Embedding Based Multilingual Atlas of Semantic Fields

Jorge Alvarado [1]

[1] *Pontificia Universidad Javeriana, Departamento de Ingeniería Industrial, Carrera 7 40-62, Bogotá, Colombia.*

Abstract

Semantic fields(domains) are an important construct in neuroscience, linguistics, psychology, and natural language processing. However, semantic field resources typically lack scalability and are not based on language usage, but on scientific and commercial taxonomies. The present project aims to create maps of semantic fields for multiple languages constructed from Word Embeddings. The clustering process is systematically described, and preliminary results for the Spanish language are presented, showing similarities and differences compared to current classifications. The present work opens up possibilities for a usage-based word classification of semantic fields and for the generation of language atlases that allow for multilingual comparison and improve the development of the aforementioned disciplines.

Keywords

Semantic domains, Word Embeddings, Distributional Semantic Models

1. Introducción y estado del arte

Semantic Domains are categories or groups of concepts, which are reflected in clusters of words whose meanings are highly related [1, 2]. Such a relationship arises in connection with reality, resulting in the set of words referring to a specific subject or topic[3].

Semantic Domains are of high importance in the fields of linguistics, neuroscience, psychology, and natural language processing. In corpus linguistics, they serve as a crucial tool for comparative language studies, both within and across languages [4]. In neuroscience, lists of words tied to a semantic domain are used to detect the locations and processes of semantic cognition [5, 6]. In psychology, norm words are used for the study of various emotional[7] and cognitive [8] phenomena. In natural language processing, supervised topic modeling is connected to the idea of the existence of a set of topics, where topics are delimited subjects/themes through a set of words[9].

Traditional methods for generating semantic domains involve a group of experts who define the existing domains a priori and then, with their expertise in linguistics, neuroscience, psychology, and/or social sciences, classify a set of words into these domains. These solutions are costly and sometimes limited to a relatively small number of words.

Distributional Semantic Models (also known as Word Embeddings) are vector representations of words based on word co-occurrence in a corpus [10]. Under the assumption of the distributional semantic hypothesis, such co-occurrence can capture the semantics of specific words[11]. Word Embeddings offer a new opportunity to create semantic domains for a language based on its current usage. In fact, one of the tasks in which Word Embeddings are often evaluated is categorization [12], described as the ability to group words into semantically related clusters

SEPLN-PD 2023: Annual Conference of the Spanish Association for Natural Language Processing 2023: Projects and System Demonstrations, September 27-29,Jaén, Spain.
EMAIL: jorge.alvarado@javeriana.edu.co (Jorge A. Alvarado)
ORCID: 0000-0001-8331-2031 (Jorge A. Alvarado)

according to a gold standard. Particularly, predictive Word Embeddings like FastText [13] and Word2vec [14] how competitive performance among various types of Word Embeddings in the categorization task [10], yielding purity results for the English language ranging from 59% to 85% [15, 16]. They also have adequate computation speed for large corpora. Specialized lexicons have already been derived from Word Embeddings in fields such as finance[17].

Although the potential of clustering semantic representations (embeddings) to generate related word groups has been extensively discussed, no formal examples of semantic domain clustering using embeddings were found in the literature search.

According to the Ugly Duckling theorem, every classification involves bias[18]. In particular, the usual ways of creating semantic domains arise from four declared biases or perspectives. First, there are semantic domains that seek to classify knowledge, such as Wikipedia, the Dewey Decimal Classification, or WordNet domains [19]. In these cases, words are grouped encyclopedically by areas of knowledge. The second group of semantic domains is based on a classification that facilitates intercultural and multilingual studies, such as the Intercontinental Dictionary series (IDS) [20] or the classification of the Summer International Institute of Linguistics [21]. In this case, the aim is to classify words into areas of interest that are fundamentally common to human experience and, therefore, can be found in all cultures. Thirdly, there is the purely linguistic classification, centered on the English language, of WordNet[22], which seeks to define potential linguistic relationships between words. Finally, some semantic domain classifications (expressed as themes or topics) are related to informational interests in the Internet world. Consequently, they are oriented towards a classification that allows commercial advertising around those informational interests, such as Twitter topics or Google topics.

