# Do Sentence Embeddings Capture Discourse Properties of Sentences from Scientific Abstracts ?

**Laurine Huber**[1], **Chaker Memmadi**[1], **Mathilde Dargnat**[2] and **Yannick Toussaint**[1],
[1] Université de Lorraine, CNRS, Inria, LORIA (UMR 7503), F-54000 Nancy, France
{laurine.huber, yannick.toussaint}@loria.fr
chaker.memmadi@outlook.fr
[2] ATILF, Université de Lorraine, CNRS (UMR 7118), Nancy, France
et Institut des Sciences Cognitives Marc Jannerod, CNRS (UMR 5304), Bron, France
mathilde.dargnat@univ-lorraine.fr

## Abstract

We introduce four tasks designed to determine which sentence encoders best capture discourse properties of sentences from scientific abstracts, namely coherence between clauses of a sentence, and discourse relations within sentences. We show that even if contextual encoders such as BERT or SciBERT encodes the coherence in discourse units, they do not help to predict three discourse relations commonly used in scientific abstracts. We discuss what these results underline, namely that these discourse relations are based on particular phrasing that allow non-contextual encoders to perform well.

## 1 Introduction

This paper compares the ability of different sentence encoders at representing coherence of sentences from scientific abstracts, and more specifically discourse relations between clauses.

Our first hypothesis is that BERT (Devlin et al., 2019) and SciBERT (Beltagy et al., 2019) models enable to produce sentence representations that take coherence into account thanks to their training done on Next Sentence Prediction. Our second hypothesis is that this training should also enable to capture discourse relations between clauses.

Discourse relations (DRs) represent the semantic and pragmatic links between discourse units (DUs), that are either clauses, sentences or groups thereof, within a hierarchical structure that represents the whole text. In this work, we focus on sentences that are defined as textual sequences separated by a period. Sentences may comprise one or more DUs, and when they have at least two DUs, the coherence links between them is represented through DRs. These DRs are either explicitly signaled, or left implicit. For example, the sentence *"By wearing a mask, we can protect the others."* conveys an `enablement` relation between the action *"wearing a mask"* and the event *"protecting the others"*, which is here lexicalized by the connective "**by**". DRs can be used to extract new knowledge, especially in scientific abstracts which are highly structured (Liddy, 1991).

Sentence embeddings (SE) represent the meaning of a sentence in a fixed-size vector space, and recent contextual approaches such as BERT have shown promising results for downstream tasks such as Semantic Textual Similarity (STS) or Natural Language Inference (NLI) (Reimers and Gurevych, 2019). When further trained on downstream tasks, these models have shown promising results. However, their performance also rely on linguistic knowledge acquired at pre-training. For example, BERT and SciBERT are trained on both Masked Language Model and Next Sentence Prediction tasks, in order to capture general linguistic properties that are then transferred to learn more specific representations.

In this work, we want to understand if discourse properties are embedded in sentence representations that are built before further training on downstream tasks.

We design probing tasks, that are classification tasks whose goals are to predict discourse properties of the sentences from their embedding. Our goal is to highlight if discourse properties of sentences are captured by those vectors without fine-tuning. We rely on the corpus SciDTB (Yang and Li, 2018) to build four datasets used to probe if embeddings capture some discourse properties of the sentences. The two first datasets probe the coherence of sentences, and the two others probe the presence of DRs. We use four different sentence encoders to produce sentence embeddings, that we then use as input vectors for two classification models. If the classifier succeeds, it means that the vectors stores the discourse property that is probed. We evaluate the classifiers, thus high-



*Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 86–95

https://doi.org/10.18653/v1/P17

lighting encoders that best encode the properties we probe.

This paper is organized as follows. We first provide background to our work. Second, we detail the tasks that we design to detect discourse properties. Third, we present our experimental setup, including the different SE that we evaluate and the choices that are specific to the corpus on which we rely. Finally, we present our results and discuss some issues.

## 2 Background

Coherence and cohesion are two key notions in the perception of a text as a unified whole.

Coherence refers to logical and semantic relations between clauses and sentences in a text, while cohesion refers to grammatical or lexical devices such as pronouns, verb tense or connectives, that form external relations of a text (Halliday and Hasan, 1976). While the former may stand without the latter and vice versa, the connectivity model of Renkema (2009) mixes coherence and cohesion cues. In this work we follow this approach, and will refer to as coherence the properties of both coherence and cohesion.

