Procesamiento del Lenguaje Natural, Revista nº 67, septiembre de 2021, pp. 139-151 recibido 10-05-2021 revisado 08-06-2021 aceptado 10-06-2021

Extraction of Terms Semantically Related to Colponyms: Evaluation in a Small Specialized Corpus

Extracción de Términos Relacionados Semánticamente con Colpónimos: Evaluación en un Corpus Especializado de Pequeño Tamaño

Juan Rojas-Garcia
University of Granada, Granada, Spain
juanrojas@ugr.es

Abstract: EcoLexicon is a terminological knowledge base on environmental science, whose design permits the geographic contextualization of data. For the geographic contextualization of named entities such as colponyms (i.e., named bays such as *Pensacola Bay*) in EcoLexicon, both count-based and prediction-based distributional semantic models (DSMs) were applied to a small-sized, English specialized corpus to extract terms related to each colponym mentioned in it and their semantic relations. Since the evaluation of DSMs in small, specialized corpora has received little attention, this study identified both parameter combinations in DSMs and five similarity/distance measures suitable for the extraction of terms which related to colponyms through the semantic relations *takes_place_in*, *located_at*, and *attribute_of*. The models were thus evaluated using three gold standard datasets. The results showed that: count-based models outperformed prediction-based ones; the similarity/distance measures performed quite similar except for the Euclidean distance; and the detection of a specific relation depended on the context window size.
Keywords: Colponym, Terminology, Knowledge Representation, Semantic Model.

Resumen: EcoLexicon es una base de conocimiento terminológica sobre el medioambiente, cuyo diseño permite la contextualización geográfica de colpónimos, esto es, bahías con nombre propio (BNP) (v.gr., *Bahía de Pensacola*). Se aplicaron modelos semánticos distribucionales (MSD), basados en recuentos y predictivos, a un corpus especializado de pequeño tamaño en inglés para extraer términos relacionados con las BNP y sus relaciones semánticas. Puesto que la evaluación de MSD en corpus especializados de pequeño tamaño ha sido menos explorada, en este artículo se identifican tanto la combinación de parámetros como las cinco medidas de similitud adecuadas para extraer términos que mantengan con las BNP las relaciones *tiene_lugar_en*, *localizado_en* y *atributo_de*. Los MSD se evalúan con tres conjuntos de datos anotados manualmente. Los resultados indican que: los modelos basados en recuentos superan a los modelos predictivos; las medidas de similitud brindan resultados semejantes, excepto la distancia euclídea; y la detección de una relación específica depende del tamaño de la ventana contextual.
Palabras clave: Colpónimo, Terminología, Representación del Conocimiento, Modelo Semántico.

1 Introduction

Although named landforms, among other named entities, are frequently found in specialized texts on the environment, their representation and inclusion in terminological knowledge bases (TKBs) have received little research attention, as evidenced by the lack of named landforms in terminological resources for the environment such as DiCoEnviro[1], GEMET[2], or the FAO Term Portal[3]. In contrast, AGROVOC[4] contains a list of named landforms with hyponymic information, whereas ENVO[5] provides descriptions with only geographic details.

The semantic representation of named landforms, such as litonyms (e.g., *Sumiyoshi Beach*), potamonyms (e.g., *River Nile*), and colponyms (e.g., *San Francisco Bay*), is barely tackled in terminological resources for two reasons, in our opinion: (1) They are considered

[1] https://cutt.ly/cbATjnQ
[2] https://www.eionet.europa.eu/gemet/en/themes/
[3] http://www.fao.org/faoterm/en/
[4] http://aims.fao.org/en/agrovoc
[5] http://www.environmentontology.org/Browse-EnvO

ISSN 1135-5948. DOI 10.26342/2021-67-12

mere instances (i.e., examples) of concepts such as BEACH, RIVER, or BAY, and their relational behavior with other concepts in a specialized knowledge domain is thus neglected and not semantically described; (2) their semantic representation depends on knowing which terms are related to each named landform, and how these terms are related to each other. This is evidently a time-consuming task taking into account that terminologists do not often resort to natural language processing (NLP) systems beyond corpus query tools such as Sketch Engine (Kilgarriff A. et al., 2004).

As a result, knowledge resources have limited themselves to representing concepts such as BAY, RIVER or BEACH, on the questionable assumption that the concepts linked to each of them are also related to all named bays, rivers and beaches in the real world. Contrary to this assumption, Rojas-Garcia J. and Faber P. (2019a and 2019b) have shown that, in specialized knowledge domains, each named landform reveals a specific conceptual structure. In other words, each named landform holds different semantic relations to specialized concepts even in the same knowledge domain. Therefore, TKBs should include the semantic representation of named landforms.

In this respect, EcoLexicon[6] is a multilingual TKB on environmental science that is the practical application of Frame-based Terminology (Faber P., 2012). The flexible design of EcoLexicon permits the representation and contextualization of data so that they are more relevant to specific subdomains, communicative situations, and geographic areas. With the ultimate goal of representing in EcoLexicon the conceptual structures underlying the usage of named landforms mentioned in a small-sized, English specialized corpus on Costal Engineering (7 million tokens), the terms related to each named landform and their semantic relations have to be manually extracted from the corpus by querying it in Sketch Engine. In this work, we focused on colponyms (Room A., 1996: 23), namely, named bays.

As such, terminologists require to extract terms which relate to each colponym, at least, by the semantic relations *takes_place_in*, *located_at*, and *attribute_of*, the most frequent relations held by named bays in the corpus. Since this is a time-consuming task, the overall aim of this study was to provide terminologists with three lists of term candidates for a colponym, one list per semantic relation, by applying distributional semantic models (DSMs).

Accordingly, this study identified both parameter combinations in DSMs and similarity/distance measures suitable for the extraction of those terms from the small specialized corpus mentioned above. Hence, the models were evaluated using gold standard evaluation data, which contained pairs of semantically related terms, manually extracted from the same corpus. One of the terms was always a colponym, and the other one was either a process (e.g., *storm surge*), an entity (e.g., *benthic geologic habitat*), or a property (e.g., *water quality*). The semantic relations that linked the terms were: (a) *takes_place_in* (e.g., STORM SURGE *takes_place_in* ESCAMBIA BAY); (b) *located_at* (e.g., BENTHIC GEOLOGIC HABITAT *located_at* GREENWICH BAY); and (c) *attribute_of* (e.g., WATER QUALITY *attribute_of* NARRAGANSETT BAY). Three gold standard datasets were thus built, one for each of the semantic relations.

