Deep Learning Methods for Identification of Multiword Flower and Plant Names

Damith Premasiri[1], **Amal Haddad Haddad**[2], **Tharindu Ranasinghe**[3] **and Ruslan Mitkov**[4]
[1]University of Wolverhampton, UK
[2]University of Granada, Spain [3]Aston University, UK [4]Lancaster University, UK
damith.premasiri@wlv.ac.uk, amalhaddad@ugr.es
t.ranasinghe@aston.ac.uk, r.mitkov@lancaster.ac.uk

Abstract

Multiword Terms (MWTs) are domain-specific Multiword Expressions (MWE) (Pajić et al., 2018) where two or more lexemes converge to form a new unit of meaning (León Araúz and Cabezas García, 2020). The task of processing MWTs is crucial in many Natural Language Processing (NLP) applications, including Machine Translation (MT) and terminology extraction. However, the automatic detection of those terms is a difficult task and more research is still required to give more insightful and useful results in this field. In this study, we seek to fill this gap by using state-of-the-art transformer models. We evaluate both BERT (Devlin et al., 2019) like discriminative transformer models and generative pre-trained transformer (GPT) (Radford et al., 2018) models on this task, and we show that discriminative models perform better than current GPT models in the identification of multiword flower and plant names for both English and Spanish. Best discriminative models perform with 94.3, 82.1 F1 scores in English and Spanish data, respectively, while ChatGPT could only return 63.3 and 47.7 F1 scores, respectively.

1 Introduction

Botany is a multidisciplinary field that encompasses different scientific disciplines such as Genetics, Ecology, Physiology, Biochemistry, Architecture, Gastronomy, Commerce, Art and Design, etc. One of the key areas in Botany is the study of flowers and plants. The market of flowers and plants is regarded as an economic engine of different economic and industrial activities. For this reason, its study and analysis are considered relevant in order to make this domain more accessible to all users, both at scientific and professional levels and also at layperson level, as flowers and plants have important national and international symbolisms and their roots are profoundly embedded in cultures and societies. The accurate identification and denomination of each plant is essential for the correct development and dissemination of science in all those multidisciplinary fields. It is also crucial for the correct communication of knowledge in different languages and also for the proper design of lexicographic resources and thesaurus.

From the point of view of applied linguistics, the identification of names of flowers and plants is relevant to language professionals. From a terminological point of view, it helps in laying the basis of term coining processes and gives insights into the underlying mechanisms of term creation. Translators also benefit from this information for the translation process.

Taking into consideration the quick development of NLP technologies and the importance of Machine translation (MT) in the dissemination of knowledge and in building new resources, it is important to extend the studies and cover new areas of research, such as Botany. The automatic identification of terms in this field helps in improving the quality of NLP applications, computer assisted translation tools and automatic translation tools (Temmerman and Knops, 2004) as well as lexicon creation, acquisition of novel terms, text classification, text indexing, machine-assisted translation and other NLP tasks (Pajić et al., 2018). For this reason, in this paper, we focus on the automatic extraction of flower and plant names, and we intend to address the shortcomings in this domain with the help of AI.

Specialised texts are rich with polylexical and monolexical terms (Estopà et al., 2000). They are both essential for efficient scientific and technical communication. Monolexical terms are formed of single lexical units, while Polylexical terms are formed of more than one lexical unit. Those last ones are also called Multiword Terms (MWT) and are defined as domain-specific Multiword Ex-

Proceedings of Recent Advances in Natural Language Processing, pages 879–887
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_095

pressions (MWE) (Pajić et al., 2018) where "two or more lexemes converge to form a new unit of meaning" (León Araúz and Cabezas García, 2020). MWTs are content-rich and are the most frequent type of lexical units in specialised discourse (Ibekwe-SanJuan and SanJuan, 2009). In this context, a term is defined as the linguistic designation of specialised concepts (Faber and Montero-Martínez, 2019).

