A WordNet View on Crosslingual Contextualized Language Models

Wondimagegnhue Tufa
CLTL Lab, VU Amsterdam
w.t.tufa@vu.nl

Lisa Beinborn
CLTL Lab, VU Amsterdam
l.beinborn@vu.nl

Piek Vossen
CLTL Lab, VU Amsterdam
p.t.j.m.vossen@vu.nl

Abstract

WordNet is a database that represents relations between words and concepts as an abstraction of the contexts in which words are used. Contextualized language models represent words in contexts but leave the underlying concepts implicit. In this paper, we investigate how different layers of a pre-trained language model shape the abstract lexical relationship toward the actual contextual concept. Can we define the amount of contextualized concept forming needed given the abstracted representation of a word? Specifically, we consider samples of words with different polysemy profiles shared across three languages, assuming that words with a different polysemy profile require a different degree of concept shaping by context. We conduct probing experiments to investigate the impact of prior polysemy profiles on the representation in different layers. We analyze how contextualized models can approximate meaning through context and examine cross-lingual interference effects.

1 Introduction

WordNet (Fellbaum, 1998) is a manually created database that relates the words of a language to concepts. Concepts are represented through synsets, based on a weak synonymy relation, whereas explicit semantic relations between synsets place these concepts in a semantic space. Words of a language can be positioned in that same space but this can become complex when they are ambiguous. A polysemous word such as ”star” can be represented in several positions of this space depending on its meaning.

Word embeddings (Mikolov et al., 2013) place words in a semantic space as well based on the dimensions of the vector that was derived when learning to predict their context words. Static word embeddings can be interpreted as an average across contexts, even when words occur with different meanings. For our example, this means that ”star” would be positioned somewhere in between *celebrity* and synonyms for the concept *celestial body* as a compromise across contexts.

More recent pre-trained Transformer-based Language Models (PTLM) such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019b) capture a more nuanced relationship between words and concepts by not only representing the vocabulary through embeddings but also distinguishing contexts: the word ”star” will be represented differently depending on the context in which it occurs. From an abstract point of view, these context-sensitive representations approximate a relation between words and concepts. Ethayarajh (2019) investigates this relationship by measuring the impact of contextualization on the representation of meaning through the layers of PTLMs, showing that representations of tokens in contextualized models deviate from their static initialization. The research by (Ethayarajh, 2019) is limited to monolingual models, which leaves open what relationship between tokens or words and concepts is captured in cross-lingual models where words and concepts are shared across languages.

In cross-lingual language models (XPTLMs) such as XLM-RoBERTa (Conneau and Lample, 2019), the challenge of contextualizing concepts is even more complex because of the additional cross-lingual ambiguity. The same word can be mapped to the same or to different concepts across languages. For example, the Dutch word ”star” is an adjective meaning *inflexible* whereas the translation for the English ”star” corresponds to ”ster” in both meanings. The Dutch language, therefore, adds ambiguity to the word-concept relationship of ”star”. As most XPTLMs use a shared vocabulary for all languages, the variation in meaning across languages can simply be interpreted as different

contexts for a word that needs to be encoded in the representations of the model.

In most multi-lingual wordnet databases, cross-lingual ambiguity is underrepresented because they are commonly build using the *expand*-method (Vossen, 1998). This means that the English representation of concepts is maintained and cross-lingual links are established by mapping the vocabulary of the new language onto the existing concept taxonomy. This approach hampers research to the universality of concepts in wordnet models (Vossen and Fellbaum, 2009) but it has been applied widely because of its clear practical advantages over the alternative *merge*-method that requires intense manual labor. XPTLMs can be constructed in different ways as well, which partially mimics the difference between the *expand* and the *merge* approach: 1) *expanding* a monolingual PTML with static lexical embeddings for target languages while freezing the other layers (Artetxe et al., 2019) or 2) training a model from texts from all languages (Conneau and Lample, 2019) so that all languages contribute conceptual representations as contexts (*a merged approach*).

In this paper, we argue that XPTLMs provide new opportunities to move beyond the conceptual limitations of multilingual wordnet databases built through the expand method. We provide empirical evidence for the impact of languages on a shared conceptual XPTLM for both the lexical and conceptual levels by measuring to what extent sharing tokens in XPTLMs has a positive or negative impact on representing concepts and to what extent the contexts in which these words occur compensate for any disturbances in the token representation. In other words: to what extent is the representation of "star" a compromise across all language meanings and to what extent is it defined by the cross-lingual contexts in which it occurs? XPTLMs use a shared vocabulary for all languages to exploit semantic commonalities across languages (cognate effects). However, cross-lingual differences caused by semantic drift (Beinborn and Choenni, 2020) can contribute to semantic interference (Lauscher et al., 2020).

