Cross-Lingual Retrieval of Organophosphorus Pesticide Names in Brazilian Research Articles

Anonymous ACL submission

Abstract

This article presents our interdisciplinary work to identify organophosphorus pesticide names in Brazilian Portuguese research articles. A combination of a state-of-the-art transformer model with expert knowledge yields promising results. However, more research is needed to get a more comprehensive overview of the different names used for organophosphorus pesticides in Brazilian Portuguese.

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 terms in Brazilian Portuguese, especially in scientific papers, 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).

To start to tackle this problem, in this paper, we bring together corpus linguists, terminology experts, chemists, and NLP researchers to take the first step toward mapping this largely uncharted terrain of Brazilian Portuguese terms referring to organophosphorus pesticides. We start out with seed terms in English as well as with an English training corpus and then use multilingual transformers to identify names of organophosphorus pesticides in Brazilian Portuguese research texts.

Our contributions to the field are twofold. First, the potential of multilingual transformers is tested on a very specialized, difficult text sort and word type. Second, we develop and make publicly accessible training data in English and Portuguese, together with a long list of Brazilian Portuguese pesticide names.

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, 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. It is difficult because of the complete lack of previous research in this area as well as the high degree of variation in terminology.

2 State of The Art

2.1 Linguistics

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.

2.2 NLP

Our approach is based on multilingual transformer-based sentence models. Our task might look similar to what is sometimes called a biomedical named entity recognition (NER), see Naseem et al. (2021). 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. In contrast, our domain of organophosphorus pesticide names in Brazilian Portuguese is entirely uncharted terrain. Furthermore, while the biomedical NER task is typically conceived as identifying string spans that name entities of various kinds as well as classifying these spans into very general categories, e.g. PROTEIN, we are only interested in a very specific kind of chemical compound, namely organophosphorus pesticides. This is why we resort to vanilla pre-trained multilingual language models without fine-tuning.

PLMs Transformer-based (Vaswani et al., 2017) pre-trained language models (PLMs) have become the state of the art in NLP. Researchers have proposed a number of highly successful natural language understanding (NLU) architectures, starting with BERT (Devlin et al., 2019), quickly followed by others, including RoBERTa (Liu et al., 2019), XLNet (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). On the word-level (or better subword-level, as transformers split up rare words into subwords), we use the multilingual RoBERTa called xlm-roberta (Conneau et al., 2020), which was trained on 2.5 terabyte 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 or distance measures such as cosine similarity.

3 Dataset

Our English training corpus consists of 210 documents that fit into the academic register, that is, scientific books, research papers, theses, and dissertations, published between 1943 and 2022. These texts were selected to represent the phosphorus chemistry domain, with a bias toward organophosphorus compounds. It has 3,472,000 tokens, of which 2,221,494 are types (i.e., similar items counted only once). Our Brazilian Portuguese corpus, on the other hand, has 172 academic documents published between 1996 and 2022. This collection gathers texts on pesticides in Brazil, mainly organophosphorus pesticides, with a token count of 1,402,237 and a type count of 1,053,438. In table 1 we can confirm our corpora biases by comparing the proportion of documents that were specifically related to organophosphorus compounds and the token count thereof with the total above-mentioned counts.

Corpus	OP-specific documents/Token count
English	95/2,109,998
Portuguese	85/836,563

Table 1: Organophosphorus-specific documents and token count.

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. With the help of an expert, we selected those items that were common pesticide names, reaching a total of 69 [CHECK] unique seeds.

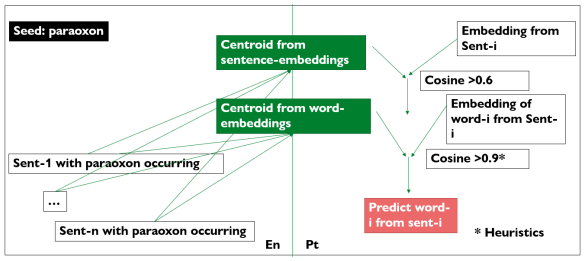

Figure 1: Illustration of our method for seed *paraoxon*.

4 Method

Given the lack of a domain-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. 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, we then work through word by word of this sentence and compare the words' cosines with each of the word centroids. All words whose embeddings pass another cosine threshold are then predicted as organophosphorus pesticide names. Compare figure 1 for an overview on the method.

In detail, our method works as follows. For the retrieval of the information from the English text:

1. We use an expert-compiled list of seed-words in English, all of them organophosphorus pesticide names, as well as a multilingual word-based model (xlm-roberta) and a multilingual sentence-based model (paraphrase-multilingual-mpnet-base-v2, for references, see above, section 2.2) and the English dataset described above (section 3). 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 sentence-based model and (2) word-embeddings using the multilingual word-based model. For the latter, we had to control for the number of subword-units into which the model chose to split 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) recorded the wordpiece-span of each of the seed words (of course, only seeds that actually occurred in the corpus were considered).

