What do Language Models know about word senses? Zero-Shot WSD with Language Models and Domain Inventories

Oscar Sainz , Oier Lopez de Lacalle , Eneko Agirre , German Rigau

HiTZ Basque Center for Language Technologies - Ixa NLP Group
University of the Basque Country UPV/EHU
{oscar.sainz, oier.lopezdelacalle, e.agirre, german.rigau}@ehu.eus

Abstract

Language Models are the core for almost any Natural Language Processing system nowadays. One of their particularities is their contextualized representations, a game changer feature when a disambiguation between word senses is necessary. In this paper we aim to explore to what extent language models are capable of discerning among senses at inference time. We performed this analysis by prompting commonly used Languages Models such as BERT or RoBERTa to perform the task of Word Sense Disambiguation (WSD). We leverage the relation between word senses and domains, and cast WSD as a textual entailment problem, where the different hypothesis refer to the domains of the word senses. Our results show that this approach is indeed effective, close to supervised systems.

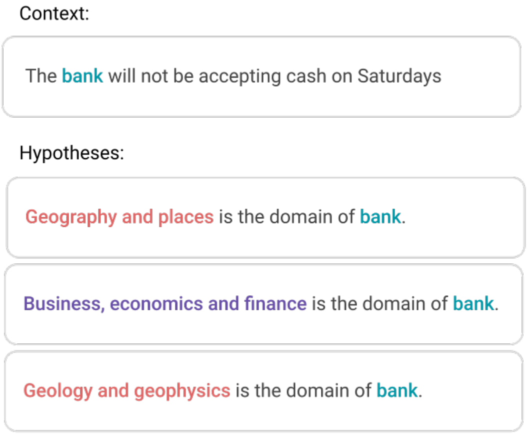
Context:
The bank will not be accepting cash on Saturdays
Hypotheses:
Geography and places is the domain of bank.
Business, economics and finance is the domain of bank.
Geology and geophysics is the domain of bank.

Figure 1: An example of the Word Sense Disambiguation task converted to Textual Entailment, where the hypothesis refer to the possible domains of word senses. To solve the task a model would be asked to select the most probable hypothesis based on the context.

1 Introduction

It is undeniable that Language Models (LM) have drastically changed the Natural Language Processing (NLP) field (Min et al., 2021). More recently, those LM have also shown to be capable of performing NLP tasks with just few examples given in the context (Brown et al., 2020), using the so called *prompting*. One of their particularities, and the key difference with previous approaches, is their contextualized token representation. Allowing the model to adopt different representations for words (tokens) depending on the context has supposed a huge advantage when sense disambiguation is required for a given inference. But, **to what extent do LM actually know about word senses?** In this work, we tried to answer that question by evaluating LMs directly on the Word Sense Disambiguation (WSD) task via prompting.

Word Sense Disambiguation is the task of identifying the correct sense of a word in a given context. Current state-of-the-art on WSD involves fine-tuning a LM on SemCor (Miller et al., 1994) to predict the correct among all possible sense glosses of the word in the given context. Other methods leverage the contextual representations of LM to perform WSD with a simple K-NN algorithm on the embedding space. Lately, the use of domain inventories was proposed to alleviate the high granularity of knowledge-bases (Lacerra et al., 2020). Recent studies that worked on zero-shot WSD refer to the task of predicting the senses of new lemmas not seeing during training as zero-shot (Lacerra et al., 2020) WSD, however we aim for a completely zero-shot evaluation, where no annotated data is available for any lemma.