In terminographic and lexicographic studies, the detection and analysis of terms are considered key to comprehending and deciphering the semantic and conceptual relations that connect one lexical unit with the other to construct meaning (Leroyer and Køhler Simonsen, 2021) properly. Those semantic and conceptual relations also have an important role in the construction of specialised domains, ontologies and terminographic resources (Faber et al., 2012). Moreover, they are also considered important for knowledge representation (Faber, 2015).

However, the detection of terms in specialised domains is not an easy task. Language users, such as professionals in specialised domains, terminologists and translators, need to acquire certain skills to be qualified to detect terms. The task is even more difficult in the cases of MWTs, as language users find it more difficult to delineate where the MWT starts and where it ends in context. Failure to detect terms leads to communicative problems, hinders the adequate construction of discourse, and provokes errors in translation processes.

Recently, Automatic Term Recognition (ATR) and Automated Term Extraction (ATE) have become more crucial to many NLP applications (Lang et al., 2021) and (Al Khatib and Badarneh, 2010). For example, those techniques are used for digital indexing, hypertext linking, text categorisation as well as in MT.

Moreover, the automatic detection of MWTs at cross-linguistic level in specialised domains is also becoming more important and its study may help in different multidisciplinary research (Temmerman and Knops, 2004). For this reason, automatic translation of all types of texts is becoming an urgent priority in all fields, and more research is still required in order to obtain more insightful results.

For this reason, and as a preliminary approach to the automatic extraction of MWTs in specialised domains, in this study, we provide the results of a case-study for the ATR and ATE in the domain of Botany in English and Spanish. To the best of our knowledge, there are no programs that could automatically identify and retrieve those terms both as single-word terms and MWTs in specialised domains, and no studies compare the already available resources in a comprehensive way. Hence, this study seeks to fill in this gap and proposes a novel method based on transformer models (Premasiri et al., 2022; Ranasinghe et al., 2021) for the automatic extraction of terms from the specialised domain of Botany[1]. At the same time, it compares the results obtained by ChatGPT to draw on conclusive results associated with their efficiency and whether they are promising to be used in further related research in different areas.

The main contributions of this study are:

1. We empirically evaluate 13 popular discriminative transformer models in MWT identification in flower and plant names in both English and Spanish.

2. We empirically compare the results with ChatGPT to explore its capabilities on the same task.

3. We release our open-source code repository[2] for the community to further research the topic.

The rest of the paper is structured as follows. Section 2 outlines related work. Section 3 describes the dataset used for our experiments, while section 4 presents the methodology. Section 5 reports the evaluation results, and finally, section 6 summarises the conclusion of this study and suggests future research.

2 Related Work

In recent years, the computational treatment of MWEs and MWTs has received considerable attention, as it is essential for NLP applications, such as MT, indexing, terminology retrieval and Translation Technologies (Monti et al., 2018). They are considered relevant and highly important due to their ubiquity in both natural language and specialised language (Ramisch and Villavicencio, 2014). Ramisch and Villavicencio (2014) highlight the importance of those terms in relation to

[1] The names of flowers and plants are considered as terms in the field of Botany by many scholars but given the differing views we have chosen the more 'neutral' wording 'Multiword Flower and Plant names'.

[2] https://bit.ly/474l9zY

NLP applications and propose including MWEs and MWTs in language technologies by means of type-based discovery, token-based identification, and MWE-aware language technology application models.

Studies such as Wang et al. (2023) show how the study of those terms may be relevant to detect synonym relations within distributional semantic models by using lexical substitution based and analogy based methods. Others such as Thanawala and Pareek (2018), show how the automatic detection of MWTs is useful in tasks related to automatic formation of compound concepts within Ontologies.

Within the field of language processing of specialised domains, previous research focused on the automatic detection of MWTs in discourse. For example, Pajić et al. (2018) used frequencies of occurrence of a text sequence in the corpus, combined with normalisation by lemmatising word by word in order to achieve the semi-automatic extraction of MWTs in the domain of Agricultural Engineering. Some authors such as Bonin et al. (2010) used the approach of identifying candidate MWTs in an automatically POS–tagged and lemmatised text, which is then weighted with the C-NC value in the domains of History of Art and in Legal domains. On the other hand, authors like Adjali et al. (2022) centred their research on the automatic extraction of MWTs from parallel corpora by using the Compositional with Word Embedding Projection (CMWEP) approach in the domain of Medicine.