For the retrieval of the organophosphorus pesticide names in the Portuguese texts, we proceed as follows.

1. For each sentence in the Portuguese target corpus, we check whether its embedding surpasses a certain cosine similarity threshold with one of the sentence-centroids retrieved;

2. if yes, we check whether any subword-span within the respective sentence also passes a cosine threshold with any of the word-centroids. Here, we use expert knowledge to privilege certain morphologies that are strongly suggestive of organophosphorus pesticide (namely "fos|phos|on$" by increasing their cosine similarity with any of the stored word-centroids by 0.1.

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 so far surpasses their Portuguese equivalents, and because the lexical variation in the English pesticide names is much smaller than in the Portuguese texts according to our expert.

We decided to proceed via sentences, as we hypothesize that this allows us to harvest more organophosphorus pesticide names than a method that directly matches individual words: While we expect the immediate context of different organophosphorus pesticides to differ substantially, in particular, because they are often produced by different companies or examined by different labs that use different terminologies, we expect the basic assertions on the sentence level to be more similar: they are all organophosphorus pesticides after all.

We compare our transformer- and sentence-based method against 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 graphematic 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. We give the results of this baseline together with the output from the transformer-based method in the next section.

5 Results

We evaluate the results of our method by asking expert annotators to categorize each prediction as either (1) a variation of the seed pesticide name, (2) another organophosphorus pesticide, (3) any kind of (non-organophosphorus) pesticide, (4) no pesticide at all. Results belonging to categories (1) and (2) are considered true positives and combined in table 2.[1] We give two different selections of transformer-based results in table 2 as well as the output from the regex-based method. Row *top 2k cosine* consists of an evaluation of the two thousand results with the highest cosine similarity to one of the word-based centroids. Row *>1 cosine* consists of only these results where the expert-suggested pattern has matched (we operationalize this by selecting these word suggestions with a cosine higher than 1, which is only possible if the original cosine is higher than 0.9 and the expert-suggested pattern matches). Row *regex-based* gives the results from our regex-based method. Given that it only suggested 194 words in total, focusing on a subset here was not necessary.

6 Discussion

We emphasize three aspects of the results of our method. First, the challenge is difficult. Unlike traditional NER-tasks, even for the biomedical domain, we tried to identify a very specific class of chemical compound, namely organophosphorus pesticides. Furthermore, we had to do so without being able to rely on a high-quality training dataset, as is common with the tasks known in the community, and our expert in the domain warned us that variation of terminology in Portuguese is extremely high (which is why we resorted to English texts to build our method). As a consequence, the precision of our method, as seen in table 2 is clearly below the industry standard. Still, the results of our combined knowledge- and transformer-based method (row *>1 cosine*) beats the regex-based baseline almost by factor 2, and it results in 5 times more true positives, which evinces that the use of transformer-based models for this task can yield genuine advantages in performance.

Selection	Org. Pest. (Count/%)	N-Org. P. (Count/%)	Total Count
Top 2k cosine	352/18%	305/15%	1981
>1 cosine	193/44%	28/6%	437
Regex-Based	37/19%	5/3%	194

Table 2: Evaluation of the results of our method.

In the same vein, we also emphasize that by using generic transformer-based models together with minimal expert input as heuristics, we managed to build a method that reaches a precision of 44% percent. While the absolute numbers of the top-2k evaluation suggest that recall suffers from this high threshold, we still managed to provide our linguistic partners with hundreds of new terms for pesticide names that were so far not recorded in any systematic way.

Third, it is obvious that our knowledge-based component, that is, increasing cosine similarity by .1 when certain morphological patterns are present, contributes substantially to finding organophosphorus pesticides and filtering out other kinds of pesticides, as row 3 of table 2 evidences.

7 Conclusion

Overall, we take our results to be encouraging. We have shown that multilingual transformers can support corpus linguistic analysis of difficult, cross-lingual challenges. In the future, we plan to build larger training datasets that allow us to fine-tune transformers to our task, and to experiment with more sophisticated matching routines.

[1] Table 3 in appendix section A presents the first twenty lines of results obtained by the transformer-based approach ordered by cosine.

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.

Maria Teresa Cabré. 1999. *La terminología: representación y comunicación: elementos parauna teoría de base comunicativa y outros artículos*. Universitaride Lingüística Aplicada, Barcelona.

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

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

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

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, Hong Kong, China. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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.

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*.

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*.

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.

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

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*.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

A Top 20 results by Cosine

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 3: Top 20 predictions issued by our method described above, section 4.