All the efforts mentioned in the previous paragraph require teams of experts and fieldwork, making the development of semantic domains a slow task, particularly for languages beyond English; they also lack the perspective that embeddings can provide. This perspective is characterized by the communicational and public use of the language in books, news, and the internet—texts that are the usual corpus of embeddings. In summary, semantic domains that arise from the clustering of embeddings can make the creation of such domains for multiple languages an automatable task, at least in part, and would reflect a perspective based on the contextual use of the language, rather than scientific or commercial classifications. It would also enable complete language mapping to observe the distances and relationships between different semantic domains.

2. The project

The Project aims to create a multilingual atlas, based on the generation of semantic domains for those languages arising from the clustering of embeddings for each respective language, with the following stages:

1. Embedding clustering model.
2. Model evaluation against existing semantic domains.
3. Extension of the clustering model and evaluation to multiple languages.
4. Comparison of the results across multiple languages.

2.1. Embeddings clustering model

The objective of this nearly completed stage was to find the best combination of clustering methods, parameters, and hyperparameters to develop semantic domains. This clustering is necessarily hierarchical, as demonstrated by various characterizations of semantic domains so far: both Wikipedia, IDS, and WordNet have semantically organized their domains due to the hierarchical nature of the language, based on hypernymy and hyponymy relationships between concepts. This stage used the Spanish language as a prototype, particularly focusing on the most frequent words with representation in the embeddings of at least 200 occurrences.

Initially, the goal was to find an appropriate number of clusters for the lowest and highest levels of the hierarchy using k-means to explore the search space. Subsequently, the self-organizing maps (SOM) and k-means algorithms were compared for generating the high and low levels of the hierarchy, and it was determined that creating an intermediate level in the hierarchy made sense. Various linkage methods within an agglomerative hierarchical cluster were tested for creating the intermediate and high levels.

Finally, the t-distributed stochastic neighbor embedding (t-SNE) technique was applied to the

lowest level of the hierarchy to reduce the dimensionality, given its ability to provide an overview and perform reductions to two dimensions for data visualization (language mapping). After the reduction with t-SNE, a DBSCAN algorithm was applied to create the higher groups in the hierarchy, yielding excellent results. A Voronoi diagram was applied to the generated clusters to produce the visual representation of the language. The most frequent part-of-speech (POS) for each word group was extracted for a final classification with the aim of visualization in maps.

2.2. Model evaluation

For the model evaluation, a comparison is being conducted with the types of semantic domains mentioned in the state of the art (particularly with IDS, WordNet, and Google Topics) and with the BLESS database[23]. This is more of a comparison than a direct evaluation since none of them serves as a gold standard, as they arise from different perspectives for classification. The comparison with BLESS allows for identifying the type of semantic relationship best identified by the clustering. The clusters were also characterized by their part-of-speech (POS), and a labeling process is underway, comparing the existing semantic domains in IDS, WordNet, and Google Topics with the generated semantic domains to assign a final name to the semantic domains and create the final language map.

2.3. Extension to multiple languages

Once the model is developed, the plan is to extend the work to multiple languages, particularly choosing languages from diverse linguistic families. The work done so far has shown that it is necessary to have speakers of the language (preferably linguists) who can support the labeling work presented in section 2.2, and they are not yet available for all languages. The project aims to expand, if possible, the work to the following languages whose embeddings are available: English, French, Malay, Japanese, Arabic, Turkish, and Yoruba, thereby attempting to cover a significant group of linguistic families.