Transformers based models have been used in previous research to detect MWTs, such as (Bechikh Ali et al., 2023). Their study focuses on detecting MWT for filtering and indexing tasks. Walsh et al. (2022) apply MWT extraction in Irish, but they show that large pre-trained models struggle to perform better in a low-resource setting. Chakraborty et al. (2020) employed transformers to evaluate MWT extraction in their own private dataset, and they could show that transformers were able to outperform the existed state-of-the-art results by greater margins. Studies have been limited because of the lack of annotated datasets, but Fusco et al. (2022) proposes an unsupervised way of annotations to combine with transformers to extract MWE.

Other studies, such as Lang et al. (2021), also use transformer-based approaches to multilingual term extraction across domains. However, they believe more research based on neural models is still required to obtain more results.

In this research, we combine the approaches to employ those methods on MWTs in the specialised domain of Botany, more specifically, on flower and plant names. In this case study, and since both MWTs detection and NER tasks are about token classification, they can be modelled by using similar models. For this reason, we are using a set of models which are used in NER for the MWT detection task, too (Rohanian et al., 2019). We seek to fill the gap by empirically evaluating multiple transformers in the task of MWT identification and extraction in the domain of Botany in English and Spanish.

3 Data

For the implementation of this case study, we extracted terms from different texts in English and Spanish corpora. With respect to the English corpora used, firstly we compiled a corpus from the Encyclopaedia of Flowers and Plants available in a digitalised editable format, published by the American Horticultural Society (Brickell, 2012). This encyclopedia contains more than 8,000 plants and 4,000 photographs and is organised in different sections to serve all users. The first section provides information on how to use the book and explains the origin of the names of plants and their etymological origins. In the second section, it has a comprehensive plant catalogue which explains the type of plants, including information on their plant life cycle, their shape and size, and whether they are trees, shrubs, roses, bulbs, etc., or if they are water or rock plants, etc. Finally, the encyclopedia offers a plant dictionary followed by an index of common names and a glossary of terms.

The advantage of annotating this encyclopedia is that the scientific names will help as a common link in all languages written with the Latin alphabet. It also has an important potential at cross-linguistic level in the field of Botany. The data was preprocessed by annotating the proper names and their condition of being MWTs or single-word terms. For example, the scientific name *Cynoglossum amabile* is annotated as MWT, while the vernacular name of this flower, *Firmament*, is annotated as a single word term.

Apart from the Encyclopaedia of Plants and Flowers (Brickell, 2012), we also compiled a corpus of other resources related to Botany in English. It consists of 437,663 words. Some of the texts are monographs, others are journal articles, and some texts are retrieved from other online resources. Those resources are Vigneron et al. (2005), Maghiar et al. (2021), Pink (2008), Blanco-Pastor et al. (2013), Ni et al. (2022). All those resources contained lists of names of plants and flowers, which were also annotated, and they all had relevant rich contexts on which we could rely to extract terms.

With respect to the Spanish dataset, we followed the same annotation criteria implemented in annotating the English dataset. The dataset in Spanish consisted of a list of flower and plant names provided in selected monographs and glossaries. Above all, we used books and articles in the domain of Botany and botanical glossaries, such as the glossaries provided in *Los Árboles en España* (de Lorenzo Cáceres, 1999), *Biología de la Conservación de Plantas en Sierra Nevada* (Peñas et al., 2019) as well as the glossary of scientific names of plants and their vernacular names provided by the Entomological Museum in Leon on the Bio-Nica webpage [3].

In order to obtain more context-rich corpora, we also used other texts in Spanish, such as Peñas and Lorite (2019), Guadalupe et al. (1985), Blanca López and Loépez Onieva (2002), Gonzáles et al. (2020), Montserrat (1960), ARMAS, Gómez García (2004) and the *Vademecum Colombiano de Plantas Medicinales* (de Salud y Protección Social de Colombia, 2008).

