairs of sentences  $(s_1, s_2)$  where  $s_2$  starts with marker  $c$ . We only consider the case where  $c$  is a single word, as detecting longer adverbial constructions is more difficult. We remove  $c$  from the beginning of  $s_2$  and call the resulting sentence  $s'_2$ . Malmi et al. (2018) make use of a list of the 80 most frequent discourse markers in the PDTB in order to extract suitable sentence pairs. We stay faithful to Rutherford and Xue (2015)’s heuristic, as opposed to Malmi et al. (2018); Jernite et al. (2017): if  $s_2$  starts with  $c$  followed by a comma, and  $c$  is an adverbial or a conjunction, then it is a suitable candidate. By limiting ourselves to sentences that contain a comma, we are likely to ensure that  $s'_2$  is meaningful and grammatical. As opposed to all the cited work mentioned above, we do not restrict the pattern to a known list of markers, but try to collect new re-

liable cues.

This pattern is decisively restrictive, since discourse markers often appear at the clausal level (e.g. *I did it but now I regret it*). But clauses are not meant to be self contained, and it is not obvious that they should be included in a dataset for *sentence* representation learning. At the same time, one could easily think of cases where  $c$  is not a discourse marker, e.g.  $(s_1, s_2) = (\text{“It’s cold.”}, \text{“Very, very cold.”})$ . However, these uses might be easily predicted with shallow language models. In the next section, we use the proposed method for the discovery of discourse markers, and we investigate whether the resulting dataset leads to improved model performance.

## 2.2 Methodology

We use sentences from the Depcc corpus (Panchenko et al., 2017), which consists of English texts harvested from commoncrawl web data. We sample 8.5 billion consecutive sentence pairs from the corpus. We keep 53% of sentence pairs that contain between 3 and 32 words, have a high probability of being English ( $> 75\%$ ) using Fast-Text langid from Grave et al. (2018), have balanced parentheses and quotes, and are mostly lowercase. We use NLTK (Bird et al., 2009) as sentence tokenizer and NLTK PerceptronTagger as part of speech tagger for adverb recognition. In addition to our automatically discovered candidate set, we also include all (not necessarily adverbial) PDTB discourse markers that are not induced by our method. Taking this into account, 3.77% of sentence pairs contained a discourse marker candidate, which is about 170M sentence pairs. An example from the dataset is shown in table 1. We only keep pairs in which the discourse marker occurs at least 10K times. We also subsample pairs so that the maximum occurrence count of a discourse marker is 200K. The resulting dataset con-

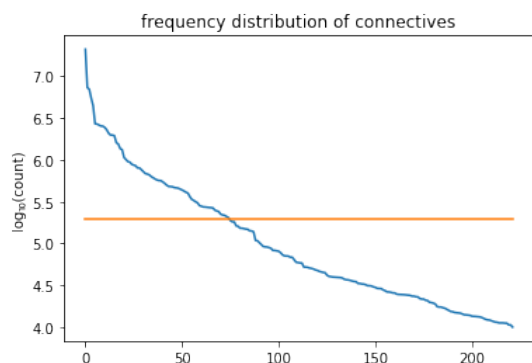


Figure 1: Frequency distribution of candidate discourse markers; the horizontal line indicates the subsampling threshold.

tains 19M pairs.

We discovered 243 discourse marker candidates. Figure 1 shows their frequency distributions. As expected, the most frequent markers dominate the training data, but when a wide range of markers is included, the rare ones still contribute up to millions of training instances. Out of the 42 single word PDTB markers that precede a comma, 31 were found by our rule. Some markers are missing because of NLTK errors, which mainly result from morphological issues.<sup>3</sup>

### 2.3 Controlling for shallow features

As previously noted, some candidates discovered by our rule may not be actual discourse markers. In order to discard them, we put forward the hypothesis that actual discourse markers cannot be predicted with shallow lexical features. Inspired by Gururangan et al. (2018), we use a Fasttext classifier (Joulin et al., 2016) in order to predict  $c$  from  $s'_2$ . The Fasttext classifier predicts labels from an average of word embeddings fed to a linear classifier. We split the dataset in 5 folds, and we predict markers for each fold, while training on the remaining folds. We use a single epoch, randomly initialized vectors of size 100 (that can be unigrams, bigrams or trigrams) and a learning rate of 0.5.