As shall be seen, the extraction of terms that hold these specific semantic relations to named bays largely depends on the context window size parameter of the DSMs, namely, 4 words for *takes_place_in*, 3 words for *attribute_of*, and 2 words for *located_at*. A similar study was also conducted for named rivers by Rojas-Garcia J. and Faber P. (2019c), but the relations frequently activated were *takes_place_in*, *located_at*, and *affects* (not *attribute_of*). Interestingly enough, for named rivers, the window size had to be 3 words to extract terms linked to rivers with the *affect* relation, whereas in the case of named bays, the same window size of 3 words was required to obtain terms that held the *attribute_of* relation. These findings led to the conclusion that it is not possible to generalize the results from named rivers to either bays or other named entities such as beaches and mountains. Hence, since each named landform is characterized by its own conceptual structure, as previously stated, this study on colponyms cannot be considered either as a "case of use" or as a "toy problem", but rather as a research objective itself.

Besides the analysis of different DSMs and similarity measures for a small-sized, specialized corpus, an important contribution of this work is the creation of both the corpus on named landforms in the Coastal Engineering domain, and the three gold standard datasets for information retrieval system evaluation.

The rest of this paper is organized as follows. Section 2 provides background on DSMs, as well as a literature review on their application and evaluation. Sections 3 and 4 explain the materials, methods, and DSMs evaluation applied in this study, and the construction of the gold standard datasets. Section 5 shows the results obtained. Finally, Section 6 discusses the results, and presents the conclusions derived from this work along with plans for future research.

[6] http://ecolexicon.ugr.es

2 Background and Literature Review

Distributional semantic models (DSMs) represent the meaning of a term as a vector, based on its statistical co-occurrence with other terms in the corpus. According to the distributional hypothesis, semantically similar terms tend to have similar contextual distributions (Miller G.A. and Charles W.G., 1991). The semantic relatedness of two terms is estimated by calculating a similarity/distance measure of their vectors, such as Euclidean distance, cosine similarity, Jaccard coefficient, Pearson correlation coefficient, or averaged Kullback-Leibler divergence, *inter alia* (see Huang A. (2008) for a detailed description of these five measures).

Depending on the language model (Baroni M. et al., 2014), DSMs are either count-based or prediction-based. Count-based DSMs calculate the frequency of terms within a term's context (i.e., a sentence, paragraph, document, or a sliding context window spanning a given number of terms on either side of the target term). Correlated Occurrence Analogue to Lexical Semantic (COALS) (Rohde D. et al., 2006) is an example of this type of model.

Prediction-based models exploit probabilistic language models, which represent terms by predicting the next term on the basis of previous terms. Examples of predictive models based on neural networks include, among others, *word2vec* (Mikolov T. et al., 2013), *fastText* (Bojanowski P. et al., 2017), and state-of-the-art transfer learning models such as BERT (Devlin J. et al., 2019). Instead, *GloVe* model (Pennington J. et al., 2014) makes predictions drawn on a regression technique.

Count-based DSMs have been amply studied (Kiela D. and Clark S., 2014; Lapesa G. et al., 2014; Sahlgren M. and Lenci A., 2016). Research shows that parameters, such as the context window size, influence the semantic relations that are captured, either syntagmatic relations or paradigmatic relations (i.e., synonymy, antonymy, hyponymy, and meronymy). The syntagmatic relations examined in much research are either phrasal associates (e.g., *help - wanted*) (Lapesa G. et al., 2014) or syntagmatic predicate preferences (Erk K. et al., 2010) in general language. The present study focused on the specific syntagmatic relations *takes_place_in*, *located_at*, and *attribute_of*, which were the most frequent relations activated by colponyms in the specialized language of Coastal Engineering in our corpus.

Count-based models and *word2vec* have also been recently compared. Baroni M. et al. (2014) contrasted them on several datasets and found that the prediction-based models provided better results. In contrast, Ferret O. (2015) found that count-based models performed better. In another study that compared the ability of both DSMs to capture paradigmatic relations (synonymy, antonymy, and hyponymy) and syntactic derivatives, Bernier-Colborne G. and Drouin P. (2016) not only observed that the semantic relations detected by the DSMs depended on the window size, but also that the values of this parameter mostly coincided in both DSMs.

Levy O. et al. (2015) yielded valuable insights, showing the following: (1) When the parameters of the models were correctly tuned, count-based and prediction-based models obtained similar accuracy; and (2) the best model depended on the task to be carried out. Nevertheless, Asr F. et al. (2016), Sahlgren M. and Lenci A. (2016), and Nematzadeh A. et al. (2017) reported that count-based models outperformed prediction-based ones on small-sized corpora of under 10 million tokens.

Work in lexical semantics and DSMs includes, *inter alia*, the identification of semantic relations (Bertels A. and Speelman D., 2014), classification of verbs into semantic groups (Gries S. and Stefanowitsch A., 2010), and the use of word vectors as features for automatic recognition of named entities in text corpora (El Bazi I. and Laachfoubi N., 2016).

3 Materials

3.1 Corpus Data

The colponyms and related terms were extracted from a subcorpus of English texts on Coastal Engineering, on which the DSMs were also built. This subcorpus, comprising roughly 7 million tokens, is composed of specialized texts (scientific articles, technical reports, and PhD dissertations), and semi-specialized texts (textbooks and encyclopedias on Coastal Engineering). It is an integral part of the EcoLexicon English Corpus (23.1 million tokens) (León-Araúz P. et al., 2018).

It is worth clarifying that we were interested in the semantic behavior of colponyms in the specialized language of Coastal Engineering. Since this behavior of colponyms, like that of all specialized terms, is different in the specialized language than it is in the general language (Pearson J., 1998; Sager J.C. et al., 1980), from an epistemological and methodological point of view, it makes no sense to expand our corpus neither with a general language corpus such as Wikipedia, nor with other specialized corpora dealing with topics other than Coastal Engineering.

Furthermore, the domain of the training corpus has an impact on the semantic relations represented by word embeddings. Hence, it is recommended using a domain-specific corpus to train word embeddings for domain-specific text mining tasks

(Chen Z. et al., 2018). Consequently, it also makes no sense to create meta-embeddings joining specialized and general pre-trained embeddings.