DRs may either explicitly signal coherence relations by discourse connectives, or left them implicit. In this latter case, the reader infers the relations based on coherence links between clauses. For example, the marker "**to**" is frequently used to links an action X and a way Y to realize X, signaling that Y is the `manner-means` to do X. However, even if some relations are explicitly signaled with discourse connectives such as "**to**" or "**by**", others are less salient. For example in *"We propose a novel extension of this work using target context information."*, the ellipsis of "**by**" make it harder to infer the relation. Some of them (e.g `enablement`, `manner-means` or `attribution`) express a logical link between the content of the clauses they relate, while others (e.g `elaboration` or `progression`) only express a continuation or additive relation. In scientific abstracts especially, DRs that convey a logical link are often lexically marked.

Several theories such as Rhetorical Structure Theory (RST) (Taboada and Mann, 2006) or Segmented Discourse Representation Theory (SDRT) (Lascarides and Asher, 2007) help to build text discourse structures, by providing both sets of DRs that are defined based on the content of the DUs and a framework for attaching the DUs by means of DRs. The RST defines a set of relations[1] that are either mononuclear (if for two DUs, one is more salient than the other) or multinuclear (if two or more DUs have the same importance). They serve to build the hierarchical discourse structure of the text, either as a constituency tree, or, more recently, with dependency trees (Morey et al., 2018), which are used to annotate scientific abstracts from SciDTB (Yang and Li, 2018).

SE have shown promising results for a wide variety of NLP tasks (Conneau and Kiela, 2018), but are not interpretable independently of the others, making it unclear what linguistic information they contain and what is the *meaning* that they represent exactly. Previous works (Shi et al., 2016; Adi et al., 2016; Conneau et al., 2018) intended to clarify what linguistic information are contained in SE by designing auxiliary tasks that take a single sentence vector as input and tries to predict a simple linguistic property of the sentence. They showed that some syntactic properties (e.g word order or syntactic tree depth) and some properties that they defined as semantic[2] (e.g the tense of the main verb or the number of the subject) are well captured by SE. We then hypothesize that the latter may also capture discourse coherence and DRs in sentences from scientific abstracts, thus being the first step to extracting information at a semantic-pragmatic level.

## 3 Probing tasks

In this section, for each probing task, we explain how we build a dataset of sentences tagged with discourse properties. They are then encoded by various methods and used to train and evaluate classifiers. We introduce two tasks designed to check if coherence between clauses of a sentence is encoded (3.1, 3.2), and two tasks to check if DRs in sentences are encoded (3.3, 3.4).

### 3.1 Swapped units detection

Ordering of clauses fully participates to sentences coherence. We evaluate if coherence is captured by SE by evaluating the ability of a classifier to distinguish between coherent sentences and incoherent ones. We produce incoherent sentences by swap-

[1] https://www.sfu.ca/rst/01intro/definitions.html

[2] They admit however that the boundary they propose between syntactic and semantic tasks is somewhat arbitrary.

ping two adjacent discourse units from sentences originally correct.

The dataset for this probing task thus consists of original multi-clausal sentences and of their incoherent equivalent obtained by swapping two random adjacent DUs. By doing so, we however do not ensure that new sentences are always incoherent. For example, swapping units 0 and 1 in the sentence *"[Word alignment] [using recency-vector based approach] [has recently become popular.]"* produces the new coherent sentence *"[Using recency-vector based approach] [word alignment] [has recently become popular.]"*.

The resulting task is a binary classification task into $\{yes, no\}$, where $yes$ corresponds to swapped sentences, and $no$ to the original sentences.

### 3.2 Scrambled sentence detection

Topic incoherence may emerge when a topic $T_1$ is involved in the context of another topic $T_2$, leading to incoherence because of the incompatibility between $T_1$ and $T_2$.

The dataset used to probe this property is built as follows. Starting from original multi-clausal sentences splitted in DUs, we replace a randomly chosen DU by another randomly chosen DU from a different document, thus changing the context of a DU. Here again, the new sentence might be still coherent. For example, replacing the second DU in the sentence "*[But these methods cannot be used][to obtain the estimates of causal effects-the quantity of interest for applied researchers.]*" produces a new coherent sentence "*[But these methods cannot be used] [using a two-phase approach.]*".

The resulting task is a binary classification task into $\{yes, no\}$, where $yes$ corresponds to scrambled sentences, and $no$ to the original sentences.

### 3.3 Relation detection

A sentence may either be composed of a single DU, or of multiple DUs. In the last case, the DUs may be links through additive or continuation DRs (in bold in Table 1), or through logical relations (in italics in Table 1). We distinguish them and evaluate if SE can be used to predict whether a sentence contains a logical DR.

