machine learning & knowledge extraction

MDPI

Article

Learning Sentence-Level Representations with Predictive Coding

Vladimir Araujo [1,2,*], Marie-Francine Moens [2] and Alvaro Soto [1]

[1] Computer Science, Pontificia Universidad Católica de Chile, Santiago 7820436, Chile
[2] Computer Science, KU Leuven, 3001 Leuven, Belgium
* Correspondence: vgaraujo@uc.cl

Abstract: Learning sentence representations is an essential and challenging topic in the deep learning and natural language processing communities. Recent methods pre-train big models on a massive text corpus, focusing mainly on learning the representation of contextualized words. As a result, these models cannot generate informative sentence embeddings since they do not explicitly exploit the structure and discourse relationships existing in contiguous sentences. Drawing inspiration from human language processing, this work explores how to improve sentence-level representations of pre-trained models by borrowing ideas from predictive coding theory. Specifically, we extend BERT-style models with bottom-up and top-down computation to predict future sentences in latent space at each intermediate layer in the networks. We conduct extensive experimentation with various benchmarks for the English and Spanish languages, designed to assess sentence- and discourse-level representations and pragmatics-focused assessments. Our results show that our approach improves sentence representations consistently for both languages. Furthermore, the experiments also indicate that our models capture discourse and pragmatics knowledge. In addition, to validate the proposed method, we carried out an ablation study and a qualitative study with which we verified that the predictive mechanism helps to improve the quality of the representations.

Keywords: deep learning; representation learning; natural language processing; language models; BERT; predictive coding

Citation: Araujo, V.; Moens, M.-F.; Soto, A. Learning Sentence-Level Representations with Predictive Coding. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 59–77. https://doi.org/10.3390/make5010005

Academic Editor: Vasile Palade

Received: 15 November 2022
Revised: 28 December 2022
Accepted: 29 December 2022
Published: 9 January 2023

1. Introduction

A natural language sentence is a set of words that makes complete sense and expresses a complete thought. Unlike words, sentences have more complex structures, such as sequential and hierarchical structures, which are essential for text understanding. In natural language processing (NLP), how to represent sentences is essential to solving many linguistic and non-linguistic problems, such as question answering, discourse understanding, translation, among others.

In recent years, various methods have been proposed, with the self-supervision approach being the main component for learning word representations [1,2] and sentence representations [3–5]. More recently, due to the introduction of the transformer model [6], new approaches adopted the transformer encoder for language modeling. Similarly to their predecessors, these approaches follow a self-supervised paradigm to learn contextualized word representations with different pre-training objectives, such as causal language modeling [7,8], masked language modeling [9], and permutation language modeling [10]. These models are also used to obtain sentence-level representations by pooling word representations or using a special token tuned through an additional sentence-level objective. However, it was shown that the resulting representations are not semantically meaningful [11] and do not capture the relationship between sentences (discourse knowledge) [12].

Diagnostics of transformer-based language models have shown that they still have limited linguistic capabilities [13–15]. This brings up the idea that the current mechanisms used to train the language models are insufficient to yield better capabilities, such

as learning general-purpose sentence representations. Unlike current language models, humans are not only trying to predict the next word of a text, but also upcoming compound linguistic information [16]. Predictive coding (PC) is a theory that postulates that our brain is continually making predictions of incoming sensory stimuli [17–20]. Studies have suggested that it plays an essential role in language development in humans [21,22], with word prediction being the main mechanism [23,24]. Furthermore, some studies speculate that the predictive process also occurs within and across utterances, fostering sentence and discourse comprehension [16,25,26].

In this work, we explore how to incorporate PC mechanisms in language models to improve sentence-level representations. Due to the success and ubiquitous use of the BERT model, we propose to augment these types of models with top-down computation. According to PC, top-down connections convey predictions from upper to lower layers, which are contrasted with bottom-up representations to generate an error signal that is used to guide the predictive process of the model. Specifically, our approach attempts to build general-purpose sentence representations that capture discourse-level relationships by continually predicting future sentences in a latent space. We perform several experiments to determine the advantages and disadvantages of our models, finding that they improve their stand-alone and discourse-aware sentence representations. Also, our models can capture document-level discourse knowledge. Furthermore, our model shows a slight improvement in some tasks that require pragmatics-based knowledge.

This paper is an extension of work originally presented at the *Conference on Empirical Methods in Natural Language Processing 2021* as a short paper [27]. The new contributions of this extended version are threefold: (1) We propose an additional method to incorporate feedback from a top-down network into a BERT-style model. Unlike [27], based on recurrent networks, this one is entirely based on transformer models; (2) We train all our models for Spanish to assess their generalization to a language other than English, which was the only language explored in [27]; (3) We extend the experimentation, including benchmarks for assessing sentence- and discourse-level representations and pragmatic knowledge, unlike [27] which only assess one benchmark for discourse knowledge. In addition, we expanded our ablation and qualitative study.

2. Predictive Coding: Overview and Motivation

The classical perception view maintains that we experience the world in a three-step bottom-up process [28]: We receive input from our environment, process input in higher levels of the brain, and respond to input accordingly. However, an alternative theory has been gaining relevance in the last decades, stating that information not only flows to higher cognitive areas but also that higher cognitive areas predict the input from our environment. This theory is known as Predictive Coding (PC) and offers a unified theory of cognition [17–19].

PC states that the brain continually generates predictions of the sensory input. A predictive model is created in higher cortical areas and communicated through top-down connections to lower sensory regions. In addition, bottom-up connections process and project an error signal, that is, the mismatch between the predicted information and the actual sensory input [17]. The predictive model is constantly updated according to this error signal and, in the process of doing so, performs learning.

In recent years, it has been suggested that PC also plays an important role in language development [29] and even the existence of a language-specific PC [30]. Some works affirm that PC is crucial for word recognition and learning in the early stages of language development [21,29]. Under this postulate, a comprehender predicts the next word of an ongoing conversation [23,24]. More recently, some work suggests that humans predict not only the following words but also sentences [16,25,26] and even their syntactic structure [31], leading to effects in sentence and discourse processing and understanding [16].

Motivated by the literature listed above, our objective is to explore if using PC mechanisms could improve what neural language models learn about textual data. Previous

work on computer vision has demonstrated the utility of this framework to enhance the learning of models when processing images and videos [32–35]. However, there is little or no presence of PC-inspired works in NLP [36]. Therefore, this is one of the first attempts to bring PC into the field of NLP to explore its usefulness for representation learning.

3. Related Work

3.1. Contextualized Word Representation Models

Several contextualization methods have been introduced in recent years, with self-supervised learning being the main component for their implementation. ELMo [2] introduced a method to contextualize representations of adjacent words using bi-directional recurrent encoders. A pooling of the word representations is usually used as the sentence representation. BERT [9] adopted a similar idea using a transformer encoder and the masked language modeling (MLM) objective to capture bi-directional context. It also proposes an additional loss called next-sentence prediction (NSP) to train a model that understands sentence relationships. Similarly, ALBERT [37] proposed a loss based primarily on coherence called sentence-order prediction (SOP). However, that kind of modeling does not generate semantically meaningful sentence embeddings [11], and removing them improves downstream task performance, as was the case of RoBERTa [38].

3.2. Sentence Representation Models

Several works have proposed models to learn general-purpose sentence representations using adjacent sentences in an unsupervised manner. Early work relied on recurrent networks [3], log-linear models [4] and matrix factorization models [5]. As for the first two, SkipThoughts [3] and FastSent [4] proposed an encoder-decoder model that encodes a sentence and tries to generate the previous and following sentences. Regarding the third, Sent2Vec [5] proposed an extension of the CBOW training objective [1] applied to sentences instead of words, yielding better results than SkipThoughts and FastSent.

More recently, due to the success of the transformer architecture in NLP, some models have been proposed. Universal Sentence Encoder [39] is trained in a multitask training setting together with an objective similar to SkipThoughts, predicting the surrounding sentences given the input sentence. As mentioned before, the BERT model was also intended to generate sentence representation using the special token [*CLS*]. CONPONO [40] incorporated a discourse-level objective on top of BERT to predict the surrounding sentences given an anchor text. SLM [41] proposed a sentence unshuffling approach for a fine understanding of the relations among the sentences.