In addition, we predict  $c$  from the concatenation of  $s_1$  and  $s'_2$  (using separate word representations for each case). One might assume that the prediction of  $c$  in this case relies on the interaction between  $s_1$  and  $s_2$ ; however, the features of  $s_1$  and  $s_2$  within Fasttext’s setup only interact additively,

<sup>3</sup>For instance, *lovely* is tagged as an adverb because of its suffix, while *besides* was never tagged as an adverb

which means that the classification most likely relies on individual cues in the separate sentences, rather than on their combination. In order to test this hypothesis, we introduce a *random shuffle* operation: for each example  $(s_1, s'_2, c)$ ,  $s'_2$  is replaced by a random sentence from a pair that is equally linked by  $c$  (we perform this operation separately in train and test sets).

Table 3 indicates that shallow lexical features indeed yield relatively high prediction rates. Moreover, the shuffle operation indeed increases accuracy, which corroborates the hypothesis that classification with shallow features relies on individual cues from separate sentences, rather than their combination.

features	accuracy (%)
majority rule	1.2
s2	18.6
s1-s2'	21.9
s1-s2' (shuffled)	24.8

Table 3: Accuracy when predicting candidate discourse markers using shallow lexical features

Tables 4 and 5 show the least and most predictable discourse markers, and the corresponding recognition rate with lexical features.

candidate marker	accuracy (%)
evidently,	0.0
frequently,	0.0
meantime,	0.0
truthfully,	0.0
supposedly,	0.1

Table 4: Candidate discourse markers that are the most difficult to predict from shallow features

candidate marker	accuracy (%)
defensively,	65.5
afterward	71.1
preferably,	71.9
this,	72.7
very,	90.7

Table 5: Candidate discourse markers that are the easiest to predict from shallow features. This shows candidates that are unlikely to be interesting discourse cues.

Interestingly, the two most predictable candidates are not discourse markers. Upon inspection of harvested pairs, we noticed that even legitimate discourse markers can be guessed with relatively simple heuristics in numerous examples. For example,  $c = \textit{thirdly}$  is very likely to occur if  $s_1$  contains  $\textit{secondly}$ . We use this information to optionally filter out such simple instances, as described in the next section.

## 2.4 Dataset variations

In the following, we call our method *Discovery*. We create several variations of the sentence pairs dataset. In *DiscoveryHard*, we remove examples where the candidate marker was among the top 5 predictions in our Fasttext shallow model and keep only the 174 candidate markers with a frequency of at least  $10k$ . Instances are then sampled randomly so that each marker appears exactly  $10k$  times in the dataset.

Subsequently, the resulting set of discourse markers is also used in the other variations of our dataset. *DiscoveryBase* designates the dataset for which examples predicted with the Fasttext model were not removed. In order to measure the extent to which the model makes use of the relation between  $s_1$  and  $s_2'$ , we also create a *DiscoveryShuffled* dataset, which is the *DiscoveryBase* dataset subjected to the *random shuffle* operation described previously. To isolate the contribution of our discovery method, the dataset *DiscoveryAdv* discards all discourse markers from PDTB that were not found by our method. Also, in order to measure the impact of label diversity, *Discovery10* uses  $174k$  examples for each of the 10 most frequent markers,<sup>4</sup> thus totalling as many instances as *DiscoveryBase*. Finally, *DiscoveryBig* contains almost twice as many instances as *DiscoveryBase*, i.e.  $20k$  instances for each discourse marker (although, for a limited number of markers, the number of instances is slightly lower due to data sparseness).

## 3 Evaluation of sentence representation learning

### 3.1 Setup

Our goal is to evaluate the effect of using our various training datasets on sentence encoding, given encoders of equivalent capacity and similar