Despite the knowledge already encoded in the LM, training data is used in one way or another to introduce knowledge about the task. To avoid drawing noisy conclusions, we evaluated the LM as they are, without further fine-tuning on or using any kind of WSD training data. To that end, we prompted LMs like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) to perform a task

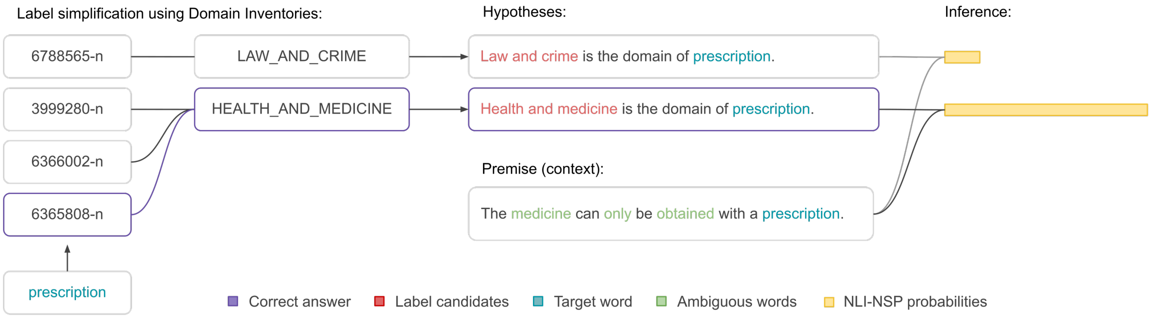

Figure 2: Graphical description of the zero-shot WSD approach using Domain Inventories.

that requires WSD knowledge to be successfully solved.

Figure 1 shows an example of how a model can be prompted to solve WSD using Textual Entailment as a proxy. On this example we consider that the word bank has senses from three different domains: *Geography and places*, *Business, economics and finance* and *Geology and geophysics*. The three possible domains are converted to hypothesis using predefined prompts. Finally, a supervised Textual Entailment model is used to perform the inference. More details on of the approach are discussed in Section 2.

In this work we first evaluated commonly used LMs as a zero-shot domain labelers with 3 different domain inventories. Then, following (Lacerra et al., 2020) we addressed the WSD using domain inventories and evaluated the LMs on them. We showed that LMs have some notion of senses as they perform zero-shot WSD significantly better than a random baseline and sometimes close to the supervised state-of-the-art. We also provided different analysis comparing different prompts and performing an error analysis over the two evaluated tasks.

2 Prompting Language Models

Since the past few years, prompting has become the *de facto* approach to probe language models (Li et al., 2022b). Min et al. (2021) defined prompting as the practice of adding natural language text, often short phrases, to the input or output to encourage pre-trained models to perform specific tasks. However, due to its wide definition, several different ways of prompting exists, such as *instruction based*, *template-based* or *proxy-task based*. For more information about prompting we encourage the reader to read the Liu et al. (2022a) survey.

In this work we focused on the *proxy-task based* approach, more precisely, we made use of the Next Sentence Prediction (NSP) and Textual Entailment (TE) tasks as a proxy. The TE is also known as Natural Language Inference (NLI), we will use both terms interchangeably. The choice of this approach was made based on previous works on zero-shot domain labelling (Sainz and Rigau, 2021).

Both, NSP and TE are sentence-pair classification tasks: the first attempts to predict whether a sentence is followed by another and the second aims to predict if an entailment relation exists between both sentences (premise and hypothesis). Figure 2 shows an example of how to perform WSD using NSP or TE models. The process can be briefly summarized as follows: (1) for each possible sense s of the target word w we obtain their corresponding domain d using a domain inventory D (domain inventories are discussed in more detail in Section 3). (2) predefined prompts are used to generate verbalizations that will serve as possible continuations (on NSP) or hypothesis (on TE) h. (3) a pretrained NSP or TE model is used to obtain a probability for each sentence/hypothesis and therefore, to each domain. Formally, for a TE model we defined the probability of word w being from domain $d_i \in D^w$ in context c as follows:

$$P(d_i|c,w) = P(\text{entailment}|c,h_{wi}) \tag{1}$$

where h_{wi} is the hypothesis generated using a predefined prompt, the domain label d_i and the word w. Similarly, for a NSP model the probability is defined as follows:

$$P(d_i|c,w) = P(\text{is\_next}|c,h_{wi}) \tag{2}$$

Table 2 shows the prompts used for probing Language Models in Domain Labelling and Word

Sense	BabelDomains	CSI	WN Domains	Gloss
00006484-n	Biology	Biology	biology	The basic structural and functional unit of all organisms; ...
02991048-n	Chemistry and mineralogy	Craft, Engineering and Technology	electronics	A device that delivers an electric current as the result of a chemical reaction.
02992529-n	Computing	Craft, Engineering and Technology	electricity telephony	A hand-held mobile radiotelephone for use in an area divided into small sections, each with its own short-range transmitter/receiver

Table 1: Example of Domain inventories for 3 senses of the word *cell*.