3.2 GeoNames Geographic Database

The automatic detection of the colponyms in the corpus was performed with a GeoNames database dump. GeoNames[7] has over 10 million proper names for 645 different geographic entities, such as bays, beaches, and rivers. For each entity, information about their normalized designations, alternate designations, latitude, longitude, and location name is stored.

3.3 Gold Standard Datasets

The DSMs, built on our domain-specific corpus, were evaluated on gold standard data. We were unable to find gold standard resources suitable for evaluating systems that link semantically related terms to a given colponym in the domain of Coastal Engineering. Consequently, the gold standard data were manually extracted from the same corpus and assessed by Terminology experts on Coastal Engineering, a common evaluation practice both in Information Retrieval (Manning C.D. et al., 2009: 164-166) and linguistic annotation in corpora (Ide N. and Pustejovsky J., 2017: 297-313). In doing so, the research community could also employ our corpus and the gold standard data as test collection for the evaluation of systems dealing with semantic relation extraction from specialized corpora.[8]

The gold standard datasets contained pairs of semantically related terms, in which the semantic relations were *takes_place_in*, *located_at*, and *attribute_of*. Three gold standard datasets were thus built, one for each of the semantic relations. The designations and meaning of these relations are those used in EcoLexicon (Faber P. et al., 2009).

The three semantic relations always linked the normalized designation of a colponym (e.g., *Josiah's Bay* and *Josiah Bay* were normalized to *Josias Bay*) to either a process, an entity, or a property expressed by a noun or noun phrase, whether monolexical (e.g., *flooding*) or multiword (e.g., *high water mark*). More specifically, the *takes_place_in* relation holds between a process (e.g., *storm surge*) and the bay where the process occurs (see Table 1). The *located_at* relation indicates the location of an entity (e.g., *inundation area*) in a bay (see Table 2). Finally, the *attribute_of* is used for terms that designate properties (e.g., *wind speed*) of a bay (see Table 3).

process	***takes_place_in***	**named bay**
storm surge	*takes_place_in*	Escambia Bay
flooding	*takes_place_in*	Pensacola Bay
geological process	*takes_place_in*	Narragansett Bay

Example from the corpus:
(1) *Within the Pensacola Bay and **Escambia Bay**, the shallow estuarine water induces significant **storm surge**...*

Table 1: Extract from the first gold standard dataset for the *takes_place_in* relation.

entity	***located_at***	**named bay**
inundation area	*located_at*	Pensacola Bay
Port Geelong	*located_at*	Port Phillip Bay
benthic geologic habitat	*located_at*	Greenwich Bay

Example from the corpus:
(1) *The **Port Geelong** located on **Port Phillip Bay** has a significant role in coastal governance arrangements.*

Table 2: Extract from the second gold standard dataset for the *located_at* relation.

property	***attribute_of***	**named bay**
water quality	*attribute_of*	Narragansett Bay
wind speed	*attribute_of*	Mobile Bay
high water mark	*attribute_of*	Pensacola Bay

Example from the corpus:
(1) *... the simulated and observed **high water marks** at six stations around **Pensacola Bay** and Escambia Bay agree...*

Table 3: Extract from the third gold standard dataset for the *attribute_of* relation.

In addition to what has been described, each of the three datasets included: (1) 100 triplets for the corresponding semantic relation, which were all used for the evaluation, therefore, the three datasets added up to 300 triplets; (2) the 50 most frequently mentioned bays in the corpus, the same 50 bays in the three datasets, since 50 information needs have usually been found to be a sufficient minimum for information retrieval system evaluation (Manning C.D. et al., 2009: 152); and (3) the two most frequent terms related to the same bay, which amounted to 100 triplets, therefore, the same bay was related to a total of six terms, two terms in each dataset.

The semantic relation annotation of the pair of terms extracted from the corpus was carried out by three terminologists from the LexiCon research group of the University of Granada (Granada, Spain), with wide experience in environmental knowledge representation. *Cohen's kappa* coefficient was used as the statistical measure of inter-annotation agreement, and the scores for all the annotator pairs stood over 90% (p-value<0.05 for all the annotator pairs).

[7] http://www.geonames.org

[8] The datasets and the corpus will be available on the website of the LexiCon research group of the University of Granada (Granada, Spain) (http://lexicon.ugr.es/).

4 Methodology

4.1 Pre-processing

The corpus texts were tokenized, tagged with parts of speech, lemmatized, and lowercased with the Stanford *CoreNLP* package (Manning C.D. et al., 2014) for R programming language. The multi-word terms stored in EcoLexicon were then automatically matched in the lemmatized corpus and joined with underscores.

In the DSMs, only terms larger than two characters were considered. Numbers, symbols, and punctuation marks were removed. Since closed-class words are often considered too uninformative to be suitable context words, stopwords were not used (i.e., determiners, conjunctions, relative adverbs, and prepositions). Additionally, the minimal occurrence frequency was set to 5 so that the co-occurrences were statistically reliable (Evert S., 2008).

4.2 Named Bay Recognition

Both normalized and alternate names of the bays in GeoNames were searched in the lemmatized corpus. The recognized designations were normalized and automatically joined with underscores. Most bays of the corpus were in GeoNames (90%), while others were identified by manual inspection (10%). Anaphoric elements referring to a bay were replaced by the corresponding colponym in the lemmatized corpus. For this task, the anaphora resolution function from *CoreNLP* package was used, and other cases were manually replaced. The 294 bays mentioned in the corpus are shown on the map in Figure 1.

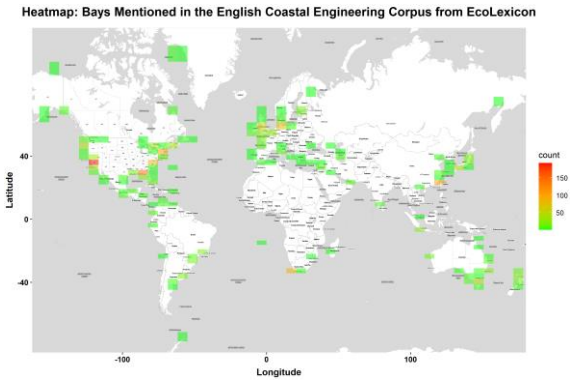

Figure 1: Heatmap with the location and color-coded frequency of the 294 named bays.