<sup>4</sup>They are: *however, hence, moreover, additionally, nevertheless, furthermore, alternatively, again, next, therefore*

setups. Thus, we follow the exact setup of In-fersent (Conneau et al., 2017), also used in the Dissent (Malmi et al., 2018) model: we learn to encode sentences into  $h$  with a bi-directional LSTM sentence encoder using element-wise max pooling over time. The dimension size of  $h$  is 4096. Word embeddings are fixed GloVe embeddings with 300 dimensions, trained on Common Crawl 840B.<sup>5</sup> A sentence pair  $(s_1, s_2)$  is represented with  $[h_1, h_2, h_1 \odot h_2, |h_2 - h_1|]$ ,<sup>6</sup> which is fed to a softmax in order to predict a marker  $c$ . Our datasets are split in 90% train, 5% validation, and 5% test. Optimization is done with SGD (learning rate is initialized at 0.1, decayed by 1% at each epoch and by 80% if validation accuracy decreases; learning stops when learning rate is below  $10^{-5}$  and the best model on training task validation loss is used for evaluation; gradient is clipped when its norm exceeds 5). Once the sentence encoder has been trained on a base task, the resulting sentence embeddings are tested with the SentEval library (Conneau et al., 2017).

We evaluate the different variations of our dataset we described above in order to analyze their effect, and compare them to a number of existing models. Table 7 displays the tasks used for evaluation. For further analysis, table 9 displays the result of *Linguistic Probing* using the method by Conneau et al. (2018). Although these tasks are primarily designed for understanding the content of embeddings, they also focus on aspects that are desirable to perform well in general semantic tasks (e.g. prediction of tense, or number of object).

### 3.2 Results

Table 6 gives an overview of transfer learning evaluation, also comparing to other supervised and unsupervised approaches. Note that we outperform DisSent on all tasks except TREC<sup>7</sup> with less than half the amount of training examples. In addition, our approach is arguably simpler and faster.

MTL (Subramanian et al., 2018) only achieves stronger results than our method on the MRPC and SICK tasks. The MTL model uses  $124M$  training examples with an elaborate multi-task setup, training on  $45M$  sentences with manual translation,  $1M$  pairs from SNLI/MNLI,  $4M$  parse trees of sentences, and  $74M$  consecutive sentence pairs.

<sup>5</sup><https://nlp.stanford.edu/projects/glove/>

<sup>6</sup> $h_1 \odot h_2 = (h_{11}.h_{21}, \dots, h_{1i}.h_{2i}, \dots)$

<sup>7</sup>This dataset is composed of questions only, which are underrepresented in our training data.

	<i>N</i>	MR	CR	SUBJ	MPQA	SST2	TREC	SICK-R	SICK-E	MRPC	AVG
InferSent	1.0	81.1	86.3	92.4	90.2	84.6	88.2	88.4	86.1	76.2	85.9
MTL	124	82.5	<b>87.7</b>	94	90.9	83.2	93	<b>88.8</b>	<b>87.8</b>	<b>78.6</b>	<b>87.4</b>
SkipThought	74	76.5	80.1	93.6	87.1	82	92.2	85.8	82.3	73	83.6
QuickThought	174	81.3	84.5	<b>94.6</b>	89.5	-	92.4	<u>87.1</u>	-	75.9	-
DisSent	4.7	80.1	84.9	93.6	90.1	84.1	<b>93.6</b>	84.9	83.7	75	85.6
DiscoveryBase	1.7	82.5	86.3	94.2	90.5	85.2	91.8	85.7	84	75.8	86.2
DiscoveryHard	1.7	81.6	86.5	93.9	90.5	84.8	90	85.4	83.2	76.5	85.8
Discovery10	1.7	81.2	85.1	93.7	90.2	83	90	85.9	83.8	75.8	85.4
DiscoveryAdv	1.4	81.4	85.8	93.8	90.5	83.4	92	86	84.3	75.7	85.9
DiscoveryShuffled	1.7	81.4	86.1	94.1	90.9	<b>85.3</b>	90.4	85.6	83.6	75.4	85.9
DiscoveryBig	3.4	<b>82.6</b>	<u>87.4</u>	94.5	<b>91.0</b>	85.2	93.4	86.4	<u>84.8</u>	<u>76.6</u>	<u>86.9</u>

Table 6: SentEval evaluation results with our models trained on various datasets. The first two models are supervised, the other ones unsupervised. All scores are accuracy percentages, except SICK-R, which is Pearson correlation percentage. InferSent is from [Conneau et al. \(2017\)](#), MTL is the multi-task learning based model from [Subramanian et al. \(2018\)](#). Evaluation tasks are described in table 7, and *N* denotes the number of examples for each dataset (in millions). Dissent is from [Nie et al. \(2017\)](#), QuickThought is from [Logeswaran and Lee \(2018\)](#) with fixed embeddings configuration. The best result per task appears in bold, the best result for unsupervised setups is underlined.