More specifically, we will address the following questions in our experiments:

- How consistent is the relationship between words and concepts with and without the influence of context for polysemous words?
- What are the effects of sharing vocabulary and contexts across languages on the representation of cross-lingual ambiguity?

In order to investigate the above questions, ideally a large sense-tagged parallel corpus would be required to identify a representative set of concepts shared across languages. Existing corpora (Bond et al., 2013) are however small and have skewed sense distributions. Another problem is that it is hard to determine the best level of granularity for identifying concepts associated with a word in contexts and they may not be distinguishable empirically through existing models (Ethayarajh, 2019). Instead of multilingual corpora with WordNet senses or all contextualized contexts, we, therefore use a controlled set of semantic classes as the representation of concepts following the work of (Zhao et al., 2020). Entity types such as *PERSON*, *ORGANIZATION*, and *LOCATION* can be seen as coarse-grained concepts for which large datasets exist. We use the XLEnt dataset (El-Kishky et al., 2021) which contains 160 million aligned entity pairs in 120 languages paired with English. We investigate how well the entity in this data are distinguished by contextualized models in contextualized layers.

Our contributions are:

- A probing method for measuring the lexical (token) and contextual (model) effects of languages within various cross-lingual models.
- Pilot results on cross-lingual interference and support effects for the typologically related languages English, German and Dutch.
- Pilot results for cross-lingual zero-shot probing for German, Dutch, Arabic, and Amharic.

The paper is further structured as follows. In the next section 2, we describe related work, especially on semantic probing of distributional models. After that, we describe in Section 3 our methodology, which is based on (Zhao et al., 2020) but applied to multilingual models. The dataset that we use is described in Section 4 and our experimental results are described in Section 5. We discuss the results and conclude in Section 6.

2 Related work

Analyzing the representational structure of contextualized models has become an essential means

towards developing more transparent and interpretable AI models, for example in the BlackboxNLP workshop which is reaching its 5th edition this year (Bastings et al., 2021). However, only a limited amount of research has been done to investigate the relationship between the vocabulary of such models and the degree of context dependency of the concepts that are associated with the words in the vocabulary. In Ethayarajh (2019), this relationship is investigated by measuring the impact of contextualization on the representation of meaning through the layers of language models. This study indicates that 1) contextualized models do enhance the meaning compared to the static initialization of the token and 2) there is no finite and discrete set of representations (thus concepts) for single tokens across concepts.

Artetxe et al. (2019) show that it is possible to transfer an English transformer to a new language by freezing all the inner parameters of the network and learning a new set of embeddings for the new language through masked language modeling. This works because the frozen transformer parameters constrain the resulting representations to become aligned with English. This approach does not adapt the concept representation established for the original language English. It only learns the token embedding using the English concept model and is thus comparable to the multilingual wordnet expand model (which uses a single English concept space and learns token mappings to another language). It is not possible to learn new concepts from another language nor adapt biases learned from the English data. Phenomena of semantic drift across languages (Beinborn and Choenni, 2020) can therefore not be captured and it remains unclear how the addition of languages affects the conceptual distribution beyond the performance on the downstream tasks.

For analyzing how a contextual language model captures the relationship between a word and a concept, we can use word sense disambiguation as a proxy task. The task evaluates model performance in associating an ambiguous word with the correct concept from the possible concept inventory. For example, the word "state" could represent a 'government' or the concept corresponding to the WordNet synset called "a way something is". One limitation of using such an approach is the granularity of the sense category. WSD categories are often too fine-grained and allow only limited abstraction (Izquierdo et al., 2009).

We opt for a task on a higher abstraction level and apply semantic class-based probing to quantify the contextualization capability of a language model using Wiki-PSE in line with Zhao et al. (2020). Wiki-PSE contains tokens used in contexts corresponding to different semantic classes. For example, the word 'apple' can refer to a technology company corresponding to the 'Organization' class or it can refer to a fruit belonging to the 'Food' class (Yaghoobzadeh et al., 2019). A concept-tagged dataset can be used to investigate relationships between a word form and a concept in a language model in a simplified setup: word forms are limited to entity names and their semantic classes define the concept inventory.

