No clues, good clues: Out of context Lexical Relation Classification

Lucía Pitarch[1], **Jorge Bernad**[1], **Licri Dranca**[2], **Carlos Bobed Lisbona**[1], **and Jorge Gracia**[1]

[1]University of Zaragoza, Zaragoza, Spain
[2] Centro Universitario de la Defensa (CUD), Zaragoza, Spain
{lpitarch,jbernad,licri,cbobed,jogracia}@unizar.es

Abstract

The accurate prediction of lexical relations between words is a challenging task in Natural Language Processing (NLP). The most recent advances in this direction come with the use of pre-trained language models (PTLMs). A PTLM typically needs “well-formed" verbalized text to interact with it, either to fine-tune it or to exploit it. However, there are indications that commonly used PTLMs already encode enough linguistic knowledge to allow the use of minimal (or none) textual context for some linguistically motivated tasks, thus notably reducing human effort, the need for data pre-processing, and favoring techniques that are language neutral since do not rely on syntactic structures.

In this work, we explore this idea for the tasks of lexical relation classification (LRC) and graded Lexical Entailment (LE). After fine-tuning PTLMs for LRC with different verbalizations, our evaluation results show that very simple prompts are competitive for LRC and significantly outperform graded LE SoTA. In order to gain a better insight into this phenomenon, we perform a number of quantitative statistical analyses on the results, as well as a qualitative visual exploration based on embedding projections.

1 Introduction

Lexical Relation Classification (LRC) is the task of predicting which lexical relation exists between two given words (e.g., ‘tall’ and ‘small’ are related by the *antonymy* relation), from a finite catalogue of lexical relations. Discovering lexico-semantic relations between words has received attention in the NLP community since Hearst’s seminal research in 1992 on the automatic acquisition of hyponyms from large text corpora based on pre-designed patterns (Hearst, 1992). Despite many recent advancements, LRC continues to be an open research topic in the NLP field (Wang et al., 2021; Ushio et al., 2021). Applications of the task are numerous: automatic thesauri creation, paraphrasing, textual entailment, sentiment analysis, ontology learning, and ontology population, among others (Weeds et al., 2014; Cimiano, 2006).

The most recent advances in LRC come with the use of pre-trained language models (PTLMs) based on the *transformers* architecture (Vaswani et al., 2017), which have been proven to capture a large amount of lexico-semantic knowledge from text successfully. One of the main benefits of the adoption of PLTMs is that, while they were trained for a general task (text generation) following a masked language model (MLM) objective in an unsupervised way, they can be easily adapted to different downstream tasks (e.g., text classification, text summarization, sentiment analysis) by introducing additional parameters and *fine-tuning* them using objective functions specific to the task. That avoids the need to train the model from scratch, still obtaining SoTA results, while decreasing computational costs and the need for very large amounts of data (Devlin et al., 2019).

More recently, the “pre-train, fine-tune” procedure is shifting in NLP tasks towards the “pre-train, prompt, and predict” paradigm (Liu et al., 2023). In that case, instead of adapting PTLMs to the downstream task via fine-tuning, the task is reformulated to look more like those solved during the original model training with the help of a textual *prompt*. Following the example in (Liu et al., 2023), when recognizing the emotion of a sentence, “I missed the bus today.”, we may continue with a prompt “I felt very”, and ask the PTLM to fill the blank with an emotion-bearing word.

A PTLM typically needs “well-formed" verbalized text to interact with it, either to fine-tune it or to exploit it via prompt engineering. While some authors claim that longer, more complex verbalizations of the input data work best for real-world text classification tasks (Schick and Schütze, 2022), or

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5607–5625
July 9-14, 2023 ©2023 Association for Computational Linguistics

relation classification (Bouraoui et al., 2020), other authors (LoganIV et al., 2022) have collected indications in the opposite direction for a wide range of NLP tasks (such as paraphrasing, textual similarity, or sentiment analysis).

We share the hypothesis that commonly used PTLMs already encode enough linguistic knowledge to allow the use of minimal (or none) textual context for some linguistically motivated tasks. In such cases, very simple prompts work almost as well or even better than hand-crafted, more complex verbalizations. Reducing the need of complex prompting notably reduces the need of human effort and the need for data pre-processing, and favors techniques that are language neutral since they do not rely on syntactic structures.

In this work[1], we explore this idea for the LRC task, and we extend it to *graded lexical entailment (LE)*, i.e., discovering the strength of the taxonomical asymmetric hyponymy–hypernymy relation between two words (Vulić et al., 2017). In previous works, other authors have explored complex verbalizations for LRC (Ushio et al., 2021) while others have essayed shorter ones (Wachowiak et al., 2020). However, there has been no systematic study on the impact of long/short prompting for LRC so far. To that end, we have experimented with different verbalizations of the training and test data in an LRC experiment. Then, we analysed which verbalization produces better predictions for at least one of the lexico-semantic relations entailed between a pair of words. We experiment with widely used benchmarks for LRC namely, CogALexV (Santus et al., 2016a), BLESS (Baroni and Lenci, 2011), EVALution (Santus et al., 2015), K&H+N (Necsulescu et al., 2015), and ROOT9 (Santus et al., 2016b). Besides, we evaluate such models with the Hyperlex (Vulić et al., 2017) dataset for *graded LE*.

Our main contributions are:

1. We show empirically that SoTA results for LRC can be reached by providing very simple verbalizations of the data or even no verbalization at all (null prompting) when fine-tuning and testing a PTLM.

2. We test the generalizability of such models trained with minimal prompting to similar tasks by testing them in graded LE, where they outperform SoTA results.

3. We provide an extensive analysis of the results (including error analysis) to further observe the strengths and limitations of minimal prompting for LRC.

4. To further understand the models' behaviour, we add a qualitative analysis of their learning process based on the visualisation of the embeddings that are built in their different layers.

Our paper is structured as follows: first, in Section 3, we formally describe both the LRC task and the LE task. Secondly, in Section 4, we describe the chosen templates for the input verbalizations, the used datasets and baselines we compare with, as well as the hyper parameter and fine-tuning setting of our models. Then, in Section 5, we analyze our results showing: a) our quantitative results, analyzing which template, model, and method work best on each dataset, b) the error analysis, checking how the distribution and linguistic characteristics of the different datasets affected the performance of our models and what examples and categories were the most difficult ones, and c) a visualization of the embedding projection, highlighting which layers are more informative for relation classification and how the model learns them through the different epochs. Finally, in Section 6, we summarize the conclusions and possible future work, stating the limitations of our work.

2 Related Work

In this section we give an overview of some related approaches that are relevant to our work.

2.1 Prompt-based Learning

In their extensive review, Liu et al. (2023) have analyzed the *prompt-based learning* paradigm, exploring different verbalization techniques used to input text to PTLMs, as a key point to reach SoTA results in few and zero-shot learning scenarios. The currently under research question is: *what kind of verbalizations work better?* Here, two different trends arise: a) automatically searched prompts (Shin et al., 2020; Liu et al., 2022; Li and Liang) and b) handcrafted prompts (Schick and Schütze, 2021, 2022). The main drawback of the first one is the necessity of additional training and computational resources to find the best prompt, and the second's major issue is the necessity of manual effort (LoganIV et al., 2022; Mahabadi et al., 2022). A third option is however possible: *null prompts* (LoganIV

[1] The code is available at: `https://github.com/sid-unizar/LRC`

et al., 2022) where the mask token is simply added to the input sentence.

Currently, no consensus has been reached on which kind of verbalizations work best, and, while authors such as Schick and Schütze (2022) obtain the best results in a variety of NLP tasks with hand-crafted verbalizations, others (LoganIV et al., 2022; Mahabadi et al., 2022) defend the advantages of short or even null prompts while still achieving competitive results. Liu et al. (2022) found different behavior for their Ptuning-v2 method depending on the task: simple classification tasks prefer shorter prompts, while hard sequence labeling tasks prefer longer ones.

