ELEXIS 2019 Transnational Research Visit Report
Encoding definitions with word embeddings for sense categorization and diachronic linguistic studies

Luis Espinosa Anke
School of Computer Science and Informatics
Cardiff University, UK

1 Motivation

This document summarizes the work carried out by Luis Espinosa-Anke during his 2-week stay at the Royal Spanish Academy in the context of an ELEXIS transnational research travel grant. His work was focused on learning concept representations with distributional (i.e., corpus-based) and lexicographic (i.e., from a concept's definition) information. These concept representations, regardless of the approach, can be used in a number of downstream applications, from contributing to the dictionary writing and querying process, to improving the performance of NLP systems requiring semantic understanding, and even for enabling corpus-driven sociolinguistic analyses. At the end of the stay, three sub projects emerged, which will be continued during the following months, aiming at disseminating the most relevant findings in NLP and/or lexicography venues.

2 Tasks

2.1 Sense categorization

Semantic clustering can improve, on one hand, information access from a resource end user perspective [3]. Moreover, the acquisition of domain-specific training data for developing downstream language technologies has yielded promising results in NLP tasks such as word sense disambiguation [11, 1], text categorization [12] or hypernymy modeling [5]. In fact, in the context of ELEXIS, this idea has also been pursued, although from a slightly different angle, as we will see[1].

Given that there is literature in the NLP domain arguing for leveraging both corpus-based statistics and dictionary definitions to improve concept representations [9, 13, 2], we propose to explore the clustering of concepts in the Dictionary of the Spanish Language (*Diccionario de la Lengua Española* or *DLE*) based on semantic criteria, i.e., by determining the extent to which a definition can be mapped to a vector representation with and without the help of recurrent neural networks. We experiment with different approaches, which we discuss as follows.

Centroid-based clustering: As a simple baseline, we simply encode a concept by computing a weighted sum (by frequency) of the bag-of-words representation of its associated dictionary definition. This has the effect of grouping together, first, terms whose definitions share the *exact same words*, and after that, those where the different expressions are semantically similar to each other. In Table 1 we can see that while this approach seems desirable when aiming at grouping synonym terms with highly similar phraseology, it may also group terms which do not share any outstanding semantic feature, but happen to be defined using the same formulaic expressions.

Autoencoding average-based definitions: After observing the perhaps undesired behaviour of the semantic clusters emerging from the centroid-based strategy, and inspired by [6], we apply a conditional autoencoder to the average-based definition embeddings. Intuitively, we are interested in training a neural network capable to accurately reconstruct the original averaged vector, but by only using the information provided in content words (and not functional or stopwords, which we argue are more linked

[1] https://elex.is/wp-content/uploads/2019/08/ELEXIS_D3_1_Lexical_semantic_analytics_for_NLP_sense_clustering_Final.pdf

CENTROID-BASED	
cuerda	Extensión o número de notas que alcanza la voz
tenor	Hombre que tiene voz de tenor
atiplar	Dicho de la cuerda de un instrumento, o de la voz: Volverse del tono grave al agudo
AUTOENCODED CENTROID-BASED	
cuerda	Extensión o número de notas que alcanza la voz
subir	Dicho de la voz o del sonido de un instrumento: Pasar a un tono más agudo
sincopado/da	Dicho del ritmo o del canto: Que tiene notas sincopadas
LSTM-BASED	
atenorado/da	Dicho de una voz: Parecida a la del tenor
atenorado/da	Dicho de un instrumento: Que tiene un sonido de timbre semejante al de la voz del tenor
atiplar	Dicho de la cuerda de un instrumento, o de la voz: Volverse del tono grave al agudo

Table 1: Most similar terms (left column) and definitions (right column) to the target term "do de pecho" (*C major*), defined as "Una de las notas más agudas a que alcanza la voz de tenor" (*one of the highest tones that a tenor's voice reaches*).

to formulaic and stylistic patterns). For this reason, our autoencoder has, at decoding time, access to *the average vector of the stopwords of the definition*. In a way, this enforces the network to not consider any formulaic information when reconstructing the original vector from a compressed representation, as this information is explicitly provided. We can see in the comparison table that slightly different clusters emerge for the same target concepts (and definitions), with the interesting byproduct that these autoencoded representations can be tuned. In [6], it was found, in a different but related distributional semantic task, that going as low as 10-dimensional vector provided a fine balance between abstraction and interpretability. The dimensionality of the definition vectors provided in Table 1 is 25 (empirically set based on qualitative analysis).