Sense Disambiguation tasks.

3 Domain Inventories

A domain inventory is a set of domain labels such as *Health and Medicine*, *Culture* or *Business and economics* that aims to cover the wider spectrum of domains as possible with a specific granularity level. Actually, these domain inventories are used to label synsets from knowledge-bases like WordNet (Fellbaum, 1998) and BabelNet (Navigli and Ponzetto, 2012). Examples of WordNet synset annotations from different domain inventories are shown in the Table 1. Recent studies (Lacerra et al., 2020) suggest to use domain inventories to address the high granularity problem that affects WSD tasks. In this section we describe the three domain inventories on which we evaluated the Language Models.

BabelDomains (Camacho-Collados and Navigli, 2017) is a unified resource that includes domain information for Wikipedia, WordNet and BabelNet. It inherits the domains from Wikipedia domains of knowledge, a total of 34 coarse labels. Although it is semi-automatically annotated, two gold standard datasets (for WordNet and Wikipedia) are provided for evaluation.

Coarse Sense Inventory (CSI) (Lacerra et al., 2020) was created to reduce the level of granularity of WordNet synsets while maintaining their expressiveness. It contains a total of 45 labels shared across the lexicon. Compared to previous alternatives, CSI provided a higher agreement among annotators. Also it was already proven to be useful for the WSD task.

WordNet Domains (Bentivogli et al., 2004) is a fine-grained domain inventory containing about 160 labels. It is organised in a hierarchical way, from global concepts such as *pure_science* to specific concepts as *oceanography*. This inventory provides a domain label to each synset in WordNet. Due to the hierarchical nature and fine granularity, in our experiments we kept only the domain labels until the third level, mapping all the labels below to the closest available domain. We end up with 60 domain labels.

4 Experimental Setup

In this section we describe the models we evaluated, and the Domain Labelling and Word Sense Disambiguation tasks we used for evaluation.

Models. For the experiments we decided to evaluate two very commonly used models: BERT and RoBERTa. We followed previous works on zero-shot domain labelling (Sainz and Rigau, 2021) for approach and model selection. As explained in Section 2 we required that the models were already fine-tuned to perform sentence pair classifications. In the case of the BERT models, we used the LM itself with the NSP head that was trained during pre-training, in the tables it is shown as NSP. For the case of RoBERTa, as it has not been pre-trained for any sentence classification task, we evaluated two checkpoints that were also fine-tuned with TE data: NLI and NLI*. The main difference between both checkpoints is the variety of data on which the models were trained. We evaluated the *large* variant of those models. The NLI variation was trained just on MultiNLI (Williams et al., 2018) dataset and NLI* variations was also trained on SNLI (Bowman et al., 2015), Fever-NLI (Thorne et al., 2018) and Adversarial-NLI (Nie et al., 2020). Both models are publicly available at HuggingFace Model Hub (Wolf et al., 2020).

Domain Labelling task is the task of classifying some text t into a set of domain labels D. In

Task	Prompt
Domain Labelling	{gloss} | The domain of the sentence is about {label}.
Word Sense Disambiguation	{context} | The domain of the sentence is about {label}.
	{context} | {label} is the domain of {word}.