Other open questions about prompting rely on the selection of the label to verbalize the mask and the order in which the mask and input are provided. Labels given in benchmark datasets are often multi-word or rare expressions consisting of more than one token, however, the mask needs to be filled by just one token (Schick and Schütze, 2022) thus there is a need to select the label either automatically or manually. The order in which input and mask are entered is also under current research (Mahabadi et al., 2022).

Previous comparisons of different prompting techniques have been mostly applied to highly context-dependent NLP tasks such as sentiment analysis, subjectivity, classification, question classification, natural language inference, question answering, word sense disambiguation or paraphrasing (LoganIV et al., 2022; Schick and Schütze, 2022; Mahabadi et al., 2022) were the input example already consists of a well-formed sentence. Yet, other NLP tasks that are less context-sensitive such as LRC, Relation Extraction, or Lexical Entailment, have received little or no attention so far in prompt comparison studies.

2.2 Lexical Relation Classification

Seminal work on LRC started exploring pattern-based techniques (Hearst, 1992), where a set of patterns that elicit the relation entailed between a pair of words is defined. A drawback of this method is that not all lexical relations are explicit in texts by a closed set of patterns. Then, the approach towards LRC shifted to distributional semantics with static embeddings, meaning one vector is given to represent each word in the embeddings space (Weeds et al., 2014; Santus et al., 2016a; Shwartz et al., 2016; Wang et al., 2019; Shwartz and Dagan, 2016). Such techniques were found beneficial to LRC tasks, in which words were normally provided without additional context (Barkan et al., 2020).

Recent work in LRC has focused on PTLMs and their dynamic embeddings, owing to their capacity to better capture polysemy than static embeddings, which led to better results (Karmakar and McCrae, 2020; Ushio et al., 2021; Wang et al., 2021). Such works have already used prompting to fine-tune PTLMs. However, none of them has focused on analyzing what kind of verbalization can be better used to extract relation information, as we do. For instance, while in (Ushio et al., 2021) the authors opted to use hand-crafted complex verbalizations motivated by previous research (Bouraoui et al., 2020; Jiang et al., 2020), Wachowiak et al. (2020) used minimal prompts, and in (Karmakar and McCrae, 2020) null prompting was used.

The focus of our work is comparing the verbalizations enumerated by Schick and Schütze (2022) in their work: *null-prompting, null-prompting with punctuation, short templates and long templates* and see how they interact with a lexical-focused task when some artificial context (i.e., not initially available in the dataset) is added to the prompt, versus when no context other than two words is provided (as in null prompting).

3 Problem Statement

Let $V = \{w_1, \dots, w_n\}$ be a set of words (our *vocabulary*), and a *sentence* s be any finite sequence of words from V. The set of all sentences over V is denoted by $\mathcal{S}$. Given a word $w \in V$, a *context* c of w is any sentence such that $w \in c$. The set of all contexts of a word w is denoted by $\mathcal{C}_w$.

A binary relation r between words is a subset of $V \times V$. Let us denote by $\mathcal{R}$ the set of all binary relations over the vocabulary V, that is, $\mathcal{R}$ is the power set of $V \times V$. We say that a set of relations, $R = \{r_1, \dots, r_k\}$, where $r_i \in \mathcal{R}$, is *mutually exclusive* if the relations in R are disjoint; and we say that R is *complete* if the union of the relations is equal to $V \times V$. Note that we can make a relation set R complete by adding a relation named unknown, which is the complementary of all the relations in R.

We consider that any context of two words induces a relation from a predefined set of relations, that is, there exists a function $f^R\colon \mathcal{P} \to R$, where $\mathcal{P} = \{c \in \mathcal{S} \mid c \in C_{w_1} \cap C_{w_2}, w_1, w_2 \in V\}$. For instance, given the set of relations $R = \{\texttt{partOf}, \texttt{unknown}\}$, the common context for the words bank and river, "I play by the bank of the river", induces the relation partOf, while "I will deposit the money in the bank beside the river" would induce the unknown relation. Thus, Relation Classification (RC) is the task of using a function $\hat{f}^R$ that estimates f^R.

Lexical Relation Classification (LRC) is a subtype of RC where the relation between words is a lexical one. The most usual and important lexical relations are hyponymy, hyperonymy, antonymy, synonymy, and meronymy. Among these relations, hyponymy and, its counterpart, hyperonymy are especially important in NLP and ontology engineering.

Finally, Lexical Entailment (LE) is the task of detecting the hyponymy relationship between two words. This task becomes *graded LE* when we have to calculate the numerical degree to which a word w_1 is a type of w_2, becoming a more challenging regression task.

4 Experimental Setup

The main goals of our experiments are: 1) to check if LRC can be conducted without adding artificial context when just a pair of words out of context is given, 2) if so, to analyze which verbalization works best for model fine-tuning, and 3) to check the generalizability of our model to other language-related tasks such as graded LE.

4.1 Chosen Verbalization

Similarly to (Schick and Schütze, 2022), we compare null prompts to punctuated ones (just the target and source words with added punctuation), and a longer template (the best performing one in (Ushio et al., 2021)). The chosen mask order and wording placement in the verbalization is the best performing one in (Mahabadi et al., 2022), inserting the mask token between both words. Table 1 presents our chosen prompts.

We explore two different options: a) adopting a sentence classification scheme, where a classification layer is added on top of the output layer (templates T1, T2, T3, and T4) to classify the CLS(special classification token) that is added at the beginning of every template, and b) instantiating the task as a fill in the blank task (templates TM1, TM2, and TM3). We use T4 as a control case to check what happens when train and test templates are different.

Template	**Id**
' W1 ' SEP ' W2 '	T1
␣W1 SEP W2	T2
Today, I finally discovered the relation between W1 and W2.	T3
Train: Today, I finally discovered the relation between W1 and W2: W1 is the LABEL of W2. **Test**: Today, I finally discovered the relation between W1 and W2.	T4
' W1 ' MASK ' W2 '	TM1
␣W1 MASK W2	TM2
Today, I finally discovered the relation between W1 and W2: W1 is the MASK of W2.	TM3

Table 1: Templates used in the experiments. Except for T4, both training and test use the same template. SEP (separator), MASK, and LABEL are substituted by special tokens, see Appendix C.

4.2 Datasets and Baselines

LRC We conducted experiments on five datasets[2]: CogALexV (Santus et al., 2016a), BLESS (Baroni and Lenci, 2011), EVALution (Santus et al., 2015), K&H+N (Necsulescu et al., 2015), and ROOT9 (Santus et al., 2016b). These datasets contain a variety of lexical relations, including hypernyms, meronyms, synonyms, antonyms, and random (equivalent to unknown relation defined in S3)[3]. For a deeper analysis (error analysis and visualization), we focus on CogALexV as it contains a subset of the most complicated examples of EVALution. To compare the performance of the different verbalizations in PTLM fine-tuning to SoTA methods, we selected the following baseline models: LexNet (Shwartz and Dagan, 2016), SphereRE (Wang et al., 2019), KEML (Wang et al., 2021), and RelBERT (Ushio et al., 2021).

Graded LE We use Hyperlex dataset (Vulić et al., 2017), which consists of 2616 pairs of words (2163 nouns and 453 verbs). Each pair was presented to at least ten human annotators to answer the question *To what degree X is a type of Y?* rang-

[2] All datasets are open source, covered by either Creative Commons 4.0 or Apache 2.0 Licences

[3] In Appendix A, a further description of the datasets, their distribution, and linguistic properties is provided

ing from 0 to 6. The final given score for each pair is the median of the human annotations. The authors of Hyperlex provide an *upper bound* of the Inter-Annotator Agreement (IAA) calculated as the average Spearman correlation of a human rater with the average of all the other raters; in particular, the annotation reaches an IAA-ρ of 0.864 (for nouns, IAA-$\rho = 0.864$, and for verbs, IAA-$\rho = 0.862$). To train supervised systems, Hyperlex is split into train/val/test datasets in two configurations: a) random split: data are randomly split into 1831/130/655 train/val/test pairs, respectively (all the words in the test split appear in the train/val splits); b) lexical split: to avoid lexical memorization, words in the test split are forced not to appear in the train/val splits, leading to fewer pairs in each split, 1133/85/269, respectively. To compare our proposal, we have considered the following SoTA models as baselines: LEAR (Vulić and Mrkšić, 2018), SDNS (Rei et al., 2018), GLEN (Glavaš and Vulić, 2019), POSTLE (Kamath et al., 2019), LexSub (Arora et al., 2020) and Hierarchy-fitting (HF) (Yang et al., 2022). Note that all these models use non-contextual embeddings; however, as far as our knowledge, there are no models in the literature that use contextual embeddings for graded LE as we do.