For example, *Los Árboles de España* includes a classification of trees in Spain. Above all, it describes their varieties, form and cultivation process and needs. It has glossaries with scientific names and family names. Other scientific articles, such as *Biología de la Conservación de Plantas en Sierra Nevada* contain tables with names of Endemic plants and flowers in the National Park of Sierra Nevada. The variety of resources allows for the list to be more inclusive. The same applies to the book *Vademecum Colombiano de Plantas Medicinales* (de Salud y Protección Social de Colombia, 2008) as it includes varieties of terms more specific to a concrete geographical area, in this case, in Colombia.

Data Preparation In general, Multi-word Term (MWT) identification tasks have been modelled as token-level classification tasks in NLP. These tasks need token-level tags which could identify the relevant parts in the sequence. We used IOB tagging for this purpose, inspired by CoNLL-2003 shared task (Tjong Kim Sang and De Meulder, 2003). Each word in a sentence has its token depending on whether the word is related to a MWT or not. B - Beginning, I - Inside and O if the word is Outside of a Multi-word term as shown in Table 1. We did IOB annotation on the corpus using an algorithm we developed based on the human annotated multi-word-term annotations on flowers and plants dataset.

Tag	B	I	O	O	O	O
Word	Blue	Moon.	Slow-growing,	compact,	clump-forming	perennial.

Table 1: Sample IOB tags for a sentence

Tagging disclosed us to the statistics of the dataset, in which we observed that the vast majority of the sentences in the corpus did not contain any multiword flower or plant names. Initial experiments showed us these sentences lead to overall poor results. This encouraged us to balance the datasets by removing a set of sentences which contained only 'O' tags. This is an important step in deep-learning-based models to balance the data with fair margins. Table 2 shows the breakdown of each dataset for train and test splits.

Dataset	Train Sentences	Test Sentences
English	1500	505
Spanish	750	250

Table 2: Breakdown of datasets

The tagged version of datasets is used for the training and testing of BERT-like models. Since we do not have enough corpus for further finetuning the GPT model to our task, we only performed testing using prompts. Therefore, we kept the sentences as is for GPT experiments.

4 Methodology

With the emergence of Transformers (Vaswani et al., 2017) and large language models (LLMs), state-of-the-art results of many NLP tasks had pushed their existing boundaries with decent margins. Attention mechanism (Vaswani et al., 2017) played a major part in these language models, which could provide a contextual understanding of the left and right sides of a text sequence at once. BERT (Devlin et al., 2019) was a prominent

[3] http://www.bio-nica.info/home/index.html

milestone in LLMs which is a variant from initial Transformers architecture. Similar LLM architectures have emerged with the differences of having different learning objectives as well as using different datasets. Having this motivation, we conduct our experiments on multiple popular transformer models to evaluate their performance on MWT extraction in flower and plant names.

Since this is a token-level classification task, we use macro averaged Precision, Recall and F1 score as our evaluation metrics.

$$Precision = TP/(TP + FP) \tag{1}$$

$$Recall = TP/(TP + FN) \tag{2}$$

$$\text{F1} = 2 * (\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall}) \tag{3}$$

The rest of this section discusses the models we used, with the categorisation of discriminative and generative models.

Discriminative Models The Original Transformer (Vaswani et al., 2017) consisted of two main parts; encoder and decoder. BERT model can be described as a stack of encoders which has been pre-trained on masked language modelling primary objective function. Generally, these discriminative models accept a sequence of tokens, and the output layer of the model can be configured such that the model is able to finetune on a downstream task such as classification. The general architecture of BERT models on token-level classification tasks is shown in Figure 1.

We used a mix of popular discriminative transformer models in our experiments with their variants as listed in Table 3.

For the experiments on the English corpus, we used all the models listed in Table 3. We considered multilingual models, mono-lingual models and different architectures like Electra and Scibert since it is specifically trained on scientific corpora.