Table 2: Prompts used for probing Language Models.

our case, the text to classify are WordNet synset glosses and the domain labels are the ones defined by the domain inventories. The task was evaluated on a small manually annotated dataset released by Camacho-Collados and Navigli (2017). The dataset consist of domain annotations for 1540 WordNet synsets using BabelDomains inventory. For those 1540 synsets we also collected the domain information from CSI and WordNet Domains. The 3 checkpoints described above were evaluated with each domain inventory. To evaluate the models on domain labelling data we used the prompts described in Table 2 to convert domain labelling examples into NLI or NSP examples. The prompt is used to generated as many hypotheses as labels are in the inventory, by replacing the *gloss* placeholder with the synset's gloss and the *label* placeholder with the corresponding label each time.

> Cell: **(biology)** the basic structural and functional unit of all organisms; ...

Figure 3: An example of WordNet gloss. The hint in the gloss is highlighted.

WordNet glosses sometimes contains domain information inside them. For example, in the gloss shown in Figure 3 the domain information is highlighted in bold. We will call them domain *hints*. As we are using those glosses as inputs to predict the domain of the synsets, the hints give a huge advantage to the models. Therefore, for the evaluation we considered two alternatives: with and without hints.

WSD task is the task of identifying the correct sense s a word w withing a context c among all its possible senses $s \in S^w$. In this case, and following recent works we reframed the task from predicting senses to more coarse set of labels (domains) (Lacerra et al., 2020). Therefore, the task aims to classify the domain of the correct sense d_s among the domains of the possible senses D^w. As senses in WordNet are very fine-grained, several senses of the same domain may coexist, after replacing them with their domain the set of possible labels might be reduced, therefore $|D^w| \leq |S^w|$. An example of two senses from the same domain is shown in Table 3. The task was evaluated on the standard commonly known SemEval (Pradhan et al., 2007; Navigli et al., 2013; Moro and Navigli, 2015) and Senseval (Edmonds and Cotton, 2001; Snyder and Palmer, 2004) datasets. For each model, we also compared two different prompts shown in Table 2: the first is the same as the one used for Domain Labelling and is used to predict the domain of the whole context; the second instead adds a reference to the target word, and is intended to focus the model to predict the domain of the given word withing the context. Finally, we report a random guessing baseline and a supervised upper-bound from Lacerra et al. (2020).

5 Results

In this section we discuss the results obtained on each experiment. First we discuss the results obtained on the Domain Labelling task. Then, we show the results from Word Sense Disambiguation. And finally we analyze the correlation between both tasks as they share the label space.

Are Language Models able to discriminate domains in sense glosses? Figure 4 shows the results obtained for the Domain Labelling task. As a general overview, the three models obtain decent results considering no data for training was provided. Comparing NLI models vs the NSP model, we can conclude that NLI based models perform better in all cases, in concordance with previous works (Wang et al., 2021a). However, additional TE data (NLI vs NLI*) does not seem to be very useful for the task. Finally, the results shows that the domain hints in the gloss affects significantly to the performance, specially in WordNet Domains, where the labels are very fine-grained.

Do Language Models know about Word Senses? Figure 5 shows the results for each of the WSD

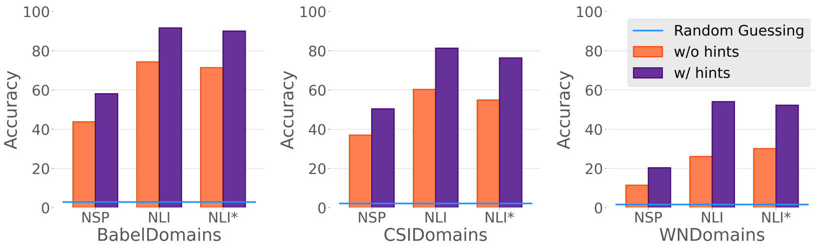

Figure 4: Results on Domain Labelling task for three different domain inventories.