The model also fine-tunes word embeddings in order to achieve a higher capacity. It is therefore remarkable that our model outperforms it on many tasks. Besides, MTL is not a direct competitor to our approach since its main contribution is its multi-task setup, and it could benefit from using our training examples.

Our best model rivals (and indeed often outperforms) QuickThought on all tasks, except relatedness (SICK-R). QuickThought’s training task is to predict whether two sentences are contiguous, which might incentivize the model to perform well on a relatedness task. We also outperform InferSent on many tasks except entailment and relatedness. Entailment prediction is the explicit training signal for InferSent.

To help the analysis of our different model variations, table 8 displays the test scores on each dataset for the original training task. It also shows the related PDTB implicit relation prediction scores. The PDTB is annotated with a hierarchy of relations, with 5 classes at level 1 (including the EntRel relation), and 16 at level 2 (with one relation absent from the test). It is interesting to see that this form of simple semi-supervised learning for implicit relation prediction performs quite well, especially for fine-grained relations, as the best model slightly beats the best current dedicated model, listed at 40.9% in [Xue et al. \(2017\)](#).

*DiscoveryHard* scores lower on its training task than *DiscoveryBase*, and it also performs worse on transfer learning tasks. This makes sense, since lexical features are important to solve the

evaluation tasks. Our initial hypothesis was that more difficult instances might force the model to use higher-level reasoning, but this does not seem to be the case. More surprisingly, preventing the encoders to use the relationship between sentences, as in *DiscoveryShuffled*, does not substantially hurt the transfer performance, which remains on average higher than [Nie et al. \(2017\)](#). Additionally, our models score well on linguistic probing tasks. They outperform InferSent on all tasks, which seems to contradict the claim that SNLI data allows for learning of *universal* sentence representations ([Conneau et al., 2017](#)). And a final interesting outcome is that the diversity of markers (e.g. using *DiscoveryBase* instead of *Discovery10*) seems to be important for good performance on those tasks, since *Discovery10* has the worst overall performance on average.

name	N	task	C
MR	11k	sentiment (movie reviews)	2
CR	4k	sentiment (product reviews)	2
SUBJ	10k	subjectivity/objectivity	2
MPQA	11k	opinion polarity	2
TREC	6k	question-type	6
SST	70k	sentiment (movie reviews)	2
SICK-E	10k	entailment	3
SICK-R	10k	relatedness	3
MRPC	4k	paraphrase detection	2
PDTB <sub>5</sub>	17k	implicit discourse relation (coarse)	5
PDTB <sub>16</sub>	17k	implicit discourse relation (fine)	15

Table 7: Transfer evaluation tasks. N is the number of training examples and C is number of classes for each task.