Since not all these models have multilingual capabilities, we used bert-base-multilingual-uncased, bert-base-multilingual-cased, xlm-roberta-base, xlm-roberta-large for Spanish experiments.

We used model training configurations shown in Table 4 on a GeForce RTX 3090 GPU hardware.

Generative Models These models took a different approach to BERT-like models, by changing the objective function to predict only the next word. This variant of transformers leverages the decoder part of the initial Transformer architecture, and a Generative Pretrained Transformer (GPT) can be introduced as a stack of decoders in terms of the architecture. ChatGPT[4] uses generative transformer architecture, and it has provided highly competitive results in conversational systems while it is capable of applying to non-conversational tasks like multi-word-terms identification.

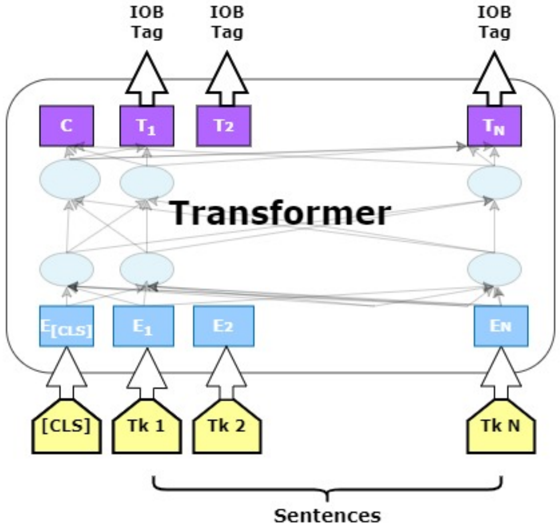

Figure 1: Transformer architecture on token level classification

ChatGPT is a human-like chatbot in which we can input a sequence of text and get an output accordingly. It is known that ChatGPT produces different results for different inputs. Therefore, finding the optimal prompt for better results is always encouraged. We tried multiple prompts to retrieve BIO tags for MWTs in the text directly, but this did not show good results since the model produced more tags than the number of tokens in the input text. After a couple of iterations, we settled for; *- Find whether there is a multi-word expression flower or plant name in the text delimited by "' - if there is no multi-word expression found in the given text; just tell 'No' - if you find a multiword expression in the given text; say yes and then give the multiword flower or plant name for example; Yes - 'Name' Text : "'{sentence}"'*

As shown in the prompt, if there is no MWT in the given sentence, we retrieve 'NO' as the output and if there is, we retrieve 'Yes - {Name}' as the output. In both cases, we post process the data using regular expressions to generate BIO tags based on the ChatGPT output. Finally, we use the generated BIO tags to evaluate the results.

[4] https://chat.openai.com/

Model Name	Size	Variants
bert (Devlin et al., 2019)	base	cased, uncased
	large	cased, uncased
	base	multilingual-cased, multilingual-uncased
xlmr (Conneau et al., 2020)	base	cased
	large	cased
xlnet (Yang et al., 2019)	base	cased
roberta (Liu et al., 2020)	base	cased
electra (Clark et al., 2020)	base	discriminator
scibert (Beltagy et al., 2019)	base	scivocab_cased, scivocab_uncased

Table 3: Model names and variants

Parameter	Value
Training Batch Size	32
Evaluation Batch Size	8
Learning Rate	4e−5
Epochs	3
Early Stopping	No

Table 4: Training configurations

For ChatGPT experiments, we used **gpt-3.5-turbo** model since it is the free version provided at the moment, and we set the temperature parameter to 0 due to reproducibility reasons. Even though the latest version of GPT is GPT4 for the time being, we did not experiment with this version since it is not freely available.