4.3 Fine-tuning Setting

We begin by briefly describing the models we use, continue by explaining how the models are fine-tuned for LRC and graded LE, and how the fine-tuned models are used for inference, and conclude the section by describing the hyperparameter setup.

Chosen PTLMs In this work, we chose to use RoBERTa and BERT, both recognized as SoTA models for general domains and tasks in English. In particular, we use both their base and large versions that can be downloaded using the Huggingface transformers library (Wolf et al., 2020)[4]. Moreover, we use the appropriate version depending on the actual underlying task we are fine-tuning, whether it is sequence classification (T1-4) or fill-in-the-mask (TM1-3). Finally, note that BERT and RoBERTa have different-sized vocabularies and treat white spaces differently; thus, we must bear in mind these differences to adapt the templates and prompts for each model.

LRC Our setup for fine-tuning a model has four components: 1) a PTLM M and its token vocabulary V_M; 2) a training set $\mathcal{T} = \{(\boldsymbol{w}_i, y_i) \mid i = 1, \dots n\}$, where $\boldsymbol{w}_i = (w_i^1, w_i^2)$ is a pair of words and $y_i \in Y$ is the label of a lexical relation ($|Y| = K$); 3) an injective function from the set of labels to the vocabulary of tokens V_M, $v\colon Y \to V_M$, called the *mask verbalizer* function; and 4) a training and a testing template, T_t and T_e, used to verbalize $\boldsymbol{w}_i$. In this context, a template T is a function, $T\colon V \times V \to \mathcal{S}$, from pairs of the word vocabulary to the set of sentences where the `CLS`, `SEP` and `MASK` special tokens of the PTLM can appear in the sentence. We denote by $T(\boldsymbol{w})_C$ and $T(\boldsymbol{w})_M$ to the `CLS` and `MASK` tokens in the sentence $T(\boldsymbol{w})$, respectively.

Depending on the template used, we adopt one of the following two training objectives: (T1-4) a classification objective to estimate the probability $P(Y = y_j | T_t(\boldsymbol{w}_i)_C)$; and (TM1-3) a mask prediction objective to estimate $P(T_t(\boldsymbol{w}_i)_M = t_j | T_t(\boldsymbol{w}_i))$, where $t_j \in V_M$ is any token in the vocabulary of the PTLM. At inference time, for a model trained with a classification objective, we use the testing template T_e to predict the label with $argmax_{y_i \in Y}\{P(Y = y_i | T_e(\boldsymbol{w})_C)\}$, and for the mask objective, $argmax_{y_i \in Y}\{P(T_e(\boldsymbol{w})_M = v(y_j) | T_e(\boldsymbol{w}))\}$. For this latter case, note that at inference time, we only use the tokens given by the mask verbalizer function v.

Graded LE In this task, we have a similar setup to the LRC one, but the training set tuples are extended with the hyponymy score for the pair of words, $s_i \in \mathbb{R}$; thus, $\mathcal{T} = \{(\boldsymbol{w}_i, s_i, y_i)\}$. We first fine-tune a model M using only the labels y_i as for the LRC task. The model M produces a logit, $l_i^j \in \mathbb{R}$ for each pair $\boldsymbol{w}_i \in \mathcal{T}$ and label y_j (token $v(y_j)$) for a model fine-tuned with a classification (masked) objective. Let us denote by $M(\boldsymbol{w}_i) = (l_i^1, \dots, l_i^K)$ the logit vector produced by the model and by $A = [M(\boldsymbol{w}_i)] \in \mathbb{R}^{n \times K}$ the matrix of logits. Then, a linear regression model is fitted to predict the scores in the training set $\{s_i \mid i = 1, \dots n\}$ with the logits A. We obtain K regression coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)$. For an unseen pair $\boldsymbol{w}$, the predicted score is the linear combination of the fitted regression coefficients and the logits produced by the model M, that is, the scalar product $score(\boldsymbol{w}) = \boldsymbol{\beta} \cdot M(\boldsymbol{w})$.

[4] Both models are open source with Apache 2.0 and MIT licenses

Hyperparameters and Fine-tuning Setup Training and evaluation were performed on a Tesla-T4 GPU through Google Colab. Overall we consumed around 850h of GPU usage. To fine-tune the models, we used the following hyperparameters: batch size of 32, Adam weight optimizer, learning rate of $2e^{-5}$, weight decay of 0.01, no warmup, 10 epochs, and 5 runs of training and evaluation to asses model's performance variability. We use the train, validation, and test splits provided by the original datasets, and, when no validation split was provided, we did not use any. We report the F1-score weighted by the support of the labels to compare ourselves with the other baselines. In the case of CogALexV, we take out the results for RANDOM before reporting the results as advised by its authors in (Santus et al., 2016a). For graded LE and Hyperlex dataset, the Spearman correlation between the median human annotators scores and our proposed score is reported. We also report the Spearman correlation restricted to nouns and verbs.

	K&H+N	BLESS	EVAL	ROOT9
RoBERTa				
T1	0.989	**0.954**	**0.764**	**0.936**
T2	0.989	**0.955**	**0.757**	**0.936**
T3	0.989	$\underline{\mathbf{0.956}}$	$\underline{\mathbf{0.771}}$	$\underline{\mathbf{0.937}}$
T4	0.312	0.133	0.087	**0.934**
TM1	0.988	0.947	**0.761**	**0.936**
TM2	0.988	0.946	**0.764**	**0.928**
TM3	0.985	**0.951**	**0.746**	**0.926**
RoBERTa base				
T1	0.983	0.949	0.745	0.931
T2	0.988	0.947	0.744	0.931
T3	0.987	0.949	**0.754**	**0.933**
T4	0.299	0.043	0.023	0.923
TM1	0.986	0.940	**0.747**	**0.926**
TM2	0.983	**0.944**	0.727	0.925
TM3	0.986	0.944	**0.729**	0.924
SoTA				
LexNET	0.985	0.893	0.600	0.813
KEML	$\underline{\mathbf{0.993}}$	0.944	0.660	0.878
SphereRE	0.990	0.938	0.620	0.861
RelBERT	0.949	0.921	0.701	0.910

Table 2: Results for K&N+N, BLESS, EVALution and ROOT9 datasets in terms of the weighted F1-score by the support of the labels.

5 Results

In this section, we report the qualitative and quantitative results of our experiments.

5.1 Quantitative Results

LRC Results We report our results[5] in Tables 2 and 3, comparing them to the SoTA[6] results. We report the mean value of the 5 runs for each measure, underlining the highest value achieved for each dataset (column-wise). Boldened numbers mark no statistical significance (at confident level $\alpha = 0.01$) to be different from the greatest mean value applying Welch's t-test. Except for KHN, we improve the F1-score in all the datasets. In some of them (EVALution and CogALexV), we outperform the baselines by almost 10 points. We hypothesize that not biasing the model by adding external artificial context might let it choose the best sense of both words. Coincidentally with (Schick and Schütze, 2022), the longer hand-crafted template (T3) obtained the best results in most datasets. However, the difference with simpler templates (T1, T2), was very small and statistically not significant in most cases. T4 reported the worst performance due to the differences between train and test which misguided the model's learning. We must point out that masked variants exhibited more stability when small models, small prompts, and small datasets are jointly used, as, in some instances with this setting, T1 and T2 did not manage to converge, entering a poor minimal local. Such situations were solved by relaunching the training.