We rely on the discourse annotations provided in SciDTB, from which we extract the subtrees corresponding to the discourse structures of the sentences. Sentences that contain only one DU and sentences made of multiple DUs whose subtree contains no logical DRs are classified as $norel$.

| | | |
|---|---|---|
| *Attribution* | *Background* | *Cause-effect* |
| *Comparison* | *Condition* | *Contrast* |
| **Elaboration** | *Enablement* | *Evaluation* |
| *Explain* | **Joint** | *Manner-means* |
| **Progression** | **Same-unit** | *Summary* |
| | *Temporal* | |

Table 1: Discourse relations in SciDTB corpus

Others are classified as $rel$ regardless of the type of the logical DR they contain.

The resulting task is a binary classification task into $\{rel, norel\}$.

### 3.4 Relation semantics detection

Being able to precisely identify which logical discourse relation is involved in a sentence may be useful as a first step toward knowledge extraction from texts. This task thus evaluates if the semantics of the DRs involved in sentences are represented in their embedding.

We rely again on the annotations to classify sentences based on the DR they contain. Sentences that contain only one DU and sentences made of multiple DU whose subtree contains no logical DRs are classified as $norel$. Others are classified based on the relation $r$ they contain, if and only if they contain only one of these relations. The sentences that contain two or more logical relations are not tackled in our approach, and are thus not considered in the dataset.

The resulting task is a K-classes classification task where K is the number of relations that are considered.

## 4 Experiments

### 4.1 Data

SciDTB[3] is a corpus of 798 scientific abstracts from ACL Anthology[4]. Abstracts are segmented into DUs in a semi-automatic way, following the guidelines of (Carlson et al., 2003), and annotated by discourse structure in dependency with DRs from the RST relations set, which was slightly modified and extended to be adapted to scientific abstracts[5].

The majority of documents contain between 5 and 7 sentences (minimum 2, maximum 14, mean

[3] https://github.com/PKU-TANGENT/SciDTB

[4] https://www.aclweb.org/anthology/

[5] We refer the reader to the paper (Yang and Li, 2018) to get explanations on the adaptations and annotation procedure. The relations set is recalled in Table 1.

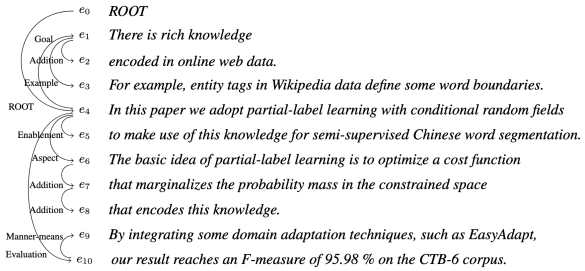


Figure 1: Example of dependency annotation from SciDTB.

6), resulting in a set of 4196 sentences in total. Among them, 787 are made of a single DU, thus containing no relation, and 3409 contain more than one DU, and thus at least one DR. We use the segmentation into DUs to produce the datasets for coherence detection tasks (swapped and scrambled). We use the inter-clausal DRs to produce the datasets for DR detection tasks (binRel and semRel)[6].

For swapped and scrambled, each sentence that has more than one DU is classified as $no$, and is used to produce a new sentence that is classified as $yes$, resulting in a dataset of 6818 sentences. For each sentence (original and modified), we ensure that punctuation and case do not bias the experiment by removing capitalization, periods and commas.

For binRel and semRel, we need to form classes that are broad enough to train the classifiers. It leads us to make choices because of the limited size of the corpus. Fig. 2 shows the number of sentences having one of the given DRs. Most of the DRs are involved in less than 200 sentences, which is not enough for training a classifier. Among the others, three relations are additive relations: `Same-Unit` links two segments of a DU broken into two parts, `Joint`, links two DUs which are in conjunction, and for `Elaboration-addition` one DU gives additional information to another DU[7]. Moreover, the latter are involved in most of sentences made of more than 3 DUs, and are thus often used together with logical DRs. We decide to ignore them, which allows us to build a sufficiently large dataset of sentences, tagged with the three relations: `enablement`, `manner-means` or `attribution`. We assume that the classifier will therefore learn to predict one of these relations.

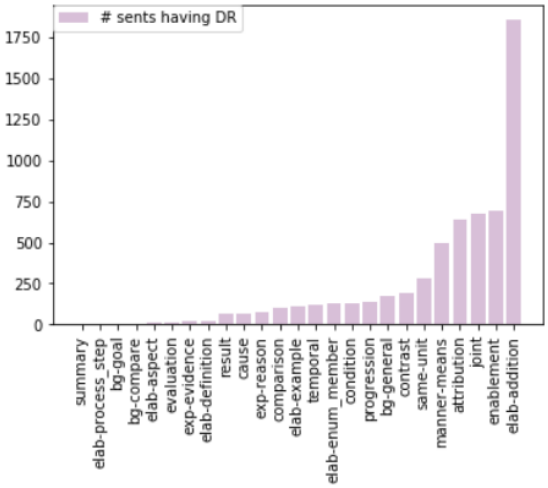


Figure 2: Distribution of relations in sentences.

