# Ask2Transformers: Zero-Shot Domain labelling with Pre-trained Language Models

**Oscar Sainz** and **German Rigau**

HiTZ Center - Ixa Group,
University of the Basque Country (UPV/EHU)
{oscar.sainz, german.rigau}@ehu.eus

## Abstract

In this paper we present a system that exploits different pre-trained Language Models for assigning domain labels to WordNet synsets without any kind of supervision. Furthermore, the system is not restricted to use a particular set of domain labels. We exploit the knowledge encoded within different off-the-shelf pre-trained Language Models and task formulations to infer the domain label of a particular WordNet definition. The proposed zero-shot system achieves a new state-of-the-art on the English dataset used in the evaluation.

## 1 Introduction

The whole Natural Language Processing (NLP) research area have been accelerated with the advent of the unsupervised pre-trained Language Models. First with ELMo (Peters et al., 2018) and then with BERT (Devlin et al., 2019) the paradigm of using pre-trained Language Models for fine-tuning on a particular NLP task has became the new standard approach, replacing the more traditional knowledge-based and fully supervised approaches. Currently, as the size of the corpus and models increase, the research community has observed that the Transfer Learning approach has the capacity to work without any or with a very small fine-tuning. Some examples of the strength of this approach are GPT-2 (Radford et al., 2019) or more recently GPT-3 (Brown et al., 2020) that shows the ability of these huge pre-trained Language Models to solve tasks for which have not even trained.

Recently, with the arrival of the GPT-3 new ways to perform zero and few shot approaches have been discovered. These approaches propose the inclusion of a small number of supervised examples in the input as a hint for the model. The model then, just by looking a small set of examples, is able to complete successfully the task at hand. Brown et al. (2020) report that they solve a wide range of NLP tasks just following the previous approach. However, this approach only looks appropriate when the model is large enough.

In this paper we exploit the domain knowledge already encoded within the existing pre-trained Language Models to enrich the WordNet (Miller, 1998) synsets and glosses with domain labels. We explore and evaluate different pre-trained Language Models and pattern objectives. For instance, consider the example shown in Table 1. Given a WordNet definition such as the one of <hospital, infirmary> and the knowledge encoded in a pre-trained Language Model, the task is to assess which is its most suitable domain label. Thus, we create an appropriate pattern in natural language adapted to the objective of the Language Model. In the example, we use a Language Model fine-tuned on a general task such as Natural Language Inference (NLI) (Bowman et al., 2015). The NLI objective is to train a model able to classify the relation between two sentences as entailment, contradiction or neutral. Having four domains such as *medicine*, *biology*, *business* and *culture*, our system performs four queries to the model, each one with one of the four domains. Each query takes as a first sentence the WordNet definition and as a second sentence *The domain of the sentence is about [domain-label].* As expected, the most suitable domain label in this example is *medicine* with a confidence of 0.77. As shown, an off-the-shelf Language Model which have been fine-tuned on a general NLI task is able to infer the most appropriate domain label for the WordNet definition without any further training. Also note that the approach can use any given set of domain labels.

Interestingly, without any training on the task at hand, the proposed zero-shot system obtains an F1 score of 92.4% on the English dataset used in the

evaluation.

All the implementation code along with the experiments is freely available on a GitHub repository [1].

After this short introduction, the next section presents previous work on domain labelling of WordNet. Section 3 presents our approach, Section 4 the experimental setup and Section 5 the results from our experiments. Finally, Section 6 revises the main conclusions and the future work.

## 2 Related Work

Building large and rich lexical knowledge bases is a very costly effort which involves large research groups for long periods of development. Starting from version 3.0, Princeton WordNet has associated topic information with a subset of its synsets. This topic labeling is achieved through pointers from a source synset to a target synset representing the topic. WordNet uses 440 topics and the most frequent one is <law, jurisprudence>.