Graded LE results The results for graded LE are shown in Table 4. We can see how models trained with a mask objective (TM1-TM3) obtain the best results, and improve the SoTA results by more than 10 points globally (*all*) and focusing only on noun pairs (*nouns*). In particular, in the lexical split, our results are about 20 points above previous proposals. Note as well that the difference of the results in the lexical split is only about 4 points less than in the random split, which is a good indicator of the generalization capabilities of our models. To the best of our knowledge, previous studies reported results just on all POS together, and some focused on nouns as well. We expand this research to verbs considering the results promising

[5] Among both models, we report here the best performing one, RoBERTa; we present the complete results table in Appendix D including BERT as well.

[6] As reported in their original papers.

	ant	hyp	part	syn	all
RoBERTa					
T1	**0.873**	**0.703**	**0.752**	**0.604**	**0.743**
T2	**0.863**	**0.682**	**0.745**	0.584	**0.728**
T3	**0.884**	**0.718**	**0.784**	**0.629**	**0.762**
T4	0.237	0.004	0.165	0.085	0.119
TM1	**0.880**	**0.709**	**0.773**	**0.599**	**0.750**
TM2	**0.871**	**0.723**	**0.787**	**0.621**	**0.758**
TM3	**0.871**	**0.718**	**0.787**	**0.616**	**0.756**
RoBERTa base					
T1	0.806	0.677	0.732	0.570	0.704
T2	0.783	0.652	0.693	0.536	0.675
T3	0.820	0.676	0.731	0.577	0.709
T4	0.027	0.000	0.102	0.092	0.044
TM1	0.809	0.678	0.743	0.561	0.706
TM2	0.801	0.673	0.742	0.556	0.701
TM3	0.815	0.679	0.730	0.561	0.705
SoTA					
LexNET	0.425	0.526	0.493	0.297	0.445
SphereRE	0.479	0.538	0.539	0.286	0.471
KEML	0.492	0.547	0.652	0.292	0.500
RelBert	0.794	0.616	0.702	0.505	0.664

Table 3: Results for CogALexV dataset.

	random	lexical
RoBERTa	Spearman ρ for all/noun/verb	
T1	**0.741/0.753/0.584**	**0.755/0.788/0.532**
T2	0.152/0.170/0.030	0.287/0.350/0.063
T3	0.774/0.790/0.631	**0.669/0.690/0.516**
TM1	**0.828/0.839/0.716**	**0.789/0.837/0.612**
TM2	**0.749/0.761/0.646**	**0.654/0.705/0.417**
TM3	**0.814/0.830/0.683**	**0.794/0.828/0.656**
RoBERTa base		
T1	0.737/0.749/0.594	0.677/0.713/0.543
T2	0.652/0.683/0.377	0.407/0.483/0.167
T3	0.742/0.757/0.637	0.626/0.693/0.391
TM1	0.796/0.811/0.639	0.736/**0.800/0.553**
TM2	0.781/0.793/0.664	0.711/0.757/0.525
TM3	0.783/0.795/0.635	**0.757**/0.807/0.634
SoTA		
LEAR	0.686/0.710/ -----	0.174/ ----- / -----
SDNS	0.692/ ----- / -----	----- / ----- / -----
GLEN	0.520/ ----- / -----	0.481/ ----- / -----
POSTLE	0.686/ ----- / -----	----- /0.600/ -----
LexSub	0.533/ ----- / -----	----- / ----- / -----
HF	0.690/ ----- / -----	----- / ----- / -----
IAA	0.864/0.864/0.862	

Table 4: Results for Hyperlex dataset. The Spearman ρ correlations for all/noun/verb are reported.

as, even if they are lower than for nouns, they show that the part of speech has influence in our models. Finally, we want to remark that our models push up the results for nouns near to the IAA given by humans (0.837 vs. 0.864).

5.2 LRC Error Analysis

Results obtained for EVALution and CogALexV datasets are noticeably lower. We hypothesize a reason for this is that EVALution is an extended version of BLESS dataset where the relations of synonyms and antonyms were added. Adding such relations makes the task of LRC more challenging as, particularly, synonyms are a very heterogeneous class difficult to be delimited even for humans. CogALexV becomes even more challenging as it consists of a selected subset of EVALution, where words were stemmed, decreasing possible morpho-semantic cues. Moreover, both EVALution and CogALexV were created to avoid lexical memorization, this meaning, they consistently use words that participate in various relations. Finally, the bigger dataset size of BLESS, ROOT09, and K&H+N should also have a beneficial impact on the results.

From now on, we focus our error analysis on EVALution and CogALexV as they contain the most challenging examples[7]. Unknown (or equivalently Random) relations and models trained with the T4 control template have been excluded from this analysis. We focused this analysis on the best-performing model in our experiments, Roberta-large, and we got two groups of word pairs, those which were well and wrongly classified with all templates. For these two groups, we analyzed different features (presented below), checking whether there was a statistically significant difference between the two groups by using χ^2-tests or Welch's t-tests. We considered that a feature had a significant impact when the p-value was below 0.05.

Relationship Type We observed that, in both datasets, all the trained models struggled correctly classifying synonyms, while they are particularly good at predicting antonyms. In comparison to previous studies with static embeddings (Etcheverry and Wonsever, 2019; Samenko et al., 2020), where

[7] For a detailed discussion of our error analysis, see Appendix B.

antonyms and synonyms were mutually confused in the classification, with our setting we overcame this problem. Yet, synonyms, in line with previous studies (Santus et al., 2016a), remain the most challenging class.

Polysemy Initially, we expected more polysemous words would be more problematic and worse predicted, as, at first sight, a wider range of categories could describe different relations between source and target words. Moreover, we expected that the lack of context (or the addition of an artificial one, not adapted to the word pair context) in our approach would make it more difficult to disambiguate between the different senses, and thus to choose the best relation. However, counterintuitively, we did not find statistical evidence that polysemy[8] affected our results.

POS When looking at the part of speech, we found out that adjectives were the best-predicted ones, compared to verbs and nouns. To extract the part of speech, the predominant part of speech annotated for the CogALexV and EVALution datasets were selected.

Semantic Domains and Prototypicality These datasets provide us for each word pair with human-annotated semantic domains[9] for both the source and target words as well as their prototypical relation. We found out that our model predicted better word pairs that contained abstract rather than concrete words, and objects better than events. Our error analysis strengthens previous studies (Necsulescu et al., 2015) that suggest LRC is sensitive to domain bias. Regarding prototypicality, as previously noted in (Santus et al., 2016a), categories more generally associated with a pair of words were the best-predicted ones (in contrast to categories where human annotators doubted the accuracy of the provided annotations).

Sampled Errors Table 5 shows a sample of the most challenging examples that failed with all our templates on all runs using CogALExV and EVALution. However, they point out the limitations of both our approach and the dataset. In the first five examples, our setting was not able to correctly capture the relation between words, as in ('cube','die') that can be either synonyms as annotated, or random as predicted (e.g., in relation to death). Polysemy might induce error in such cases. On the other hand, the last five examples show that some of the original annotations were misleading and our model predicted more sensible relationships.

Pair	Annotated	Predicted
(purpose, goal)	IsA	Random
(law, theory)	PartOf	Antonym
(boy, man)	IsA	Antonym
(cube, die)	Synonym	Random
(city, build)	HasA	IsA
(fish, animal)	Antonym	IsA
(sand, beach)	Synonym	PartOf
(orange, fruit)	PartOf	IsA
(england, great britain)	IsA	PartOf
(rabbit, animal)	PartOf	IsA

Table 5: Examples of pairs failed by our models. The first five show errors in our approach, while the five below ones would be caused by dataset issues.

5.3 Embedding Projection Visualization

In Figure 1, we can observe the learning process of the network represented by the distribution of the embeddings with Principal Component Analysis (PCA)(F.R.S., 1901) across layers and epochs. We show the test embedding projections using the TensorFlow embedding projector platform[10] for RoBERTa base fine-tuned model with template T2 for the CogALexV dataset. Each type of relation is represented by a color, and each point represents a pair of words. Highlighted pairs of words represent the embeddings for the word pairs containing the word "dollar". Lexical memorization (Levy et al., 2015) seems to happen in epoch 5, where the network already clusters by lexical relations (upper figures in every row) but also by words (lower ones). However, in epoch 10, the embedding projection shows how word pairs are now distributed throughout the whole vectorial space. Thus, it seems that the model is indeed learning the relation entailed between different pairs of words without pairing a particular word to a relation. Visualization supports the idea that our model avoids the lexical memorization problem (similar distributions were seen when using the other templates).