Probing has been established as a tool to test whether linguistics information is encoded in language model representations (Adi et al., 2016; Belinkov et al., 2017b; Tenney et al., 2019). Adi et al. (2016) train a classifier to predict sentence characteristics such as length, semantic information, and word order from sentence representation. Higher performance in the classification task indicates that information about the measured property is encoded in the embedding. Liu et al. (2019a) extend the probing tasks to a wider range of linguistic phenomena such as coreference, semantic relations, and entity information. Tenney et al. (2019) introduced edge probing and establish a standard format to quantify the availability of linguistic structure in pre-trained language models using various NLP benchmark tasks.

Our work follows Zhao et al. (2020) in that we use sentence probing to measure the relationship between a word, its context, and the corresponding concept. We extend this approach to various multilingual models instead of English BERT. We present pilot experiments to explore the utility of using semantic class probing with these multilingual models.

3 Methodology

To analyze how language models capture the relationship between words and concepts, we identify words that illustrate edge cases for the relation between concepts and contexts: 1) a monosemous (**mono**) relation between a word and a single concept, 2) **balanced** polysemous relations between a word and multiple concepts, and 3) **skewed** polysemous relations where one concept is dominant

in language use. We expect that the patterns in concept distribution are reflected in the probing performance of the cross-lingual models.

Our approach represents only a rough approximation of the set of concepts related to a word as well as the distribution of concepts in language use. The actual range of concepts is unknown and is the result of the pretraining of the model. Our pilot experiments, therefore, explore whether the large-scale annotations of XLEnt can serve as a proxy for probing word-concept relationships in multilingual models. We assume that such data provide sufficient information on the relation between words and concepts to measure the degree of ambiguity and the capability of models to identify concept relations from contexts. We hypothesize that our observations for a selected set of words can be generalized to a larger sample, which should be tested in future research.

In the probing setup, the model representation built during pretraining is not changed and can be tested for its capacity to represent a concept in target contexts at different layers. We assume that the lexical initialization in the first layer will reflect the prior ambiguity of the word in the pretraining data and that the integration of context will adjust the representation toward the target concept in higher layers. We expect the following observations for the respective profiles:

1. mono: only minor differences between the lexical initialization level and higher contextual levels
2. skewed:
   (a) **matching** distribution for test cases: same as mono
   (b) **diverging** distribution for test cases: low probing accuracy on the lexical level, strong indications of concept sensitivity in higher levels
3. balanced: low probing accuracy at the lexical level, improved concept knowledge in higher levels in all cases but not as strong as for diverging

In our experiments below, we report on the results for skewed and balanced ambiguous words in English and across the language English, Dutch, and German. Our code is publicly available at https://github.com/cltl/probing-cross-linqual-model.

4 Data set and Experiment

For our experiments, we use entities and their respective semantic class as a proxy for a more general notion of words and concepts due to data available for many languages with a controlled number of concepts in the form of entity types as semantic classes. Specifically, we select a sample from XLEnt which contains 160 million entity mentions annotated with 10 classes in 120 languages (El-Kishky et al., 2021). We describe the selection procedure in the following subsections.

4.1 Pre-processing and Sampling

We include English, German, and Dutch in our analysis.[1] Table 1 shows the statistical summary of the total available data.

	EN	NL	DE
Sentences	17,942,551	12,429,622	5,512,929
Entities	4,219,046	6,737,100	2,917,688
Unique Entities	59,054	60,777	38,930
LOC	512,219	744,024	329,030
ORG	1,690,244	3,282,967	1,580,477
PER	2,016,583	2,710,109	1,008,181

Table 1: Statistics of entities distribution in XLEnt for English, Dutch, and German.

For each of these languages, we selected sentences from one of the three semantic classes: Location, Organization, or Person. We selected these semantic classes because they correspond to clearly-distinct classes which cannot easily be used interchangeably in the same sentence, as opposed to clear metonymically-related classes such as Organization and Product.

The distribution across language and semantic classes in XLEnt varies. To maintain similar distribution across our target languages, we, therefore, sampled an equal number of sentences for each semantic class.

From the total set of entity names, we selected a sample of clear cases with monosemous, balanced polysemous, and skewed polysemous relations. Furthermore, the selected names should occur as tokens in the English, Dutch, and German data set. This results in a subset of 21 names related to the concepts of Person, Organisation, and Location. In the appendix B, the complete list of names

[1] The main reason for choosing these three languages is that we have native and up-to-native knowledge of these languages. In future research, we will also apply the same tests to other languages.

is given with the distributions and the division over the three polysemy profiles: mono, skewed and balanced. Table 2 shows a few examples of entities that are shared across languages. From these examples, *Tasman* and *Aquarias* are skewed towards a location interpretation, whereas *Chimera* is skewed towards an organization and *Prana* is balanced. *Sirius* is underrepresented towards Person.