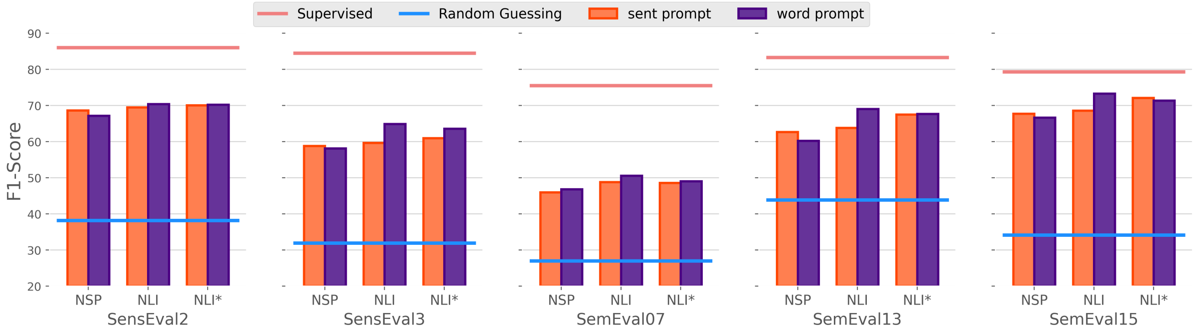

Figure 5: Word Sense Disambiguation results for the three systems in the 5 evaluation datasets. The red line indicates the state-of-the-art supervised scores and the blue line the scores obtained by random guessing.

datasets along with random and supervised baselines. In general, the results suggest that **in fact the Language Models know about senses**. While still far from a supervised upper-bound, the three models have shown significantly better performance than a random classifier. Moreover, for the SemEval-15 task the models achieve a performance close to the upper-bound. Comparing the NSP model against the NLI models, the same pattern as in the Domain Labelling task occur, the NLI models are better in all scenarios. If we compare both TE models, both perform similarly when the *sentence prompt* is used, for the *word prompt* instead the NLI model shows slightly better results. Overall, the best combination is NLI model with the *word prompt*.

Model	Noun	Adj	Verb	Adv	All
Random	40.7	48.4	23.7	59.1	38.8
Sentence prompt					
NSP	60.3	84.9	50.4	86.6	62.6
NLI	64.3	86.2	54.8	86.4	66.1
NLI*	65.0	85.9	55.0	85.3	66.4
Word prompt					
NSP	59.4	84.8	50.2	86.4	61.9
NLI	**66.2**	**86.8**	**57.0**	**87.3**	**67.8**
NLI*	65.3	85.5	55.7	85.5	66.8

Table 3: F1-Scores per word category

Do Language Models perform differently depending on the word category? To answer this question we report the results grouped by the word category in the Table 3. The table reports the same results as Figure 5 except for the supervised upper-bound which has not been reported by Lacerra et al. (2020) under this setting. We also report the *micro-averaged* F1-Score for all categories, allowing us to clearly compare all the systems. Considering the results, the NLI model with the *word prompt* is again the best performing system across all word categories. Comparing the NLI$_{\text{word}}$ model against the random baseline we can observe a high correlation in the scores, which suggest that the errors on each category depend more on the task difficulty rather than specific language model issues.

To what extent does the performance on Domain Labelling affects WSD? As we are framing WSD as a Domain Labelling problem, it is intuitive to think that the performance on Domain Labelling can affect the performance on WSD. The evaluation we carried out in both tasks have a common label space, and therefore, we can compute the correlation between label scores. For each label,

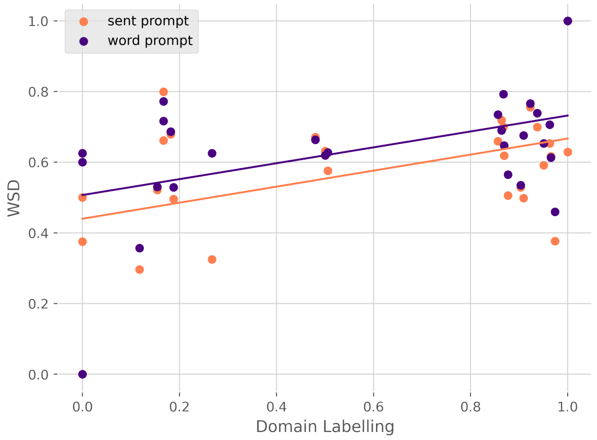

Figure 6: F1 correlation between Domain Labelling and WSD tasks.