In order to reduce the manual effort required, a few semi-automatic and fully automatic methods have been applied for associating domain labels to synsets. For instance, WordNet Domains[2] (WND) is a lexical resource where synsets have been semi-automatically annotated with one or more domain labels from a set of 165 hierarchically organized domains (Magnini, 2000; Bentivogli et al., 2004). The uses of WND include the possibility to reduce the polysemy degree of the words, grouping those senses that belong to the same domain (Magnini et al., 2002). But the semi-automatic method used to develop this resource was far from being perfect. For instance, the noun synset <diver, frogman, underwater diver> defined as *some-one who works underwater* has domain *history* because it inherits from its hypernym <explorer, adventurer> also labelled with *history*. Moreover, many synsets have been labelled as *factotum* meaning that the synset cannot be labelled with a particular domain. WND also provides mappings to WordNet Topics and also to Wikipedia categories.

eXtended WordNet Domains[3] (XWND) (Gonzalez-Agirre et al., 2012; González et al., 2012) applied a graph-based method to propagate the WND labels through the WordNet structure.

Domain information is also available in other lexical resources. For instance, IATE[4], a European Union inter-institutional terminology database. The domain labels of IATE are based on the Eurovoc thesaurus[5] and were introduced manually.

More recently, BabelDomains[6] (Camacho-Collados and Navigli, 2017) propose an automatic method that propagates the knowledge categories from the Wikipedia to WordNet by exploiting both distributional and graph-based clues. As domains of knowledge, BabelDomains opted for domains from the Wikipedia featured articles page[7]. This page contains a set of thirty-two domains of knowledge. When labelling WordNet synsets with these domains, BabelDomains reports a precision of 81.7, a recall of 68.7 and an F1 score of 74.6. Unfortunately, as these numbers suggest not all WordNet synsets have been labelled with a domain. For instance, the synset <hospital, infirmary> with a gloss definition *a health facility where patients receive treatment* has no Babeldomain assigned.

It is worth to note that all these methods depart from a particular set of domain labels (or categories) manually assigned to a set of WordNet synsets (or Wikipedia pages). Then, these labels are propagated through the WordNet structure following automatic or semi-automatic methods. In contrast, our zero-shot method does not require an initial manual annotation. Furthermore, it is not designed for a particular set of domain labels. That is, it can be applied to label from scratch any dictionary or lexical knowledge base (or wordnet) with distinct sets of domain labels.

## 3 Using pre-trained LMs for domain labelling

Recent studies such as the one of GPT-3 (Brown et al., 2020) shows that when increasing the size of the model, the capacity to solve different tasks with just a few positive examples also increases (few-shot learning). However, very large Language Models also have important hardware requirements (i.e. large RAM GPUs). Thus, we decided to keep the size of the models used manage-

[1] https://github.com/osainz59/Ask2Transformers

[2] http://wndomains.fbk.eu/

[3] https://adimen.si.ehu.es/web/XWND

[4] http://iate.europa.eu/

[5] https://op.europa.eu/en/web/eu-vocabularies/th-dataset/-/resource/dataset/eurovoc

[6] http://lcl.uniroma1.it/babeldomains/

[7] https://en.wikipedia.org/wiki/Wikipedia:Featured_articles

| Definition: | hospital: a health facility where patients receive treatment. | | |
|---|---|---|---|
| Pattern: | The domain of the sentence is about | **medicine** | **0.77** |
| | | biology | 0.08 |
| | | business | 0.04 |
| | | culture | 0.02 |

Table 1: An example of domain labelling.

able with small hardware requirements.

The task where we focused on is the domain labelling of WordNet glosses. This task consist in the following. Given a WordNet gloss $g$ to predict the corresponding domain $d$ of the WordNet concept defined. In this paper, the domains are taken from BabelDomains (Camacho-Collados and Navigli, 2017). Supervised domain labelling can be solved as any other multiclass problem, where the output of the model is a class probability distribution. In our zero-shot experiments we did not modify any of the pre-trained models. We just reformulate the domain labelling task to match with the LMs training objective.

