Low-Resource, But High Specificity: A Novel, Sentence-Based Method for Cross-Lingual Retrieval of Pesticide Names in Research Articles

Anonymous ACL submission

Abstract

In this article, we present the results from our interdisciplinary work to identify pesticide names in research articles primarily in Brazilian Portuguese, but also in Spanish, French, and Italian. We proceed cross-lingually, extracting information from a large, high-quality corpus in English, which we then apply to the lower-resource languages. We show that a combination of a state-of-the-art multilingual transformer models, sentence-based similarity metrics, and expert knowledge yields the best results in our low-resource task. It yields twice as many true positives as the November 2023 version of gpt-4, and it decisively outperforms other baselines, including a classical NER-model fine-tuned on our training data. Our approach offers a promising start and might be transferable to other similarly demanding tasks in low-resource contexts.

1 Introduction

In 2020/21, Brazil produced 137 million metric tons (mmt) of soybeans, and 83 mmt of them were exported worldwide, consolidating the country's leadership as both a producer and an exporter of grains (Kamrud et al., 2022). Consequently, it has also become the largest consumer of pesticides in the world. Given this context, Chemistry researchers seek to study and create a set of environmentally sustainable methodologies for pesticide degradation, since they are extremely harmful to the health of the general population. However, one problem that previous studies (Pinto and Lima, 2018) have found is the lack of terminological standardization of pesticide names in Brazilian Portuguese, especially in scientific papers from the field of organophosphorus pesticides, which may lead to the misinterpretation of product labels as well as hinder lawmaking on the issue.

One example that illustrates this problem is the term 'malathion", which is a pesticide common name usually translated to Brazilian Portuguese as *malationa* (adapted to the morphology of the language, but not representative of the pesticide's most important chemical group), or most appropriately, as *malation* (indicating the correct chemical group). Other possible translations are *malatiom* (a spelling variant of the former one) and *malatião* (commonly used in European Portuguese) (Souza et al., 2022). In this situation, the method presented in this paper pursues the following main goal: *given a Brazilian Portuguese text belonging to the genre of research publications (broadly conceived) and focusing on a topic around agriculture and pest or disease control, we would like to identify as many pesticide names as possible*. In other words, our priority is recall (returned true positives divided by total true positives in the data) rather than precision (true positives divided by returned positives).

To start tackling this problem, in this paper, we bring together a multilingual group of corpus linguists, terminology experts, chemists, and NLP researchers to take the first step toward mapping this largely uncharted terrain, primarily in Brazilian Portuguese, but also in Italian, French, and Spanish. What is novel about our approach is that we include sentence similarity, in addition to similarity of individual tokens, to find new referents to pesticides. Theoretically, this is grounded in the so-called priority of the proposition, practically, it seems tailored to the requirements of our low-resource setting.

Our contributions to the field are twofold. First, we present a novel, sentence-based approach to named-entity classification that has potential for other multilingual, low-resource settings (we will transfer it to Italian, Spanish, and French in a probing study, with encouraging results). Second, we develop and make publicly accessible our code and classifiers, together with a long list of pesticide names in the five languages mentioned.

Making progress in this area is both important

and difficult. It is important because, without a comprehensive view of the existing terminology in pesticide research in Brazilian Portuguese, but also in the other three languages, researchers might not be aware of other ongoing research in the field, and government bodies might not be aware of the harmful effects of certain pesticides. More generally speaking, there are numerous other cases where the relevance of the topic is high, but the available resources, especially expensively labelled datasets, are not available. It is difficult because of the very specific kind of entity in focus, the lack of standard resources such as high-quality annotated training datasets as well as the high degree of variation in terminology. Ultimately, the results of this study are being used in the development of a multilingual glossary of pesticide names.

2 State of The Art

2.1 Terminology Research & Semantics

Terminology Research Although the International Union of Pure and Applied Chemistry (IUPAC) has encouraged scholars to follow an internationally standardized terminology, it is still regional, unlike its symbology, which is universal. Even though the conditions for technical communication are, to a certain extent, more controlled, the terminology used is dynamic and chosen by its users in a subjective manner (Azenha Jr., 1999). In this sense, variation has been an inherent part of specialized language which is widely described by authors on different levels (Cabré, 1999; Faulstich, 2001). Although the intersection between Translation and Terminology is undeniable, very little has been studied about the characteristics and motivations for this relation, and even less has been considered about the limits between them both. In Brazil, the language direction of translated texts has long been from English to Portuguese, nevertheless, the international business exchange has increased significantly, making it necessary for translators to work with the other language direction and often create neologisms or even paraphrased terms (Krieger and Finatto, 2004). Terminology has long provided the necessary aids for the translating process, however, in the area of Pesticide Chemistry, there is still a wide gap to be filled in since this is a young and fast-changing area.