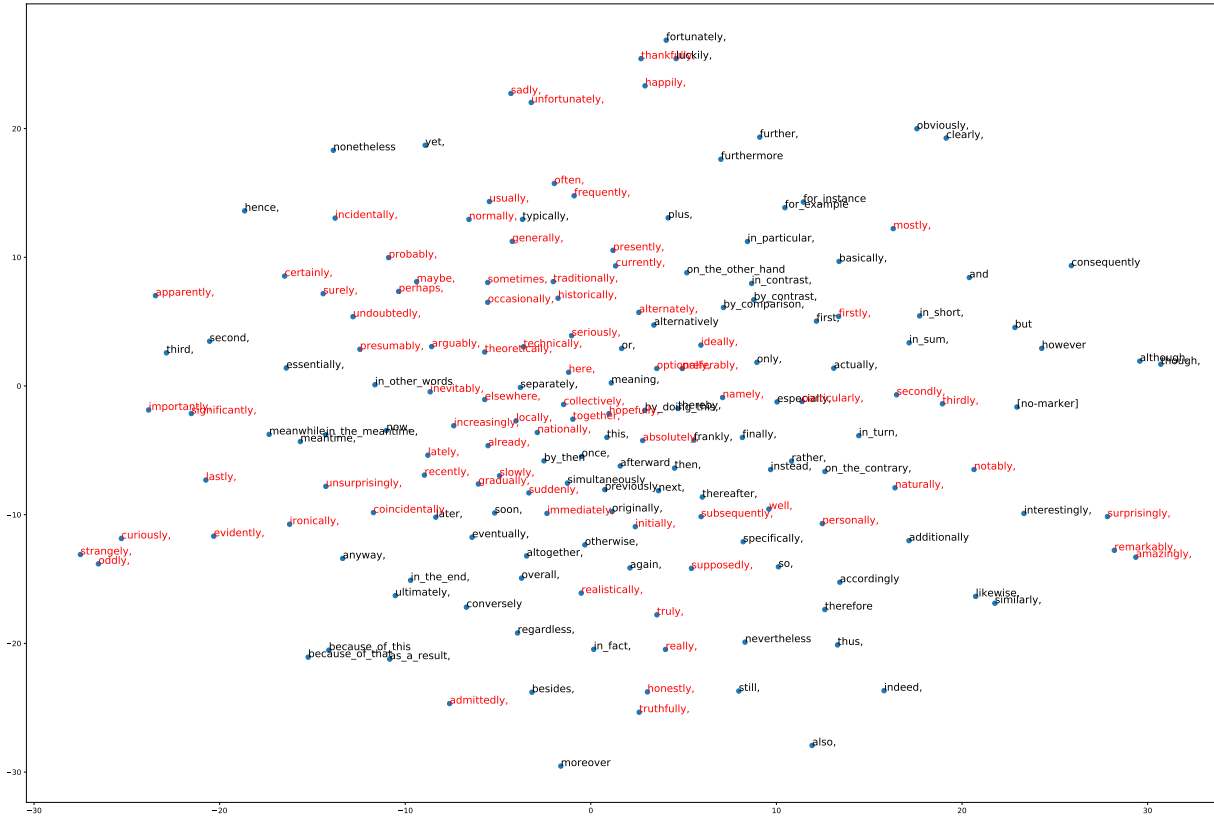


Figure 2: TSNE visualization of the softmax weights from our *DiscoveryBig* model for each discourse marker, after unit norm normalization. Markers discovered by our method (e.g. absent from PDTB annotations) are colored in red.

	PDTB <sub>5</sub> coarse	PDTB <sub>16</sub> fine	$\mathcal{T}$
InferSent	46.7	34.2	-
DisSent	48.9	36.9	-
DiscoveryBase	<b>52.5</b>	40.0	20.6
DiscoveryHard	50.7	39.8	9.3
Discovery10	48.3	37.7	51.9
DiscoveryAdv	49.7	37.6	26.1
DiscoveryShuffled	51.0	39.5	11.5
DiscoveryBig	51.3	<b>41.3</b>	22.2

Table 8: Test results (accuracy) on implicit discursive relation prediction task (PDTB relations level 1 and 2, i.e coarse-grained and fine-grained) and training tasks  $\mathcal{T}$ . Note that scores for  $\mathcal{T}$  are not comparable since the test set changes for each version of the dataset.

### 3.3 Visualisation

The softmax weights learned during the training phase can be interpreted as embeddings for the markers themselves, and used to visualize their relationships. Figure 2 shows a TSNE (van der Maaten and Hinton, 2008) plot of the markers’ representations. Proximity in the feature space seems to reflect semantic similarity (e.g. *usu-*

*ally/normally*). In addition, the markers we discovered, colored in red, blend with the PDTB markers (depicted in black). It would be interesting to cluster markers in order to empirically define discourse relations, but we leave this for future work.

## 4 Related work

Though discourse marker prediction in itself is an interesting and useful task (Malmi et al., 2017), discourse markers have often been used as a training cue in order to improve implicit relation prediction (Marcu and Echiabi, 2001; Sporleder and Lascarides, 2005; Zhou et al., 2010; Braud and Denis, 2016). This approach has been extended to general representation learning by Jernite et al. (2017)—although with empirically unconvincing results, which might be attributed to an inappropriate training/evaluation set-up, or the use of a limited number of broad categories instead of actual discourse markers. Nie et al. (2017) used the more standard InferSent framework and obtained better results, although they were still outperformed by QuickThought (Logeswaran and Lee, 2018),