3.3. Predictive Coding in Deep Learning

The PC framework has been used in machine learning and deep learning, primarily in computer vision research [32–35], to benefit two areas: training methods and latent representations. Regarding the first one, the algorithm of error backpropagation is commonly used to update the parameters of networks. However, the update rule is biologically implausible because of the non-locality of the updates. Some works have proposed using a PC variant to approximate the backpropagation error leading to similar results [42–44]. As for the second, some approaches to modify the networks were proposed, including mechanisms to predict current or future input. As a result, latent representations are enhanced, leading to better performances on downstream tasks. Some works take inspiration from PC theory to build models for accurate [32,33] and robust [34] image classification. Also, a network capable of predicting future frames in a video sequence by making local predictions using top-down connections was proposed [35]. More recent works have used PC intending to implement general latent representations for images [36,45], text [36], speech [36,46], and video [47]. Our work is located in this area, as the objective is to improve the text embeddings at the sentence level.

4. Proposed Method

Our main objective is to propose a method to implement PC on top of a vanilla BERT-style model. Under this framework, a BERT-style model could be like a bottom-up network that generates a latent representation at each layer given a sentence as input. We propose to augment such models with a top-down network to allow them to predict an incoming sentence representation at each layer and use the prediction error as an additional training objective. We hypothesize that this process could enable models to improve sentence-level representations by learning relationships between sentences.

In this work, we present two methods to implement the top-down network. The first is based on recurrent networks (first proposed in our previous work [27]), and the second is based on a transformer decoder. Below are the details of our proposal. We first present the general architecture of our approach and then detail the two proposed variations.

4.1. Architecture Description

Figure 1 shows the overview of our method. Given a sequence of contiguous coherent sentences $s_1, s_2, \ldots, s_n$ from a document. A bottom-up network g_{bu} generates a representation z_t^l at time t for each layer l (from 1 to L) of a sentence s_t.

$$z_t^l = g_{bu}(s_t);\ where\ z_t^l, s_t \in \mathbb{R}^{d_{model}} \tag{1}$$

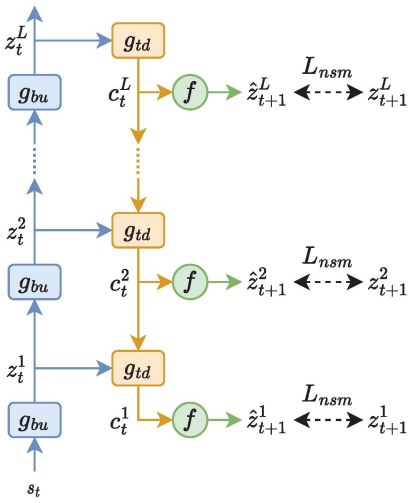


Figure 1. General scheme of the proposed method. A BERT-type model (g_{bu}) is augmented with a top-down network (g_{td}) which generates a contextual vector c at each layer to predict the next sentence representation $\hat{z}_{t+1}$ through a predictive layer (f). The predicted representation $\hat{z}_{t+1}$ is then contrasted with the actual one z_{t+1} to compute $\mathcal{L}_{nsm}$ to optimize the complete model.

Then, a top-down network g_{td} produces a context vector c_t^l for each l given each layer's sentence representation z_t^l and the context vector of the upper layer c_t^{l+1}. Note that this network processes in the opposite way to the bottom-up network from layer L to layer 1.

$$c_t^l = g_{td}\left(z_t^l, c_t^{l+1}\right);\ where\ c_t^l \in \mathbb{R}^{d_{model}} \tag{2}$$

Finally, a predictive function f is introduced to predict a future sentence representation $\hat{z}_{t+1}^l$ given the contextual vector c_t^l for each layer.

$$\hat{z}_{t+1}^l = f\left(c_t^l\right) \tag{3}$$

In the spirit of Seq2Seq [48], the whole process can be repeated for N steps (where N is a hyper-parameter) enabling the model to predict the next sentences sequentially.

4.2. Top-Down Networks Alternatives

Our bottom-up network g_{bu} by default is a pre-trained BERT-style model. For the top-down network g_{td}, we explore two different architectures.

The first is based on the GRU model [49] (Figure 2a), a type of recurrent network that employs a gated process to control and manage the flow of information. GRU facilitates capturing dependencies from sequential data. With this network, the sequence of sentences is processed one by one. At each step, the input is the concatenation ($\oplus$) of the sentence representation z_t^l with the context vector of the upper layer c_t^{l+1} and a hidden state, which in this case is the context vector of the previous step c_{t-1}^l. Note that the concatenation operation results in a vector of dimension $\mathbb{R}^{2\times d_{model}}$; however, the GRU output vector is set to $\mathbb{R}^{d_{model}}$.

$$c_t^l = GRU\left(z_t^l \oplus c_t^{l+1}, c_{t-1}^l\right) \tag{4}$$

(**a**) GRU-based top-down layer

(**b**) Transformer-based top-down layer. Skip connections were not illustrated for simplicity.

Figure 2. Diagram of the alternative top-down networks: (**a**) based on GRU, processes the sequences step by step, (**b**) based on transformer, processes the sequences as a whole.

The second option uses a transformer model (Figure 2b). This architecture avoids recursion by processing the inputs as a whole using attention mechanisms. Specifically, we use a transformer decoder which includes a cross-attention layer that mixes two different embedding sources. Unlike the recurrent version, the transformer-based top-down network needs to access all sentence representations and context vectors and not one by one. Therefore, the inputs for this model are the context vectors of the upper layer C^{l+1} and the sentence representations Z^l for all $t_{\geq}$.

$$C^l = Transformer\left(C^{l+1}, Z^l\right);\ where\ C^l \in \mathbb{R}^{d_{model}\times N} \tag{5}$$

Note that the transformer's dot-product attention mechanism computes the key K_C and value V_C vectors from the context vectors and the query Q_Z from the sentence representations. The dot products are scaled down by the dimensionality of the keys $\sqrt{d_k}$.

$$CrossAttn(Q_Z, K_C, V_C) = softmax\left(\frac{Q_Z K_C^\top}{\sqrt{d_k}}\right)V_C;\ where\ Q_Z, K_C, V_C \in \mathbb{R}^{d_k\times N} \tag{6}$$

4.3. Loss Functions

Following previous work [40], we keep the masked language model objective $\mathcal{L}_{mlm}$ for our model's training because it helps maintain a good token-level representation needed for the sentence-level language modeling. This task consists of masking tokens in a sequence with a masking token and directing the model to fill that mask with an appropriate token [9], allowing the model to obtain contextual representations.

We then introduce an additional loss function, which we refer to as next-sentence modeling $\mathcal{L}_{nsm}$. This loss is based on the InfoNCE loss [36], which constructs a binary task

where the goal is to classify one positive sample among many negative samples. Given the ground truth representations z_i^j and the predicted representations $\hat{z}_i^j$ where i denotes the temporal index and j is the layer index, a dot product computes the similarity between the predicted and ground truth pair. Note that ground truth representations are computed with the same backbone on the fly. A cross-entropy loss is then used to distinguish the positive pair from all other negative pairs. In this way, the model encourages the predicted representation $\hat{z}$ to be close to the ground truth z.

$$\mathcal{L}_{nsm} = -\sum_{i,j}\left[\log \frac{\exp\left(\hat{z}_i^{j^\top} \cdot z_i^j\right)}{\sum_m \exp\left(\hat{z}_i^{j^\top} \cdot z_m^j\right)}\right] \tag{7}$$

For a predicted sentence $\hat{z}_i^j$, there is only one positive pair $(\hat{z}_i^j, z_i^j)$ which are the features at the same time step and same layer. The remaining pairs $(\hat{z}_i^j, z_m^j)$ are negative pairs, where $(i, j) \neq (m, j)$. In practice, negative samples are drawn from both batch and time dimensions in the minibatch. Unlike our previous work [27], we do not extract negative samples for the layer dimension, as we found that it degrades the pre-training.

Our loss function is used in conjunction with the BERT masked language model loss ($\mathcal{L}_{mlm}$) to minimize:

$$\mathcal{L} = \mathcal{L}_{nsm} + \mathcal{L}_{mlm} \tag{8}$$

4.4. Technical Details

We use the BERT [9] and ALBERT [37] models as the backbone for three reasons: (1) they use MLM with an additional sentence-level objective that we replace with our own, (2) they originally use the same corpus for pre-training, and (3) they are available for English and Spanish [50,51]. As a result, our augmented models are PredBERT and PredALBERT with the suffix -R or -T to refer to the GRU and transformer versions, respectively.