Semantics Our approach is based on what has recently been called the priority of the proposition principle: the primary locus of meaning is in entire propositions, expressed in sentences, not in concepts, expressed in individual words. This means that to identify occurrences of certain classes of pesticides in text, it might be more promising to compare sentences than to compare words. The founders of this position of the priority of the proposition (over concepts) are no others than Immanuel Kant (Kant, 1998 [1781/1787]) and Gottlob Frege (Frege, 1892), the latter being the inventor of modern predicate logic. More recently, the position has been defended by Quine (1974); Brandom (1994); Frápolli (2019).

2.2 NLP

Our task might look similar to what is sometimes called a biomedical named entity recognition (NER), see Naseem et al. (2021). Sometimes, this task is also called "concept recognition", or "entity mention extraction" (Tseytlin et al., 2016). For this domain, there are challenges and benchmarks for languages other than English, in particular, Spanish (see the PharmaCoNER task, Gonzalez-Agirre et al. 2019). For instance, Hakala and Pyysalo (2019) use multilingual BERT, an earlier multilingual language model that has been outperformed by xlm-roberta used here. They can rely on almost 4000 annotated samples for fine-tuning.

While there is very scarce research specific to pesticide NER, let alone on multilingual pesticide NER, there is an active research interest in pest recognition (see, e.g., (Liu et al., 2020; Rodríguez-García et al., 2021; Hernández-Castillo et al., 2019)). Still, however, this research is exclusively mono-lingual, and predominantly focused on English.

Closer to our use-case, G et al. (2023), is one of the few NER methods that contain an umbrella class for pesticides. It is, however, as is typical, exclusively focusing on English. Furthermore, they use a sophisticated ensemble method comprising an open information extraction module as well as recognition modules specific to every single kind of entity that they wish to recognize. Our approach, in contrast, is designed to be as language-agnostic as possible to allow for a maximally efficient application to new languages.

Looking beyond the category of pesticides, there are approaches using a knowledge base (KB), such as Wang et al. (2021), which are once more tailored to English, and do not focus on pesticides. Furthermore, these KB-based approaches predominantly rely on Wikipedia as their knowledge base.

For specialized domains such as pesticides, however, Wikipedia is barely useful because it lacks the domain-specific entities and relationships. This is already true for English, and the scarcity of data in such specialized domains only increases for other languages.

For the Chinese language, finally, there are some pesticide-specific approaches, such as Ji et al. (2023), which, however, are once more designed with one single language in mind, in this case, Chinese.

A final class of approaches leverages existing descriptions (as opposed to knowledge bases) of the entity class in focus. It extracts contextualized word embeddings from these descriptions and then measures cosine similarity of these embeddings with embeddings in the target texts. If the similarity surpasses a certain threshold, it is classified as belonging to that entity class. Wu et al. (2020) and Logeswaran et al. (2019) both use variations of this approach.

Our approach is similar to these description-based approaches as we also compare embeddings using cosine similarity. However, in our low-resource setting, there are no entity descriptions available. Furthermore, these approaches are also strictly focused on one single language, while our interest is multi-lingual from the start. Finally, while the typical NER setting aims at a balance between precision and recall, we are involved in a larger research project where recall is clearly more important than precision.

Technically, our approach is based on multilingual transformer-based sentence models. Transformer-based (Vaswani et al., 2017) pretrained language models (PLMs) have become the state of the art in NLP. Based on the transformer's encoder, researchers have proposed a number of highly successful NLU architectures, starting with BERT (Devlin et al., 2019a), quickly followed by others, including RoBERTa (Liu et al., 2019), XL-Net (Yang et al., 2019), DeBERTa (He et al., 2020), and smaller versions such as DistilBERT (Sanh et al., 2019) and Albert (Lan et al., 2019). Additionally, a number of sequence-to-sequence architectures have been proposed that are more similar to the original transformer than to BERT in that they directly try to transform one sequence to another, much like the basic set-up of neural machine translation. These include T5 (Raffel et al., 2019) and BART (Lewis et al., 2020).