To classify the distribution of an entity as balanced or skewed, we first normalized the frequency distribution between 0 and 1 using the total frequency across all types. We then applied a threshold value to categorize it into balanced and skewed. For a threshold value of 0.95 (95%) or higher, we classified an entity as skewed to a particular semantic class. If an entity occurs in more than one semantic class in a comparative way (at 0.35 or higher), we classify it as a balanced case.

Shared Entity	LOC	ORG	PER
Tasman	13	5	5
Prana	12	19	16
Sirius	391	481	42
Chimera	10	85	17
Aquarius	124	11	59

Table 2: Sample of names for entities with sufficient coverage and different polysemy profiles in English, Dutch, and German.

Using the same threshold, we further distinguish between cases where Dutch and German have similar distributions as English and cases with different distributions. We applied a similar approach to compare the distribution of entities across languages by comparing the normalized frequency distribution of entities. We assume that similar cross-lingual distributions result in better representation for a target language, whereas diverging distributions confuse the model and result in poorer representations. Note that the words are shared across these languages and get the same lexical initialization.

Our predictions should generalize over the sampled names per polysemy profile. Our probing framework can be used to test any language model that covers these words and the languages from the dataset. The results tell us to what extent pretraining resulted in a bias for the lexical initialization and to what extent the model can correct for this using the context. Below, we apply our probing methodology to XLM-RoBERTa and mBERT as a cross-lingual model to capture the relation between a word and concepts. We also apply the test to English BERT itself for comparison. We can easily extend the test to others models that include the probing words in the vocabulary.

4.2 Probing Experiment

For our probing experiment, we use a simple one-layer perceptron (MLP) similar to (Zhao et al., 2020). We designed a three-class classifier by taking each of the three distinct semantic classes. Figure 1 shows the architecture of our probing classifier.

For the experiments, we use the list of entities, a set of context sentences where these entities occur, and the semantic class associated with the entity for each context. In our probing, we first take the target sentence and pass it through a cross-lingual language model to generate the contextual representation associated with the target entity and the sentence which contains the entity word. From the language model output, we use the representation from the input layer (layer-0), the middle layer (layer-3), and the last layer(layer-12) as input for our classifier.[2] We use layer-0 as the baseline since it is initialized with the lexical token representation of the language model and should exhibit a prior ambiguity profile. In the middle and last layers, we get representations of our target words that are modified by the context. We train and test our probing model with these representations to detect the semantic class for the names in context.

4.3 Baseline

One of the core challenges of a probing method is how to interpret the results of a probing classifier. Previous works compare the result of the classifier with different approaches including majority baselines (Belinkov et al., 2017a; Conneau et al., 2018), static word embeddings (Belinkov, 2022; Tenney et al., 2019) and a random baseline by training the probing classifier on a randomized version of the input feature (Zhang and Bowman, 2018; Tenney et al., 2019). In our work, we include three baselines to compare and interpret the result of our probing model.

5 Results

We first examine our probing setup for resolving conceptual ambiguity in English entities and next

[2]We choose the third layer because it gave the best performance in most of our experiments

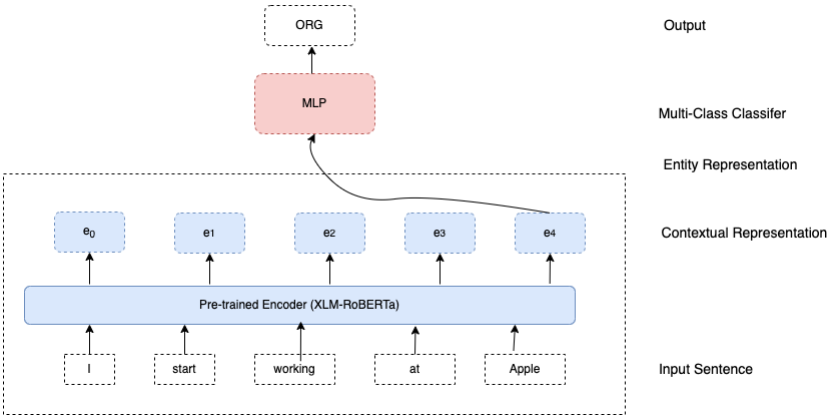


Figure 1: Architecture of our probing classifier

conduct additional analyses to examine the effect of shared tokens across multiple languages on conceptual ambiguity. Lastly, we conduct tests across typologically related and distant languages to check if a relationship learned between context and concept in one language is relevant for another language.