### 3.1 Masked Language Modeling

The Masked Language Modeling (MLM) is a pre-training objective followed by models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). This objective works as follows. Given a sequence of tokens $s = [t_1, t_2, ..., t_n]$, the sequence is first perturbed by replacing some of the tokens $t$ with an special token [MASK]. Then, the model is trained to recover the original sequence $s$ given the modified sequence $\hat{s}$. This denoising objective can be seen as an evolution for the contextual embeddings of the previous CBOW from word2vec (Mikolov et al., 2013).

For domain labelling, we have replaced the input for the model following the next pattern:

$s$: Context: [context] Topic: [MASK]

where we introduce the input sentence replacing the [context] tag. Then, we let the model predict the most probable token for the [MASK] tag. For instance, given the biological definition of *cell*, the model returns the following topics: *Biology*, *evolution*, *life*, etc.

This approach has been used to explore the knowledge of the model without any predefined set of domain labels in Section 5.7.

### 3.2 Next Sentence Prediction

Along with the MLM the Next Sentence Prediction (NSP) is the training objective used by the BERT models. Given a pair of sentences $s_1$ and $s_2$, this objective predicts whether $s_1$ is followed by $s_2$ or not.

To adapt the BERT objective to the domain labelling task, we propose the next strategy inspired in the work from Yin et al. (2019). We use the following input pattern:

$s_1$: [context]
$s_2$: Domain or topic about [domain-label]

where $s_1$ encodes a WordNet gloss as a context and $s_2$ is formed by a *template* and a domain-label. In order to make the classification, we run as many times as domain labels and then apply a softmax over the positive class outputs. We hypothesize that, no matter if any of the $s_2$ can really follow the given $s_1$, the most probable one should be the $s_2$ formed by the correct label. For instance, recall the *hospital* example shown in Table 1.

### 3.3 Natural Language Inference

In this case, we use a pre-trained LM that has been fine-tuned for a general inference task which is the Natural Language Inference (Williams et al., 2018a). Given two sentences in the form of a premise $s_1$ and an hypothesis $s_2$, the NLI task consists on redicting whether the $s_1$ *entails* or *contradicts* $s_2$ or if the relation between both is *neutral*.

We also used the input pattern shown in the previous NSP approach to adapt the NLI models to the domain labelling task. In this case, we just use the predictions of the *entailment* class. The predictions of the *c*ontradiction and *neutral* are not used. As in the previous case, no matter if any of the $s_2$ hypothesis entails the premise $s_1$ or not, the most probable entailment should be the correct domain label. For example, consider again the example

presented in Table 1.

## 4 Experimental setting

This section describes our experimental setup. We introduce the pre-trained Language Models and the dataset used. For the case of the Language Models, we have tested BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and BART (Wang et al., 2019). For the dataset, we have used the one released by Camacho-Collados et al. (2016) based on WordNet.

### 4.1 Pretrained models

All the Language Models have been obtained from the Huggingface Transformers library (Wolf et al., 2019).

**MLM** For the objective we have used *roberta-large* and *roberta-base* checkpoints. These models have obtained state-of-the-art results on many NLP tasks and benchmarks.

**NSP** For this objective we use the BERT models as they are the only ones trained on that objective. For the sake of comparing the performance of more than one model of each objective we have selected the *bert-large-uncased* and *bert-base-uncased* checkpoints. They only differ on the size of the Language Model.

**NLI** For this objective we used a checkpoint based on RoBERTa *roberta-large-mnli* which have been fine-tuned with MultiNLI (Williams et al., 2018b). We also include *bart-large-mnli* for testing a generative model.

### 4.2 Dataset

We evaluate our approaches on a dataset derived from WordNet which have been annotated with Babeldomain labels (Camacho-Collados et al., 2016). This dataset consist of **1540** synsets manually annotated with their corresponding Babeldomain label. The distribution of domain labels in the dataset is shown in Figure 1. Note that the dataset is quite unbalanced. In fact, some important domains such as *Transport and travel* or *Food and drink* have no single labelled example. As our system is unsupervised, we use the whole dataset for testing.