With regard to multilingual transformer models, mBERT, based on the BERT architecture (Devlin et al., 2019b) was among the first multilingual pretrained models. It was quickly followed by a variety of other methods (see the overview on Doddapaneni et al. 2021) as well as theoretical work on how best to model cross-lingual transfer of information (Chi et al., 2021). On the word-level (or better sub-word, or token-level, as transformers split up rare words into sub-words, usually represented by bite-pair encoding, Sennrich et al. 2015), we use multilingual xlm-roberta-large (Conneau et al., 2019), which was trained on 2.5 terabytes of text from 100 different languages, including about 100GB in Portuguese. For the sentence-embeddings, we use a multilingual SBERT-Model (Reimers and Gurevych, 2019), namely paraphrase-multilingual-mpnet-base-v2, originally proposed by Song et al. (2020). SBERT-Models are optimized for sentence-level comparison of embeddings via geometric similarity measures such as cosine similarity.

3 Method

Extraction of Information from Training Corpus Given the lack of a domain- and task-specific annotated training corpus, we decided to rely on a fully self-supervised approach using multilingual transformer models combined with minimal knowledge-based input. For each seed word, we retrieve sentence-embeddings of sentences where the seed occurs from an English corpus, then we compute the centroid of these embeddings. In the same way, we also retrieve the centroid of all occurrences of the words.

In detail, our retrieval method works as follows.

1. We use an expert-compiled list of seed-words in English (see the next section for details on seed-word collection), all of them (organophosphorus) pesticide names, as well as a multilingual word-based model (xlm-roberta-large) and a multilingual sentence-based model (paraphrase-multilingual-mpnet-base-v2, for references, see above, section 2.2) and the English dataset described below (section 4). For each of the seed words, we search for occurrences in sentences from this dataset. If there is a match, we retrieve (1) sentence-embeddings using the multilingual sentence-based model and (2) word-embeddings using the multilingual word-based model. For the latter, we had to control for the number of sub-word-units into which the model chose to split

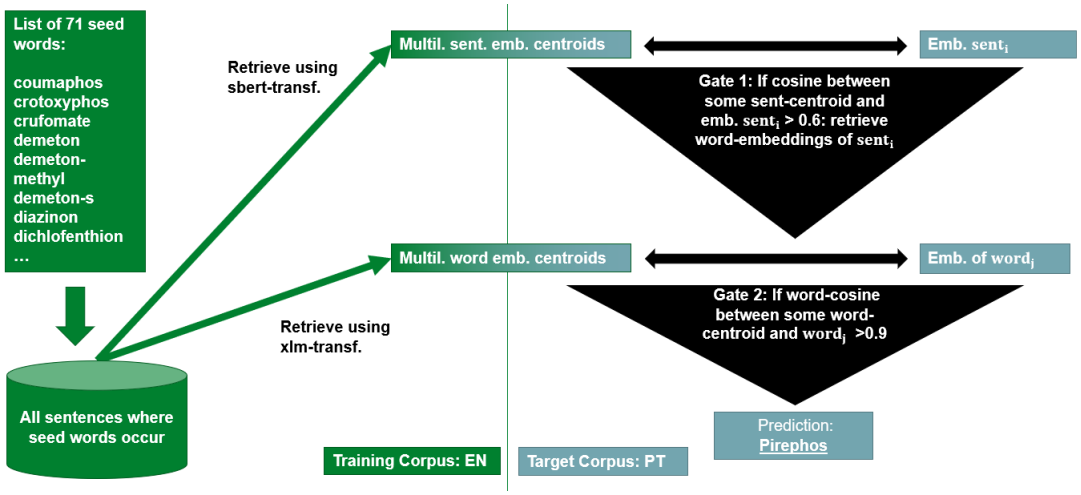

Figure 1: Illustration of our method for target language Portuguese.

the original pesticide name. This is a common procedure for transformer-based models, but it is particularly relevant for our case, as the pesticide names are predominantly rare words that the models have chosen not to represent integrally.

2. Then, we compute (1) one word-based centroid per seed, (2) one sentence-based centroid per seed, and (3) record the token-span of each of the seed words (of course, only seeds that actually occurred in the corpus were considered).

Prediction in Target Corpus Then, we measure the cosine similarity between each sentence embedding in the target corpus with all the sentence-centroids obtained.