5 Results and Discussion

English Table 5 shows the results for MWT identification of flower and plant names in English. It is noticeable that all the discriminative transformer models have produced highly competitive results, while bert-large-cased model performs 94.3127 F1 score as the best performer. The least successful discriminative model is xlm-roberta-large, but even this model scored 91.5564 showing that transformers are highly able to identify MWTs in flower and plant names. In comparison to discriminative models, ChatGPT has performed less, marking 63.3183 F1 score. Given the fact that we did not fine-tune the GPT model, we believe this is a very good score. Even though ChatGPT is leading in conversational AI models, there could be more areas, like MWT extraction in flower and plant names, where ChatGPT falls behind. We think there could be multiple reasons for this. One possibility could be that the GPT model does not see the words from both sides. Instead, it uses the left-side sequence only to predict the next token. Typically, this approach is good in general, but we feel that it does not perform equally well in multi-word term identification setting. However, extensive experiments will need to confirm this.

Spanish Similar to English results, Transformers show significant results on Spanish as highest F1 score of 82.1733 by bert-base-multilingual-cased model. Similar to English experiments, discriminative models showed very competitive results, but the difference between the highest performer and lowest performer increased by 7.6647. However, ChatGPT does not do well with 47.7925 F1 score. This confirms that ChatGPT is also capable of identifying Spanish MWTs, but there is still a long way to go.

6 Conclusions

Detection of terms is an important research area for many NLP applications and is considered a challenging task, above all when the task involves MWTs besides single-word terms. The automatic identification of terms helps in improving the quality of NLP applications, such as computer assisted translation tools and automatic translation tools, as well as lexicon creation, knowledge representation, ontology building, text classification, text indexing, creation of terminographic resources and other NLP tasks.

Those NLP applications need to be developed in all fields of study in order to widen the scope of NLP applications and be more inclusive. Botany is no exception. Moreover, there is a need to fill this void as Botany is one of the important interdisciplinary areas which is intertwined with many other activities and areas of research. Within the scope of Botany, we focus on the automatic extraction of

Model	Precision	Recall	F1
bert-base-uncased	95.5851	92.6156	94.0379
bert-base-cased	95.1363	92.8490	93.9485
bert-large-uncased	95.6642	92.4974	94.0190
bert-large-cased	95.1992	93.4754	**94.3127**
bert-base-multilingual-uncased	95.3530	93.1413	94.1751
bert-base-multilingual-cased	94.9715	92.5637	93.7326
xlm-roberta-base	93.2733	91.3631	92.2856
xlm-roberta-large	92.0389	91.1048	91.5564
xlnet-base-cased	94.1032	91.6107	92.7907
roberta-base	93.2224	92.2400	92.7225
google/electra-base-discriminator	95.6244	91.3245	93.3517
allenai/scibert_scivocab_uncased	95.3931	93.0981	94.1983
allenai/scibert_scivocab_cased	95.8673	92.4853	94.0875
ChatGPT	70.4278	59.6787	63.3183

Table 5: Results for multiword flower and plant names identification in English

Model	Precision	Recall	F1
bert-base-multilingual-uncased	81.7597	75.9625	78.6295
bert-base-multilingual-cased	81.8485	82.5835	82.1733
xlm-roberta-base	76.2378	73.3251	74.5086
xlm-roberta-large	83.4353	79.9430	81.5646
ChatGPT	58.8073	44.3087	47.7925

Table 6: Results for multiword flower and plant names identification in Spanish

terms of names of flowers and plants.

We empirically show that general transformer models can produce very good results in Multiword Term identification of flower and plant names tasks for both English and Spanish. Further, we comparatively show that ChatGPT is not performing as well as the other discriminative models.

The results obtained from this experiment can be relevant for the comprehension of term formation processes and may be helpful for the design of new lexicographic resources related to new term formation in languages with low resources.

In future research, we would like to explore more specialised domains and involve more languages and bigger datasets, and extend the study to multilingual parallel corpora.

7 Acknowledgements

Part of this research was carried out within the framework of the projects the projects PID2020-118369GB-I00 and A-HUM-600-UGR20, funded by the Spanish Ministry of Science and Innovation and the Regional Government of Andalusia. Funding was also provided by an FPU grant (FPU18/05327) given by the Spanish Ministry of Education.

References

Omar Adjali, Emmanuel Morin, and Pierre Zweigenbaum. 2022. Building comparable corpora for assessing multi-word term alignment. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3103–3112, Marseille, France. European Language Resources Association.