4.3 Construction of the DSMs

Our experiment involved a comparative evaluation of three types of DSM for a small-sized, specialized corpus, namely, count-based, prediction-based, and pre-trained models. The model types produced the vector representation of a term based on the contexts in which it appeared in our corpus. For this study, the contexts of a target term (i.e., a colponym) were the terms that co-occurred with it inside a sliding context window, which spanned a certain number of terms on either side of the target term.

The count-based and prediction-based DSMs have various parameters that must be set to build the models. The parameters impinge on both the term representations and the accuracy of the similarity scores between term vectors when the models are compared (Baroni M. et al., 2014). Therefore, to assess the influence of the parameters of both DSMs on their ability to capture the three semantic relations targeted in this study, various settings for each parameter were tried, and the combinations of these parameter settings were evaluated.

4.3.1 Parameter Setting of the Count-based Models

The first model type evaluated was a count-based model, also called bag-of-words (BOW) model. The BOW model was built with the R package *quanteda* (Benoit K. et al., 2018) for text mining.

To build a BOW model, a term-term matrix of co-occurrence frequencies was first computed, according to a specific size for the sliding context window. Then, the matrix was subjected to a specific weighting scheme, namely, an association measure that increases the importance of the context terms that are more indicative of the meaning of the target term. The 1,000 most frequent terms were used, which included all the colponyms and terms stored in the three evaluation datasets.

Regarding the context window, we tested size values ranging from 1 to 10 words on either side of the target term, and the context window was allowed to span sentence boundaries. The context window shape was always rectangular (i.e., the increment added to the co-occurrence frequency of a pair of terms was always 1, regardless of the distance between the two terms inside the context window). The frequencies observed on the left and right of a target term were added.

With respect to the weighting schemes, three association measures, defined in Evert S.'s (2008) work on collocation, were tested: (1) statistical log-likelihood; (2) positive pointwise mutual information (PPMI); and (3) *t*-score. Log-likelihood and PPMI are widely used in computational linguistics, whereas *t*-score is popular in computational lexicography (Evert S. et al., 2017).

Research in computational linguistics reveals that log-likelihood is able to capture syntagmatic and paradigmatic relations (Lapesa G. et al., 2014), and perform better for medium- to low-frequency data than other association measures (Alrabia M. et al.,

2014). PPMI and *t*-score, on the other hand, have been found to work adequately for different applications in previous research when compared to other association measures (Baroni M. et al. 2014; Kiela D. and Clark S., 2014).

Finally, following Lapesa G. et al. (2014), the association scores were transformed to reduce skewness in this way: log-likelihood and PPMI scores were both transformed by adding 1 and calculating then the natural logarithm (ln), whereas *t*-scores were transformed by calculating the square root (sqrt).

The settings tested for each of the two parameters were:

1. Size of the context window: 1-10 words.
2. Weighting scheme: ln(log-likelihood + 1), ln(PPMI + 1), sqrt(*t*-score).

4.3.2 Parameter Setting of the Prediction-based Models

Three prediction-based models were evaluated, namely, the *word2vec* (Mikolov T. et al., 2013), the *fastText* (Bojanowski P. et al., 2017), and the *GloVe* (Pennington J. et al., 2014) models. In *word2vec* (W2V), the term vectors are learned by training a neural network on a corpus according to two different architectures. The continuous bag-of-words (CBOW) architecture predicts the target term based on its context terms, while the skip-gram architecture predicts the context terms of a target term. The W2V model was built with the original *word2vec* package.[9]

For W2V, five hyperparameters were examined, the same as those tested by Bernier-Colborne G. and Drouin P. (2016) for paradigmatic relations and syntactic derivatives. The first one was the architecture used to learn the term vectors. The second one was the training algorithm, either using a hierarchical softmax function, or by sampling negative examples, in which case the number of negative samples must be selected. The third hyperparameter was the subsampling threshold for frequent terms, namely, some occurrences of those terms whose relative frequency in the corpus is greater than a threshold, are randomly deleted before the model is trained. Finally, the dimensionality of the term vectors, and the size of the context window were the other hyperparameters.

The settings tested for each of the five hyperparameters were:

1. Architecture: CBOW or skip-gram.
2. Negative samples: 5, 10 or none (in this case, hierarchical softmax is used).
3. Subsampling threshold: low (10^{-5}), high (10^{-3}) or none.
4. Dimensionality of term embeddings: 100 or 300.
5. Size of context window: 1-10 words.

In the *fastText* model (FTX), which is essentially an extension of the W2V model, each word is treated as composed of subwords, namely, all the substrings contained in a word between a minimum and a maximum size. Hence, the vector for a word is made of the sum of these subword vectors. The FTX model was built with the original *fastText* package.[10] For FTX, the same five hyperparameters as those for W2V were probed. All the subwords between 3 and 6 characters were taken (default values for the model).

The *GloVe* model optimizes the likelihood of term probabilities, based on context, to learn term representation as in CBOW, but uses ratios of co-occurrence probabilities as the basis for learning. The model was built with the original *GloVe* package,[11] and two hyperparameters were explored:

1. Dimensionality of term embeddings: 100 or 300.
2. Size of context window: 1-10 words.

In addition, for both GloVe, W2V, and FTX, the number of epochs was fixed to 10, and the learning rate to 0.05.

4.3.3 Pre-trained Models

Pre-trained word vectors, estimated from exceptionally large, general corpora, typically improve the performance of NLP systems (Baroni, M. and Lenci A., 2010). For that reason, we also assessed the pre-trained *word2vec* and *fastText* models (Mikolov T. et al., 2018),[12] and the pre-trained *GloVe* model,[13] all of them trained on the Common Crawl corpus (600-840 billion tokens) with 300-dimension vectors. The pre-trained BERT deep learning model was also considered.

The parameter values of the pre-trained models were already set in the pre-training phase. For instance, the context window size of the pre-trained *word2vec* and *fastText* models was fixed to 15 words, and that of the pre-trained *GloVe* model was fixed to 10 words. Consequently, the window size of these three pre-trained models could not be modified for our evaluation. This was deemed to be a drawback with respect to the overall goal of this study, since it aimed to provide terminologists with three lists of term candidates for a colponym, one list per semantic relation. Instead, a pre-trained model could only extract a single list of term candidates for a colponym.