Our approach processes contiguous sentences one at a time. The original BERT model was optimized for long lines (512 tokens), and using shorter sequences tends to harm the MLM performance [52]. For that reason, we join 3 adjoining sentences to create a long sequence. Furthermore, as we proposed in our original work [27], we use an overlapping sentence between sequences, since it improves the predictive capacity of the models. As an example, if we have a paragraph $s_1, s_2, \ldots, s_9$, the first sequence is s_1, s_2, s_3, the second sequence is s_3, s_4, s_5, and so on.

For pre-training, we use the BookCorpus [53] and Wikipedia datasets for English and the Spanish Unannoted Corpora [50] for Spanish. We set $N = 2$ to reach two future predictions. We initialize the bottom-up networks with the corresponding checkpoints of BERT or ALBERT and the top-down networks with random weights. The initial (upper) context vector c is initialized to zero. We consider dynamic masking for the MLM, where the masking pattern is generated every time we feed a sequence to the model. We mask 10% of all tokens in each sequence at random to avoid heavily corrupted input such as sentence representations. Finally, the models are trained for 1M steps with batch size 8 and using Adam optimizer with a learning rate of 5×10^{-5}. These hyperparameters follow the standard values used for fine-tuning BERT-style models.

In order to use our models as feature extractors or sentence encoders, we discard the top-down connections and the prediction layer keeping only the backbone. In this way, we obtain a model equivalent to BERT in terms of parameters and processing speed.

5. Experiments

Our central experimentation focuses on sentence-level evaluation. However, due to the architecture modification and the inter-sentence objective proposed in the model, it is expected that it captures more complex knowledge and relationships. For this reason, we conduct an evaluation focused on discourse and pragmatics knowledge encoding.

Next, we will describe the baselines we use to compare our approach. Then the description and the results of the benchmark are presented. Note that we use resources available in both English and Spanish, which allows us to verify the consistency and generalization of our models.

5.1. Baselines

Our models follow a self-supervision paradigm to obtain the general-purpose representations of the text, so we compare them with models with a similar approach. We evaluate non-transformer-based and transformer-based models publicly available in English and Spanish.

- **Sent2Vec** [5] is a bi-linear model to compose sentence embeddings using word vectors along with n-gram embeddings. We use the official implementation for the English (github.com/epfml/sent2vec, accessed on 14 November 2022) version and a released Spanish (github.com/BotCenter/spanish-sent2vec, accessed on 14 November 2022) model.
- **ELMo** [2] is a general approach for learning deep context-dependent representations from bidirectional language models. We use the ELMo library for many languages (github.com/HIT-SCIR/ELMoForManyLangs, accessed on 14 November 2022), including English and Spanish.
- **BERT** [9] is a transformer-based model designed to pre-train deep bidirectional representations from the unlabeled text. We use the original model in English and the version in Spanish BETO [50].
- **ALBERT** [37] is a lite version of BERT that incorporates two parameter reduction techniques. We use the official English model and ALBETO [51] the version in Spanish.

5.2. General-Purpose Sentence Representation Evaluation

5.2.1. Description

A well-known methodology for evaluating text representation is probing tasks, consisting of a classifier based on embeddings generated by sentence encoders for specific downstream tasks. In this way, it is possible to evaluate the representational quality of embeddings in particular tasks. We rely on SentEval (github.com/facebookresearch/SentEval, accessed on 14 November 2022) [54] and DiscoEval (github.com/ZeweiChu/DiscoEval, accessed on 14 November 2022) [55] and their Spanish versions (github.com/OpenCENIA/Spanish-Sentence-Evaluation, accessed on 14 November 2022) [56] for this evaluation.

On the one hand, SentEval includes stand-alone sentence and sentence pair tasks modeled by logistic regression or classification. We group the tasks into five groups. Sentence Classification (SC) which includes tasks like sentiment analysis, subjective/objective, and question-type classification. Pair Classification (SPC), which includes entailment and paraphrasing tasks. Supervised Semantic Similarity (SSS) and Unsupervised Semantic Similarity (USS), which include textual similarity tasks based on cosine similarity. Linguistic probing tasks (PT), which include probing tasks to evaluate individual linguistic properties.

On the other hand, DiscoEval includes tasks to evaluate discourse-related knowledge in the same way that SentEval does. It has five groups of tasks. Sentence Position (SP) to evaluate the ability of a model to order ideas in a paragraph. Binary Sentence Ordering (BSO) to determine if the order of two sentences is correct. Discourse Coherence (DC) to determine if a sequence of six sentences forms a coherent paragraph. Sentence Section Prediction (SSP) to determine the section of a given sentence. Discourse Relations (DR) to evaluate the relations between sentences based on Rhetorical Structure Theory (RST) [57] and Penn Discourse Treebank (PDTB) [58].

5.2.2. Results

We use the default SentEval and DiscoEval settings for our experimentation. To evaluate Sent2Vec and ELMo, we use the pooling of the representations. For transformer-based models, we use each layer's average of the special tokens [*CLS*] as a representation of

sentences. Table 1 shows the results of our models across all tasks and languages. We ran the experiment 10 times with different seeds and reported the average accuracy (Pearson's correlation for SSS and USS) and standard deviations.

Table 1. Results in SentEval and DiscoEval benchmarks in English and Spanish. The metric is accuracy except for SSS and USS, which is Pearson's correlation. We report standard deviations by running the evaluations 10 times with different seeds.

(a) Results for models in English.

	SentEval						DiscoEval			
	SC	**SPC**	**SSS**	**USS**	**PT**	**SP**	**BSO**	**DC**	**SSP**	**DR**
Sent2Vec	79.22 ± 0.1	72.69 ± 0.4	75.36 ± 0.1	**63.95**	66.68 ± 0.1	44.85 ± 0.2	62.65 ± 0.2	54.63 ± 0.3	76.99 ± 0.5	41.24 ± 0.4
ELMo	76.87 ± 0.2	73.10 ± 0.2	70.76 ± 0.1	54.70	72.29 ± 0.1	45.73 ± 0.3	63.72 ± 0.2	59.43 ± 0.3	78.44 ± 0.5	44.32 ± 0.6
BERT	80.28 ± 0.2	71.10 ± 0.2	62.83 ± 0.2	30.14	73.88 ± 0.1	**53.18** ± 0.3	**68.36** ± 0.2	59.38 ± 0.5	**80.43** ± 0.2	47.94 ± 0.3
ALBERT	80.46 ± 0.4	70.02 ± 0.3	59.75 ± 0.4	21.68	70.17 ± 0.1	52.18 ± 0.2	67.91 ± 0.2	52.85 ± 1.4	80.10 ± 0.4	43.50 ± 0.5
PredBERT-R	80.03 ± 0.1	73.68 ± 0.2	71.71 ± 0.1	47.89	**75.36** ± 0.1	50.49 ± 0.2	66.84 ± 0.1	**62.33** ± 0.4	80.16 ± 0.5	**47.96** ± 0.4
PredALBERT-R	**80.48** ± 0.2	**74.12** ± 0.2	**76.51** ± 0.1	43.17	74.76 ± 0.1	49.85 ± 0.2	66.77 ± 0.2	61.46 ± 0.2	79.01 ± 0.2	46.82 ± 0.6
PredBERT-T	78.40 ± 0.1	73.94 ± 0.1	72.00 ± 0.1	49.47	74.55 ± 0.1	48.56 ± 0.1	66.17 ± 0.1	61.69 ± 0.3	78.25 ± 0.4	46.35 ± 0.5
PredALBERT-T	79.28 ± 0.2	73.56 ± 0.2	75.18 ± 0.1	30.42	74.47 ± 0.1	49.30 ± 0.3	66.57 ± 0.3	60.88 ± 0.2	79.01 ± 0.4	45.26 ± 0.4

(b) Results for models in Spanish.