5.1 Probing Ambiguous Entities in English

In the first experiment, we focus on ambiguous entities in English and their representation in XML-RoBERTa. Entities are ambiguous if they have annotations for all three semantic classes in the data, either balanced or skewed towards one type as explained above. Table 3 shows the details of the distributions and the experimental results for the balanced and skewed cases respectively. Note that the train and test cases are randomly selected from the data and exhibit a similar distribution of balanced and skewed distribution. However, the test results are differentiated among them. For the balanced cases in Table 3, we see that layer-0 results are lowest, layer-3 are highest and layer-12 results are in between for all three concepts. Furthermore, location performs slightly better than organization and person. Looking at the skewed cases in Table 3, we see a similar pattern that results are lowest in layer-0, best in layer-3, and go down in layer-12. Overall, the results are better for skewed cases than for balanced cases at all levels, except for location. Remarkably, location performs lower than organization and person for the skewed cases.

The first conclusion we can draw here is that layers do correct for confusion at the lexical level by the context but some of this is lost in the higher levels. We can only partially confirm the prediction that balanced distributions are harder than skewed distributions. The prediction holds for organizations and persons, which perform lower for balanced than for skewed at all levels but not for location at layer-0 and layer-12. Apparently, the skewed cases are poorly represented for location at layer-0, which is correct in layer-3 (outperforming the balanced cases) but drops considerably in layer-12.

	LOC	**ORG**	**PER**
#Train	1506	1490	1504
#Test	494	510	496
#Single-Token Entity	252	779	1088
#Multi-Token Entity	1748	1221	912
Balanced			
#Test	417	334	314
Layer-0	0.65	0.58	0.52
Layer-3	0.81	0.78	0.79
Layer-12	0.78	0.75	0.75
Skewed			
#Test	77	176	182
Layer 0	0.61	0.75	0.76
Layer 3	0.86	0.87	0.9
Layer 12	0.78	0.82	0.86

Table 3: F1 scores for probing the different layers of XLM-RoBERTa on ambiguous entities. We run the experiment five times with seed from (0,1,2,3,4) Results are averaged over five runs. We observe a standard deviation between 0.003 and 0.009 For entities that are split into sub-tokens during tokenization, we took the mean of each of the vector embeddings

To investigate the impact of dominance on a concept at the lexical level, we differentiate the results for the skewed names into test cases that match the bias and cases that do not match. The results are shown in Table 4. We perform targeted analysis of the quantitative performance by explicitly distinguishing the dominant semantic classes. As can be expected, the probing performance for detecting

the location concept for names that predominantly occur with this concept is already very high already at layer-0 and increases further at layer-3 and layer-12. We observe the same pattern for the other two classes. We also see that the layer-0 performances for the non-dominant concepts are very low (from 0 to max .38), while the probing performance increases slightly in layer-3 and layer-12. The integration of context in the higher layers thus balances out the bias towards the dominant concept during initialization but not completely. The fact that the final scores are significantly lower shows that the lexical layer initialization does matter for obtaining optimal results. This also implies that confusion in a cross-lingual model created by sharing tokens across languages could result in poorer initialization in layer-0 that needs more repairing in the context-sensitive layers. We investigate the impact of such token or vocabulary sharing in the next subsection.

	LOC	ORG	PER
Skewed to LOC			
Layer-0	0.82	0.38	0.25
Layer-3	0.9	0.63	0.73
Layer-12	0.9	0.53	0.67
Skewed to ORG			
Layer-0	0.24	0.81	0.34
Layer-3	0.85	0.93	0.75
Layer-12	0.59	0.85	0.5
Skewed to PER			
Layer-0	0	0	0.97
Layer-3	0.67	0.29	0.97
Layer-12	0.67	0.25	0.97

Table 4: Result of probing the different layers of XLM-RoBERTa on entities that are skewed toward a specific semantic class. Result evaluated on F1-Score averaged over five runs

5.2 Probing Shared Entities across English, Dutch, and German

In the second experiment, we specifically probe entity names that are shared across the English, Dutch, and German data. We first select names that occur in all three languages. In the second step, we filter entities that are ambiguous across the three target classes. From these shared ambiguous entities, we identify two subcategories: 1) entities that have a similar type distribution in all three languages, and 2) entities that clearly exhibit a deviating distribution in both Dutch and German compared to English. For the first category, we expect that the shared distribution should improve the probing accuracy for English, and in the second category, we expect cross-lingual interference. Table 5 shows the details of the distribution and the experimental results. We observe the same consistent pattern of lowest probing performance on the lexical layer, highest performance for layer-3, and intermediate performance on layer-12. Our analyses indicate an impact of sharing tokens across languages. When Dutch and German have similar type distributions the results are substantially higher than when they have a different distribution. This holds for most results except for the organization class in layer-3 and layer-12.