Another downside to the pre-trained *word2vec*, *fastText*, and *GloVe* models was found. Despite the

[9] https://code.google.com/archive/p/word2vec/

[10] https://github.com/facebookresearch/fastText

[11] https://nlp.stanford.edu/projects/glove/

[12] https://fasttext.cc/

[13] https://nlp.stanford.edu/data/glove.840B.300d.zip

considerable size of the training corpus and vocabulary in the three pre-trained models, they had less terminology coverage than the domain-specific models evaluated in this work. This pitfall has been already reported by Nooralahzadeh F. et al. (2018), and it is hardly surprising given that previous studies have observed that multi-word terms account for more than 90% of the terms of a specialized knowledge domain (Krieger MG. and Finatto MJB., 2004; Nakov P., 2013; Nguyen N.T.H. et al., 2017; Sager J.C. et al., 1980). As a consequence, since the pre-trained models did not contain most of the multi-word terms used in our specialized corpus and evaluation data (96% of the terms in the gold standard data are multi-word units), we calculated the missing multi-word term vectors by applying a compositional semantic model called Basic Additive Model (BAM) (Mitchell J. and Lapata M., 2008). BAM computes the vector of a multi-word term by adding its component single-word vectors. The compositional pre-trained models are henceforth referred to as pt-W2V-BAM, pt-FTX-BAM, and pt-GloVe-BAM.

The pre-trained BERT model (Devlin J. et al., 2019) was also evaluated. Context-free models such as W2V, FTX, and GloVe produce a single, fixed embedding representation for each word in a corpus. Instead, BERT is a contextual deep learning model which generates as many representations for a target word as the number of times it appears in a corpus, since each representation is based on the other words that accompany the target word in each sentence.

We employed the uncased version of the BERT-Base model in Python,[14] with 768-dimension vectors. This model has 12 encoder layers, 768 hidden units in the feed-forward networks, and 12 self-attention heads. The terms of our corpus were added to the vocabulary file of the model. Each of the contextualized embeddings for a term was obtained by adding up the vectors from the last four encoder layers, a procedure already applied by Devlin J. et al. (2019). Nevertheless, for the model evaluation, we used a single, averaged embedding for each term, which resulted from the average of all the different contextualized embeddings for the same term. As in the case of GloVe, W2V, and FTX, the number of epochs was fixed to 10, and the learning rate to 0.05. In addition, the parameter for the maximum sentence length was set to 64 because: (1) It is one of the values recommended by Devlin J. et al. (2019); and (2) the maximum sentence length of our corpus was 57 words.

Although there exists the pre-trained SciBERT model (Beltagy I. et al., 2019), based on BERT but trained on a large corpus of scientific texts, SciBERT was not used because the training corpus consisted of papers from the computer science and biomedicine domains, which are far from being related to the Coastal Engineering domain of our corpus.

In summary, we applied and evaluated eight different DSMs: BOW, W2V, FTX, GloVe, pre-trained BERT, and the three compositional pre-trained models.

4.4 Evaluation of the DSMs

First, for each bay included in the gold standard datasets, a sorted list of neighbours was obtained by computing a similarity/distance measure between the bay's vector and the vectors of all other context terms. Then, these context terms were sorted in descending order of magnitude. As such, for each bay, a list of ranked retrieval results was compiled.

Subsequently, the sorted lists of neighbours were evaluated on the whole gold standard dataset constructed for each of the three semantic relations. The measure used to evaluate the models was *Mean Average Precision* (MAP) (Manning C.D. et al., 2009: 158-162). Unlike the Precision, Recall, and F-score measures, which are computed using unordered sets of items, MAP is more appropriate for the evaluation of ranked retrieval results, such as ours. MAP provides a single-figure measure of quality across recall levels, and so it is roughly the average area under the precision-recall curve for a set of queries. Additionally, MAP has been shown to have especially good discrimination and stability (*ibidem*, p. 160). This measure tells us the overall accuracy level of the sorted lists of neighbours obtained for all bay queries, based on the rank of the related terms according to the gold standard. The nearer the related terms are to the top of the list for each bay, the higher the MAP.

The evaluation process delineated above was repeated for each of the five similarity/distance measures computed between a bay's vector and the vectors of all other context terms. The five measures evaluated in this study were Euclidean distance, cosine similarity, Jaccard coefficient, Pearson correlation coefficient, and averaged Kullback-Leibler divergence. For space constraints, we refer readers to Huang A. (2008) for a detailed description of the properties and formulas of these measures.

5 Results

The eight models were compared by observing the MAP of each model on the three datasets. Regarding the similarity/distance measures, it was found that, except for the Euclidean distance, which performed the worst, the other four measures had comparable effectiveness for all the DSMs and semantic relations, according to the results of the ANOVA tests, run to

[14] https://github.com/google-research/bert

determine the significance of the performance-wise differences amongst the similarity/distance measures. MAP scores were used as the basis of comparison.

This behavior is in line with previous research on similarity measure comparison by Huang A. (2008), and Strehl A. et al. (2000). For that reason, Table 4 shows the maximum MAP achieved by each model when applied cosine similarity, since this measure is widely used in NLP systems.

Dataset	BOW model		
	Maximum MAP	Weighting scheme	Window size
takes_place_in	**0.552** (0.395 ± 0.080)	LL	4
located_at	**0.410** (0.308 ± 0.054)	LL	2
attribute_of	**0.339** (0.197 ± 0.052)	LL	3
Dataset	**GloVe model**		
	Maximum MAP		**Window size**
takes_place_in	0.522 (0.395 ± 0.077)		4
located_at	0.381 (0.278 ± 0.050)		2
attribute_of	0.302 (0.190 ± 0.042)		3
Dataset	**FTX model**		
	Maximum MAP		**Window size**
takes_place_in	0.482 (0.284 ± 0.107)		4
located_at	0.339 (0.223 ± 0.061)		2
attribute_of	0.274 (0.136 ± 0.057)		3
Dataset	**W2V model**		
	Maximum MAP		**Window size**
takes_place_in	0.349 (0.312 ± 0.031)		4
located_at	0.209 (0.183 ± 0.014)		2
attribute_of	0.170 (0.111 ± 0.032)		3
Dataset	**Uncased BERT-Base model**		
	Maximum MAP (single value)		
takes_place_in	0.355		
located_at	0.213		
attribute_of	0.173		
Dataset	**pt-GloVe-BAM model**		
	Maximum MAP (single value)		**Fixed window size**
takes_place_in	0.264		10
located_at	0.151		10
attribute_of	0.109		10
Dataset	**pt-FTX-BAM model**		
	Maximum MAP (single value)		**Fixed window size**
takes_place_in	0.231		15
located_at	0.114		15
attribute_of	0.072		15
Dataset	**pt-W2V-BAM model**		
	Maximum MAP (single value)		**Fixed window size**
takes_place_in	0.199		15
located_at	0.089		15
attribute_of	0.046		15

Table 4: Maximum MAP of the models on each dataset when applied cosine similarity. Average and standard deviation are shown in brackets, LL stands for the log-likelihood weighting scheme.