Khalid Al Khatib and Amer Badarneh. 2010. Automatic extraction of arabic multi-word terms. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, pages 411–418. IEEE.

CRISTINA ARMAS. Facilitación de las especies almohadilladas y cambio global en las comunidades alpinas del parque nacional de sierra nevada.

Chedi Bechikh Ali, Hatem Haddad, and Yahya Slimani. 2023. Multi-word terms selection for information retrieval. *Information Discovery and Delivery*, 51(1):74–87.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

G Blanca López and Mariéa Rosa Loépez Onieva. 2002. *Flora amenazada y endémica de Sierra Nevada*. Junta de Andalucía, Consejería de Medio Ambiente.

JL Blanco-Pastor, M Fernández-Mazuecos, and P Vargas. 2013. Past and future demographic dynamics of alpine species: limited genetic consequences despite dramatic range contraction in a plant from the s panish s ierra n evada. *Molecular Ecology*, 22(16):4177–4195.

Francesca Bonin, Felice Dell'Orletta, Giulia Venturi, Simonetta Montemagni, et al. 2010. A contrastive approach to multi-word term extraction from domain corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 19–21. Malta Valetta.

Christopher Brickell. 2012. Encyclopedia of plants and flowers. In *Encyclopedia of plants and flowers*, Santa Fe, New Mexico, USA. Dorling Kindersley.

Sritanu Chakraborty, Dorian Cougias, and Steven Piliero. 2020. Identification of multiword expressions using transformers.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Ministerio Ministerio de Salud y Protección Social de Colombia. 2008. *Vademecum colombiano de plantas medicinales*. El Ministerio de Salud y Protección Social de Colombia, Bogotá, colombia.

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.

Rosa Estopà, Jordi Vivaldi, MT Cabré, and Rambla Santa Mónica. 2000. Extraction of monolexical terminological units: requirement analysis. *Paper submitted to: Computational Terminology for Medical and Biological Applications, Patras, Greece. http://www. iula. upf. es/iulaterm.*

Pamela Faber. 2015. Frames as a framework for terminology. *Handbook of terminology*, 1(14):14–33.

Pamela Faber and Silvia Montero-Martínez. 2019. Terminology. In *The Routledge Handbook of Spanish Translation Studies*, pages 247–266. Routledge.

Pamela B Faber et al. 2012. *A cognitive linguistics view of terminology and specialized language*. De Gruyter Mouton Berlin, Boston.

Francesco Fusco, Peter Staar, and Diego Antognini. 2022. Unsupervised term extraction for highly technical domains. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1–8, Abu Dhabi, UAE. Association for Computational Linguistics.

Daniel Gómez García. 2004. Flora y vegetación de la jacetania.

Paúl Gonzáles, Asunción Cano, Tiina Särkinen, Zoë Goodwin, Niels Valencia, Inés Sachahuamán, and JL Marcelo-Peña. 2020. Las plantas comunes del bosque seco del marañón: Biodiversidad para las comunidades locales. *Lima: Paúl Henry Gonzáles Arce (Editor).*

M López Guadalupe, C Sierra Ruiz de la Fuente, and G Marín Calderón. 1985. Comunidades, hábitat y tipos de suelos sobre los que se desarrolla la manzanilla de sierra nevada. *Ars Pharmaceutica (Internet)*, 26(4):255–263.

Fidelia Ibekwe-SanJuan and Eric SanJuan. 2009. Use of multiword terms and query expansion for interactive information retrieval. In *Advances in Focused Retrieval: 7th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2008, Dagstuhl Castle, Germany, December 15-18, 2008. Revised and Selected Papers 7*, pages 54–64. Springer.

Christian Lang, Lennart Wachowiak, Barbara Heinisch, and Dagmar Gromann. 2021. Transforming term extraction: transformer-based approaches to multilingual term extraction across domains. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3607–3620.

Pilar León Araúz and Melania Cabezas García. 2020. Term and translation variation of multiword terms. *MonTI, 2020, Special Issue 6*.