[8] Polysemy was estimated by obtaining the product of the number of WordNet synsets associated with both words in the relation.

[9] Semantic domains were annotated by anonymous raters through crowdsourcing. As recommended by EVALution, we only perform our analysis when two or more people tagged a word with the same domain.

[10] Accessible in: https://projector.tensorflow.org/

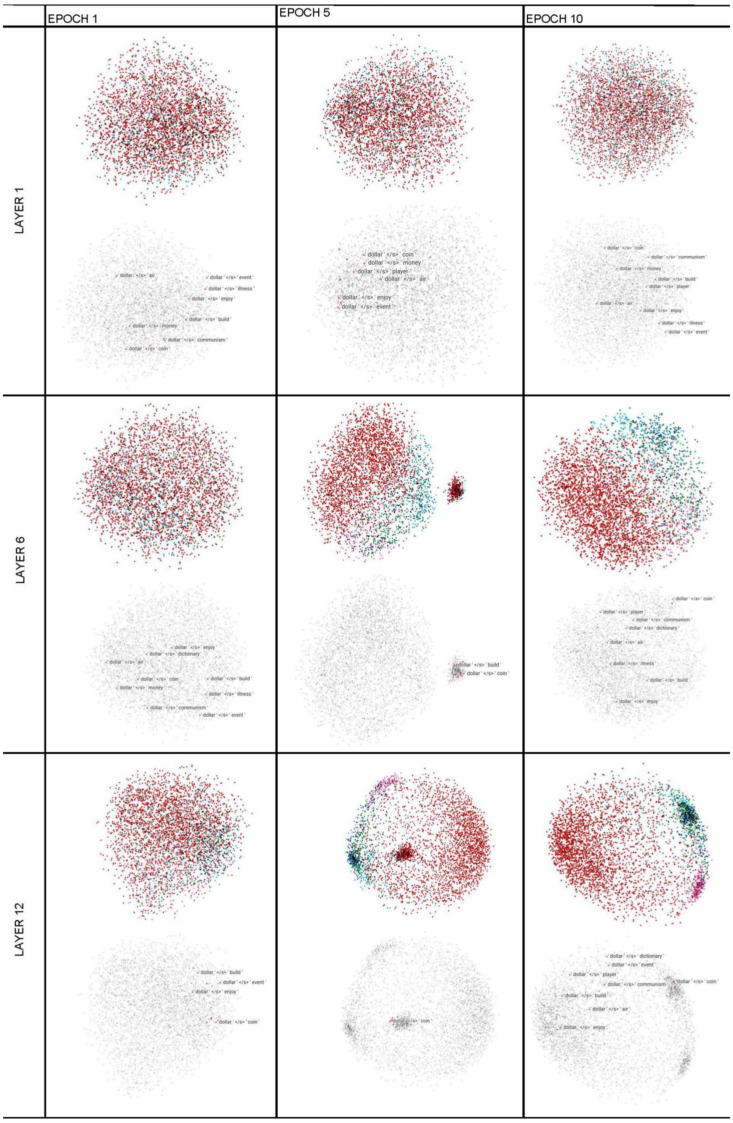

Figure 1: CogALexV embedding projection when finetuning RoBERTa with T2 template. Black and White figures highlight relations containing ‘dollar’.

In the visualization of the embedding projections, we annotated our data with some linguistic features such as polysemy, word frequency, and linguistic register (formal vs colloquial and geographical differences) extracted from WordNet to check whether any clear clusters appeared for the unattested relations group. Yet, in this initial exploration, we could not find any clear clustering.

6 Conclusions and Future Work

Our experiments show that minimal prompts work equally well to more complex ones for the LRC task, thus, allowing less human effort and computational cost, and following a language-neutral approach. Moreover, we show that minimal prompting outperforms SoTA results in graded LE. We conducted an extensive error analysis showing that: synonymy remains the hardest category to classify, there is some domain and POS bias, and polysemy was proven to be an issue. We highlight the need of crafting more balanced datasets in terms of POS and domain, with finer-graded annotations for the different types of synonyms. As future work, we would like to a) address LRC as a multilabel classification task to alleviate the polysemy challenge, b) check the approach with other languages, c) extend the study to other semantic relations, and d) gain insights in why null prompting improves the SoTA for LRC and if this line of research could be generalized to other relations, or if not, what characterizes Lexico-Semantic relations to fit this well the null prompting approach.

7 Limitations

1. **Computational cost**: For our experiments, we used almost 850h of GPUs. In future research, we could try to lower this cost by experimenting with prompting for LRC task in few-shot scenarios, which would also help when conducting the task for low-researched languages.

2. **Language**: Our experiments were conducted just for the English language. Thus, and with the advantage derived from minimal prompting of being language independent, in further research we would like to expand our experiments to multilingual datasets such as the ones from (Wachowiak et al., 2020).

3. **Original dataset limitations**: In line with (Lang et al., 2021), we found some misleading annotations in CogALexV dataset. This not only decrease the performance of the model but can also lead to hard-to-detect biases. Once again, few-shot tuning would decrease the annotation cost, making it possible to train with, although less, better-annotated examples. Additionally, synonymy remains the most difficult relation to capture, a more fine-graded annotation of the different kinds of synonyms could improve their classification.

4. **Domain dependence**: The limitation spotted by (Necsulescu et al., 2015) is persistent in our model. A richer domain annotation would be advised to better research domain bias in the LRC task.

Aknowledgements

Supported by the Spanish project PID2020-113903RB-I00 (AEI/FEDER, UE), by DGA/FEDER, by the *Agencia Estatal de Investigación* of the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the "Ramón y Cajal" program (RYC2019-028112-I), and by the EU research and innovation program HORIZON Europe 2021 through the "4D PICTURE" project under grant agreement 101057332.

References

Kushal Arora, Aishik Chakraborty, and Jackie C. K. Cheung. 2020. Learning lexical subspaces in a distributional vector space. *Transactions of the Association for Computational Linguistics*, 8:311–329.

Oren Barkan, Avi Caciularu, and Ido Dagan. 2020. Within-between lexical relation classification.

Marco Baroni and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from bert. volume 34, pages 7456–7463.

Philipp Cimiano. 2006. *Ontology learning and population from text: Algorithms, evaluation and applications*. Springer US.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. volume abs/1810.04805, pages 4171–4186.

Mathias Etcheverry and Dina Wonsever. 2019. Unraveling antonym's word vectors through a Siamese-like network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3297–3307, Florence, Italy. Association for Computational Linguistics.

Karl Pearson F.R.S. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Goran Glavaš and Ivan Vulić. 2019. Generalized tuning of distributional word vectors for monolingual and cross-lingual lexical entailment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4824–4830, Florence, Italy. Association for Computational Linguistics.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. volume 2, page 539. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8.

Aishwarya Kamath, Jonas Pfeiffer, Edoardo Maria Ponti, Goran Glavaš, and Ivan Vulić. 2019. Specializing distributional vectors of all words for lexical entailment. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 72–83, Florence, Italy. Association for Computational Linguistics.

Saurav Karmakar and John P. McCrae. 2020. Cogalex-vi shared task: Bidirectional transformer based identification of semantic relations. pages 65–71. ACL.

Christian Lang, Lennart Wachowiak, Barbara Heinisch, and Dagmar Gromann. 2021. Cogalex 2.0: Impact of data quality on lexical-semantic relation prediction.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *North American Chapter of the Association for Computational Linguistics*.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. pages 4582–4597.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55:1–35.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. pages 61–68. Association for Computational Linguistics.

Robert LoganIV, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2022. Cutting down on prompts and parameters: Simple few-shot learning with language models. pages 2824–2835. Association for Computational Linguistics.