For binRel, each sentence which has no DR or only DRs from {`joint`, `elaboration-addition`, `same-unit`} is classified as $norel$, and the others as $rel$. For semRel, the $norel$ classification is done in the same way, and sentences that involve one and only one relation from {`attribution`, `enablement`, `manner-means`} are classified according to the relation they contain. Table 2 summarizes the datasets used for each probing task, where the classes for each task are balanced.

| | classes | # sent | # sent / class |
|---|---|---|---|
| binRel | $rel$<br>$norel$ | 2894 | 1447 |
| semRel | $attribution$<br>$enablement$<br>$manner-means$<br>$norel$ | 1432 | 358 |
| swapped | $yes$<br>$no$ | 6818 | 3409 |
| scrambled | $yes$<br>$no$ | 6818 | 3409 |

Table 2: Datasets description.

### 4.2 Sentence embeddings

In this section, we present the different SE that we study, most of them being obtained by using pretrained models available on HuggingFace's (Wolf et al., 2019) website[8]. We compare SE obtained

[6]We make the resulting datasets available at https://gitlab.com/laurinehu/scidtb_expe/-/tree/master/probing_datasets.

[7]We admit however that saying that `Elaboration-addition` does not contain an interesting semantics is somehow arbitrary, but can be understood by comparison with the semantics of other DRs.

[8]https://huggingface.co/

by averaging non-contextual word vectors (bag-of-vectors), and SE obtained by using more recent language contextual models based on transformers that have shown promising results for various tasks. Comparing the last to bag-of-vectors enables us to determine whether coherence or DRs are only captured because of linguistic cues, or because of implicit links between clauses.

Global word Vectors (Pennington et al., 2014) are word representations obtained from aggregated global word-word co-occurrence statistics from a large crawled corpus. Training objective is to learn word vectors such that their dot product is the log of the word's probability of co-occurrences. We use the largest pre-trained model available, constituted of a vocabulary of 2.2M words, and containing vectors of dimension 300. We build sentence vectors by computing the mean of the word vectors, thus producing bag-of-vectors (BoV) of dimension 300.

Google USE (Cer et al., 2018) uses a Deep Averaging Network that is first trained in an unsupervised way as in Skip-Thoughts (Kiros et al., 2015), and whose output is transferred to be further trained on a supervised way on Natural Language Inference task. This encoder was the state-of-the-art before more recent contextual approaches with which we compare it. We use the largest model, which encodes sentences into vectors of dimension 512.

BERT (Devlin et al., 2019) is the current state-of-the-art for most NLP tasks. It produces contextual word representations, from training a bi-directional encoder on two tasks: the first called Masked LM (MLM) predicts a masked word from its left and right context, and the second called Next Sentence Prediction (NSP) predicts whether from two sentences $A$ and $B$, $B$ is the actual sentence that follows $A$. BERT thus represents the meaning of a word according to its immediate context, but also on the basis of relationships between sentences. We use BERT-base model, from which we recover the last hidden state after having processed the sentence in the model, thus producing vectors of dimension 768.

SciBERT (Beltagy et al., 2019) is a BERT model specifically trained on full papers from the corpus of Semantic Scholar, thus seeking to provide a better representation of scientific vocabulary. We use SciBERT-base and proceed in the same way as for BERT, obtaining vectors of the same dimension.

For GloVe, BERT and SciBERT, we calculate the embedding e(S) of the sentence S by calculating the mean of each word vector as defined in 1.

$$e(S) = \frac{1}{|S|} \sum_{w \in S} e(w) \quad (1)$$

### 4.3 Classification

Because we deal with datasets of different size and vectors made of latent variables, we train and evaluate two different classifiers for each task, namely a Logistic Regression (LR) and a Multi-Layer-Perceptron (MLP) with one hidden layer and three hidden units. On the one hand, the datasets are small in size, which can make the training of the MLP hard and a good LR performance possible. On the other hand, the properties that we probe may not be linearly separable, which would make them hard to tackle with a LR classifier. The comparison of two classifiers thus gives us a fine-grained analysis of both the results obtained and the way in which the properties are encoded.