For the sentences whose embeddings pass a certain cosine threshold (0.6, a figure that has been empirically determined), we then work through them with a carefully designed token-by-token-method. The issue here is that, given the low-resource context, it is to be expected that most of the interesting words, i.e., previously unknown pesticide names in one of the target languages, are going to be split up into a number of sub-word units by xlm-roberta's tokenizer. This is particularly clear given the fact that with xlm-roberta, the space for tokens has to be shared across all languages on which it has been trained. Even though Conneau et al. (2019) use a large vocabulary size of 250k, given the variety of target languages, this is still a very small vocabulary space. We therefore implement a look-ahead method that uses some simple heuristics, such as whether the next token starts with a lowercase character, to determine whether we should add another token to the current span of tokens that might represent a pesticide name.

At this point, we add a heuristic step, where we privilege certain sequences that are highly indicative of a Brazilian Portuguese pesticide name by increasing the respective cosine by 0.1, which means that the resulting cosine figure could be higher than 1.0.

We then extract the mean embedding over all the embeddings of the tokens identified and compare it with all the word-embeddings extracted from the training corpus. As we use multilingual models throughout, this can all happen in the same embedding space, spanned by the respective transformer. All spans of tokens whose embeddings pass another cosine threshold are then predicted as pesticide names. Compare figure 1 for an overview on the method.

We decided to rely on English seed terms and source texts for three reasons: because the sheer amount as well as the specificity of the texts available in English is not even remotely matched in another language, and because the lexical variation in the English pesticide names is much smaller than in the Portuguese, French, Italian, and Spanish texts according to our expert.

Proceeding by comparing sentence similarity is theoretically grounded in the priority of the proposition. We hypothesize that this allows us to harvest more pesticide names than a method that directly matches individual words: It adds a dimension of comparison to the method that might exploit features that are invisible on the token-level.

4 Datasets

Since our goal was to tackle variation in pesticide names in academic papers, we used Google Scholar to find research articles, theses, dissertations, reports, books, and book chapters that represented research in the domain of organophosphorus chemistry. To achieve that, our search query consisted of the keywords 'organophosphorus pesticides/compounds' and 'organophosphates' in English; for the other languages, we used 'organofosforados' (Portuguese and Spanish), 'organofosforici' (Italian), and 'organophosphorés' (French).

We display statistics for our five datasets in table 1. We give all the references to all the individual texts in a separate supplementary document. We chose not to enclose it in the appendix of the article, as it is 62 pages long. Our English training corpus consists of documents published between 1943 and 2022, the Portuguese corpus ranges from 1996 to 2022, the French one from 1960 to 2019, the Spanish one from 1982 to 2023, and the Italian one from 1981 to 2023.

Corpus	Doc. count	Token count
English	210	3,5M
Portuguese	172	1,4M
French	52	711,8k
Italian	37	622,3k
Spanish	38	494,6k

Table 1: Basic statistics on our five datasets.

Finally, the seed words were extracted from the aforementioned English corpus by means of a keyword extraction method named *simple maths* (Kilgarriff, 2009), whose output consists of a list of items that can be used to understand the corpus's main topics. This method allows the user to change the value of its smoothing parameter in order to retrieve rarer words (for lower values) or more common words (for higher values). Previous tests showed that some pesticide names would only appear in 'rarer' lists while others only in 'more common' lists. To deal with this issue, we generated 6 keyword lists with different smoothing parameter values (0.001 – 0.01 – 0.1 – 1 – 10 – 100) and a maximum of 4,000 tokens each. Those lists were then merged and duplicates were deleted, resulting in a single list of 7,995 items. With the help of two experts, we selected those items that were pesticide names, reaching a total of no more than 84 unique seeds.

These seeds occur 2781 times in the training corpus in total, with 6 occurring only 1 times and 43 occurring less than 10 times. The three most frequent names, parathion, paraoxon, and demethon, occur 1300 times, almost half of total occurrences. This shows that the corpus is less than ideal for the task at hand. In consequence, if we can achieve decent performance based on these training data, we can confidently infer that our method will also generalize to other use cases.

5 Experiments

To test the promise of our novel approach, we conduct three experiments.

Experiment 1 First, we compare our transformer- and sentence-based method with three other approaches. First, a regular-expression (regex) based one that includes expert knowledge on the make-up of (organophosphorus) pesticide names. In brief, this method functions by cutting the final syllable from each seed word and then matching any word that begins with the resulting cropped seed. We have tried to make sure that this regex-based method can serve as a genuine baseline and not merely as a straw-man. As a consequence, we used preprocessing with natural language toolkit (nltk, see Bird 2006) to match only nouns (as opposed to adverbs and other parts of speech), we applied transformation rules for the most common graphemic variants, and we used expert knowledge to define final syllables that must not be cut because they are central for the meaning of the terms.