The results indicated that the BOW model obtained the best performance in terms of MAP on the three semantic relations when its parameters were correctly tuned. They also showed that the *takes_place_in* relation was the most accurately captured by all models when they were tuned for this relation, followed by the *located_at* and *attribute_of* relations.

The greater accuracy of *takes_place_in* may be due to the large number of instances in specialized texts in Coastal Engineering which express the processes that occur in named bays. As for the *located_at* and *attribute_of* relations, these texts frequently mention the entities in named bays and the properties of these landforms. However, it seems that the number of instances of both semantic relations in the whole corpus is not large enough for the DSMs to represent them as accurately as *takes_place_in* instances.

Table 4 also shows that the maximum MAP of the BOW model was achieved when:

1. The statistical association measure for the three semantic relations was log-likelihood, transformed by adding 1 and calculating the natural logarithm.
2. The window size for the *takes_place_in* relation was 4 words.
3. The window size for the *attribute_of* relation was 3 words.
4. The window size for the *located_at* relation was 2 words.

Strikingly, the BERT and the three compositional pre-trained models performed the worst of all DSMs. Various factors are known to be associated with this behavior. Firstly, in NLP systems for specialized domains, the performance of domain-specific term vectors is higher than that of pre-trained embeddings, even when the size of the specialized corpus is considerably smaller (Nooralahzadeh F. et al., 2018). Secondly, domain-specific terms are inefficiently represented in pre-trained embeddings since there are few statistical clues in the underlying general-domain corpora for these words (Bollegala D. et al., 2015; Pilehvar M.T. and Collier N., 2016). Thirdly, BAM models tend to perform worse in comparison to their non-compositional counterparts that learn multi-word term vectors (Nguyen N.T.H. et al., 2017).

Interestingly, in each dataset, the maximum MAP of the BOW, GloVe, FTX, and W2V models was reached when the window size was the same. For that reason, to assess the impact of the window size on the accuracy of the DSMs, the average MAP for each setting of this parameter (i.e., for each window size between 1 and 10 words) is illustrated in Figure 2. The average MAP was used, instead of the maximum, because it allowed us to determine which window-size settings consistently produced satisfactory results, regardless of the settings used for the other parameters.

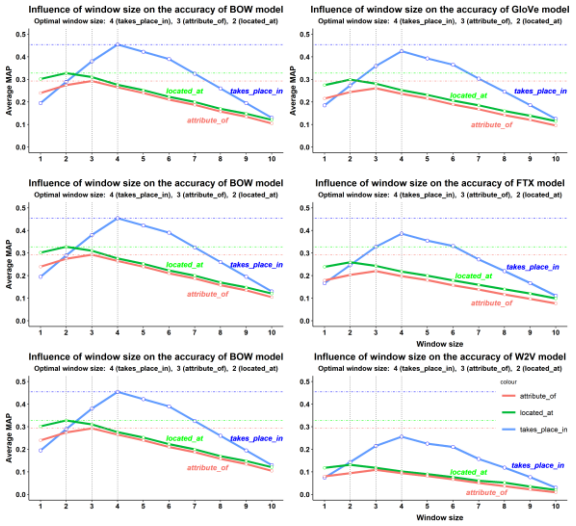

Figure 2: Average MAP of BOW (always left), GloVe (upper right), FTX (middle right), and W2V (bottom right) w.r.t. window size.

In Figure 2 we can observe that, in the four DSMs (BOW, GloVe, FTX, and W2V), the optimal window size was 4 words for the *takes_place_in* relation, 3 words for *attribute_of*, and 2 words for *located_at*. The compositional pre-trained models were not shown owing to their extremely suboptimal performance and their fixed window sizes.

Since the count-based model BOW notably outperformed predictive models on the three datasets, for the sake of simplicity, the setting influence of the other four hyperparameters of FTX and W2V are succinctly reported because they did not lead to substantial accuracy improvements on either dataset. As such, for both predictive models, settings can be summarized as follows: (1) The neural network architecture skip-gram worked, on average, better than CBOW; (2) a negative sampling of 10 samples reached a larger MAP than the hierarchical softmax; (3) the subsampling threshold was not conducive to significant gains; and (4) the optimal setting for the dimensionality of the term embeddings was 300 dimensions.

These optimal settings for the predictive models FTX and W2V were thus in line with previous research (Chiu B. et al., 2016). Moreover, FTX seemed to perform markedly better that W2V for the three semantic relations. This behavior may be linked to the fact that, as FTX exploits character-level similarities between terms, it is able to model low-frequency terms more effectively, thereby achieving better performance for small-sized corpora (Bojanowski P. et al., 2017: 140-141).

Regarding GloVe, with 300-dimension vectors, it was the only predictive model whose performance reached values similar to those of BOW. There is some evidence that the generalization ability of neural network-based models, such as FTX, W2V, and BERT, decreases when they learn on a limited amount of data (Collobert R. et al., 2011). Accordingly, since GloVe is not implemented with neural networks, the model did not seem unduly affected by the reduced corpus size.

In order to verify our observations on the behavior of the BOW, GloVe, FTX, and W2V models, statistical tests were run to determine the significance of the performance-wise differences amongst the models. MAP scores were used as the basis of comparison. As they did not deviate from the normal distribution according to Shapiro-Wilk test results (p-value>0.05), parametric statistical tests were thus carried out. For each semantic relation, we conducted the independent measures one-way ANOVA test, followed by post-hoc multiple pairwise comparisons between the models. For the multiple testing correction, we employed false discovery rate using the Benjamini-Hochberg procedure (Benjamini Y. and Hochberg Y., 1995).