**Experimental set-up** We use the evaluation toolkit Senteval[9] (Conneau and Kiela, 2018), and keep the parameters as defined in it, namely the optimizer is RmsProp (Tijmen and Geoffrey, 2012), the batch size is 164 and the loss function is cross-entropy. Because the number of sentences that we have is quite small, we proceed with 5-fold cross validation. We compute the mean of accuracies at each fold. To compute the test accuracy, we keep the best model and test it on the testing set (10%) of the data. We compare both results to ensure that the model is not overfitting.

## 5 Results

In this section, we comment the results obtained for each probing tasks. Table 3 shows the test accuracy and the mean accuracy of 5-fold cross validation. We only comment the latter, as it is more relevant due to the small datasets we have.

The LR classifier obtains better results than the MLP for all tasks. Although losses in accuracy when using MLP can be explained by the lack of data, LR gives good results with BERT and SciBERT for binRel, swapped and scrambled, suggesting that those properties may be encoded linearly in the vectors.

**Relation detection** DR are almost as well encoded by BoV than by BERT and SciBERT, which

[9] https://github.com/facebookresearch/SentEval

| Model | Encoder | binRel | | semRel | | swapped | | scrambled | |
|---|---|---|---|---|---|---|---|---|---|
| | | Test | Mean 5-fold | Test | Mean 5-fold | Test | Mean 5-fold | Test | Mean 5-fold |
| *LR* | | | | | | | | | |
| | BoV-Glove | **73.45** | 73.83 | **67.36** | 61.3 | 50 | 51 | 52.2 | 51.9 |
| | Google USE | **71.22** | 71.03 | **47.92** | 50.07 | 58.8 | 63.3 | 50.4 | 59.8 |
| | Bert-base uncased | 70.69 | 74.97 | 61.81 | 59.8 | 75.22 | 76.49 | **78.59** | 77.53 |
| | cased | **71.38** | 74.84 | **61.11** | 55.65 | **73.31** | 77.33 | 78.3 | 76.30 |
| | SciBert uncased | 76.55 | 77.98 | 50.69 | 60.16 | 77.71 | 77.74 | **80.06** | 79.77 |
| | cased | **78.28** | 76.76 | **63.19** | 59.22 | **77.71** | 79.64 | 81.09 | 79.53 |
| *MLP* | | | | | | | | | |
| | BoV-Glove | 72.76 | 71.53 | 60.42 | 51.82 | 49.85 | 50.87 | 52.49 | 52.06 |
| | Google USE | 67.59 | 66.64 | 48.61 | 45.67 | 60.26 | 60.82 | 56.01 | 57.31 |
| | Bert-base uncased | 69.31 | 68.47 | 36.81 | 28.16 | 77.57 | 67.71 | 75.22 | 64.27 |
| | cased | 50 | 52.62 | 25 | 25.85 | 78.15 | 67.85 | 75.66 | 54.39 |
| | SciBert uncased | 73.45 | 74.13 | 38.89 | 37.23 | 75.81 | 68.49 | 79.91 | 65.76 |
| | cased | 71.38 | 68.72 | 43.75 | 33.83 | 78.01 | 70.09 | 78.01 | 64.59 |

Table 3: Test and 5-fold mean accuracies of LR and MLP classifiers on each task for each sentence representation.

can be explained by the fact that they are most of the time lexicalized. Contextual embeddings slightly improve the results for binRel (from $\approx$ 73% for BoV to $76.4 \pm 1.5$), but not for semRel (61.3% for BoV to $57.9 \pm 2.2$).

By comparing two by two the results of BoV, BERT, and SciBERT encoders for binRel, we highlight that those models nearly give the same predictions (see Table. 4). Among the 290 sentences used as test, 94 are predicted differently by BERT and BoV, 85 by SciBERT and BoV, and 67 by SciBERT and BERT. Among the sentences that are similarly predicted, less than 20% correspond to common mistakes, and the others to common good predictions. For BERT and SciBERT, around 70% of the predictions are similar, showing that the models capture similar information for this task. For BERT and SciBERT with BoV respectively, more than 50% are common, but still they make a lot of different predictions, showing that they do not capture same aspects of the sentence.

For semRel however, the models make very different predictions. Among the 144 examples used as test, 47 are predicted differently by BERT and BoV, 78 by SciBERT and BoV, and 71 by SciBERT and BERT. The results of Table 5 are consonant with this variation, showing that all encoders have in fact different abilities for the prediction.