## 5 Evaluation and Results

This section presents a quantitative and qualitative evaluation. One the one hand, the quantitative evaluation has been done incrementally in order to obtain the best-performing system. First, we have evaluated the different alternative models using the same objective pattern. Then, once the best approach was selected we have explored alternative patterns using the best model. When the best performing pattern was discovered we have focus on finding a better label representation. Finally, we have compared our best system against the previous state-of-the-art methods.

On the other hand, as one of our system is based on a generative approach (MLM) the applied restrictions may not show the real performance of the method. So, we decided to at least do an small qualitative review of the approach.

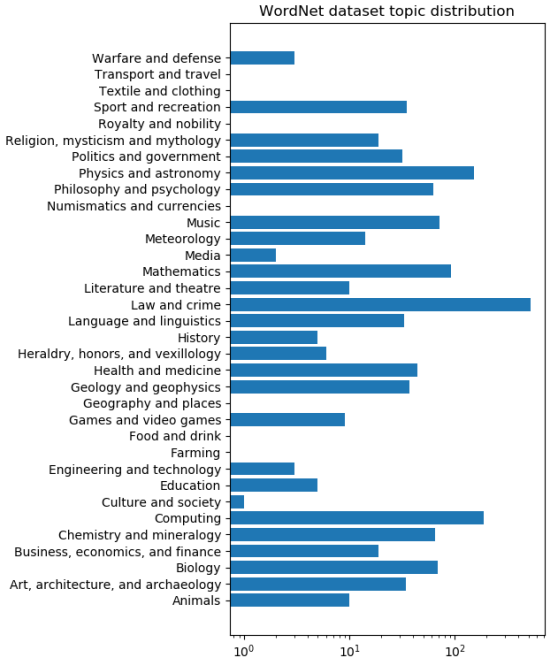


Figure 1: Distribution of domains in the WordNet dataset.

| Method | Top-1 | Top-3 | Top-5 |
|---|---|---|---|
| MNLI (roberta-large-mnli) | **78.44** | **87.46** | **89.74** |
| MNLI (bart-large-mnli) | 61.81 | 79.85 | 87.59 |
| NSP (bert-large-uncased) | 2.07 | 8.57 | 16.49 |
| NSP (bert-base-uncased) | 2.85 | 10.32 | 16.88 |

Table 2: Top-K accuracy of different approaches.

### 5.1 Approach comparison

Table 2 shows the Top-1, Top-3 and Top-5 accuracy of each system when using the same objective pattern. To understand better the behaviour of the systems we also present in the Figure 2 the Top-K

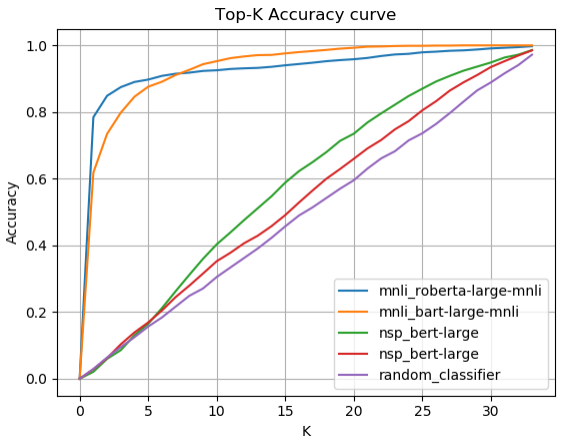


Figure 2: Top-K accuracy curve of the different approaches and a random classifier baseline.

accuracy curve comparing all the approaches and a random baseline. As expected the systems that follow the same approaches perform similarly and share a similar curve. The best performing system is the MNLI based *roberta-large-mnli*, followed by the *bart-large-mnli* checkpoint. We observe a large difference between the different models. For instance, the models pre-trained on the NLI task perform much better than those pre-trained on the general NSP task. The NSP approaches perform slightly better than the random classifier which can be a signal of a non appropriated objective model to use.