	Dom Lab.	WSD_{sent}	WSD_{word}
Dom Lab.	1.00	0.32	0.41
WSD_{sent}	0.32	1.00	0.81
WSD_{word}	0.41	0.81	1.00

Table 4: Spearman's correlation of F1-Scores between tasks using shared labels. The scores correspond to the NLI model.

we compared the F1-score obtained on Domain Labelling and WSD tasks. Figure 6 shows the per-domain F1 scores on Domain Labelling and WSD tasks, each point represents the F1 obtained on a specific label. In the figure, we included the F1 for both *sentence prompt* and *word prompt* systems. The results shows **very little correlation** between both tasks. The Table 4 shows the Spearnman's correlation for each task pair. The results again shows that both tasks are poorly correlated, even when we use the same prompt. However, this comparison might not be completely fair, there are 2 main reasons that could affect the results: the Domain Labelling glosses have a particular structure and different from WSD contexts, also, on WSD the system needs to predict the correct among **possible** labels rather than all the label space as in Domain Labelling. We should take into consideration those differences at the time of interpreting the results.

6 Related Work

Word Sense Disambiguation Approaches to WSD range from supervised that only use annotated data (Agirre et al., 2014; Hadiwinoto et al., 2019; Bevilacqua and Navigli, 2019) to knowledge-based (Moro et al., 2014; Agirre et al., 2014; Scozzafava et al., 2020), as well as approaches that combine supervised and knowledge-based approaches (Kumar et al., 2019; Bevilacqua and Navigli, 2020; Blevins and Zettlemoyer, 2020; Conia and Navigli, 2021; Barba et al., 2021).

Knowledge-based approaches employ graph algorithms on a semantic network (Moro et al., 2014; Agirre et al., 2014; Scozzafava et al., 2020), in which senses are connected through semantic relations and are described with definitions and usage examples. Unfortunately, their independence from annotated data comes at the expense of performing worse than supervised models (Pilehvar and Navigli, 2014).

Supervised approaches frame the task as a classification problem and use available annotated data to learn mapping the words in context to senses. Before supervised neural models emerged as state of the art in NLP, the task of supervised WSD was performed based on a variety of lexico-syntantic and semantic feature representations that are fed to a supervised machine learning classifier (Zhong and Ng, 2010). Instead, current state-of-the-art supervised models rely on the use of pretrained Transformers as core architecture of the model. Among these models we can find approaches that exclusively use annotated data to learn effective representations of the target word in context and feed it to some classification head (Raganato et al., 2017; Hadiwinoto et al., 2019; Bevilacqua and Navigli, 2019; Conia and Navigli, 2021).

Some approaches have shown that an effective way to improve sense representation is to exploit the glosses provided by the sense inventories. Gloss representation are then incorporated to the sense embedding (Peters et al., 2018), in which the most probable sense is retrieve according to the similarity with the given context. Multiple works have been shown effective in WSD such as LMSS (Loureiro and Jorge, 2019), SensEmBERT (Scarlini et al., 2020a), ARES (Scarlini et al., 2020b), SREF (Wang and Wang, 2020), EWISE (Kumar et al., 2019) and EWISER (Bevilacqua and Navigli, 2020), among many others. Glosses have also been exploited in sequence-tagging approaches (Huang et al., 2019; Yap et al., 2020), where the task is framed as sequence classification problem (Barba et al., 2021). In a similar manner, (Bevilacqua and Navigli, 2020) propose a generative approach to cast WSD as sequence classification problem. In adition to glosses, other approaches presented ways to make use of the knowledge encoded in KBs such as WordNet. For instance, (Loureiro and Jorge, 2019; Wang and Wang, 2020) propagate sense embeddings using WordNet as a graph. Please refer to (Bevilacqua et al., 2021) to obtain further details of the recent trends in WSD.