We show both precision and recall obtained for SemRel in Table 5. For `attribution` detection, BoV, BERT-uncased and SciBERT-cased gives the best recall (81%), for `enablement` it is BERT-uncased (67%), for `manner-means` it is BoV (78%), and for `norel` it is SciBERT-uncased (75%). These good recalls obviously result in lower precisions. However, as we want to find as many instances of a relation R, rather than maximizing the number of instances that are correctly found, we are still satisfied.

**Coherence detection** Representations build from BoV fail at predicting if units have been swapped (51%[10]), and perform badly for predicting if units have been scrambled ($\approx$ 51.9%). Google Universal Sentence Encoder improves swapped only by 10% and scrambled by 4%. Representations built from contextual embeddings improve accuracy for swapped by almost 26% for BERT and 27.5% for SciBERT and for scrambled by almost 25% for BERT and 27.8% for SciBERT. This shows that training on STS, on which Google USE is trained, is not enough to capture coherence links between clauses, and that the "Next Sentence Prediction" (NSP) task on which both BERT and SciBERT are trained seems indeed to help a lot. This task enables somehow to account for the order of the units in the sentence. Even if BERT and SciBERT are not trained at clause level, the corpus on which they are trained is large and therefore contains a large number of sentences that are made of a single clause. This enables the model to learn clause level contingency relations, which are in fact the task that we probe here.

[10]We consider that an accuracy close or less than 50% corresponds to random prediction, because a classifier predicting each class with a probability of 0.5 would have an accuracy of 50%.

| | binRel | | | semRel | | |
|---|---|---|---|---|---|---|
| | BERT-BoV | SciBERT-BoV | SciBERT-BoV | BERT-BoV | SciBERT-BoV | SciBERT-BoV |
| # common predictions | 196 | 205 | 223 | 97 | 66 | 73 |
| # different predictions | 94 | 85 | 67 | 47 | 78 | 71 |
| # common errors | 162 | 175 | 180 | 64 | 54 | 73 |

Table 4: Two-by-two comparison of the predictions obtained by different encoders for binRel and semRel.

## 6 Discussion

In this section, we discuss potential biases in the probing tasks that we designed, how we control them, and the possible improvements that could be further done.

**Relation prediction** A bias that could affect DR detection is the length of the sentence. In particular, Conneau et al. (2018) showed that BoV obtain 66.6% accuracy for predicting the length of the sentence, and get up to 99% with other more elaborated encoders such as BiLSTM or gated convolutional networks. We took that into account when selecting the different sentence sets. By including sentences that have more than one DU (sentences that only have relations with weak semantics) in the `norel` class , we somehow control the size of the sentences and ensure that the distributions of words per sentence are not specific to each class. The distributions are given in Fig. 3 , and are close enough to guarantee that the Sentence Length does not bias our predictions for binRel.

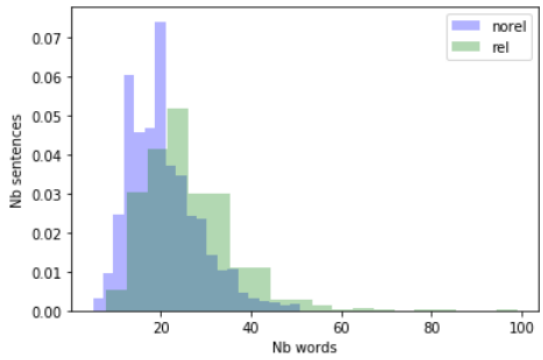


Figure 3: Distribution of the number of words per sentences for binRel.

Similarly, we ensured that the length of the sentence does not affect the detection of the relation semantics. The distributions which are given in Fig. 4 are close and, so, cannot be the only parameter involved in the good quality of the prediction.

However, good performances of BoV for this task are influenced by the fact that most of them are linguistically signalled. Even if this bias is in fact

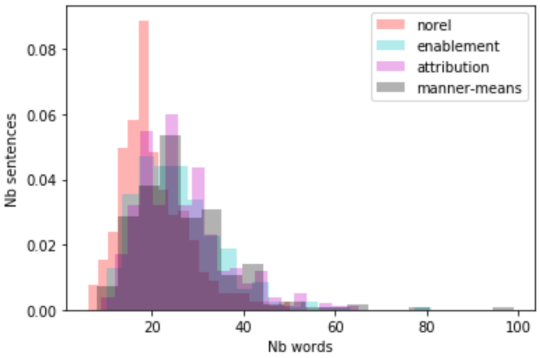


Figure 4: Distribution of the number words per sentences for semRel.

related to the specific genre of the corpus that we study, we discuss here the linguistic properties that play a role in the various encodings and predictions.