### 5.2 Input representation

Once selected the pre-trained Language Model, we evaluate different input patterns for the *roberta-large-mnli* checkpoint. As mentioned before, the MNLI approaches follow the same structure as NSP, where $s_1$ is the gloss of the synset and $s_2$ the sequence formed by a textual template plus the label.

Table 3 shows the results obtained by testing different textual patterns. Very short patterns obtain low results. The best performing textual template is obtained with *The domain of the sentence is about [label].*

### 5.3 Label descriptors / Mapping

As important as the input patterns is the set of domain labels used. Actually, BabelDomains uses labels that refers to one or several specific domains. For instance, *Art, architecture and archaeology*. Although these coarse-grained labels can be useful when clustering close-related domains, we also implemented a two-step labelling procedure taking into account those specific domains. First, we run the system over a set of specific domains or descriptors. Second, we apply a function that maps the descriptors to the original BabelDomains.

**Descriptors** The descriptors defined in this work are quite simple. Given a composed domain label such us *Art, architecture and archaeology*, we define the set of descriptors as each of the components of the label. For instance, in this case *Art*, *Architecture* and *Archaeology*. In the case of labels that consist on a single domain, the descriptors are just the labels. For example, in the case of *Music* the descriptor is also *Music*.

**Mapping function** The mapping function that we use in this work consists on taking the maximum result of the descriptors as the result of the original domain label, i.e. $l_i = \max(d_{i1}, d_{i2}, ..., d_{in})$.

### 5.4 Training a specialized student

The inference time increases linearly with the number of labels. That is, for each example we need to test all the different domain labels. To speed-up the labelling process we annotate automatically the rest of WordNet glosses (around 79.000 glosses) using our best zero-shot approach. Then, we use that automatically annotated dataset to train a much smaller Language Model for the task. For instance, to label new definitions or new lexicons. We have fine-tuned two different models, the first one based with DistilBert (Sanh et al., 2019) which is 5 times smaller than the *roberta-large-mnli* and a XLM-RoBERTa (Conneau et al., 2020) *base* which is 2 times smaller and is trained in a multilingual fashion. We called them $\text{A2T}_{\text{FT-small}}$ and $\text{A2T}_{\text{FT-xlingual}}$ respectively. The first one achieve a **x425** faster inference (5 times smaller and 85 times less inferences) while the second one a speed boost of **x170**.

### 5.5 Results

In order to know how good is our final approach we compare our new systems with the previous ones. The results are reported on the Table 4 in terms of Precision, Recall and F1 for comparison purposes. We also include the results from two previous state-of-the-art systems. As we can see, the new systems based on pre-trained Language Models obtain much better performance (from a

| Input pattern | Top-1 | Top-3 | Top-5 |
|---|---|---|---|
| Topic: [label] | 59.61 | 69.48 | 74.02 |
| Domain: [label] | 58.50 | 67.40 | 72.27 |
| Theme: [label] | 59.67 | 73.96 | 81.36 |
| Subject: [label] | 60.58 | 69.74 | 74.35 |
| Is about [label] | 73.37 | 87.72 | 91.94 |
| Topic or domain about [label] | 78.44 | 87.46 | 89.74 |
| The topic of the sentence is about [label] | 80.71 | 92.92 | 95.77 |
| The domain of the sentence is about [label] | **81.62** | **93.96** | **96.42** |
| The topic or domain of the sentence is about [label] | 76.62 | 88.63 | 91.23 |

Table 3: Some of the explored *input patterns* for the MNLI approach and their Top-1, Top-3 and Top-5 accuracy.

previous best result with an F1 of 74.6 to the new one of 82.10). We also obtain an small improvement when establishing a threshold to decide whether a prediction is taken into consideration or not. Our system performs slightly better with a confidence score greater than 5% ($A2T_{(> 0.05)}$). Figure 3 reports the Precision/Recall trade-off of the A2T system. As mentioned before labels composed of multiple domains can make the prediction harder for the zero-shot system. As a result, a simple system using the label descriptors boosts the performance of the system reaching a final **92.14** F1 score ($A2T_{+ \text{ descriptors}}$). Finally, we also include the results of both the fine-tuned student versions which still obtain very competitive results while drastically reducing the inference time of the original models.