2.4. Comparison across languages

After the creation of semantic domains and maps for each of the languages (the proposed atlas in the title), both the level of alignment between the clustering of embeddings and other sources of semantic domains (IDS, WordNet) and the similarity of the generated groups will be compared. An interesting comparison will arise from the relative size of the named geographical entities within each language as an approximation of the proximity of linguistic cultures. However, the metrics and comparison mechanisms are yet to be defined.

3. Preliminary results

The initial results for the Spanish language show that the k-means algorithm and the hierarchical clustering method with Ward linkage were superior at all hierarchical levels. The evaluation metric used was the adjusted random index (ARI). The number of clusters finally selected for the three managed hierarchical levels was 1024 for the low hierarchy, 481 for the middle hierarchy, and 64 for the high hierarchy. The assignment of each word to the three hierarchical clusters, along with a descriptive labeling based on the most frequent words in each cluster, can be found in Zenodo [24]. The level of agreement compared to the IDS classification showed agreement levels between 15% and 65%, with higher agreement in categories such as *Animals* and *Kinship*.

Figure 1 shows a preliminary map of the Spanish language generated through the entire process. In the results, the classification into large groups initially relied on part-of-speech as a fundamental element, but it was also necessary to use language, named entities, and the distinction between the natural world and the social world as sources for the "continents" of the graph.

In Figure 1, certain similarities and differences with the scientific and commercial classifications are notable, and I will highlight some related to IDS:

1. Some IDS semantic domains are not directly visible in the new semantic domains as they dissolve into multiple places. For example, *The body* does not appear to be a distinct semantic domain based on language usage but rather divides into several others like health, crime, and nearby sensory objects. Similar

occurrences are observed with domains like *clothing* and *social relations*.

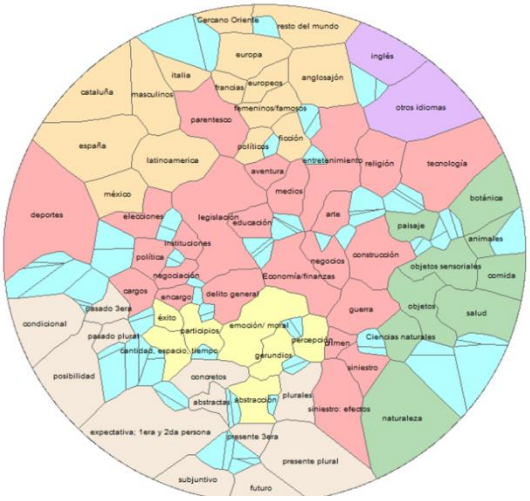

Figure 1: Preliminary Map of Spanish. Colors indicate the major word groups. Labels and regions mark clustering locations with DBSCAN of the two-dimensional location of the 1024 clusters obtained with t-SNE. Orange = named entities; Violet = other languages; Red = social world; Green = natural world; Light brown = verbs; Yellow = other parts of speech; Blue = isolated categories.

2. On the other hand, some IDS domains appear very clearly in the new usage-based semantic domains, such as *food* and *warfare*.
3. Domains not specified in other classifications emerge as relevant from usage, such as success, natural sciences, and crime or disaster in general.
4. As part of the language, the semantic domains clearly include named entities, and in the geographical context, they seem to reflect the level of relationship that the language has with specific geographies.

While it is known that embeddings generate separations by parts of speech, it is important to question whether it is usage that generates these domains. Among verbs, the distinction between concrete and abstract is not a traditional grammatical distinction, yet it appears when creating the semantic domains.

4. Conclusion

The preliminary results of the project show that building semantic domains through the clustering of embeddings is a promising path for a new understanding of language and for supporting the creation of word lists for semantic domains. It is expected that the multilingual extension of the project will broaden the scope to different cultures and reveal linguistic and cultural similarities and differences.

5. Funding and research groups

This work is funded by Pontificia Universidad Javeriana, and the ZENTECH (Improvement and Technology) research group from the same university is participating

6. References

[1] B. Nerlich and D. D. Clarke, "Semantic fields and frames: Historical explorations of the interface between language, action, and cognition," Journal of Pragmatics, 32. 2 (2000): 125-150.