The conclusions drawn from the statistical test results can be outlined as follows, and apply to the three semantic relations: (1) The performance of BOW was not significantly better than that of GloVe (p-value>0.05), whereas both models significantly outperformed the remaining DSMs (p-values<0.05); (2) there was no significant difference between the performance of the FTX and W2V models (p-value>0.05), despite the maximum MAP values for FTX were higher than those for W2V.

The evaluation indicated, in all models, MAP scores that could initially be regarded as quite low. These results are striking, given that the models were specially tuned to work in the specified scenario and with three semantic relations. To truly appreciate the value of this work and the difficulty of the task involved, we compare our results to those of two other studies that addressed similar scenarios, and which also compared count-based and prediction-based DSMs.

Bernier-Colborne G. and Drouin P. (2016) compared the ability of both types of DSM to capture relations from the web-crawled PANACEA Environment English monolingual corpus (Prokopidis P. et al., 2012),[15] with a size of over 50 million tokens. The authors reported maximum MAP figures ranging from 0.199 to 0.544. These values are surprisingly similar to those found in our study for the BOW, Glove, FTX, and W2V models (from 0.170 to 0.552, according to Table 4), although the size of our corpus is much smaller (7 million tokens).

[15] http://catalog.elra.info/en-us/repository/browse/ELRA-W0063/

On the other hand, Nguyen N.T.H. et al. (2017), among other objectives, aimed to extract, with both types of DSM, scientific and vernacular names synonymous to plant species from the English subset of the Biodiversity Heritage Library (BHL) (Gwinn N. and Rinaldo C., 2009),[16] an open-access repository containing millions of digitized pages of legacy literature on biodiversity. The enormous corpus size of the English subset of BHL amounts to around 49 gigabytes of data. Nonetheless, the authors reported moderate maximum MAP scores, ranging from 0.283 to 0.621. In contrast, Table 4 shows that the maximum MAP values obtained by the BOW model varied from 0.339 to 0.552. These are extremely promising measures, especially considering the tiny size of our corpus compared to that of BHL corpus.

Overall, the MAP values of our BOW model are striking because they are quite high despite the small size of the corpus.

Finally, the error analysis revealed that the terms in the gold standard datasets with the lower number of mentions in the corpus systematically occupied lower positions in the lists of ranked retrieval results compiled for each DSM. Thus, this fact negatively affected the MAP scores.

6 Conclusions

The representation in EcoLexicon of the conceptual structures (Faber P., 2012) that underlie the usage of colponyms in a small-sized, English Coastal Engineering corpus requires terminologists to manually extract from the corpus the terms which relate to each colponym through the semantic relations *takes_place_in*, *located_at*, and *attribute_of*, the three most frequent relations held by named bays in the corpus. Since this is a time-consuming task, the overall aim of this study was to provide terminologists with three lists of term candidates for a colponym, one list per semantic relation, by applying DSMs.

Accordingly, count-based and prediction-based DSMs, pre-trained models, and five similarity measures were applied to the corpus. Since the construction of DSMs is highly parameterized, and their evaluation in small specialized corpora has scarcely received attention, this study identified both parameter combinations in DSMs and similarity measures suitable for the extraction of terms which related to colponyms through the abovementioned semantic relations. The models were thus evaluated using three gold standard datasets.

Count-based models, with the log-likelihood association measure, showed the best performance for the three semantic relations. These results reinforce the findings of previous research that states, on the one hand, that count-based DSMs surpass prediction-based ones on small-sized corpora of under 10 million tokens (Asr F. et al., 2016; Sahlgren M. and Lenci A., 2016; Nematzadeh A. et al., 2017), and on the other hand, that log-likelihood achieves greater accuracy for medium- to low-frequency data than other association measures (Alrabia M. et al., 2014). In this respect, research on the application of DSMs in small specialized corpora, such as ours, is particularly scarce, compared to the plethora of work that analyzes DSMs in large general corpora. Hence, more studies of this type are needed so that further insights can be gained into the efficient representation of small specialized corpora in DSMs.

For both count-based and prediction-based DSMs, the optimal window size depended on the semantic relation that was to be captured, and the specific values coincided in both types of DSM, namely, a window size of 4 words for the *takes_place_in* relation, 3 words for *attribute_of*, and 2 words for *located_at*. The dependence of the window size on the specific semantic relation is in line with the findings by Bernier-Colborne G. and Drouin P. (2016).

It was also found that the *takes_place_in* relation was the most accurately represented by the DSMs, followed by *located_at* and *attribute_of*. This was possibly due to the insufficient number of instances of both semantic relations in the corpus for the DSMs to represent them as accurately as *takes_place_in* instances.

The pre-trained models *GloVe*, *word2vec*, f*astText*, and *BERT* performed the worst of all DSMs. In addition, they only provided a single list of term candidates for a colponym, which became less meaningful because it was not clear the relation of the listed terms to the colponym.

Regarding the similarity measures, it was found that, except for the Euclidean distance, which performed the worst, the other four measures had comparable effectiveness for all the DSMs and semantic relations. This behavior is in agreement with previous research on similarity measure comparison by Huang A. (2008), and Strehl A. et al. (2000).

Finally, an extension of this work will include testing the same DSMs and similarity/distance measures on gold standard datasets for named beaches.

Acknowledgements

This research was carried out as part of project PID2020-118369GB-I00, Transversal Integration of Culture in a Terminological Knowledge Base on Environment (TRANSCULTURE), funded by the Spanish Ministry of Science and Innovation.

[16] https://www.biodiversitylibrary.org/

References

Alrabia, M., N. Alhelewh, A. Al-Salman, and E. Atwell. 2014. An empirical study on the Holy Quran based on a large classical Arabic corpus. *International Journal of Computational Linguistics*, 5(1): 1-13.

Asr, F., J. Willits, and M. Jones. 2016. Comparing predictive and co-occurrence-based models of lexical semantics trained on child-directed speech. In A. Papafragou, D. Grodner, D. Mirman, and J. Trueswell (eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society (CogSci)*, Philadelphia (Pennsylvania), 1092-1097.

Baroni, M., G. Dinu, and G. Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 1, 238-247.

Baroni, M., and A. Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4): 673-721.

Beltagy, I., K. Lo, and A. Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, 3615-3620.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1): 289-300.

Benoit K., K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, and A. Matsuo. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30): 774.

Bernier-Colborne, G., and P. Drouin. 2016. Evaluation of distributional semantic models: a holistic approach. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm)*, Osaka (Japan), 52-61.