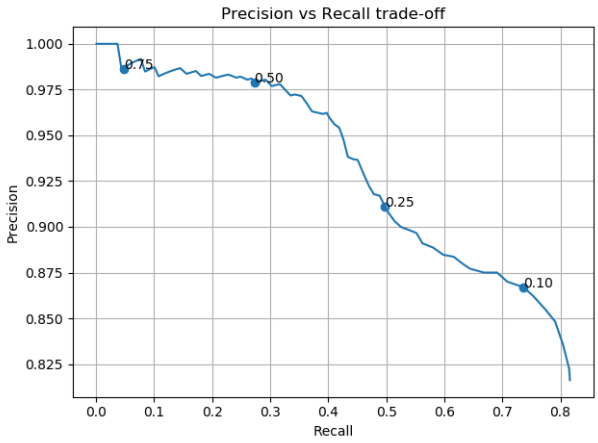


Figure 3: Precision/Recall trade-off of A2T system. Annotations indicates the probability thresholds.

| Method | Precision | Recall | F1 |
|---|---|---|---|
| Distributional | 84.0 | 59.8 | 69.9 |
| BabelDomains | 81.7 | 68.7 | 74.6 |
| A2T | 81.62 | 81.62 | 81.62 |
| $A2T_{(> 0.05)}$ | 83.20 | 81.03 | 82.10 |
| $A2T_{+ \text{ descriptors}}$ | **92.14** | **92.14** | **92.14** |
| $A2T_{\text{FT-small}}$ | 91.42 | 91.42 | 91.42 |
| $A2T_{\text{FT-xlingual}}$ | 90.58 | 90.58 | 90.58 |

Table 4: Micro-averaged precision, recall and F1 for each of the systems. Distributional (Camacho-Collados et al., 2016) and BabelDomains (Camacho-Collados and Navigli, 2017) measures are the ones reported by them.

### 5.6 Error analysis

Figure 4 presents the confusion matrix of our best system. The matrix is row wise normalized due to the imbalance of the dataset label distribution. Looking at the figure there are 4 classes that are misleading. The "Animals" domain is confused with the related domains "Biology" and "Food and drink". For instance, this is the case of the synset <diet> with the definition *the usual food and drink consumed by an organism (person or animal)* which is labelled by our system as "Food and drink". The "Games and video games" domain is confused with the related domain "Sport and recreation". For example the sense referring to *game: a single play of a sport or other contest; "the game lasted two hours"* which is labelled by our system as "Sport and recreation". The third one, "Heraldry, honors and vexillology" is also confused with a very close domain "Royalty and nobility". Obviously, close-related domains can be very difficult to distinguish even for humans. For example, the sense <audio cd, audio compact disc> annotated in the gold standard as "Music" is labelled by our system as "Media". Finally,

| Synset | cell | phase space | rounding error | wipeout |
|---|---|---|---|---|
| Label | Biology | Physics and astronomy | Mathematics | Sports and Recreation |
| Top predictions | **Biology** | EOS | rounding | sports |
| | EOS | physics | EOS | EOS |
| | biology | **Physics** | math | sport |
| | evolution | geometry | taxes | accident |
| | life | relativity | **Math** | **Sports** |

Table 5: Top predictions of the MLM approach using the *roberta-large* checkpoint.