LSTM-based clustering: Recent literature in computational lexicography and computational semantics has argued for taking advantage of the mapping power of recurrent neural networks to produce a vector representation of a definition such that it resembles as much as possible the vector representation of the definiens (i.e., the term being defined). This is useful because one single neural network architecture can be trained over a full dictionary resource, and from such model the hidden representation of the last LSTM [10] state can be thought of encoding the semantics of the definition, as well as the mapping history from previous ⟨term, definition⟩ pairs. As we can see in the selected examples in Table 1, this approach is better at downweighting formulaic (and semantically void) definitional expressions, while at the same time retaining the most salient semantic features of each definition.

2.2 Thematic category classification

The DLE provides a manually constructed thematic classification, which can be thought of a grouping of dictionary entries by their belonging to domains of knowledge[2]. The Spanish Academy, on the other hand, currently holds a vast number of unannotated resources, ranging from diachronic to to domain-specific (e.g., legal) dictionaries. As a proof of concept, we trained a CBLSTM classifier (convolutional layer followed by a bidirectional LSTM), which has proven effective in previous work for definition modeling [7]. We trained it on the coarsest-grained categorization of thematic labels on the DLE (which contains 8 classes), and generated a thematically-tagged Wiktionary-ES version. Because we have used a set of cross-lingual word embeddings [4], such thematic categorization could easily be ported to other languages. This constitutes an interesting case of transfer learning, where English can be considered the target language, and Spanish, due to the availability of the Academy's resources, the (resource-rich)

[2]See a user guide at `https://enclave.rae.es/pdfs/Guia_de_uso\%20Enclave_RAE.pdf`.

source language. To get an idea of the difficulty of the task, we show below in Figure 1 a PCA projection of the centroid-based representation of the DLE definitions and their different thematic categories. We can see, for example, that definitions categorized as 'el mundo' (*the world*) are easier to classify from the rest, whereas 'vida humana' (*human life*) and 'ciencias humanas' (*human science*) are almost undistinguishable from each other.

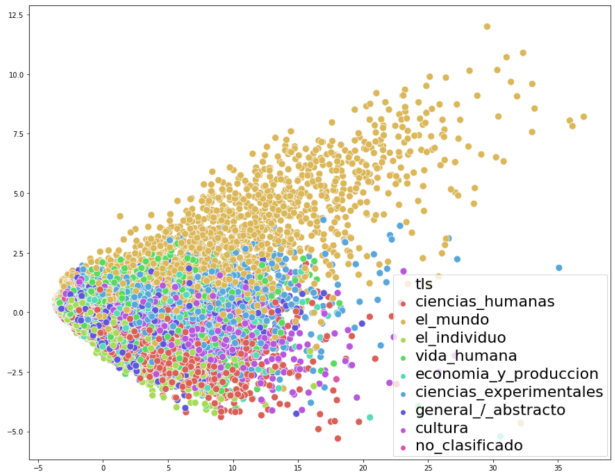

Figure 1: PCA visualization of DLE definitions and their grouping according to their thematic categorization.