	BShift	CoordInv	Depth	ObjNum	SubjNum	OddM	Tense	TC	WC	AVG
InferSent	56.5	65.9	37.5	79.9	84.3	53.2	87	78.1	95.2	70.8
SkipThought	<b>69.5</b>	69	39.6	83.2	86.2	54.5	<b>90.3</b>	82.1	79.6	72.7
QuickThought	56.8	70	40.2	79.7	83	55.3	86.2	80.7	90.3	71.4
DiscoveryBase	63.1	70.6	45.2	83.8	87.2	57.3	89.1	83.2	94.7	74.9
DiscoveryHard	62.7	70.4	44.5	83.4	<b>88.1</b>	57.3	89.5	82.8	94.1	74.8
Discovery10	61.3	69.7	42.9	81.8	86.7	55.8	87.8	81.4	<b>96.1</b>	73.7
DiscoveryAdv	61.5	70	43.9	82.6	86.2	56.2	89.1	82.8	<b>96.1</b>	74.3
DiscoveryShuffled	62.6	<b>71.4</b>	45.3	<b>84.3</b>	88	<b>58.3</b>	89.3	82.8	93.4	75
DiscoveryBig	63.3	<b>71.4</b>	<b>46.0</b>	84.1	87.8	57.1	89.4	<b>84.2</b>	96	<b>75.5</b>

Table 9: Accuracy of various models on linguistic probing tasks using logistic regression on SentEval. BShift is detection of token inversion. CoordInv is detection of clause inversion. ObjNum/SubjNum is prediction of the number of object resp. subject. Tense is prediction of the main verb tense. Depth is prediction of parse tree depth. TC is detection of common sequences of constituents. WC is prediction of words contained in the sentence. OddM is detection of random replacement of verbs/nouns by other verbs/nouns. AVG is the average score of those tasks for each model. For more details see [Conneau et al. \(2018\)](#). SkipThought and Infersent results come from [Perone et al. \(2018\)](#), QuickThought results come from [Brahma \(2018\)](#).

which uses a much simpler training task. Both of these rely on pre-established lists of discourse markers provided by the PDTB, and both perform a manual annotation for each marker—[Nie et al. \(2017\)](#) uses dependency patterns, while [Jernite et al. \(2017\)](#) uses broad discourse categories. Our work is the first to automatically discover discourse markers from text.

More generally, various automatically extracted training signals have been used for unsupervised learning tasks. Hashtags ([Felbo et al., 2017](#)) have been successfully exploited in order to learn sentiment analysis from unlabelled tweets, but their availability is mainly limited to the microblogging domain. Language modeling provides a general training signal for representation learning, even though there is no obvious way to derive sentence representations from language models. BERT ([Devlin et al., 2018](#)) currently holds the best results in transfer learning based on language modeling, but it relies on sentence pair classification in order to compute sentence embeddings, and it makes use of a simple sentence contiguity detection task (like QuickThought); this task does not seem challenging enough since BERT reportedly achieves 98% detection accuracy. [Phang et al. \(2018\)](#) showed that the use of SNLI datasets yields significant gains for the sentence embeddings from [Radford \(2018\)](#), which are based on language modeling.

For the analysis of our models, we draw inspiration from critical work on Natural Language Inference datasets ([Dasgupta et al., 2018](#); [Levy et al., 2018](#)). [Gururangan et al. \(2018\)](#); [Poliak et al. \(2018\)](#) show that baseline models that disregard

the hypothesis yield good results on SNLI, which suggests that the model does not perform the high level reasoning we would expect in order to predict the correct label. They attribute this effect to bias in human annotations. In this work, we show that this issue is not inherent to human labeled data, and propose the *shuffle* perturbation in order to measure to what extent the relationship between sentences is used.

## 5 Conclusion

In this paper, we introduce a novel and efficient method to automatically discover discourse markers from text, and we use the resulting set of candidate markers for the construction of an extensive dataset for semi-supervised sentence representation learning. A number of dataset variations are evaluated on a wide range of transfer learning tasks (as well as implicit discourse recognition) and a comparison with existing models indicates that our approach yields state of the art results on the bulk of these tasks. Additionally, our analysis shows that removing ‘simple’ examples is detrimental to transfer results, while preventing the model to exploit the relationship between sentences has a negligible effect. This leads us to believe that, even though our approach reaches state of the art results, there is still room for improvement: models that adequately exploit the relationship between sentences would be better at leveraging the supervision of our dataset, and could yield even better sentence representations. In future work, we also aim to increase the coverage of our method. For instance, we can make use of more lenient patterns that capture an even wider



range of discourse markers, such as multi-word markerse.