P Leroyer and H Køhler Simonsen. 2021. Reconceptualizing lexicography: the broad understanding. *EURALEX XIX*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

José Manuel Sánchez de Lorenzo Cáceres. 1999. *Los Árboles en España: Manual de Identificación*. Mundi-Prensa, Spain.

Lăcrămioara M Maghiar, Ilie A Stoica, and Andrew J Tanentzap. 2021. Integrating demography and distribution modeling for the iconic leontopodium alpinum colm. in the romanian carpathians. *Ecology and Evolution*, 11(18):12322–12334.

Johanna Monti, Violeta Seretan, Gloria Corpas Pastor, and Ruslan Mitkov. 2018. Multiword expressions in machine translation and translation technology. In *Multiword Expressions in Machine Translation and Translation Technology*, pages 1–37. John Benjamin Publishers.

Pedro Montserrat. 1960. La flora del pirineo.

Lianghong Ni, Weitao Li, Zhili Zhao, Dorje Gaawe, and Tonghua Liu. 2022. Migration patterns of gentiana crassicaulis, an alpine gentian endemic to the himalaya–hengduan mountains. *Ecology and Evolution*, 12(3):e8703.

Vesna Pajić, Staša Vujičić Stanković, Ranka Stanković, and Miloš Pajić. 2018. Semi-automatic extraction of multiword terms from domain-specific corpora. *The Electronic Library*, 36(3):550–567.

Julio Peñas, Eva Cañadas, and Jesús Del Río. 2019. Fitogeografía de sierra nevada e implicaciones para la conservación. *Biología de la conservación de plantas en Sierra Nevada: Principios y retos para su preservación*, pages 81–116.

Julio Peñas and Juan Lorite. 2019. *BIOLOGÍA DE LA CONSERVACIÓN DE PLANTAS EN SIERRA NEVADA*. UNIVERSIDAD DE GRANADA.

Alfred Pink. 2008. *Dictionary of flowers and plants for gardening*. Teresa Thomas Bohannon.

Damith Premasiri, Tharindu Ranasinghe, Wajdi Zaghouani, and Ruslan Mitkov. 2022. DTW at qur'an QA 2022: Utilising transfer learning with transformers for question answering in a low-resource domain. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 88–95, Marseille, France. European Language Resources Association.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Carlos Ramisch and Aline Villavicencio. 2014. Computational treatment of multiword expressions.

Tharindu Ranasinghe, Diptanu Sarkar, Marcos Zampieri, and Alexander Ororbia. 2021. WLV-RIT at SemEval-2021 task 5: A neural transformer framework for detecting toxic spans. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 833–840, Online. Association for Computational Linguistics.

Omid Rohanian, Shiva Taslimipoor, Samaneh Kouchaki, Le An Ha, and Ruslan Mitkov. 2019. Bridging the gap: Attending to discontinuity in identification of multiword expressions. 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 2692–2698, Minneapolis, Minnesota. Association for Computational Linguistics.

Rita Temmerman and Uus Knops. 2004. The translation of domain specific languages and multilingual terminology management. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, 3.

Pratik Thanawala and Jyoti Pareek. 2018. Mwtext: automatic extraction of multi-word terms to generate compound concepts within ontology. *International Journal of Information Technology*, 10:303–311.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jean Pol Vigneron, Marie Rassart, Zofia Vértesy, Krisztián Kertész, Michaël Sarrazin, László P Biró, Damien Ertz, and Virginie Lousse. 2005. Optical structure and function of the white filamentary hair covering the edelweiss bracts. *Physical review E*, 71(1):011906.

Abigail Walsh, Teresa Lynn, and Jennifer Foster. 2022. A BERT's eye view: Identification of Irish multiword expressions using pre-trained language models. In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, pages 89–99, Marseille, France. European Language Resources Association.

Yizhe Wang, Béatrice Daille, and Nabil Hathout. 2023. Exploring synonymy relation between multi-word terms in distributional semantic models. In *10th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguisitics (LTC'23)*, pages 331–336.

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.