The second competitor in this first experiment is xlm-roberta-base fine-tuned to NER using the very same training English training dataset that also serves our novel method. To make our competitor as strong as possible, we run fine-tuning three times and compare our approach against the best-performing model of the three fine-tune-runs. For further details of the fine-tuning procedure, see the appendix, section A.

The third competitor is a version of GPT-4, namely gpt-4-1106-preview, the version of OpenAI's gpt-4 that was available in preview-mode at the beginning of November 2023, when we ran our experiments.[1] We give details of the prompt in the

[1]We have decided not to reference any publication by OpenAI on this model, as their publications deliberately sidestep basic scientific norms of transparency and reproducibility, which means that they should be considered corporate com-

	Accuracy				Count True Positives				Total Pos.	CK
	100	300	1k	2k	100	300	1k	2k		
Ours	0.72	0.6	0.44	0.33	72	180	447	657*	3053	0.89
gpt-4	0.81	0.84	–	–	81	253	315	315	370	0.75
Best FT	0.62	–	–	–	73	121	121	121	181	–
Regex	0.21	–	–	–	21	42	42	42	195	–

Table 2: Results from our first experiment. With our approach and gpt-4, we always report the result of a (at least) two-thirds majority vote of two out of our three annotators, except for the 2k figure with our method, where only one annotator annotated the second kilo. With Regex and Best FT, we report the results of our only annotator assigned to these outputs. Best performance per setting is underlined.

appendix, section B. Importantly, while we were able to print the sentence where the positive was registered with the first three approaches, that is, ours, the regex- and the finetuned-ner one, we were unable to devise a prompt that would get gpt-4 to do that. While this probably has not influenced prediction, it might have had consequences in the evaluation of the results.

Experiment 2 In our second experiment, we explore the recall of our method as well as of the benchmarking methods that we apply. To this end, we ask one annotator to go through four particularly promising texts in Portuguese and extract all pesticide names that they find there. This yields a total of 121 Pesticide name occurrences, belonging to 44 pesticide types. We then run all four methods over these four articles and then count how many pesticide types they are able to extract.

Experiment 3 Our third experiment broadens the perspective from Brazilian Portuguese to three other Romance languages, namely French, Spanish, and Italian. As our method also relies on specific knowledge-based heuristics, which we leave unchanged, we expect the performance of our method to decrease in proportion to the distance in terms of morphology between Brazilian Portuguese and the three languages. In other words, we expect a decrease in performance from Spanish, to Italian, and finally to French.

6 Results

Experiment 1 We give the results of experiment 1 in table 2. On the first row, we give our results. As we sort our results descending by cosine similarity to the centroid that caused the prediction, accuracy decreases with increasing size of the data partition considered. We see that it yields a total of 657 true positives, with a little more than 3k positives returned. gpt-4 only yields a total of 370 positives, but its accuracy is unmatched with 0.81 and 0.84 respectively.

For the fine-tuning approach, we report the figures of the best model resulting from three fine-tuning runs; as mentioned previously, in our context, the emphasis is on recall, so we choose the model with the highest relative recall, even if it is not performing best in terms of accuracy. Hence, we report performance of this third fine-tune run. The regex-baseline, finally, yields a total of 194 positives, only 42 being true positives.

To measure inter-annotator reliability, we ask three different annotators to annotate the results of the two competitive methods, of ours and of gpt-4, allowing us to compute Cohen's Kappa there. We report a 2/3-majority vote for results of our method and for gpt-4, while we report the results of our single annotator for the two other, less competitive approaches. Cohen's Kappa is at 0.89 for our method and at 0.75 for gpt-4. According to Greve and Wentura (1997, 111), any figure at or above 0.75 signifies very good agreement. We therefore judge our results to be very dependable. Furthermore, the annotators reported that annotating the output from gpt-4 was more challenging, as they were missing the sentence where the prediction occurred there, which might explain the lower figure for Cohen's Kappa with the results of gpt-4.

Table 3 shows the results of experiment 2. The figures also result from a majority vote among the three annotators which, given the specific values for Cohen's Kappa, is usually also a unanimous vote. It was important to us to measure inter-annotator agreement here, as the first language of our anno-

munication and hence not cited in a scientific study.