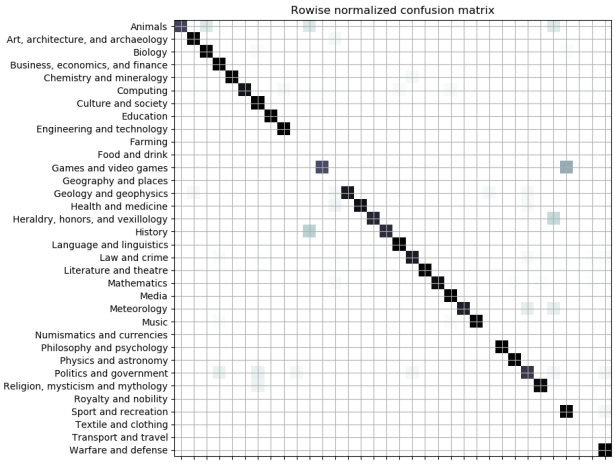


Figure 4: Rowise normalized confusion matrix of the $A2T_{+\text{ descriptors}}$ system.

sometimes the ”History” domain is confused with ”Food and drink”. A curious example of this case is the sense referring to the history event <Boston tea party> that is labelled as ”Food and drink”.

### 5.7 Qualitative analysis

Table 5 shows some of the top predictions obtained by a Masked Language Model (MLM) and the real label for 4 different synsets. In this case, the system is guessing its best predicted domain. That is, the system is not restricted to a select the best label from a pre-defined set of domain labels. Now, the system is free to return the word that best fit the masked term.

We can see in the table that the predictions of the model are close to the correct label although not always equal. Sometimes because of a different case. They can also be seen as fine-grained domains or domain keywords of the real domain.

## 6 Conclusions and Future Work

In this paper we have explored some approaches for domain labelling of WordNet glosses by exploiting pre-trained LM in a zero-shot manner. We have presented a simple approach that achieves a new state-of the art on the Babeldomain dataset.

Even if we have focused on domain labelling of WordNet glosses, our method seems to be robust enough to be adapted to work on tasks such as Sentiment Analysis or other type of text classification. In particular, we think that the approach can be very useful when no annotated data is available.

For the future, we have considered three main objectives. First, we plan to apply this approach to other sources of domain information such as WordNet topics and WordNet Domains. We will also explore how to deal with definitions with generic domains (with no BabelDomains labels or with WordNet Domains factotum label). Second, we also aim to explore the cross-lingual capabilities of pre-trained Language Models for domain labelling of non-English wordnets and other lexical resources. Finally, we also plan to explore the utility of these findings in the Word Sense Disambiguation task.

## Acknowledgments

This work has been funded by the Spanish Ministry of Science, Innovation and Universities under the project DeepReading (RTI2018-096846-B-C21) (MCIU/AEI/FEDER,UE) and by the BBVA Big Data 2018 “BigKnowledge for Text Mining (BigKnowledge)” project. We also acknowledge the support of the Nvidia Corporation with the donation of a GTX Titan X GPU used for this research.

## References


Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising the wordnet domains hierarchy: semantics, coverage and balancing. In *Proceedings of the workshop on multilingual linguistic resources*, pages 94–101.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jose Camacho-Collados and Roberto Navigli. 2017. BabelDomains: Large-scale domain labeling of lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 223–228, Valencia, Spain. Association for Computational Linguistics.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36 – 64.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

Aitor González, German Rigau, and Mauro Castillo. 2012. A graph-based method to improve wordnet domains. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 17–28. Springer.

Aitor Gonzalez-Agirre, Mauro Castillo, and German Rigau. 2012. A proposal for improving wordnet domains. In *LREC*, pages 3457–3462.

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

B Magnini. 2000. G. cavagli a. integrating subject field codes into wordnet. In *Proceedings of LREC-2000, 2nd International Conference on Language Resources and Evaluation*, pages 1413–1418.

Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, and Alfio Gliozzo. 2002. The role of domain information in word sense disambiguation. *Natural Language Engineering*, 8(4):359–373.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Liang Wang, Wei Zhao, Ruoyu Jia, Sujian Li, and Jingming Liu. 2019. Denoising based sequence-to-sequence pre-training for text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4003–4015, Hong Kong, China. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018a. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018b. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.