Rabeeh Karimi Mahabadi, Luke Zettlemoyer, James Henderson, Lambert Mathias, Marzieh Saeidi, Veselin Stoyanov, and Majid Yazdani. 2022. Prompt-free and efficient few-shot learning with language models. volume 1, pages 3638–3652. Association for Computational Linguistics.

Silvia Necsulescu, Sara Mendes, David Jurgens, Núria Bel, and Roberto Navigli. 2015. Reading between the lines: Overcoming data sparsity for accurate classification of lexical relationships. pages 182–192. Association for Computational Linguistics.

Marek Rei, Daniela Gerz, and Ivan Vulić. 2018. Scoring lexical entailment with a supervised directional similarity network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 638–643, Melbourne, Australia. Association for Computational Linguistics.

Igor Samenko, Alexey Tikhonov, and Ivan P. Yamshchikov. 2020. Synonyms and antonyms: Embedded conflict. *ArXiv*, abs/2004.12835.

Enrico Santus, Anna Gladkova, Stefan Evert, and Alessandro Lenci. 2016a. The cogalex-v shared task on the corpus-based identification of semantic relations. pages 69–79.

Enrico Santus, Alessandro Lenci, Tin Shing Chiu, Qin Lu, and Chu Ren Huang. 2016b. Nine features in a random forest to learn taxonomical semantic relations. ROOT09<br/>.

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. 2015. Evalution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. pages 64–69. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. pages 255–269. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2022. True few-shot learning with prompts—a real-world perspective. *Transactions of the Association for Computational Linguistics*, 10:716–731.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. pages 4222–4235. Association for Computational Linguistics.

Vered Shwartz and Ido Dagan. 2016. Cogalex-v shared task: Lexnet-integrated path-based and distributional method for the identification of semantic relations.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. pages 2389–2398. Association for Computational Linguistics.

Asahi Ushio, Jose Camacho-Collados, and Steven Schockaert. 2021. Distilling relation embeddings from pretrained language models. pages 9044–9062. Association for Computational Linguistics.

Ashish Vaswani, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. pages 6000–6010.

Ivan Vulić and Nikola Mrkšić. 2018. Specialising word vectors for lexical entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1134–1145, New Orleans, Louisiana. Association for Computational Linguistics.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. Hyperlex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43:781–835.

Lennart Wachowiak, Christian Lang, Barbara Heinisch, and Dagmar Gromann. 2020. Cogalex-vi shared task: Transrelation - a robust multilingual language model for multilingual relation identification. pages 59–64.

Chengyu Wang, Xiaofeng He, and Aoying Zhou. 2019. Spherere: Distinguishing lexical relations with hyperspherical relation embeddings. pages 1727–1737. Association for Computational Linguistics.

Chengyu Wang, Minghui Qiu, Jun Huang, and Xiaofeng He. 2021. Keml: A knowledge-enriched meta-learning framework for lexical relation classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:13924–13932.

Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. 2014. Learning to distinguish hypernyms and co-hyponyms. pages 2249–2259.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Dongqiang Yang, Ning Li, Li Zou, and Hongwei Ma. 2022. Lexical semantics enhanced neural word embeddings. *Knowledge-Based Systems*, 252:109298.

A Datasets Description

All the five datasets used for LRC, except K&H+N, are to some extent expansions and modified versions of the BLESS dataset. BLESS aimed to provide pair of words to conduct research on distributional semantics through analogies. This first dataset used the McRae norms, Wordnet and ConceptNet as sources. They used single words instead of multiwords and crowdsourced random words to create noise in the dataset at the same time that they assured no relation between them was entailed. They tried to avoid ambiguities, and relied on prototypical terms to stay as 'little controversial as possible'. As categories, they study meronyms and hyponyms, excluding synonyms due the alleged problematic description and heterogeneity.

EVAlution was developed as an expansion of BLESS, to which synonyms and antonyms were added, containing IsA (hypernymy), antonymy, synonymy, meronymy (part of, member of, and made of), entailment, hasA(possession), has property (attribution) relations with heterogeneous distribution of them. Complementary linguistic data is also provided, as for example the domain[11]. CogALexV dataset was provided at the ACL lexical relation classification workshop in 2016 as a challenging subset of Evalution, where words were stemmed. ROOT9 is an expansion of CogALexV.

K&+N is an expansion of Kozareva and Hongs, 2010 dataset, which extracted its original data from hyponymy and hypernymy relations in Wordnet, for animal, plant and vehicle domains. In the current K&H+N dataset, cohyponyms and meronyms were added. As in the previous datasets, multiwords were avoided.

Most datasets, by being descendants of BLESS, contain the same limitations, being mostly the elusion of rare vocabulary and ambiguous words.

For graded LE, in the original Hyperlex dataset, the hyponym pairs are annotated in four levels, namely hyp-i, $1 \leq$i≤ 4, where i is the path length in the WordNet hierarchy. We collapse all labels hyp-i to hyp in our experiments. The same rationale is applied to the hyperonym labels r-hyp-i.

In Table 6, we show the number of pairs for relation in the train/validation/test splits.

B Detailed Error Analysis

To conduct the error analysis, we take the easiest and the most difficult examples to classify trained with RoBERTa (large) for CogaALexV and EVALution datasets. We take two groups of pairs: those which were well and wrongly classified in all of the 5 runs and all templates, except for template $T4$. We test if there is statistical evidence that some features influence the well/wrongly classified pairs. We have a total of 1527 pairs, 586 from CogALexV and 941 from EVALution, divided into $1359/168$ well/wrongly predicted pairs.

The first studied feature is the relation between the words, that is, we ask if there is some lexical relation that it is easier/harder to predict. Figure 2 contains a visualization of the contingency tables of the well/wrongly predicted pairs by relation. In both datasets, applying a χ^2-test, there is statistical evidence that the relation type influences the prediction (p-values$<< 0.05$). In particular, there is a great difference in the predictions for antonyms and synonyms, the former being better predicted than the latter.

We check if the pairs containing polysemous words are more difficult to predict. We use WordNet to obtain the number of synsets for each word, and we consider that the polysemous level of a pair is the product of the number of synsets of the words in the pair. Although the mean of the polysemous level is less for well-predicted pairs, 108.5 vs. 120.6, performing a Welch's t-test to evaluate if the means are different, we find that there is no statistical evidence, with a high p-value equal to 0.40.

We also study if the part of the speech (POS) influences the predictions. CogALexV and EVALution datasets are also annotated with the predominant POS and a list of the different possible POS of each word. We restrict our POS study to the well/wrongly predicted pairs where both words in the pairs have the same predominant POS or there is only one POS in the intersection lists of possible POS. As it is appreciated in the contingency table (Figure 3), adjectives are easier to predict than nouns and verbs.

The domain of the words in CogALexV and EVALution were annotated by humans. We get pairs with common domains, and we restrict the study to the most common domains: abstract, concrete, event and object domains. The visualization of the contingency table can be seen in Figure 4.

[11] Domain information was crowdsourced and not always reliable, thus, authors advised to only take domains as valid when two or more raters annotated the word as belonging to the same domain

	K&H+N	BLESS	EVALution	ROOT9
Unknown	18,319/1,313/6,746	8,529/609/3,008	-	4,479/327/1,566
Hypoym	3,048/202/1,042	924/63/350	1,327/94/459	2,232/149/809
Co-hyponym	18,134/1,313/6,349	2,529/154/882	-	2,222/162/816
Meronym	755/48/240	2,051/146/746	218/13/86	-
Attribute	-	1,892/143/696	903/72/322	-
Antonym	-	-	1,095/90/415	-
Synonym	-	-	759/50/277	-
Has a	-	-	377/25/142	-
Event	-	2,657/212/955	-	-

	CogALexV	Hyperlex (lexical)	Hyperlex (random)
Unknown	2,228/3,059	112/10/35	202/14/74
Hyponym	255/382	563/39/119	849/63/243
Co-hyponym	-	111/8/26	209/7/16
Meronym	163/224	115/10/22	166/14/61
Antonym	241/360	39/3/15	73/6/19
Synonym	167/235	72/4/20	13/10/53

Table 6: Datasets statistics: Number of pairs for each relation in the train/validation/test splits.