	C.>1	Top 500	Top 1k	CK
ES	0.48	0.42	0.29	0.88
IT	0.49	0.39	0.26	0.89
FR	0.26	0.3	0.20	0.84

Table 3: Accuracies of the predictions of our method in experiment 2 (CK: Cohen's Kappa; number of samples with Cosine >1: 395 ES, 327 IT, 581 FR).

	Ours	gpt-4	FT	Regex
Recall	0.7	0.64	0	0.18

Table 4: Results of our third experiment, measuring the recall over four articles.

tators is Brazilian Portuguese, hence, it was not antecedently clear that they would perform well at annotating other Romance languages. However, as the CK values testify, they did perform very well, with CK never falling below 0.84.

Table 4 shows the results of our small recall-probing. It shows that, with regard to the four articles considered in this experiment, the best performance is achieved by our method, closely followed by gpt-4. With regard to these four articles, the fine-tuned NER model did not yield any true positives, while the regex method achieved a recall score below 0.2.

We also note that, given the context of the task, we were generous regarding the ability of all methods to correctly separate punctuation symbols from their prediction. In our context, finding a previously unknown variant of a pesticide name with a semicolon attached to it is just as valuable as without the semicolon.

7 Discussion

We emphasize four aspects of the results of our experiments. First, we wish to emphasize that this is not a task comparable to mainstream NER settings, as it differs in terms of resources (no knowledge base or high-quality annotated data), target class (too specific for mainstream NER), and overarching goal, as it prioritizes recall over precision. That it is not a classical NER task also clearly shows in the performance of our fine-tuning method. Our approach yields five times more true positives and it is, to the extent to which it is comparable, also superior in terms of accuracy. Established methods in the field that would yield excellent precision and recall in mainstream settings cannot help us too much in this task.

Manual inspection of the results by our method shows that, as the figures would suggest, the quality of the predictions decreases with decreasing cosine similarity both on the sentence- and on the word level. Furthermore, our method by and large manages very well to focus on entire words, and very often on noun words as well, which is surprising because we have never explicitly injected any part-of-speech information. This must therefore originate from the similarity to the centroid. Furthermore, we find that the variation of the same pesticide name is indeed quite large, as table 5 shows.

We give the top 10 predictions by our method in the appendix, section C. Towards the end of the top 2k results considered here, the quality of the prediction decreases substantially. For instance, the method predicts proper names such as "Brunner" or non-pesticide chemicals, such as "Sacarose".

The second aspect that we would like to point out is the quite fascinating comparison with gpt-4. On the one hand, it is genuinely impressive that gpt-4 correctly identifies no less than 315 pesticide names in Brazilian Portuguese research literature, and that it manages to do so with a precision of more than 0.8. Given that it has never been trained for this task, and given the very demanding setting of it, this is excellent. On the other hand, we are happy to point out that our method yielded more than twice as many pesticide names, and six and fifteen times as many as our other two benchmarks. We take this to show that ascending on the sentence-level and developing a method that is relatively lightweight, but which has some expert knowledge as well as an excellent topic-specific corpus can still decisively outperform very large state-of-the-art language models. We hypothesize that it might be a consequence of gpt-4's reinforcement-learning process that it, as it were, tends to err on the side of caution and only predicts a pesticide if it is really sure because it has been conditioned to behave this way. This, however, as most aspects about gpt-4, is little more than speculation.

Manual inspection shows that gpt-4 makes some rather obvious blunders as well, for instance predicting "Record" or "Moser et al., 2006" as a pesticide name. While the former is simply an English noun, the second is easily identified as a reference to an academic text due to its form and would never

be mistaken for a pesticide name by a human being. As usual with gpt-4, it is close to impossible to understand precisely why it predicted the way that it did.

We are now moving to the third aspect that we wish to raise. Considering the results form experiment 2 on table 3, we can see that the results we receive for Portuguese are no outliers. Even with the heuristics that are tailored to Brazilian Portuguese, we obtain results in French, Italian, and Spanish that are of great help in charting the terminological landscape there. In particular, it is remarkable that these heuristics, resulting in cosines higher than 1.0, perform well with the morphologically similar languages Spanish and Italian, but not with French. We emphasize that absolutely no adaptation of our method was needed for this transfer to take place. We used the same centroids that had been extracted from the same English training corpus, and we used the very same embeddings obtained from the very same multilingual model that was applied successfully to Portuguese also to Spanish, Italian, and French. As a consequence, and considering the database on which xlm-roberta was trained (Conneau et al., 2019), applying our method to texts relating to pesticides in dozens of other languages should be straightforward. All that is additionally required is the texts.