	Spanish SentEval						Spanish DiscoEval			
	SC	**SPC**	**SSS**	**USS**	**PT**	**SP**	**BSO**	**DC**	**SSP**	**DR**
Sent2Vec	75.18 ± 0.2	59.55 ± 0.2	78.01 ± 0.1	**75.08**	66.78 ± 0.2	36.61 ± 0.2	55.10 ± 0.1	55.40 ± 0.6	69.96 ± 0.7	37.28 ± 1.2
ELMo	72.34 ± 0.2	61.73 ± 0.2	75.53 ± 0.1	57.49	71.77 ± 0.2	37.06 ± 0.3	55.12 ± 0.3	58.89 ± 1.1	73.55 ± 0.9	42.58 ± 1.1
BETO	76.21 ± 0.3	58.78 ± 0.3	63.09 ± 0.1	52.74	69.60 ± 0.2	41.79 ± 0.4	57.16 ± 0.4	60.56 ± 0.9	76.53 ± 0.8	47.53 ± 0.9
ALBETO	69.28 ± 0.3	52.46 ± 0.6	53.12 ± 0.2	29.61	66.54 ± 0.1	**43.56** ± 0.2	57.42 ± 0.2	56.26 ± 0.5	**79.85** ± 0.8	42.78 ± 0.8
PredBETO-R	**76.65** ± 0.3	61.37 ± 0.1	77.75 ± 0.1	65.31	**73.58** ± 0.1	43.42 ± 0.2	57.22 ± 0.3	**64.04** ± 0.6	77.26 ± 0.9	**49.33** ± 0.9
PredALBETO-R	72.52 ± 0.3	60.67 ± 0.4	78.39 ± 0.1	71.74	70.49 ± 0.2	41.56 ± 0.3	56.96 ± 0.4	63.37 ± 0.7	77.27 ± 0.9	43.88 ± 0.9
PredBETO-T	75.75 ± 0.3	**61.77** ± 0.2	**78.70** ± 0.2	66.67	73.54 ± 0.1	41.77 ± 0.2	57.36 ± 0.2	60.66 ± 0.6	75.87 ± 0.7	48.04 ± 0.7
PredALBETO-T	74.01 ± 0.3	61.37 ± 0.2	76.87 ± 0.1	67.10	71.72 ± 0.2	42.43 ± 0.2	**57.83** ± 0.2	62.36 ± 0.4	79.10 ± 0.8	45.53 ± 0.9

In the case of SentEval, we see that our models outperform the baselines for all tasks except USS. Regarding the English models, we found that PredALBERT-R is the best model improving its baseline ALBERT by ∼9.39 points. The overall improvement of our models and their direct baselines averages ∼6.72 points. Among the Spanish models, we found that PredBETO-T is our best model, surpassing its baseline BETO by ∼9.20 points. Interestingly, for the Spanish language, the overall improvement of our models from their direct baselines is even higher than the English version, averaging ∼12.65 points. The results also show that for both languages, the best version of the model is the one that used recurrent networks during pre-training. Furthermore, the USS group does not outperform the Sent2Vec and ELMo models. We suspect this is due to the inadequacy of the [*CLS*] token for semantic similarity purposes [11] of the vanilla BERT model that is transferred to our model. However, we found an average improvement of ∼16.83 for English and ∼26.53 for Spanish of our models with respect to the transformer baselines. Also, the supervised version (SSS) improves performance even over non-transformer baselines, supporting our assumption.

Concerning DiscoEval, our models improve two groups of tasks with respect to their direct baselines. The performances of the DC and DR tasks are improved by ∼5.48 and ∼0.87 on average, respectively, for the English language. In the case of Spanish, we found improvement in three groups of tasks. DC is improved by ∼4.20 points while DR by ∼1.54 on average. Additionally, the Spanish model slightly improved the BSO task by ∼0.05 points. Furthermore, we found that our models outperform all the non-transformer baselines by ∼3.49 for English and ∼5.11 for Spanish. Regarding the better performance of the baselines on SP, BSO, and SSP tasks, we hypothesize that it could be due to the optimization goals of the baselines: ALBERT with sentence order prediction and BERT with topic prediction, which are different from our next sentence prediction based objective that may be promoting the capture of discourse relationships.

5.3. Document-Level Discourse Evaluation

5.3.1. Description

Due to the results of our model on DiscoEval, we extend the discourse-based evaluation to discover to what extent our model is acquiring discourse knowledge. In [59], a benchmark to assess how much discourse structure language models based on pre-trained transformers have and whether they generalize across languages was proposed. It introduces seven document-level discourse-probing tasks for several languages, including English and Spanish. The tasks included are: Next Sentence Prediction (NSP), to predict the next one among 4 candidates by giving 2–8 sentences as context. Sentence Ordering, to reproduce the original order of 3–7 shuffle sentences. Discourse Connective prediction, to identify an appropriate discourse marker given 2 clauses. RST Nuclearity prediction, to predict the nucleus/satellite for a given ordered pair of elementary discourse units (EDU). RST Relation prediction, to predict the relation that holds an ordered pairing of EDUs. RST EDU Segmentation, to chunk a sequence into its component EDUs. Cloze story test, to predict the best ending from 2 options given a 4-sentence story.

5.3.2. Results

We use the original benchmark implementation (github.com/fajri91/discourse_probing, accessed on 14 November 2022) with its default settings. Figure 3 shows the probing task performance in English (7 tasks) and Spanish (5 tasks) at each layer. It is possible to see that for the 5 common tasks (Next sentence prediction, Sentence ordering, Nuclearity prediction, Relation prediction, and EDU segmentation), the English and Spanish models behave similarly.

In the case of the Next sentence prediction task, the BERT and ALBERT models perform well ($\sim$99%) since the 5th layer, while our models reach that performance since the 11th layer. This is to be expected as models tend to specialize the last layers to the specific task they perform [60], in this case, our next sentence prediction optimization goal.

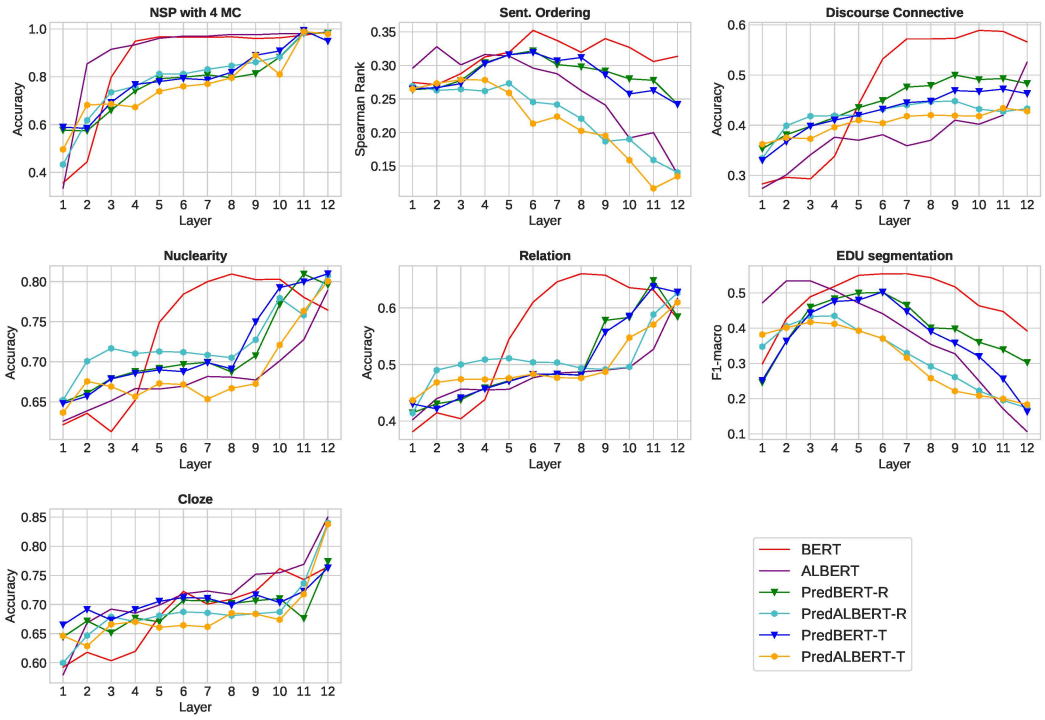
Regarding the Sentence ordering task, all models, both in English and Spanish, perform similarly on the lower layers, but our models' performance drops on the upper layers. This task is similar to BSO of DiscoEval, so these results confirm that our models are not acquiring the capacity to order because the optimization objective does not induce it explicitly.

Discourse connective results are only available for English models. It can be seen that the performance of the lower layers increases and that of the upper layers is competitive. This means that our model is implicitly learning about discourse markers.