Prompting Language Models has changed the paradigm of how Language Models can be used to extract even more potential from them. Initially with very large LM like GPT-3 (Brown et al., 2020) and later with smaller ones (Gao et al., 2021) prompts allowed the models to perform zero or few-shot classifications with simple natural language. This ability also allowed models to improve performance on data-scarce problems by large margin (Le Scao and Rush, 2021; Min et al., 2021; Liu et al., 2022a). These prompts can be discrete (Gao et al., 2021; Schick and Schütze, 2021a,b,c) close to natural language or continuous (Liu et al., 2022b) close to other efficient deep learning methods like Adapters (Pfeiffer et al., 2020). Closer to our work, Textual Entailment (Dagan et al., 2006) has been used as a source of external supervision to solve several text classification tasks (Yin et al., 2019, 2020; Wang et al., 2021b; Sainz and Rigau, 2021; McCann et al., 2018; White et al., 2017), Named Entity Recognition (Li et al., 2022a; Poliak et al., 2018; Yang et al., 2022), Relation Extraction (Levy et al., 2017; Sainz et al., 2021), Event Extraction (Lyu et al., 2021), Event Argument Extraction (Sainz et al., 2022a,b), Intent Classification (Xia et al., 2021), Aspect-based Sentiment Analysis (Shu et al., 2022) and many more.

Domain Inventories. Domain information was added to Princeton WordNet (Fellbaum, 1998) since version 3.0. In total 440 topics were represented as a synsets in the graph. The topic label assignment was achieved through pointers from source synsets to target synsets. Being the most frequent topic is LAW, JURISPRUDENCE. However, the manual assignment of topic labels to synsets in WordNet is very costly. As a consequence, semi-automatic methods were developed. For instance, WordNet Domains (Bentivogli et al., 2004) is a semi-automatically annotated domain inventory that labels WordNet synsets with 165 hierarchically organised domains. The use of domain inventories such as WordNet Domains, allowed to reduce polysemy degree of WordNet synsets by grouping those that belong to the same domain (Magnini et al., 2002). However, far from being perfect, many synsets were labelled as FACTOTUM, meaning that the synset cannot be labelled with a particular domain. Several works were proposed to improve WordNet Domains, such as eXtended WordNet Domains (González-Agirre et al., 2012; González et al., 2012), that applied graph-based methods to propagate the labels through the WordNet structure.

Domain information is not only available in WordNet, for example IATE[1] is a European Union inter-institutional terminology database. The domain labels of IATE are based on the Eurovoc thesaurus[2] and were introduced manually. More recently, several new domain inventories appeared, such as BabelDomains (Camacho-Collados and Navigli, 2017) or Coarse Sense Inventory (Lacerra et al., 2020).

7 Conclusions

In this work we present an evaluation approach to test Language Models on the tasks of Domain Labelling and Word Sense Disambiguation without annotated data requirements. For the WSD task we followed Lacerra et al. (2020) to reduce the granularity level. Our results showed that the Language Models we tested here **have some notion of word senses**. They easily outperformed the baseline, and sometimes almost reached to supervised systems performance. In addition, our further analysis shows that there is very low error propagation from Domain Labelling to WSD as their errors are poorly correlated. For the future, we plan to evaluate larger Language Models on the task to try to understand to what extent scaling these LMs affects to sense recognition.

Acknowledgments

Oscar is funded by a PhD grant from the Basque Government (PRE_2020_1_0246). This work is based upon work partially supported via the IARPA BETTER Program contract No. 2019-19051600006 (ODNI, IARPA), DeepKnowledge (PID2021-127777OB-C21) project funded by MCIN/AEI/10.13039/501100011033 and by FEDER Una manera de hacer Europa,

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

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

AWARE (TED2021-131617B-I00) project funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGeneration EU/ PRTR, and by the Basque Government (IXA excellence research group IT1570-22).

References

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. ESC: Redesigning WSD with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online. Association for Computational Linguistics.

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, Geneva, Switzerland. COLING.

Michele Bevilacqua and Roberto Navigli. 2019. Quasi bidirectional encoder representations from transformers for word sense disambiguation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 122–131, Varna, Bulgaria. INCOMA Ltd.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online. Association for Computational Linguistics.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Samuel R. 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, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

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.