The fourth aspect that we wish to emphasize is the fact that our sentence-based method yielded promising results for a task that one might not primarily associate with the sentence-level, namely the identification of expressions on the word-level referring to certain kinds of things. Here, it seems natural to operate on the word-level using models specifically designed for that purpose, or knowledge bases which are of course again confined to the word level. This suggests exploiting the sentence-level-information also for tasks that are more genuinely aligned with this level, such as topic extraction, semi-automatic grading of student's answers to open question, or relation extraction.

8 Conclusion

Overall, we take our results to be very encouraging. We have shown that multilingual transformers can support corpus linguistic analysis of difficult, cross-lingual challenges and clearly outperform both regex-baselines and fine-tuning approaches. In the future, we plan to apply our approach to other tasks in similar low-resource contexts, e.g., to NER of an entirely different class, where we also have nothing more than a list of seed terms in a high-resource-language, which we leverage using careful extraction of relevant information by means of multilingual transformers.

Seed	temephos	fenthion
Var 1	temephos	fenthion
Var 2	temefos	fention
Var 3	temefós	fentiona

Table 5: Examples of two seed words in English and their various forms in (Brazilian) Portuguese research articles.

Limitations

We see two main limitations of this work. First, we have only applied it to a very specific task, namely the identification of pesticide names in research literature. It is not clear how well it generalizes to other tasks with other text types. Second, we only work with Indo-European, even Romance languages. It is conceivable that the performance of our method would suffer when applying it to a non-Indo-European language, even if they are covered well by xlm-roberta's training dataset.

Ethics Statement

As the product of our research consists in lists of pesticide names and centroids in vector spaces, we see no risk of accidentally publishing personally protected information, offensive material, or biases that could discriminate against marginalized groups.

References

João Azenha Jr. 1999. *Tradução técnica e condicionantes culturais: primeiros passos para um estudo integrado*. Humanitas, São Paulo.

Steven Bird. 2006. NLTK: The natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Robert Brandom. 1994. *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard university press.

Maria Teresa Cabré. 1999. *La terminología: representación y comunicación: elementos parauna teoría*

de base comunicativa y outros artículos. Universitaride Lingüística Aplicada, Barcelona.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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.

Sumanth Doddapaneni, Gowtham Ramesh, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2021. A Primer on Pretrained Multilingual Language Models.

Enilde Faulstich. 2001. Aspectos de terminologia geral e terminologia variacionista. *Tradterm*, 7:11–40.

María José Frápolli. 2019. Propositions first: Biting geach's bullet. *Royal Institute of Philosophy Supplements*, 86:87–110.

Gottlob Frege. 1892. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50.

Veena G, Vani Kanjirangat, and Deepa Gupta. 2023. AGRONER: An unsupervised agriculture named entity recognition using weighted distributional semantic model. *Expert Systems with Applications*, 229:120440.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10.

Werner Greve and Dirk Wentura. 1997. *Wissenschaftliche Beobachtung: Eine Einführung*. Saarländische Universitäts-und Landesbibliothek.

Kai Hakala and Sampo Pyysalo. 2019. Biomedical named entity recognition with multilingual BERT. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 56–61.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*.

Carlos Hernández-Castillo, Héctor Hiram Guedea-Noriega, Miguel Ángel Rodríguez-García, and Francisco García-Sánchez. 2019. Pest recognition using natural language processing. In *International Conference on Technologies and Innovation*, pages 3–16. Springer.

Wenqing Ji, Yinghua Fu, and Hongmei Zhu. 2023. Multi-Feature Fusion Method for Chinese Pesticide Named Entity Recognition. *Applied Sciences*, 13(5):3245.

Gwen Kamrud, William W. Wilson, and David W. Bullock. 2022. Logistics competition between the u.s. and brazil for soybean shipments to china: An optimized monte carlo simulation approach. *Journal of Commodity Markets*, page 100290.

Immanuel Kant. 1998 [1781/1787]. *Kritik der reinen Vernunft*. Hamburg: Meiner.

Adam Kilgarriff. 2009. Simple maths for keywords. In *Proceedings of Corpus Linguistics Conference 2009*, University of Liverpool, UK.