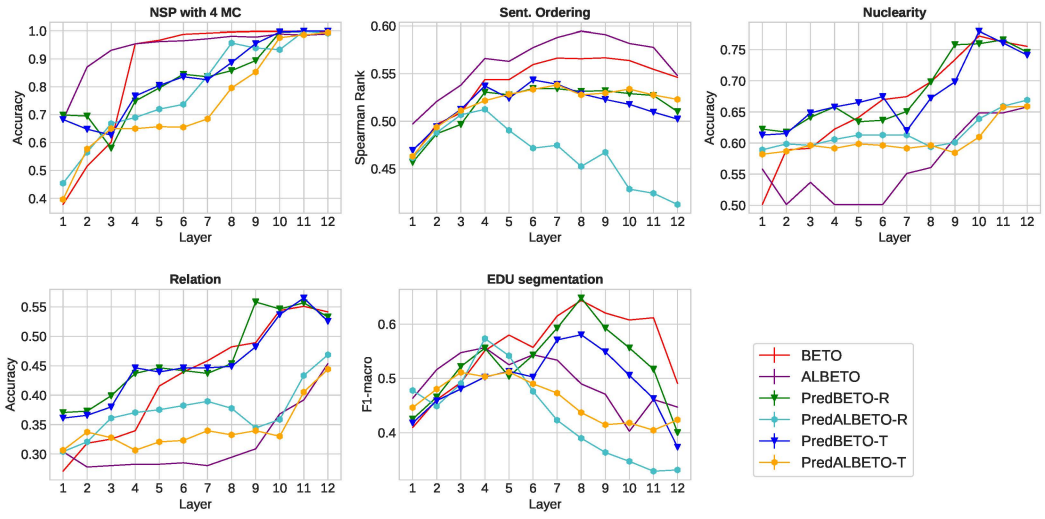
Concerning the Nuclearity and Relation prediction tasks, we found our models perform better than the baselines both in English and Spanish. The performance is almost always superior to the baseline across the layers, and the maximum performance is reached at layers 11th or 12th. These tasks are similar to DR from DiscoEval, and these results confirm that our models' pre-training induces the capture of discourse relations.

In the case of EDU segmentation, we found that our models perform similarly to the baselines in the lower layers. However, the performance is comparable or even slightly lower in the upper layers for both languages. EDU segmentation is a complex task, and we suppose the performance of our model is not superior because the pre-training was at the sentence level and not at the elementary discourse unit level.

Finally, the Cloze task is only available in English. This task is similar to next sentence prediction but a bit harder as it requires understanding commonsense. We found that our model performs slightly better than the baselines demonstrating that our method is not inducing commonsense knowledge but can predict better story's ends.



(**a**)



(**b**)

Figure 3. Document-level discourse probing task performance for each of the five tasks. We perform the evaluation three times with different initial seeds. (**a**) Results for models in English; (**b**) Results for models in Spanish.

5.4. Pragmatics-Focused Evaluation

5.4.1. Description

Linguistic work has shown that comprehenders generate expectations about the direction a discourse is likely to take [16]. This is because, in addition to processing structural and semantic relationships, humans also take advantage of pragmatic relationships to understand the language better. We suspect that our models are gaining some pragmatic knowledge because of the anticipation mechanisms leading to an improvement in discourse-related tasks. For that reason, we propose to test our models on a recently published pragmatics-focused benchmark.

PragmEval [61] is a recent benchmark that compiles eleven pragmatics-focused tasks (available only for English). The aim is to complement semantics-focused evaluation that only focuses on the literal content of sentences or shallow aspects of a document structure (e.g., sentence ordering). Unlike the previous benchmarks, this one follows a fine-tuned process instead of relying on sentence embeddings and fixed compositions. The benchmark presents three groups of tasks: (1) Speech act classification tasks; (2) Persuasiveness; and (3) Detecting emotional content, verifiability, formality, informativeness, or sarcasm.

5.4.2. Results

For this experimentation, we use the original implementation provided by the authors (github.com/sileod/pragmeval, accessed on 14 November 2022) with default settings. Table 2 shows the results of our English models and their corresponding baselines.

Table 2. Results in the PragmEval benchmark for our models in English. The metric is accuracy for all tasks. We perform the evaluation three times with different initial seeds.

	PDTB	**STAC**	**GUM**	**Emerg.**	**SwitchB.**	**MRDA**	**Persua.**	**Sarcasm**	**Squinky**	**Verif.**	**EmoB.**
BERT	51.24	57.13	42.33	79.92	64.25	45.05	70.61	77.82	88.71	84.57	76.29
ALBERT	52.81	56.82	50.80	82.62	65.02	45.42	73.14	80.59	88.28	84.85	75.90
PredBERT-R	49.76	56.44	38.30	82.23	64.25	45.81	70.92	77.18	**88.86**	83.95	76.34
PredALBERT-R	46.35	52.60	37.50	77.22	**65.79**	45.21	**73.81**	77.61	87.48	84.69	73.44
PredBERT-T	47.65	55.98	37.10	78.38	64.64	**46.03**	71.25	77.61	87.71	84.45	**76.68**
PredALBERT-T	45.16	50.23	34.27	76.06	64.71	44.31	73.01	76.76	87.35	84.16	75.01

In the speech act classification tasks, our models improve performance on SwitchBoard and MRDA dealing with utterance intention detection. Interestingly, our models do not perform better on PDTB, STAC, GUM, and Emergent, which deal with discourse relation prediction with varying domains and formalisms. One possible reason is that these tasks are more complex than the previous benchmarks. Also, the fine-tuning process might be interfering with the resulting embeddings, so a grid search might be needed to get the best hyperparameters.

Regarding the persuasiveness task, we found that our models outperform all the baselines. This means that our models, to some extent, can measure how well a sentence can achieve its intended goal.

Finally, detecting emotional content, verifiability, formality, informativeness, or sarcasm allow us to figure out in what realm communication is occurring. We found that our models improve the performance of emotion detection (EmoBank) and detection of formality, informativeness, and implicature (Squinky) while remaining competitive for the other tasks.

6. Further Experimentation

We carry out further exploration to better understand our model. First, we perform an ablation study to check the influence of the PC mechanism on the resulting representation. We then explore the quality of our model representations by analyzing the embedding space and performing sentence retrieval.

6.1. Ablation Study

The ablation study allows us to understand which components or configurations have the most impact on the performance of a model. In this case, we are interested in analyzing the impact of the proposed PC mechanism in the pre-training of our models. We use PredALBERT-R and PredALBETO-R as the default models and carry out two ablations using SentEval and DiscoEval (see Table 3): (1) remove top-down connections, and (2) remove recurrence processing.

Table 3. Results of ablation study in SentEval and DiscoEval. PredALBERT-R, both in English and Spanish, was used as the default model.

	English		Spanish	
	SentEval	**DiscoEval**	**SentEval**	**DiscoEval**
Default Model	69.81 ± 0.1	60.78 ± 0.3	70.76 ± 0.1	56.61 ± 0.7
# TD Layers = 6	69.89 ± 0.1	60.61 ± 0.4	70.87 ± 0.1	56.26 ± 0.7
# TD Layers = 1	65.59 ± 0.1	60.03 ± 0.4	70.22 ± 0.1	55.58 ± 0.7
w/o TD connections	63.75 ± 0.1	57.99 ± 0.3	68.70 ± 0.1	53.89 ± 0.7

By default, the number of top-down (TD) layers is 12 since the BERT-style models we use 12 layers. For the first ablation, we changed the number of TD layers to 6 and 1. Note that the number of layers is counted from the top to the bottom. If we set it to 6, the BERT-style model is augmented with the PC mechanism from the 12 to 6 layer. We see that for both Spanish and English, the final performance is slightly reduced on DiscoEval. However, the performance is a bit better on SentEval. We hypothesize that this is because the representations of the upper layers are the ones that provide more information to solve tasks at the sentence level [56,62], so improving these representations leads to better performance. When the TD layers are equal to 1, the mechanism is only at the final layer of the BERT-style model. As expected, model performance drops, supporting the importance of the top-down network to improve final embeddings.

It is unclear whether the top-down connections that carry information from higher to lower layers actually influence the quality of the representations. For that reason, as a second ablation, we remove such connections, which means that the model predicts the next sentence representations in each layer with only the context of the previous steps of the sequence and not of the layers above. In particular, we found that the performance of all tasks and languages dropped. This indicates that both components, prediction and top-down connections, are important during the pre-training process to improve sentence-level representations.