[2] T. T. Hills, P. M. Todd, and M. N. Jones, "Foraging in Semantic Fields: How We Search Through Memory," Topics in Cognitive Science, 7. 3 (2015): 513-534.

[3] L. Brinton and D. M. Brinton. Workbook: The linguistic structure of modern english. John Benjamins Publishing Company. https://benjamins.com/sites/z.156/exercise/c6q4.

[4] B. Thompson, S. G. Roberts, and G. Lupyan, "Cultural influences on word meanings revealed through large-scale semantic alignment," Nature Human Behaviour, 4.10 (2020): 1029-1038.

[5] L. Chen, M. A. Lambon Ralph, and T. T. Rogers, "A unified model of human semantic knowledge and its disorders," Nature Human Behaviour, 1.3 (2017).

[6] O. Azad, "The Analysis of Semantic Field in Persian-Speaking Patients With Wernicke's Aphasia," Iranian-Rehabilitation-Journal, 18.3(2020): 257-262. doi: 10.32598/irj.18.3.378.3.

[7] P. N. Johnson-laird and K. Oatley, "The language of emotions: An analysis of a semantic field," Cognition and Emotion 3.2 (1989): 81-123.

8] N. Segalowitz and R. G. de Almeida, "Conceptual Representation of Verbs in Bilinguals: Semantic Field Effects and a Second-Language Performance Paradox," Brain and Language 81.1 (2002): 517-531.
[9] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit," Information Processing & Management 57.2 (2020).
[10] A. Lenci, M. Sahlgren, P. Jeuniaux, A. Cuba Gyllensten, and M. Miliani, "A comparative evaluation and analysis of three generations of Distributional Semantic Models," Language Resources and Evaluation (2022): 1-45.
[11] A. Lenci, "Distributional semantics in linguistic and cognitive research," Italian Journal of Linguistics 20.1 (2008) 1-31.
[12] A. Bakarov, "A survey of word embeddings evaluation methods," arXiv preprint arXiv:1801.09536, 2018.
[13] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, Enriching word vectors with subword information, in: Transactions of the association for computational linguistics, vol. 5, 2017, pp. 135-146.
[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representations in vector space, in :1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings, 2013. [Online].
[15] T. Schnabel, I. Labutov, D. Mimno, and T. Joachims, Evaluation methods for unsupervised word embeddings,in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 298-307.
[16] B. Wang, A. Wang, F. Chen, Y. Wang, and C. C. J. Kuo, Evaluating word embedding models: methods and experimental results, in: APSIPA transactions on signal and information processing, vol. 8, 2019.
[17] S. R. Das, M. Donini, M. B. Zafar, J. He, and K. Kenthapadi, "FinLex: An effective use of word embeddings for financial lexicon generation", The Journal of Finance and Data Science, 8 (2022) : 1-11.
[18] S. Watanabe, "Knowing and Guessing a Quantitative Study of Inference and Information," 1ST edition, Wiley, New York, NY, 1969.
[19] L. Bentivogli, P. Forner, B. Magnini, and E. Pianta, Revising the wordnet domains hierarchy: semantics, coverage and balancing, in: Proceedings of the Workshop on Multilingual Linguistic Resources, 2004, pp. 94-101.
[20] M. R. Key and B. Comrie (eds). The Intercontinental Dictionary Series, 2023. URL: https://ids.clld.org
[21] R. Moe, "Compiling dictionaries using semantic domains," Lexikos 13(2003) 215-223.
[22] G. A. Miller, WordNet: a lexical database for English in: Communications of the ACM, vol. 38, no. 11, 1995, pp. 39-41.
[23] M. Baroni and A. Lenci, How we BLESSed distributional semantic evaluation, in: Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, 2011, pp. 1-10.
[24] J. Alvarado. Spanish Semantic Fields, doi: https://doi.org/10.5281/zenodo.7620794.