Maria da Graça Krieger and Maria José Bocorny Finatto. 2004. *Introdução à Terminologia: teoria e prática.* Contexto, São Paulo.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *ArXiv*, abs/1910.13461.

Wenjie Liu, Guoqing Wu, Fuji Ren, and Xin Kang. 2020. DFF-ResNet: An insect pest recognition model based on residual networks. *Big Data Mining and Analytics*, 3(4):300–310.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A Robustly Optimized Bert Pretraining Approach. *arXiv preprint arXiv:\-1907.11692*.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-Shot Entity Linking by Reading Entity Descriptions.

Usman Naseem, Matloob Khushi, Vinay Reddy, Sakthivel Rajendran, Imran Razzak, and Jinman Kim. 2021. Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Paula Tavares Pinto and Marcela de Freitas Lima. 2018. A tradução na área de química orgânica: da adaptação à tradução literal. *Estudos Linguísticos (São Paulo. 1978)*, 47(2):573–585.

Willard Van Orman Quine. 1974. *The Roots of Reference*. Open Court Publishing Co.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Miguel Ángel Rodríguez-García, Francisco García-Sánchez, and Rafael Valencia-García. 2021. Knowledge-based system for crop pests and diseases recognition. *Electronicsweek*, 10(8):905.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv preprint arXiv:1910.01108*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*.

José Victor de Souza, Paula Tavares Pinto, and Marcela Marques de Freitas Lima. 2022. Malationa, malation ou malatiom? a variação denominativa no processo de criação de um glossário bilíngue da área de química de pesticidas. *Acta Scientiarum. Language and Culture*, 44(11):e55894–e55894.

Eugene Tseytlin, Kevin Mitchell, Elizabeth Legowski, Julia Corrigan, Girish Chavan, and Rebecca S Jacobson. 2016. Noble–flexible concept recognition for large-scale biomedical natural language processing. *BMC bioinformatics*, 17(1):1–15.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Xuan Wang, Vivian Hu, Xiangchen Song, Shweta Garg, Jinfeng Xiao, and Jiawei Han. 2021. ChemNER: Fine-Grained Chemistry Named Entity Recognition with Ontology-Guided Distant Supervision. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5227–5240, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems*, 32.

A Details on Fine-Tuning

In this section, we give the fine-tuning hyperparameters used for our three fine-tune-runs. The method builds on a Colab Notebook (located here), which in turn builds on this Github Repository.

The fine-tuning occurred on one GPU of a DGX-2 computing cluster, taking about 2 hours per fine-tune-run.

The hyper-parameters are:

- MAX_LEN = 512
- TRAIN_BATCH_SIZE = 4
- VALID_BATCH_SIZE = 2
- EPOCHS = 1
- LEARNING_RATE = 1e-05
- MAX_GRAD_NORM = 10

B Details on Evaluating GPT-4

We used Langchain's Python Interface (details here), and we gave gpt-4 the following prompt:

> "You are a specialized named-entity recognition system, your task is to find all organophosphorous pesticides in the following text and print them on one line, encosed by ampersand (&), with nothing added, just the bare list of organophosphorous pesticides that actually occur in the text, one Pesticide per line, for instance: "& Malathion & Parathion &""

C Top 20 results by Cosine

Table 6 shows the top 20 results of our method when applied to the Brazilian Portuguese corpus.

Sentence-key (en)	Candidate (ptbr)	Cosine
azinphosmethyl	phosphamidon,	1,094672322
demeton-methyl	methylazinphos.	1,094556451
glyphosate	Glyphosate	1,094305277
oxydemeton-methyl	clorpirifos-oxon.	1,094272614
methamidophos	methamidophos	1,094191313
azinphos-methyl	clorfenvinfos,	1,093927383
oxydemeton-methyl	Clorpirifos-oxon.	1,093624115
oxydemeton-methyl	azinfos-metílico,	1,093557715
methylparathion	methylbromphenvinphos	1,093243122
azinphosmethyl	monocrotophos,	1,092985988
temephos	temefos	1,09279871
chlorpyrifos	quinalphos,	1,092792988
methylparathion	diflubenzuron	1,092625618
dicrotophos	fensulfotion	1,092409134
diazinon	Baysiston	1,092201591
temephos	temephos	1,092031837
diazinon	Neguvon	1,091972828
crotoxyphos	fosforamidato,	1,09195435
crotoxyphos	mevinfos,	1,091821551
methylparathion	fosforotioatos	1,091743708

Table 6: Top 20 predictions issued by our method described above, section 3.