Table 5 also shows that we can apply the same probing to other models such as BERT and mBERT, in this case only testing on English target sentences. We observe exactly the same patterns as for XLM-RoBERTa and even the scores are very similar, even for the BERT which was pre-trained on English data only.

Our results confirm that the representation in contextualized language models varies across layers. Concepts can be identified less well at the lexical level (layer-0) unless they match the dominant meaning, while higher levels integrate contextual information for further disambiguation. This indicates that lexical biases get repaired and that we can measure the degree to which this happens in line with the findings by Ethayarajh (2019). Our pilot experiment provides a proof of concept for analyzing the effect of the shared vocabulary on conceptual representations in cross-lingual contextualized language models. In future work, we hope to use this insight to improve such models for languages that are most affected by sharing vocabulary.

5.3 Cross-Lingual Evaluation

In this part of the experiment, we evaluated a probing model trained on an English dataset with test data from German, Dutch, Amharic, and Arabic. We first select monosemous and polysemous words by using the frequency distribution of entities and their types. Based on these distributions, we classify a word as monosemous if it belongs to one semantic class frequently. We applied a threshold value in such a way that if a word occurs 90% of the time as a single semantic class, we consider it a monosemous word. If a word occurs in two or more classes, we consider it a polysemous word.

	LOC	ORG	PER						
#Train	589	600	611						
#Test	199	212	189						
	XLM-RoBERTa			**BERT**			**mBERT**		
Similar	LOC	ORG	PER	LOC	ORG	PER	LOC	ORG	PER
Layer-0	0.76	0.67	0.78	0.75	0.66	0.77	0.74	0.59	0.79
Layer-3	0.83	0.83	0.84	0.84	0.85	0.86	0.82	0.83	0.84
Layer-12	0.81	0.78	0.85	0.82	0.83	0.86	0.85	0.83	0.87
Diverging									
Layer-0	0.57	0.53	0.42	0.53	0.51	0.4	0.54	0.51	0.45
Layer-3	0.78	0.82	0.68	0.82	0.84	0.71	0.76	0.81	0.72
Layer-12	0.75	0.78	0.66	0.8	0.81	0.77	0.81	0.79	0.78

Table 5: Result of probing the different layers of XLM-RoBERTa, BERT, and mBERT on entities that are shared between English, Dutch, and German with similar/diverging distribution across types. Results are evaluated in F1-Score and are averaged over five runs.

We then applied three filtering criteria: (a) We focus on single-word entities instead of multi-word entities to control ambiguity that might be introduced by multi-word entities. (b) We only include sentences with a single target entity to control contextual information that might be associated with an additional entity. (c) We restrict our selection to entities that are labeled as one of the four semantic classes LOCATION, ORG, PERSON, and EVENT since these can barely be used interchangeably.

Zero-shot probing We train a multi-class probing classifier using the English dataset and the setting discussed in Section 4.2 and test it on randomly sampled sentences from each of the four semantic classes that adhere to the specified criteria. We distinguish two categories of target languages. In the first category, we sampled test sentences from Dutch and German which are typologically related to English and share the same script. In the second category, we sampled test data from Arabic and Amharic which are typologically distant from English and use a different script.

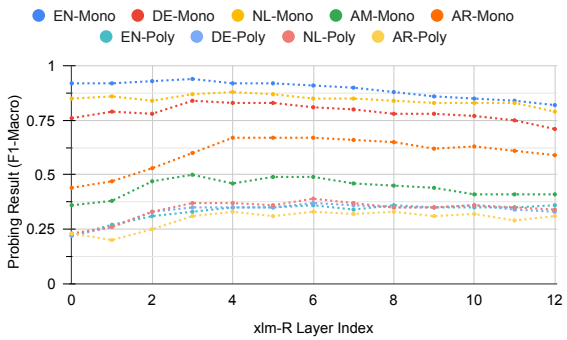
We distinguish between the following conditions: the model can be trained on English monosemous data or on English polysemous data. The test data is sampled from Dutch and German (category 1) and from Amharic and Arabic (category 2). For each language, we further distinguish between monosemous and polysemous test data. Figure 2 shows the result of evaluating the English probing model.

Results In the monosemous condition, we observe higher results for German and Dutch than for Arabic and Amharic. In a standard Zero-shot evaluation where a pre-trained language model is fine-tuned in a downstream task in a source language and evaluated on a target language, it has been widely reported that cross-lingual transfer yields better results for related languages (Pires et al., 2019; Wu and Dredze, 2019). As we probe the cross-lingual representation directly, we show that transfer occurs even before a pre-trained model is fine-tuned on a downstream task. Our results show that to a smaller extent transfer effects can even be observed for Arabic and Amharic although they are typologically different from English and use another script.