## References

- Regina Barzilay and Mirella Lapata. 2008. [Modeling Local Coherence: An Entity-Based Approach](#). *Computational Linguistics*, 34(1):1–34.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, volume 43.
- Siddhartha Brahma. 2018. Unsupervised learning of sentence representations using sequence consistency. *CoRR*, abs/1808.04217.
- Chloé Braud and Pascal Denis. 2016. [Learning connective-based word representations for implicit discourse relation identification](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 203–213, Austin, Texas. 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 SIGDIAL 2001 Workshop, The 2nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, Saturday, September 1, 2001 to Sunday, September 2, 2001, Aalborg, Denmark*.
- Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, 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. Association for Computational Linguistics.
- Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J. Gershman, and Noah D. Goodman. 2018. [Evaluating Compositionality in Sentence Embeddings](#). (2011).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Bjarke Felbo, Alan Mislove, Anders Sogaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *EMNLP*.
- Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. [Breaking NLI Systems with Sentences that Require Simple Lexical Inferences](#). *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, (3):1–6.
- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. [Annotation Artifacts in Natural Language Inference Data](#).
- Yacine Jernite, Samuel R. Bowman, and David Sontag. 2017. [Discourse-Based Objectives for Fast Unsupervised Sentence Representation Learning](#).
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. [Bag of tricks for efficient text classification](#). Cite arxiv:1607.01759.
- Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.
- Omer Levy, Samuel R Bowman, and Noah A Smith. 2018. Annotation Artifacts in Natural Language Inference Data. *Proceedings of NAACL-HLT 2018*, pages 107–112.
- Lajanugen Logeswaran and Honglak Lee. 2018. [An efficient framework for learning sentence representations](#). pages 1–16.
- Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. 2016. [Sentence Ordering using Recurrent Neural Networks](#). pages 1–15.
- Laurens van der Maaten and Geoffrey Hinton. 2008. [Visualizing data using t-SNE](#). *Journal of Machine Learning Research*, 9:2579–2605.
- Eric Malmi, Daniele Pighin, Sebastian Krause, and Mikhail Kozhevnikov. 2017. [Automatic Prediction of Discourse Connectives](#).
- Eric Malmi, Daniele Pighin, Sebastian Krause, and Mikhail Kozhevnikov. 2018. Automatic Prediction of Discourse Connectives. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Daniel Marcu and Abdessamad Echihabi. 2001. [An unsupervised approach to recognizing discourse relations](#). *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, (July):368.

- Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 368–375.
- Allen Nie, Erin D. Bennett, and Noah D. Goodman. 2017. [DisSent: Sentence Representation Learning from Explicit Discourse Relations](#).
- Alexander Panchenko, Eugen Ruppert, Stefano Faralli, Simone P Ponzetto, and Chris Biemann. 2017. Building a Web-Scale Dependency-Parsed Corpus from Common Crawl. pages 1816–1823.
- Christian S. Perone, Roberto Silveira, and Thomas S. Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *CoRR*, abs/1806.06259.
- Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. [Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks](#). *CoRR*, abs/1811.01088.
- Emily Pitler and Ani Nenkova. 2009. [Using syntax to disambiguate explicit discourse connectives in text](#). In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore, Short Papers*, pages 13–16.
- Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. [Easily identifiable discourse relations](#). In *Coling 2008: Companion volume: Posters*, pages 87–90. Coling 2008 Organizing Committee.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*, (1):180–191.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltasakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). [Http://www.lrec-conf.org/proceedings/lrec2008/](http://www.lrec-conf.org/proceedings/lrec2008/).
- Alec Radford. 2018. Improving language understanding by generative pre-training.
- Attapol Rutherford and Nianwen Xue. 2015. [Improving the inference of implicit discourse relations via classifying explicit discourse connectives](#). In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 799–808.
- Caroline Sporleder and Alex Lascarides. 2005. Exploiting linguistic cues to classify rhetorical relations. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, Bulgaria.
- Caroline Sporleder and Alex Lascarides. 2008. [Using automatically labelled examples to classify rhetorical relations: an assessment](#). *Natural Language Engineering*, 14(3):369–416.
- Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. *International Conference on Learning Representations*.
- Nianwen Xue, Vera Demberg, and Attapol Rutherford. 2017. [A systematic study of neural discourse models for implicit discourse relation](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 281–291.
- 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.