Simone Conia and Roberto Navigli. 2021. Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3269–3275, Online. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Toulouse, France. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. 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 González-Agirre, Mauro Castillo, and German Rigau. 2012. A proposal for improving WordNet domains. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3457–3462, Istanbul, Turkey. European Language Resources Association (ELRA).

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5297–5306, Hong Kong, China. Association for Computational Linguistics.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. 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 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681, Florence, Italy. Association for Computational Linguistics.

Caterina Lacerra, Michele Bevilacqua, Tommaso Pasini, and Roberto Navigli. 2020. CSI: A coarse sense inventory for 85% word sense disambiguation. In *Proceedings of the 34th Conference on Artificial Intelligence*, pages 8123–8130. AAAI Press.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Bangzheng Li, Wenpeng Yin, and Muhao Chen. 2022a. Ultra-fine entity typing with indirect supervision from natural language inference. *Transactions of the Association for Computational Linguistics*, 10:607–622.

Jiaoda Li, Ryan Cotterell, and Mrinmaya Sachan. 2022b. Probing via prompting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1144–1157, Seattle, United States. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2022a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* Just Accepted.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022b. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

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.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy. Association for Computational Linguistics.

Qing Lyu, Hongming Zhang, Elior Sulem, and Dan Roth. 2021. Zero-shot event extraction via transfer learning: Challenges and insights. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 322–332, Online. Association for Computational Linguistics.

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.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado. Association for Computational Linguistics.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

Roberto Navigli and Simone Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Matthew E. 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.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Roberto Navigli. 2014. A large-scale pseudoword-based evaluation framework for state-of-the-art word sense disambiguation. *Computational Linguistics*, 40(4):837–881.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167, Copenhagen, Denmark. Association for Computational Linguistics.

Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. 2022a. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2439–2455, Seattle, United States. Association for Computational Linguistics.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oscar Sainz, Haoling Qiu, Oier Lopez de Lacalle, Eneko Agirre, and Bonan Min. 2022b. ZS4IE: A toolkit for zero-shot information extraction with simple verbalizations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 27–38, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Oscar Sainz and German Rigau. 2021. Ask2Transformers: Zero-shot domain labelling with pretrained language models. In *Proceedings of the 11th Global Wordnet Conference*, pages 44–52, University of South Africa (UNISA). Global Wordnet Association.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020a. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8758–8765.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020b. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 390–402, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021c. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Federico Scozzafava, Marco Maru, Fabrizio Brignone, Giovanni Torrisi, and Roberto Navigli. 2020. Personalized PageRank with syntagmatic information for multilingual word sense disambiguation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–46, Online. Association for Computational Linguistics.

Lei Shu, Hu Xu, Bing Liu, and Jiahua Chen. 2022. Zero-shot aspect-based sentiment analysis.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Ming Wang and Yinglin Wang. 2020. A synset relation-enhanced framework with a try-again mechanism for word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6229–6240, Online. Association for Computational Linguistics.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021a. Entailment as few-shot learner.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021b. Entailment as few-shot learner.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. 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, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Congying Xia, Wenpeng Yin, Yihao Feng, and Philip Yu. 2021. Incremental few-shot text classification with multi-round new classes: Formulation, dataset and system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1360, Online. Association for Computational Linguistics.

Zeng Yang, Linhai Zhang, and Deyu Zhou. 2022. SEE-few: Seed, expand and entail for few-shot named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2540–2550, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Boon Peng Yap, Andrew Koh, and Eng Siong Chng. 2020. Adapting BERT for word sense disambiguation with gloss selection objective and example sentences. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 41–46, Online. Association for Computational Linguistics.

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.

Wenpeng Yin, Nazneen Fatema Rajani, Dragomir Radev, Richard Socher, and Caiming Xiong. 2020. Universal natural language processing with limited annotations: Try few-shot textual entailment as a start. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8229–8239, Online. Association for Computational Linguistics.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83, Uppsala, Sweden. Association for Computational Linguistics.