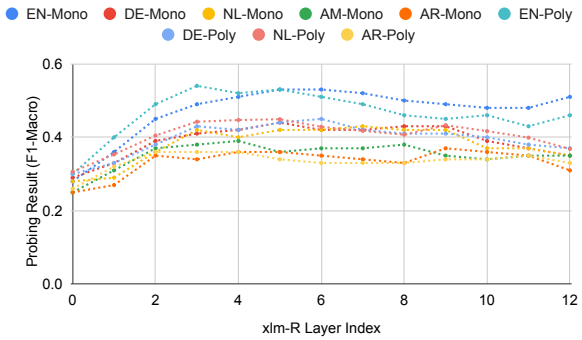
In the more difficult case of the polysemous condition, the performance of the classifier on correctly labeling the ambiguous semantic class is lower in comparison to the monosemous condition across all languages but outperforms a lexical baseline. With a closer look at the result per layer, we observe that the performance improves for representations extracted from higher layers. Remarkably, the differences across the related and unrelated languages got smaller in the polysemous condition. Apparently, there is a lower bound of performance at which the performances clutter together as a result of the complexity of the task and there are less differences for the languages.

6 Conclusions

In this paper, we investigated to what extent polysemous profiles play a role in establishing a relation between words and concepts. We focused on English but we also investigated words shared across languages in cross-lingual pre-trained language models. We selected representative cases for concept distributions from a large dataset of entity mentions as ambiguity profiles. Our prob-



(a) Result for the English monosemous model



(b) Result for the English polysemous model

Figure 2: F1-scores for the different conditions macro-averaged across four classes. Mono refers to monosemous test data in the corresponding language. Poly refers to polysemous test data in the corresponding language. The result of the baseline experiment and detailed results per layer are presented in Appendix A.

ing experiments indicate that prior probabilities of polysemy profiles are reflected in the lexical initialization and that context is integrated for disambiguation in higher layers. Our cross-lingual results indicate that sharing of tokens and contexts across languages has an influence on probing accuracy.

Our experiments are restricted to five languages: English, Dutch, German, Arabic, and Amharic. In future work, we will extend our experiments to more languages. We plan to investigate the impact of optimizing the probing classifier with cross-lingual training data. Training on the data of other languages extends the fund of concepts in the classifier, which is comparable to an expand model for multilingual wordnets.

Our method is limited by the annotations in contexts. It is therefore difficult to extend it to other words and concepts than entity names. Nevertheless, the entity results can be seen as a proof of concept to develop more sophisticated methods for analyzing concept relations in multilingual models. When more sense-tagged data becomes available, this method can also be applied to other words and concepts.

Acknowledgements

We thank the anonymous reviewers for their insightful feedback and suggestions. W. Tufa's research was supported by Huawei Finland through the DreamsLab project. All content represented the opinions of the authors, which were not necessarily shared or endorsed by their respective employers and/ or sponsors. L. Beinborn's research was supported by the Dutch National Science Organisation (NWO) through the projects ClariahPlus (CP-W6-19-005) and VENI (Vl.Veni.211C.039).

References

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

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Jasmijn Bastings, Yonatan Belinkov, Emmanuel Dupoux, Mario Giulianelli, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad, editors. 2021. *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Punta Cana, Dominican Republic.

Lisa Beinborn and Rochelle Choenni. 2020. Semantic Drift in Multilingual Representations. *Computational Linguistics*, 46(3):571–603.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language*

Processing (Volume 1: Long Papers), pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Francis Bond, Shan Wang, Eshley Huini Gao, Hazel Shuwen Mok, and Jeanette Yiwen Tan. 2013. Developing parallel sense-tagged corpora with wordnets. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 149–158.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ahmed El-Kishky, Adi Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. XLEnt: Mining cross-lingual entities with lexical-semantic-phonetic word alignment. In *Preprint*, Online.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database and some of its applications.

Rubén Izquierdo, Armando Suárez, and German Rigau. 2009. An empirical study on class-based word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 389–397, Athens, Greece. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019a. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Piek Vossen. 1998. A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers. doi*, 10:978–94.

Piek Vossen and Christiane Fellbaum. 2009. Universals and idiosyncrasies in multilingual wordnets. *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, 200:319–346.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Yadollah Yaghoobzadeh, Katharina Kann, T. J. Hazen, Eneko Agirre, and Hinrich Schütze. 2019. Probing for semantic classes: Diagnosing the meaning content of word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5740–5753, Florence, Italy. Association for Computational Linguistics.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.