Bertels, A., and D. Speelman. 2014. Clustering for semantic purposes: Exploration of semantic similarity in a technical corpus. *Terminology*, 20(2): 279-303.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5: 135-146.

Bollegala, D., T. Maehara, Y. Yoshida, and K. Kawarabayashi. 2015. Learning word representations from relational graphs. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, Palo Alto, 2146-2152.

Chen, Z., Z. He, X. Liu, and J. Bian. 2018. Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases. *BMC Medical Informatics and DecisionMaking*, 18(Suppl 2): 65.

Chiu, B., G. Crichton, A. Korhonen, and S. Pyysalo. 2016. How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical NLP*, Berlin, 166-174.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug): 2493-2537.

Devlin, J., M.W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.04805v2*.

El Bazi, I., and N. Laachfoubi. 2016. Arabic named entity recognition using word representations. *International Journal of Computer Science and Information Security*, 14(8): 956-965.

Erk, K., S. Padó, and U. Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4): 723-763.

Evert, S. 2008. Corpora and collocations. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*. Berlin: Walter de Gruyter, 1212-1248.

Evert, S., P. Uhrig, S. Bartsch, and T. Proisl. 2017. E-VIEW-alation – A large-scale evaluation study of association measures for collocation identification. In *Proceedings of the eLex 2017 Conference*, Leiden, 531-549.

Faber, P. (ed.). 2012. *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/Boston: De Gruyter Mouton.

Faber, P., P. León-Araúz, and J.A. Prieto. 2009. Semantic relations, dynamicity, and terminological knowledge bases. *Current Issues in Language Studies*, 1: 1-23.

Ferret, O. 2015. Réordonnancer des thésaurus distributionnels en combinant différents critères. *TAL*, 56(2): 21-49.

Gries, S., and A. Stefanowitsch. 2010. Cluster analysis and the identification of collexeme classes. In S. Rice, and J. Newman (eds.), *Empirical and Experimental Methods in*

Cognitive/Functional Research. Stanford (California): CSLI, 73-90.

Gwinn N., and C. Rinaldo. 2009. The Biodiversity Heritage Library: Sharing biodiversity with the world. *IFLA Journal*, 35(1):25-34.

Huang, A. 2008. Similarity measures for text document clustering. In *Proceedings of the New Zealand Computer Science Research Student Conference 2008*, Christchurch, 49-56.

Ide, N., and J. Pustejovsky (eds.). 2017. *Handbook of Linguistic Annotation*. Dordrecht: Springer.

Kiela, D., and S. Clark. 2014. A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, Gothenburg, 21-30.

Kilgarriff, A., P. Rychlý, P. Smrz, and D. Tugwell. 2004. The Sketch Engine. In *Proceeding of the 11th EURALEX International Congress*, Lorient, 105-115.

Krieger, M.G., and M.J.B. Finatto. 2004. *Introdução à Terminologia: teoria & prática*. São Paulo: Contexto.

Lapesa, G., S. Evert, and S. Schulte im Walde. 2014. Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In *Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics*, Dublin, 160-170.

León-Araúz, P., A. San Martín, and A. Reimerink. 2018. The EcoLexicon English corpus as an open corpus in Sketch Engine. In *Proceedings of the 18th EURALEX International Congress*, Ljubljana, 893-901.

Levy, O., Y. Goldberg, and I. Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the ACL*, 3: 211-225.

Manning, C.D., P. Raghavan, and H. Schütze. 2009. *Introduction to Information Retrieval*. Cambridge (England): Cambridge University Press.

Manning, C.D., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, 55-60.

Mikolov, T., E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, Miyazaki, 52-55.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop Proceedings of International Conference on Learning Representations*. Scottsdale.

Miller, G.A., and W.G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1): 1-28.

Mitchell, J., and M. Lapata. 2008. Vector-based models of semantic composition. In *Proceeding of ACL-08*, Columbus (Ohio), 236-244.

Nakov, P. 2013. On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, 19: 291–330.

Nematzadeh, A., S.C. Meylan, and T.L. Griffiths. 2017. Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, London, 859-864.

Nguyen, N.T.H, A.J. Soto, G. Kontonatsios, R. Batista-Navarro, and S. Ananiadou. 2017. Constructing a biodiversity terminological inventory. *PLoS ONE*, 12(4): e0175277.

Nooralahzadeh, F., L. Øvrelid, and J.T. Lønning. 2018. Evaluation of domain-specific word embeddings using knowledge resources. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, Miyazaki, 1438-1445.

Pearson, J. 1998. *Terms in context*. Amsterdam: John Benjamins.

Pennington, J., R. Socher, and C.D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods for Natural Language Processing (EMNLP)*, Doha (Qatar), 1532-1543.

Prokopidis, P., V. Papavassiliou, A. Toral, M.P. Riera, F. Frontini, F. Rubino, and G. Thurmair. 2012. *Final report on the corpus acquisition & annotation subsystem and its components*. Technical Report WP-4.5, PANACEA Project.

Pilehvar, M.T., and N. Collier. 2016. Improved semantic representation for domain-specific entities. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, Berlin, 12-16.

Rohde, D., L. Gonnerman, and D. Plaut. 2006. An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*, 8: 627-633.

Rojas-Garcia J., and P. Faber. 2019a. Extraction of terms for the construction of semantic frames for named bays. *Argentinian Journal of Applied Linguistics*, 7(1): 27-57.

Rojas-Garcia J., and P. Faber. 2019b. Extraction of terms related to named rivers. *Languages*, 4(3): 46.

Rojas-Garcia J., and P. Faber. 2019c. Evaluation of distributional semantic models for the extraction of semantic relations for named rivers from a small specialized corpus. *Procesamiento del Lenguaje Natural*, 63: 51-58.

Room, A. 1996. *An Alphabetical Guide to the Language of Name Studies*. Lanham/London: The Scarecrow Press.

Sahlgren, M., and A. Lenci. 2016. The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin (Texas), 975-980.

Sager, J.C., D. Dungworth, and P.F. McDonald. 1980. *English Special Languages. Principles and Practice in Science and Technology*. Wiesbaden: Brandstetter Verlag.

Strehl, A., J. Ghosh, and R. Mooney. 2000. Impact of similarity measures on web-page clustering. In *AAAI-2000: Workshop on Artificial Intelligence for Web Search*, Austin, 58-64.