Verbs such as "*we show*" or "*we demonstrate*" typically occur in sentences which contain an `attribution` relation. Conjunctions such as "**to**" or "**in order to**" express `enablement` relations. Finally, `manner-means` are often signalled by verb forms such as "(**by**) +**ing**". We think that these signals play an important role in the good performances of BoV especially for `attribution` and `manner-means`.

**Manner-means** The sentence *"Most sentence embedding models typically represent each sentence only using word surface, which makes these models indiscriminate for ubiquitous homonymy and polysemy"* is a typical example of the sentences whose relation is predicted well with BoV encoders but neither by BERT, which predicts an `attribution`, nor by SciBERT, which predicts `noRel`. The verb "using" seem to be responsible for this good prediction. This underlies the good recall that BoV gets for the prediction of this DR, which gets a very bad recall for BERT and SciBERT.

**Attribution** The sentence *"[Most studies on statistical Korean word spacing do not utilize the information]_1 [provided by the input sentence]_2 [and assume]_3 [that it was completely concate-*

| | | BoV | Google USE | BERT uncased | BERT cased | SciBERT uncased | SciBERT cased |
|---|---|---|---|---|---|---|---|
| **Precision** | **Attribution** | 66 | 62 | 78 | 85 | 96 | 74 |
| | **Enablement** | 70 | 44 | 48 | 46 | 100 | 71 |
| | **Manner-Means** | 65 | 36 | 73 | 55 | 43 | 58 |
| | **NoRel** | 70 | 69 | 55 | 70 | 38 | 53 |
| **Recall** | **Attribution** | **81** | 78 | **81** | 78 | 61 | **81** |
| | **Enablement** | 58 | 19 | **67** | 58 | 14 | 47 |
| | **Manner-Means** | **78** | 69 | 53 | 64 | 53 | 58 |
| | **NoRel** | 53 | 25 | 47 | 44 | **75** | 67 |

Table 5: Precision and Recall for SemRel.

*nated ]_4."* is well predicted by all encoders, except SciBERT. Here, the captured relationship is a `manner-means` between units 3 and 4, which is in fact the deepest DR in the full structure of the sentence, and thus the less important. The second part (3+4) of the sentence is in conjunction with the first (1+2) and refers to it. The `manner-means` between 3 and 4 is thus not the more salient in the RST, but is however well predicted. This suggests that the choices that we did to ignore non-logical DRs work well and that the training made on sufficient data suffice to make the classifier focus on the relations that we consider.

**Enablement** With BERT, most of the sentences that contain the connective "**to**" are well detected as an `enablement` relation. Among those that are not well classified, most of them are classified as `attribution` or `norel`. One common characteristic to the sentences that are classified as `norel`, is that they are often long and contain more than 2 DUs and several discourse connectives, including "**to**" or "**for**". It is thus harder for the model to understand and distinguish the role of the connectives, as they are drowned in a lot of lexical information, leading to errors because of the difficulty of the task. For sentences that are wrongly classified as `attribution`, we observe a similar behaviour. The sentence *"[Recent work has shown success] [in using continuous word embeddings] [to improve supervised NLP systems.]"* contains an `enablement` between segments 1 and 2, but is classified as an `attribution`. This errors seems to be due to the verb "shown" that appears a lot in sentences that contain an attribution. This shows that even if BERT improves a sort of coherence detection, it in fact relies a lot on lexical cues for detecting the semantics of relations.

**Swapped and scrambled** We admit that the tasks designed for probing coherence are somehow arbitrary, and produces different biases that could be better controlled. We present two major issues, and propose solutions for future work.

The first issue concerns the swapped task, where modifications of sentences may produce other coherent sentences, which are thus tagged as incoherent in the training set and used as it by the classifier. Those cases specifically correspond to sentences that are formed by two DUs and linked by a particle (such as "**by**", "**for**" or "**to**"), which are thus swapped and still coherent. For example, the sentence "*[By incorporating textual information,][RCM can effectively deal with data sparseness problem.]*" becomes "*[RCM can effectively deal with data sparseness problem][by incorporating textual information.]*". We expect such sentences to be poorly predicted, but it is not the case, showing that BERT and SciBERT probably rely on other indices.

Conneau et al. (2018) also created acceptable sentences for three of their probing tasks, and filtered them by crowd-sourcing, asking people to rate sentences according to their acceptability. This method would be the best suited to solve this issue, as it may be hard to control coherence automatically with syntactic parsers.