Mengjie Zhao, Philipp Dufter, Yadollah Yaghoobzadeh, and Hinrich Schütze. 2020. Quantifying the contextualization of word representations with semantic class probing. *arXiv preprint arXiv:2004.12198*.

A Result of English Monosemous and Polysemous Model With Baseline

	EN-Mono	DE-Mono	NL-Mono	AM-Mono	AR-Mono	EN-Poly	DE-Poly	NL-Poly	AR-Poly
Majority-Vote	0.13	0.12	0.11	0.16	0.12	0.15	0.15	0.15	0.15
Word Embeddings	0.87	0.13	0.10	NA	NA	0.30	0.19	0.21	NA
Tf-Idf	0.53	0.28	0.28	0.21	0.16	0.46	0.28	0.29	0.18
Layer-0	0.92	0.76	0.85	0.36	0.44	0.22	0.22	0.23	0.23
Layer-1	0.92	0.79	0.86	0.38	0.47	0.27	0.26	0.26	0.2
Layer-2	0.93	0.78	0.84	0.47	0.53	0.31	0.33	0.33	0.25
Layer-3	0.94	0.84	0.87	0.5	0.6	0.33	0.35	0.37	0.31
Layer-4	0.92	0.83	0.88	0.46	0.67	0.35	0.35	0.37	0.33
Layer-5	0.92	0.83	0.87	0.49	0.67	0.35	0.35	0.36	0.31
Layer-6	0.91	0.81	0.85	0.49	0.67	0.36	0.37	0.39	0.33
Layer-7	0.9	0.8	0.85	0.46	0.66	0.34	0.36	0.37	0.32
Layer-8	0.88	0.78	0.84	0.45	0.65	0.36	0.35	0.35	0.33
Layer-9	0.86	0.78	0.83	0.44	0.62	0.35	0.35	0.35	0.31
Layer-10	0.85	0.77	0.83	0.41	0.63	0.35	0.36	0.36	0.32
Layer-11	0.84	0.75	0.83	0.41	0.61	0.35	0.34	0.35	0.29
Layer-12	0.82	0.71	0.79	0.41	0.59	0.36	0.33	0.34	0.31
Majority-Vote	0.13	0.12	0.11	0.16	0.12	0.15	0.15	0.15	0.15
Word Embeddings	0.87	0.13	0.10	NA	NA	0.30	0.19	0.21	NA
Tf-Idf	0.53	0.28	0.28	0.21	0.16	0.46	0.28	0.29	0.18
Layer-0	0.92	0.76	0.85	0.36	0.44	0.22	0.22	0.23	0.23
Layer-1	0.92	0.79	0.86	0.38	0.47	0.27	0.26	0.26	0.2
Layer-2	0.93	0.78	0.84	0.47	0.53	0.31	0.33	0.33	0.25
Layer-3	0.94	0.84	0.87	0.5	0.6	0.33	0.35	0.37	0.31
Layer-4	0.92	0.83	0.88	0.46	0.67	0.35	0.35	0.37	0.33
Layer-5	0.92	0.83	0.87	0.49	0.67	0.35	0.35	0.36	0.31
Layer-6	0.91	0.81	0.85	0.49	0.67	0.36	0.37	0.39	0.33
Layer-7	0.9	0.8	0.85	0.46	0.66	0.34	0.36	0.37	0.32
Layer-8	0.88	0.78	0.84	0.45	0.65	0.36	0.35	0.35	0.33
Layer-9	0.86	0.78	0.83	0.44	0.62	0.35	0.35	0.35	0.31
Layer-10	0.85	0.77	0.83	0.41	0.63	0.35	0.36	0.36	0.32
Layer-11	0.84	0.75	0.83	0.41	0.61	0.35	0.34	0.35	0.29
Layer-12	0.82	0.71	0.79	0.41	0.59	0.36	0.33	0.34	0.31

B Distribution of Selected Ambiguous Entities

Entity	**LOC**	**ORG**	**PER**
Mercury	562	215	26
Sirius	391	481	42
Olympus	177	3	11
Uranus	385	7	169
Reich	12	16	266
Cloud	22	63	
Ceres	191	49	21
Aquarius	124	11	59
Chimera		85	17
Vesta	75	9	29
Quartz	12	73	7
Regulus	8	23	67
Terra	42	21	66
Sol	26	64	56
Lab	16	58	7
Triton	16	51	12
Solaris	9		24
Tyre	7		28
Electra	9	28	17
Beguinage	23	7	8
Prana	12	19	16