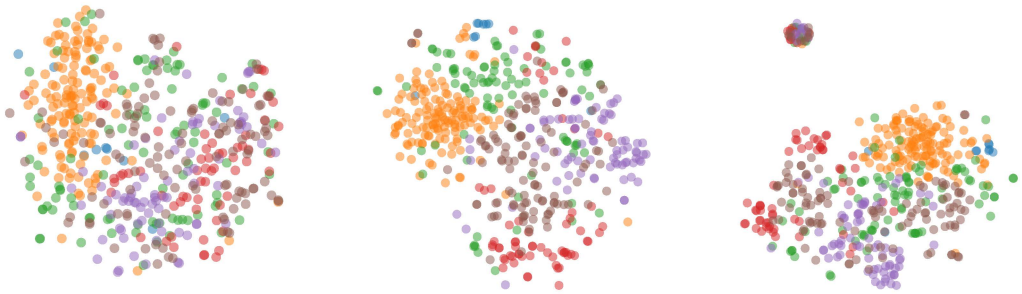
6.2. Learned Sentence Embedding Space

We applied t-SNE [63] to the vectors extracted from our models as additional experimentation. It allows us to visualize the embedding space learned by the models and analyze their separability in low dimensionality.

For this experiment, we use two equivalent tasks from SentEval, TREC [64] for English and SQC for Spanish [65]. Both tasks consist of a question-type classification with six different classes. We extracted the vectors from the best model version for each task, PredALBERT for TREC and PredBETO for SQC.

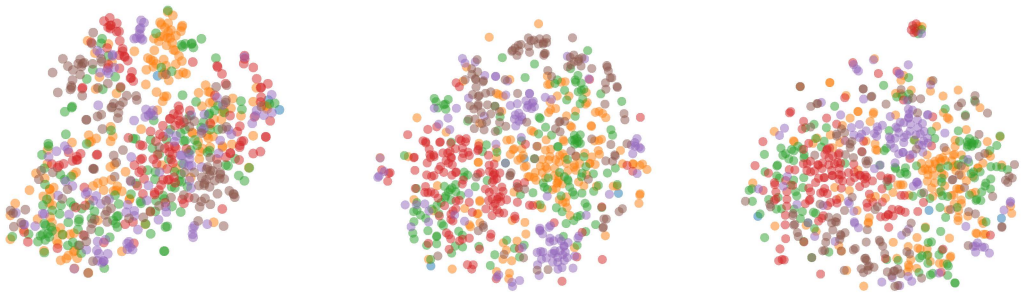
Figure 4 show all the visualizations for TREC. It is possible to see that PredALBERT-R (Figure 4b) groups the orange, red, and blue clusters a bit more compared to its reference ALBERT (Figure 4a), in which most of the points are spread out in space. A similar result is found for PredALBERT-T (Figure 4c); however, in this case, it can be seen that the red cluster is not fully grouped, and there is a very separate cluster that mixes several samples from different semantic categories. This could explain the reduced performance of this model compared to PredALBERT-R.

Interestingly, for the Spanish language, we find similar behaviors. The BETO (Figure 5a) baseline shows an overlap of all the different categories. While PredBETO-R (Figure 5b) and PredBETO-T (Figure 5c) group the orange, red and brown clusters a little better. Once again, it can be seen that there is an additional cluster with points of different categories for the PredBETO-T model. We assume this is because transformer models can capture very complex relationships and can represent some shared semantic information between those examples.



(**a**) ALBERT (**b**) PredALBERT-R (**c**) PredALBERT-T

Figure 4. TSNE visualization of the embedding space learned on TREC [64] for ALBERT and PredALBERT in English. Each color indicates a ground truth semantic category.



(**a**) BETO (**b**) PredBETO-R (**c**) PredBETO-T

Figure 5. TSNE visualization of the embedding space learned on SQC [65] for BETO and PredBETO in Spanish. Each color indicates a ground truth semantic category.

In general, we see that the clusters are not well separated and even mixed, which means that the classes are not easily separable. However, it is possible to observe that some clusters improved with the proposed mechanism, which may be causing an improvement in the final performance of the model.

6.3. Retrieval of Discourse Relations

To explore whether our models' resulting representations actually represent a sentence's role in its discursive context, we perform a sentence retrieval task. Following the methodology proposed in [41], we use labeled sentences with discourse relations as queries to retrieve the top 3 most similar sentences from an unlabeled corpus.

Specifically, we used the representations of annotated sentences from the MIT Discourse Relations Annotation Guide (bit.ly/3z45IG2, accessed on 14 November 2022) as queries to compute the cosine similarity of sentences from the Gutenberg corpus. For Spanish, we manually translated the queries and used the MLSUM dataset [66] to retrieve the sentences. This process allows us to empirically verify that similar vectors share identical or equivalent discourse relations.

Table 4 shows the retrieval results for both languages, using the four major types of discourse relationships. It is possible to see that the retrieved sentences share syntactic, semantic, and discursive aspects with the given query. Note that the model was not fine-tuned, demonstrating the quality of our model representations for the retrieval task.

Table 4. Sentences retrieved from an unannotated corpus given an annotated query with discourse relations. Words highlighted in red and blue show discursive connectors and semantic relationships, respectively.

Relation Type	**Query and Nearest Neighbors**	
	English	**Spanish**
Contingency	**Query:** I refused to pay the cobbler the full $95 because he did poor work. ✓(1) I cannot owe money to the little village cobbler who mends my shoes, because he demands and receives his payment when his job is done. ✗(2) He offers me money–not paid money down, which would have certain allurements. ✓(3) And I had to give Theodore fifty francs on the transaction, as he threatened me with the police when I talked of giving him the sack.	**Query:** Me negué a pagarle al zapatero los $95 completos porque hizo un mal trabajo. ✗(1) Yo tenía casi asumido que los secuestradores iban a cobrar el dinero y que luego me iban a pegar un tiro. ✓(2) Dejé de hacerlo porque me ponía la cabeza como un bombo. ✓(3) Le dije que si la aumentaba al 100%, me iba al minuto.
Temporal	**Query:** He knows a tasty meal when he eats one. ✗(1) He knows where he can get a good dinner. ✓(2) He is an ass who talks when he might eat. ✓(3) He says he keeps a biscuit in his pocket to eat before going into a sick house.	**Query:** Él reconoce una sabrosa comida cuando come una. ✓(1) Le sale una voz aniñada cuando canta en español. ✗(2) Una chica le ofrecía una botella de agua para aliviarla. ✗(3) El hombre aparece cuando más le necesitan.
Comparison	**Query:** IBM's stock price rose, but the overall market fell. ✗(1) The stock markets of the world gambled upon its chances, and its bonds at one time were high in favor. ✓(2) Tommy's heart beat faster, but his casual pleasantness did not waver. ✓(3) And the discrepancy has greatly increased as young States have been added to the Union, while the old States have increased in population.	**Query:** El precio de las acciones de IBM subió, pero el mercado en general cayó. ✗(1) La capacidad de compra se desplomó y la economía mundial se derrumbó. ✗(2) La maniobra gustó a Wall Street, donde las acciones de la firma subieron un 13%. ✓(3) El problema de la lluvia ácida no desapareció, pero se redujo considerablemente.
Expansion	**Query:** I never gamble too far; in particular, I quit after one try. ✗(1) I never gamble, and therefore no other scrape can be awful. ✗(2) I much prefer to do some of the talking myself, but I seldom get a chance. ✓(3) I have to break off and betake myself to lighter employments; such as the biographies of great men.	**Query:** Nunca apuesto demasiado; en particular, renuncio después de un intento. ✗(1) No en vano, es conocido por su facilidad para llorar. ✓(2) Otros tenían más suerte y comentaban contentos que sus vacaciones acababan de comenzar. ✓(3) Es decir, nadie es culpable.

The first relation is *Contingency*, which indicates causal influences. We see that for both languages a sentence was retrieved each with the same relationship as the query (1,2), which is *Cause-reason* type using the *because* connector. Furthermore, sentence (3) retrieved in English shows the same relationship but using the *as* connector. Interestingly, sentence (3) in Spanish shows a *Condition* type relation as it uses the conditional connector *si* which is also part of the contingency category.