Secondly, we did not ensure that the property captured does not rely on syntax, because sentence modifications may also produce sentences that are syntactically incorrect and that we did not filter out. Here the problematic cases mostly come from two clauses sentences containing a conjunction such as "**that**" or "**and**". For example the sentence "*[We present a human judgments dataset] [and an adapted metric for evaluation of Arabic machine translation.]*" becomes "*[And an adapted metric for evaluation of Arabic machine translation] [we present a human judgments dataset.]*", thus being obviously syntactically incorrect, and possibly captured as incoherent based on this syntactic property. A possible solution to that issue could be to forbid the swap of the first and last DU,

thus inducing another different bias and in our case reducing the size of the corpus. Another possibility is to introduce rules based on the syntactic tree of new sentences in order to filter those cases. In this preliminary work, we introduced the second coherence detection task (scrambled) with the aim of reducing those biases.

## 7 Conclusion

We introduced four tasks for probing the discourse properties involved in sentences from scientific abstracts. We evaluated the ability of a classifier to predict a discourse property of a sentence from its embedding. The performance of the classifier highlights the extent to which discourse properties are encoded in those representations.

We showed that coherence links are captured by vectors made from contextual models, as well as DRs, but that those models in fact do not encode the semantics of those DRs. We highlighted that BoV embeddings perform nearly as good as contextual embeddings for both DRs detection tasks, highlighting the fact that these relations are most of the time explicitly signalled in scientific abstracts.

This confirms our hypothesis that BERT and SciBERT training suffice to encode coherence to some extent. We think that the Next Sentence Prediction task is the reason for this performance, as it allows the model to learn adjacency properties of sentences, and thus of clauses. The second hypothesis however is not confirmed, as we concluded from the experiments that semantics of DRs are as well captured by BoV, as BERT or SciBERT.

The present work opens various possibilities, for improving the tasks as well as the analysis of the results. We decided here to focus on the detection of three specific relations that occur within sentences. We adopted strategies to make this possible, for example by ignoring DRs that are of continuation or additive and not logical. We could however go further by considering all relations involved in the sentences. A multi-label classification could be a solution to the problem of predicting all DRs of a sentence from its embedding. A Sequence-to-Sequence model could be a solution to the problem of predicting the discourse structure (or sequence of relations) from its embedding. For coherence prediction, we also plan to determine to what extent the tasks we introduced (swapped and scrambled) depend on syntactic properties. A statistical analysis of the syntactic structures involved could help to clarify the possible biases coming from the syntax. We also plan to further check the datasets that we introduce for those tasks, as we have raised that our method created coherent sentences, labeled as incoherent in our training dataset.

Other corpora such as RST-DT (Carlson and Okurowski, 2002) or PDTB (Webber et al., 2005), which could be combined, could be used to produce other datasets for probing DRs, enabling to evaluate how other DRs (such as `contrast` or `reason`) are encoded in sentence embeddings.

## Acknowledgments

We want to thank the reviewers for their constructive comments and suggestions.

This work was supported partly by the french PIA project "Lorraine Université d'Excellence", reference ANR-15-IDEX-04-LUE.

## References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *CoRR*, abs/1608.04207.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Daniel Marcu Carlson, Lynn and Mary Ellen Okurowski. 2002. Rst discourse treebank.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. *Current Directions in Discourse and Dialogue*, pages 85–112.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

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 Language Resources Association (ELRA).

Alexis Conneau, German 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.

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.

M.A.K Halliday and Ruqaiya Hasan. 1976. Cohesion in english. *Longman*.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. *CoRR*, abs/1506.06726.

Alex Lascarides and Nicholas Asher. 2007. Segmented Discourse Representation Theory: Dynamic Semantics With Discourse Structure. In Harry Bunt and Reinhard Muskens, editors, *Computing Meaning*, volume 3. Springer Netherlands, Dordrecht.

Elizabeth DuRoss Liddy. 1991. The discourse-level structure of empirical abstracts: An exploratory study. *Information Processing & Management*, 27(1):55–81.

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2018. A dependency perspective on RST discourse parsing and evaluation. *Computational Linguistics*, 44(2):197–235.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Jan Renkema. 2009. *The Texture of Discourse*. John Benjamins Publishing Company.

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.

Maite Taboada and William C Mann. 2006. Rhetorical structure theory: Looking back and moving ahead. *Discourse studies*, 8(3):423–459.

Tieleman Tijmen and Hinton Geoffrey. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4:26–31.

Bonnie Webber, Aravind K Joshi, Eleni Miltsakaki, Rashmi Prasad, Nikhil Dinesh, Alan Lee, and Katherine Forbes. 2005. A short introduction to the penn discourse treebank. *Copenhagen Working Papers in Language and Speech Processing*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*.

An Yang and Sujian Li. 2018. SciDTB: Discourse dependency TreeBank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449, Melbourne, Australia. Association for Computational Linguistics.