There is statistical evidence (p-value$<<0.5$) that the domain influences the correctness of the prediction: words in the abstract and object domains are better predicted.

Finally, CogALexV and EVALution were annotated by humans with the prototypicality of the annotated relation. The pairs of words in the datasets were exposed to five humans to answer to what extent they agreed with the annotated relation (from 0-strongly disagree to 5-strongly agree). So, it is interesting to check if the prototypicality is higher for well-predicted pairs. We perform a Welch's t-test to test if the prototypicality means for well/wrongly predicted pairs are equal. We get that well/wrongly means are 4.63/4.51 with p-value$<<0.05$, so they are different. Although the means seem quite similar, take into account that about 90% of the prototypicality in the datasets range from 4 to 5.

C Mask Verbalizer

In Table 7 it is shown the used tokens to verbalize the mask token in templates TM1, TM2 and TM3.

D Complete Results

We present the results for BERT and RoBERTa (large and base) models. Table 8 contains the mean of the weighted by the support labels of precision of the 5 runs, recall and F1-score. The greatest value for each measure (column) is underlined. A value is boldened if there is no statistical evidence to be different from the greatest one performing a Welch's t-test for the mean values. A similar rationale is applied for Table 9, with the complete results for CogALexV dataset and Table 10 for Hyperlex dataset.

Relation label					Chosen verb. mask token	
BLESS	EVALution	CogALexV	KH&+N	ROOT9	BERT	RoBERTa
event					event	␣event
	hasa				contains	␣contains
	madeof				material	␣material
mero	partof	part_of	mero		part	␣part
random		random	false	random	random	␣random
coord			sibl	coord	coordinated	␣coordinated
	synonym	syn			synonym	␣equivalent
	antonym	ant			contrary	␣contrary
attri	hasproperty				attribute	␣attribute
hyper	isa	hyper	hypo	hyper	minor	␣subclass

Table 7: Verbalization of the relation label for the mask token for each dataset and model in templates TM1, TM2 and TM3. Each word in the "Verb. token mask" column corresponds with one token in the vocabulary of the model. The underscores in the RoBERTa mask verbalizations are to emphasize that the tokens have a white space in front of them.

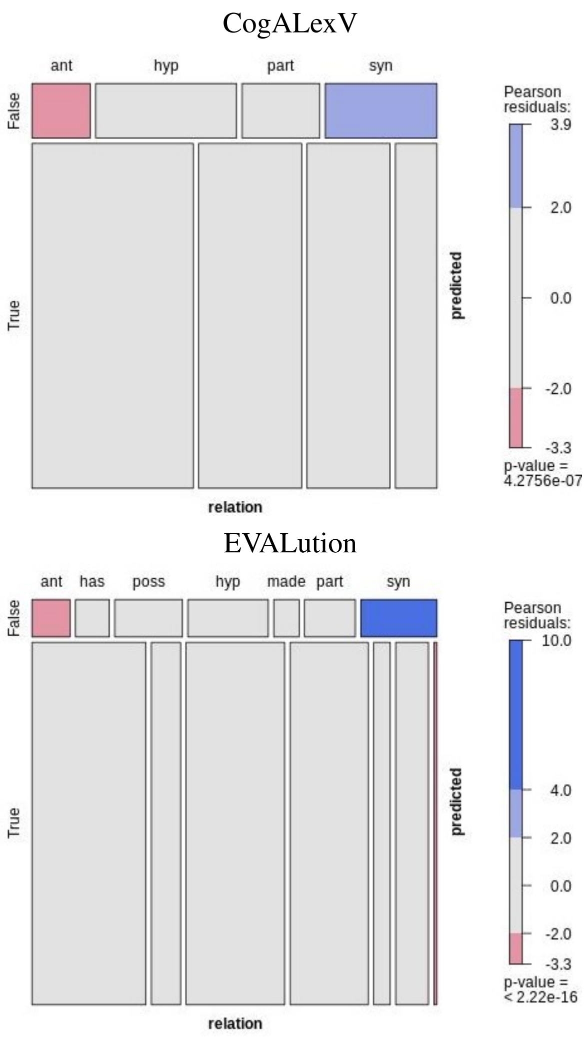

Figure 2: Visualization of the contingency tables for bad/wrongly-predicted pairs regarding the relation type.

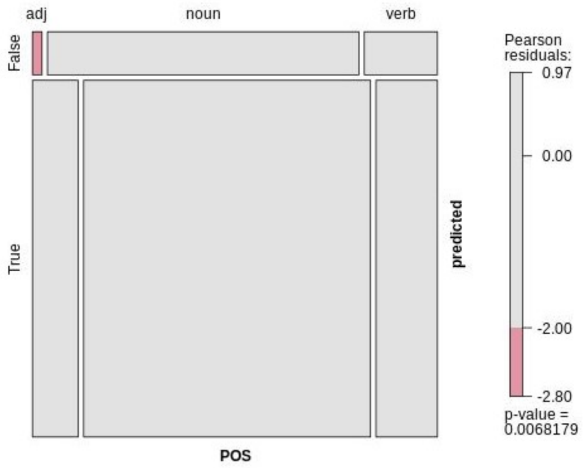

Figure 3: Visualization of the contingency table for bad/wrongly-predicted pairs regarding the POS.

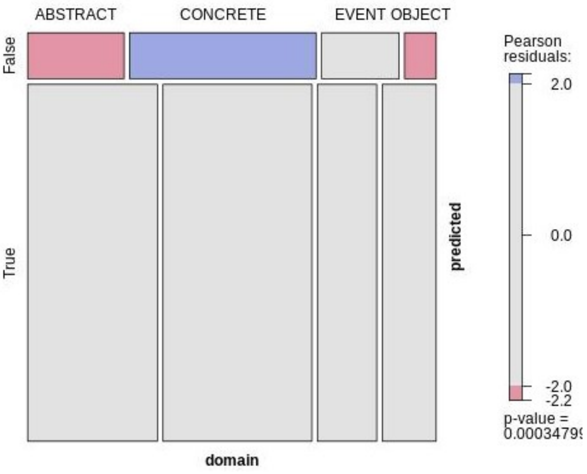

Figure 4: Visualization of the contingency table for bad/wrongly-predicted pairs regarding the domains of the words.