Temporal is the second relationship and indicates situations that are temporally related. As in the previous case, a sentence in each language shares the same relationships as the query (2,1), which is a relation of type *Synchronous* that uses the *when* connector. Sentence (3) in English shows an *Asynchronous* relationship type instead. Sentence (3) in Spanish has the connector *cuando* as the query; however, it shows a *Contingency* relationship.

The third relation is *Comparison*, used to highlight prominent differences between the two situations. For English, two retrieved sentences contain a *Contrast* type relationship. The (2), similar to the query, uses the *but* connector, while (3) uses the *while* connector. In the case of Spanish, one sentence (3) contains a *Contrastive* relationship with the connector *pero* as the query, while the others only share a semantic similarity.

Finally, the *Expansion* relation expands the discourse and moves its narrative or exposition forward. In the case of English, sentence (3) is the only one that has an *Expansion* relation of the *Instantiation* type but uses a *such as* connector instead of *in particular*. For Spanish, sentence (2) shows an *Expansion* relation using the *y* connector. Sentence (3) has the connector *es decir*, used for the type *Restatement*; however, the sentence is incomplete. This is because *Expansion* relations are more common between two sentences, and we focus mainly on single-sentence retrieval.

7. Discussion

The results show that our models improve the quality of sentence-level representations using a top-down network with a next-sentence modeling objective. Our findings complement and extend existing work on deep learning for computer vision that showed that image [36,45] and video [47] representations could be improved using the PC framework. The fact that the proposed method helps to improve the representations of BERT-style models supports the idea that the PC could be one of the central mechanisms of language learning in humans [16,25,26]. Furthermore, the consistent results in English and Spanish demonstrated that the mechanism is language-independent, which is consistent with work in psycholinguistics that demonstrated that monolingual and multilingual people show similar predictive processing [67]. Finally, the ablation and qualitative study demonstrate that the top-down connections are actually responsible for enhancing the feature space, which additionally allows the embeddings to be helpful for the retrieval processes.

Although improvements were obtained on some of the discourse and pragmatics benchmark tasks, the negative results elucidate that the method is insufficient to allow the model to acquire complete knowledge of discourse and pragmatics. Some of the limitations of this work open several opportunities for future work. Our next-sentence modeling objective only allows the model to capture relationships between sentences. A way to induce other skills, such as sentence ordering or EDU segmentations, to improve the models' discourse knowledge would be desirable. We hypothesized that the model's escalation in batch size and the number of prediction steps used during pre-training beyond the academic budget could generate other emerging abilities in the model [68]. Likewise, the efficacy of our transformer-based models could be ensured by using longer prediction steps, as transformer models have been shown to benefit from long sequences [52,69]. Furthermore, improving the InfoNCE loss function in terms of the number of negative samples could also help, for instance, by dynamically adjusting the negative sampling ratio [70]. Another important direction is to implement the method on the multilingual model, as in this work, we only demonstrate the consistency of the models for English and Spanish separately.

8. Conclusions

In this work, we explored improving sentence-level representations of BERT-style models by taking inspiration from the PC theory. Specifically, we use a BERT-style model as a bottom-up network to extend it with a top-down network which could be a recurrent network or a transformer. Also, we propose a next-sentence modeling objective to encourage the models to learn the relation between sentences. We perform extensive experimentation to evaluate the quality of the resulting sentence representations. Also, we explore the discourse and pragmatics-based knowledge acquired by our models.

Our models improved the sentence-level representations for stand-alone sentence and sentence pair tasks concerning their direct baselines. Also, our models demonstrate an improvement of discourse-aware sentence representations, mainly for discourse relation

and coherence detection tasks. Furthermore, our experiments on document-level discourse and pragmatics-based assessment showed that they elicited that kind of knowledge to some extent.

Finally, we perform an ablation study and qualitative experimentation. On the one hand, we found that the top-down connections of the PC mechanism are indeed helping to improve the model representations. However, having this mechanism at all the layers of the BERT-style model is not necessary to obtain improvements. On the other hand, we found that some clusters tend to be more organized by visualizing the representations in low dimensionality. Furthermore, the representations are good enough for retrieving sentences with discursive relationships given a sentence annotated as a query.

Author Contributions: Conceptualization, V.A.; methodology, V.A., M.-F.M. and A.S.; software, V.A.; validation, V.A.; formal analysis, V.A.; investigation, V.A.; data curation, V.A.; writing—original draft preparation, V.A.; writing—review and editing, V.A., M.-F.M. and A.S.; visualization, V.A.; supervision, M.-F.M. and A.S.; funding acquisition, V.A., M.-F.M. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Research Council Advanced Grant 788506, FONDECYT grant 1221425, and the National Center for Artificial Intelligence CENIA FB210017 - Basal ANID.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2013; Volume 26.
2. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. 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), New Orleans, LA, USA, 1–6 June 2018. [CrossRef]
3. Kiros, R.; Zhu, Y.; Salakhutdinov, R.R.; Zemel, R.; Urtasun, R.; Torralba, A.; Fidler, S. Skip-Thought Vectors. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Cambridge, MA, USA, 2015; Volume 28.
4. Hill, F.; Cho, K.; Korhonen, A. 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, San Diego, CA, USA, 12–17 June 2016. [CrossRef]
5. Pagliardini, M.; Gupta, P.; Jaggi, M. Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1–6 June 2018. [CrossRef]
6. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
7. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. OpenAI Blog. 2018. Available online: https://paperswithcode.com/paper/improving-language-understanding-by (accessed on 14 November 2022).
8. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. OpenAI Blog. 2019. Available online: https://paperswithcode.com/paper/language-models-are-unsupervised-multitask (accessed on 14 November 2022).
9. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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), Minneapolis, MN, USA, 2–7 June 2019. [CrossRef]
10. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32, pp. 5753–5763.

11. Reimers, N.; Gurevych, I. 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), Hong Kong, China, 3–7 November 2019. [CrossRef]
12. Huber, L.; Memmadi, C.; Dargnat, M.; Toussaint, Y. Do sentence embeddings capture discourse properties of sentences from Scientific Abstracts? In Proceedings of the First Workshop on Computational Approaches to Discourse, Online, 7–13 August 2020; Association for Computational Linguistics: Stroudsburg, PA, USA. [CrossRef]
13. Ettinger, A. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 34–48. [CrossRef]
14. Aspillaga, C.; Carvallo, A.; Araujo, V. Stress Test Evaluation of Transformer-based Models in Natural Language Understanding Tasks. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020.
15. Rogers, A.; Kovaleva, O.; Rumshisky, A. A Primer in BERTology: What We Know About How BERT Works. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 842–866. [CrossRef]
16. Rohde, H. Coherence-Driven Effects in Sentence and Discourse Processing. Ph.D. Thesis, University of California, San Diego, CA, USA, 2008.
17. Rao, R.P.N.; Ballard, D.H. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **1999**, *2*, 79–87. [CrossRef]
18. Friston, K. A theory of cortical responses. *Philos. Trans. R. Soc. B Biol. Sci.* **2005**, *360*, 815–836. [CrossRef] [PubMed]
19. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **2013**, *36*, 181–204. [CrossRef] [PubMed]
20. Hohwy, J. *The Predictive Mind*; Oxford University Press: Oxford, UK, 2013. [CrossRef]
21. Ylinen, S.; Bosseler, A.; Junttila, K.; Huotilainen, M. Predictive coding accelerates word recognition and learning in the early stages of language development. *Dev. Sci.* **2016**, *20*, e12472. [CrossRef] [PubMed]
22. Zettersten, M. Learning by predicting: How predictive processing informs language development. In *Patterns in Language and Linguistics*; Busse, B., Moehlig-Falke, R., Eds.; De Gruyter: Berlin, Germany; Munich, Germany; Boston, MA, USA, 2019; pp. 255–288. [CrossRef]
23. Berkum, J.J.A.V.; Brown, C.M.; Zwitserlood, P.; Kooijman, V.; Hagoort, P. Anticipating Upcoming Words in Discourse: Evidence From ERPs and Reading Times. *J. Exp. Psychol. Learn. Mem. Cogn.* **2005**, *31*, 443–467. [CrossRef]
24. Kuperberg, G.R.; Jaeger, T.F. What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* **2015**, *31*, 32–59. [CrossRef]
25. Kandylaki, K.D.; Nagels, A.; Tune, S.; Kircher, T.; Wiese, R.; Schlesewsky, M.; Bornkessel-Schlesewsky, I. Predicting "When" in Discourse Engages the Human Dorsal Auditory Stream: An fMRI Study Using Naturalistic Stories. *J. Neurosci.* **2016**, *36*, 12180–12191. [CrossRef]
26. Pickering, M.J.; Gambi, C. Predicting while comprehending language: A theory and review. *Psychol. Bull.* **2018**, *144*, 1002–1044. [CrossRef]
27. Araujo, V.; Villa, A.; Mendoza, M.; Moens, M.F.; Soto, A. Augmenting BERT-style Models with Predictive Coding to Improve Discourse-level Representations. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online, 7–11 November 2021; Association for Computational Linguistics: Punta Cana, Dominican Republic, 2021. [CrossRef]
28. von Helmholtz, H. *Treatise on Physiological Optics Vol. III*; Dover Publications: Menasha, WI, USA, 1867.
29. Casillas, M.; Frank, M. The development of predictive processes in children's discourse understanding. In Proceedings of the Annual Meeting of the Cognitive Science Society, Austin, TX, USA, 2013; Volume 35. Available online: https://www.mpi.nl/publications/item1796081/development-predictive-processes-childrens-discourse-understanding (accessed on 14 November 2022).
30. Shain, C.; Blank, I.A.; van Schijndel, M.; Schuler, W.; Fedorenko, E. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia* **2020**, *138*, 107307. [CrossRef] [PubMed]
31. Ferreira, F.; Qiu, Z. Predicting syntactic structure. *Brain Res.* **2021**, *1770*, 147632. [CrossRef] [PubMed]
32. Wen, H.; Han, K.; Shi, J.; Zhang, Y.; Culurciello, E.; Liu, Z. Deep Predictive Coding Network for Object Recognition. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Dy, J., Krause, A., Eds.; Volume 80, pp. 5266–5275.
33. Han, K.; Wen, H.; Zhang, Y.; Fu, D.; Culurciello, E.; Liu, Z. Deep Predictive Coding Network with Local Recurrent Processing for Object Recognition. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018; Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31.
34. Huang, Y.; Gornet, J.; Dai, S.; Yu, Z.; Nguyen, T.; Tsao, D.; Anandkumar, A. Neural Networks with Recurrent Generative Feedback. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 535–545.
35. Lotter, W.; Kreiman, G.; Cox, D.D. Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
36. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.

37. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
38. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
39. Cer, D.; Yang, Y.; Kong, S.Y.; Hua, N.; Limtiaco, N.; John, R.S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; et al. Universal Sentence Encoder. *arXiv* **2018**, arXiv:1803.11175.
40. Iter, D.; Guu, K.; Lansing, L.; Jurafsky, D. Pretraining with Contrastive Sentence Objectives Improves Discourse Performance of Language Models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020. [CrossRef]
41. Lee, H.; Hudson, D.A.; Lee, K.; Manning, C.D. SLM: Learning a Discourse Language Representation with Sentence Unshuffling. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 7–13 August 2020. [CrossRef]
42. Whittington, J.C.R.; Bogacz, R. An Approximation of the Error Backpropagation Algorithm in a Predictive Coding Network with Local Hebbian Synaptic Plasticity. *Neural Comput.* **2017**, *29*, 1229–1262. [CrossRef]
43. Millidge, B.; Tschantz, A.; Buckley, C.L. Predictive Coding Approximates Backprop Along Arbitrary Computation Graphs. *Neural Comput.* **2022**, *34*, 1329–1368. [CrossRef] [PubMed]
44. Salvatori, T.; Song, Y.; Xu, Z.; Lukasiewicz, T.; Bogacz, R. Reverse Differentiation via Predictive Coding. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 8150–8158. [CrossRef]
45. Dora, S.; Pennartz, C.; Bohte, S. A Deep Predictive Coding Network for Learning Latent Representations. *bioRxiv* **2018**. [CrossRef]
46. Jati, A.; Georgiou, P. Neural Predictive Coding Using Convolutional Neural Networks Toward Unsupervised Learning of Speaker Characteristics. *IEEE/ACM Trans. Audio Speech Lang. Proc.* **2019**, *27*, 1577–1589. [CrossRef]
47. Han, T.; Xie, W.; Zisserman, A. Video Representation Learning by Dense Predictive Coding. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Seoul, Korea, 27 October–2 November 2019.
48. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Cambridge, MA, USA, 2014; Volume 27.
49. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 26–28 October 2014. [CrossRef]
50. Cañete, J.; Chaperon, G.; Fuentes, R.; Ho, J.H.; Kang, H.; Pérez, J. Spanish Pre-Trained BERT Model and Evaluation Data. In Proceedings of the PML4DC at ICLR 2020, Addis Ababa, Ethiopia, 26 April–1 May 2020.
51. Cañete, J.; Donoso, S.; Bravo-Marquez, F.; Carvallo, A.; Araujo, V. ALBETO and DistilBETO: Lightweight Spanish Language Models. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022.
52. Press, O.; Smith, N.A.; Lewis, M. Shortformer: Better Language Modeling using Shorter Inputs. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Bangkok, Thailand, 1–6 August 2021. [CrossRef]
53. Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 19–27. [CrossRef]
54. Conneau, A.; Kiela, D. SentEval: An Evaluation Toolkit for Universal Sentence Representations. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan, 7–12 May 2018.
55. Chen, M.; Chu, Z.; Gimpel, K. Evaluation Benchmarks and Learning Criteria for Discourse-Aware Sentence Representations. 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), Hong Kong, China, 3–7 November 2019. [CrossRef]
56. Araujo, V.; Carvallo, A.; Kundu, S.; Cañete, J.; Mendoza, M.; Mercer, R.E.; Bravo-Marquez, F.; Moens, M.F.; Soto, A. Evaluation Benchmarks for Spanish Sentence Representations. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022.
57. Mann, W.C.; Thompson, S.A. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-Interdiscip. J. Study Discourse* **1988**, *8*, 243–281. [CrossRef]
58. Prasad, R.; Dinesh, N.; Lee, A.; Miltsakaki, E.; Robaldo, L.; Joshi, A.; Webber, B. The Penn Discourse TreeBank 2.0. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 28–30 May 2008.
59. Koto, F.; Lau, J.H.; Baldwin, T. Discourse Probing of Pretrained Language Models. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021. [CrossRef]
60. Kovaleva, O.; Romanov, A.; Rogers, A.; Rumshisky, A. Revealing the Dark Secrets of BERT. 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), Hong Kong, China, 3–7 November 2019. [CrossRef]

61. Sileo, D.; Muller, P.; Van de Cruys, T.; Pradel, C. A Pragmatics-Centered Evaluation Framework for Natural Language Understanding. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022.
62. Liu, N.F.; Gardner, M.; Belinkov, Y.; Peters, M.E.; Smith, N.A. 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), Minneapolis, MN, USA, 2–7 June 2019. [CrossRef]
63. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
64. Voorhees, E.M.; Tice, D.M. Building a Question Answering Test Collection. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, 24–28 July 2000; Association for Computing Machinery: New York, NY, USA, 2000; pp. 200–207. [CrossRef]
65. García Cumbreras, M.Á.; Ureña López, L.A.; Martínez Santiago, F. BRUJA: Question Classification for Spanish. Using Machine Translationand an English Classifier. In Proceedings of the Workshop on Multilingual Question Answering, Trento, Italy, 3–7 April 2006.
66. Scialom, T.; Dray, P.A.; Lamprier, S.; Piwowarski, B.; Staiano, J. MLSUM: The Multilingual Summarization Corpus. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020. [CrossRef]
67. Contemori, C.; Dussias, P.E. Prediction at the Discourse Level in Spanish–English Bilinguals: An Eye-Tracking Study. *Front. Psychol.* **2019**, *10*, 956. [CrossRef] [PubMed]
68. Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent Abilities of Large Language Models. *arXiv* **2022**, arXiv:2206.07682.
69. Popel, M.; Bojar, O. Training tips for the transformer model. *arXiv* **2018**, arXiv:1804.00247. [CrossRef]
70. Wu, C.; Wu, F.; Huang, Y. Rethinking InfoNCE: How Many Negative Samples Do You Need? In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, Vienna, Austria, 23–29 July 2022; Raedt, L.D., Ed.; International Joint Conferences on Artificial Intelligence Organization: Vienna, Austria, 2022; pp. 2509–2515. [CrossRef]