	K&H+N			BLESS			EVALution			ROOT9		
	pre	rec	F1	pre	rec	F1	prec	recl	F1	pre	rec	F1
BERT												
T1	0.989	0.989	0.989	**0.952**	**0.951**	**0.951**	0.748	0.748	0.747	0.927	0.926	0.926
T2	0.989	0.989	0.989	**0.95**	**0.948**	**0.948**	0.739	0.739	0.737	0.93	0.929	0.929
T3	0.99	0.99	0.99	**0.953**	**0.952**	**0.952**	**0.753**	0.75	0.751	0.931	0.931	0.931
T4	0.741	0.588	0.51	0.244	0.2	0.088	0.116	0.149	0.053	0.929	0.928	0.928
TM1	0.987	0.987	0.987	0.942	0.941	0.941	**0.755**	0.744	0.745	0.927	**0.925**	0.925
TM2	0.987	0.987	0.987	0.946	0.944	0.945	0.738	0.729	0.722	**0.925**	**0.925**	**0.925**
TM3	0.986	0.986	0.985	**0.948**	**0.947**	**0.947**	0.73	0.726	0.724	0.927	0.924	0.924
RoBERTa												
T1	0.989	0.989	0.989	**0.955**	**0.954**	**0.954**	**0.769**	**0.765**	**0.764**	**0.937**	**0.936**	**0.936**
T2	0.989	0.989	0.989	**0.955**	**0.954**	**0.955**	**0.759**	**0.759**	**0.757**	**0.936**	**0.936**	**0.936**
T3	0.989	0.989	0.989	**0.956**	**0.955**	**0.956**	**0.773**	**0.771**	**0.771**	**0.938**	**0.937**	**0.937**
T4	0.603	0.326	0.312	0.511	0.194	0.133	0.23	0.191	0.087	**0.936**	**0.934**	**0.934**
TM1	0.989	0.989	0.988	0.948	0.946	0.947	**0.772**	**0.762**	**0.761**	**0.936**	**0.936**	**0.936**
TM2	0.988	0.988	0.988	0.947	0.945	0.946	**0.771**	**0.765**	**0.764**	**0.93**	**0.929**	**0.928**
TM3	0.986	0.985	0.985	**0.951**	**0.95**	**0.951**	0.774	**0.754**	**0.746**	**0.926**	**0.926**	**0.926**
BERT base												
T1	0.988	0.988	0.988	0.944	0.942	0.942	0.69	0.691	0.689	0.926	0.924	0.924
T2	0.987	0.987	0.987	0.943	0.941	0.941	0.675	0.672	0.672	0.919	0.918	0.918
T3	0.987	0.987	0.987	0.944	0.942	0.942	0.696	0.694	0.694	0.922	0.921	0.921
T4	0.548	0.429	0.316	0.37	0.228	0.165	0.213	0.218	0.119	0.921	0.919	0.919
TM1	0.986	0.986	0.986	0.939	0.936	0.936	0.707	0.7	0.698	0.917	0.917	0.917
TM2	0.985	0.986	0.985	0.94	0.939	0.94	0.69	0.686	0.684	0.918	0.917	0.917
TM3	0.985	0.985	0.985	0.941	0.939	0.939	0.697	0.692	0.686	0.918	0.915	0.915
RoBERTa base												
T1	0.983	0.984	0.983	0.95	0.949	0.949	0.749	0.744	0.745	0.932	0.931	0.931
T2	0.988	0.988	0.988	0.948	0.947	0.947	0.746	0.744	0.744	0.931	0.931	0.931
T3	0.987	0.987	0.987	0.95	0.949	0.949	**0.756**	0.753	**0.754**	**0.934**	**0.933**	**0.933**
T4	0.66	0.455	0.299	0.504	0.139	0.043	0.121	0.095	0.023	0.924	0.923	0.923
TM1	0.987	0.986	0.986	0.941	0.94	0.94	**0.758**	0.745	**0.747**	**0.927**	**0.926**	**0.926**
TM2	0.983	0.983	0.983	**0.946**	**0.944**	**0.944**	0.74	0.724	0.727	0.926	0.926	0.925
TM3	0.986	0.986	0.986	0.946	0.944	0.944	0.74	0.737	**0.729**	0.924	0.924	0.924
SoTA												
LexNET	0.985	0.986	0.985	0.894	0.893	0.893	0.601	0.607	0.6	0.813	0.814	0.813
KEML	**0.993**	**0.993**	**0.993**	0.944	0.943	0.944	0.663	0.66	0.66	0.878	0.877	0.878
SphereRE	0.99	0.989	0.99	0.938	0.938	0.938	0.62	0.621	0.62	0.86	0.862	0.861
RelBERT	-	-	0.949	-	-	0.921	-	-	0.701	-	-	0.91

Table 8: Complete results for K&H+N, BLESS, EVALution and ROOT9 datasets.

	ant	hyp	part	syn	all
BERT					
T1	0.77	0.68	0.715	0.564	0.69
T2	0.769	0.675	0.728	0.528	0.683
T3	0.789	0.681	0.736	0.566	0.7
T4	0.119	0.044	0.078	0.0	0.063
TM1	0.798	0.682	0.746	0.585	0.709
TM2	0.782	0.688	0.742	0.56	0.7
TM3	0.779	0.682	0.742	0.563	0.698
RoBERTa					
T1	**0.873**	**0.703**	**0.752**	**0.604**	**0.743**
T2	**0.863**	**0.682**	**0.745**	0.584	**0.728**
T3	**0.884**	**0.718**	**0.784**	**0.629**	**0.762**
T4	0.237	0.004	0.165	0.085	0.119
TM1	**0.88**	**0.709**	**0.773**	**0.599**	**0.75**
TM2	**0.871**	**0.723**	**0.787**	**0.621**	**0.758**
TM3	**0.871**	**0.718**	**0.787**	**0.616**	**0.756**
BERT base					
T1	0.554	0.591	0.657	0.361	0.546
T2	0.529	0.544	0.61	0.278	0.499
T3	0.565	0.605	0.684	0.375	0.562
T4	0.081	0.0	0.101	0.006	0.044
TM1	0.645	0.625	0.707	0.431	0.607
TM2	0.57	0.622	0.685	0.393	0.573
TM3	0.636	0.648	0.721	0.43	0.615
RoBERTa base					
T1	0.806	0.677	0.732	0.57	0.704
T2	0.783	0.652	0.693	0.536	0.675
T3	0.82	0.676	0.731	0.577	0.709
T4	0.027	0.0	0.102	0.092	0.044
TM1	0.809	0.678	0.743	0.561	0.706
TM2	0.801	0.673	0.742	0.556	0.701
TM3	0.815	0.679	0.73	0.561	0.705
SoTA					
LexNET	0.425	0.526	0.493	0.297	0.445
SphereRE	0.479	0.538	0.539	0.286	0.471
KEML	0.492	0.547	0.652	0.292	0.5
RelBert	0.794	0.616	0.702	0.505	0.664

Table 9: Complete results for CogALexV dataset.

	random	lexical
BERT	all/noun/verb	
T1	0.644/0.654/0.525	0.686/0.737/0.499
T2	0.577/0.586/0.432	0.402/0.433/0.286
T3	0.728/0.742/0.551	0.747/**0.781/0.623**
TM1	0.8/0.822/0.577	**0.766**/0.807/0.672
TM2	0.778/0.804/0.553	0.657/0.717/0.478
TM3	0.794/0.817/0.578	0.741/**0.781/0.633**
RoBERTa		
T1	**0.741/0.753/0.584**	**0.755/0.788/0.532**
T2	0.152/0.17/0.03	0.287/0.35/0.063
T3	0.774/0.79/0.631	**0.669/0.69/0.516**
TM1	**0.828/0.839/0.716**	**0.789/0.837/0.612**
TM2	**0.749/0.761/0.646**	**0.654/0.705/0.417**
TM3	**0.814/0.83/0.683**	**0.794/0.828/0.656**
BERT base		
T1	0.643/0.666/0.426	0.471/0.557/0.173
T2	0.626/0.657/0.306	0.374/0.446/0.116
T3	0.638/0.669/0.375	0.614/0.691/0.312
TM1	0.719/0.747/0.428	0.597/0.68/0.38
TM2	0.707/0.743/0.366	0.575/0.656/0.277
TM3	0.685/0.717/0.417	0.584/0.6650.356
RoBERTa base		
T1	0.737/0.749/0.594	0.677/0.713/0.543
T2	0.652/0.683/0.377	0.407/0.483/0.167
T3	0.742/0.757/0.637	0.626/0.693/0.391
TM1	0.796/0.811/0.639	0.736/**0.8/0.553**
TM2	0.781/0.793/0.664	0.711/0.757/0.525
TM3	0.783/0.795/0.635	**0.757**/0.807/0.634
SoTA		
LEAR	0.686/0.71/-	0.174/-/-
SDNS	0.692/-/-	-/-/
GLEN	0.52/-/-	0.481/-/-
POSTLE	0.686/-/-	-/0.60/-
LexSub	0.533/-/	-/-/
HF	0.69/-/-	-/-/
IAA	0.864/0.864/0.862	

Table 10: Complete results for Hyperlex dataset.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7. Limitations.

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
S1. Introduction.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

S4. Experimental setup.

☑ B1. Did you cite the creators of artifacts you used?
S4. Experimental setup.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
S4. Experimental setup.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
S1,2 and 4. Introduction, Related work and Experimental setup.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
S4 and Appendix A.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Results and appendix.

C ☑ Did you run computational experiments?

S4. Experimental setup.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
S4. Experimental setup.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
S4. Experimental setup.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix B and S5. Results.

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No, but they are provided in the source code for our experiments in Github.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.