2.3 Semantic Change

An interesting application of vector representations of words is that they can be used to assess the extent to which a word's meaning changes (or drifts), as this change is correlated with different contexts in which the word is used, and therefore, different co-occurrence statistics will emerge, yielding a different position in the vector space. More importantly, the nearest neighbours of the word vector at different timeframes will be different, as these are a reflection of its meaning. This intuition was explored using corpora of different centuries in [8]. During our ELEXIS project, and due to the availability of diachronic corpora from the 1970s of the largest newspaper in the Spanish language (El País[3]), we started working on a piece of software for exploring the semantic drift of target words in the Spanish language, and decided empirically to do this over a 5-year window from the earliest to the latest available documents (effectively, from 1977 to 2011). An illustrative example of some words which have clearly changed meaning in these three decades are shown in Table 2. We see that for 'rescate', the criminal sense (a rescue of hostages) was more predominant decades ago, while in its modern form is more associated with financial rescuing or bailout. Another interesting example is the word 'entertainment' (*entretenimiento*), which in the past was more associated to TV, and now it is more linked to digital entertainment (with similar words like 'interactivo' (*interactive*) and 'online'.

2.4 Conclusions and future work

This research stay has had three major impacts:

- **Strong collaboration**. A strong line of collaboration between the Academy and Cardiff University has been established, and we plan to turn it into one or more submissions to NLP or Lexicography forums. In fact, the Academy is currently considering bringing into their regular pipeline (both for lexicographers and for user queries) access to semantic clusters emerging from word embeddings and neural network-based encodings.

- **Insights on the Spanish language**. We have learned a lot about how the Spanish language works,

[3] www.elpais.es.

RESCATE		CARTERA		ENTRETENIMIENTO	
77-81	*07-11*	*77-81*	*07-11*	*77-81*	*07-11*
salvamento	salvamento	pedidos	carteras	instrumento	interactivo
operación	Grecia	carteras	diversificación	programas	online
secuestradores	emergencia	Industrias	inversión	vídeo	televisivo
pagado	pecuniario	Agricultura	activos	ocio	diversión
reparación	rescates	Finanzas	pedidos	visual	ocio

Table 2: Nearest neighbours for three target words: 'rescate' (*rescue* or *bailout*), 'cartera' (*wallet* or *portfolio*) and 'entretenimiento' (*entertainment*).

both in terms of how its words are defined in its reference dictionary, but also how they are classified and how their meaning has changed over time.

- **Researcher growth**. Personally, I am very satisfied by the opportunities for development that this grant has provided. I am confident that I am now a better researcher and that I know more about language and lexicography than I did before the stay, and I am looking forward to a future of collaborations.

For the future, in addition to turning these initiated projects into tangible submissions, we plan to extend our work to the cross-lingual setting, and to focus on objective and measurable success criteria for the clustering algorithms. Finally, we would like to incorporate a dimension of *relational similarity*, which due to lack of time and unclear straightforward application, was not tackled during the stay.

References

[1] Eneko Agirre, Oier Lopez De Lacalle, and Aitor Soroa. Knowledge-based wsd on specific domains: performing better than generic supervised wsd. In *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[2] Tom Bosc and Pascal Vincent. Auto-encoding dictionary definitions into consistent word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1532, 2018.

[3] Jose Camacho-Collados and Roberto Navigli. Babeldomains: Large-scale domain labeling of lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 223–228, 2017.

[4] Yerai Doval, Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert. Improving cross-lingual word embeddings by meeting in the middle. *arXiv preprint arXiv:1808.08780*, 2018.

[5] Luis Espinosa-Anke, Horacio Saggion, Francesco Ronzano, and Roberto Navigli. Extasem! extending, taxonomizing and semantifying domain terminologies. In *AAAI*, 2016.

[6] Luis Espinosa-Anke and Steven Schockaert. Seven: Augmenting word embeddings with unsupervised relation vectors. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2653–2665, 2018.

[7] Luis Espinosa-Anke and Steven Schockaert. Syntactically aware neural architectures for definition extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 378–385, 2018.

[8] William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*, 2016.

[9] Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30, 2016.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[11] Bernardo Magnini and Gabriela Cavaglia. Integrating subject field codes into wordnet. In *LREC*, pages 1413–1418, 2000.

[12] Roberto Navigli, Stefano Faralli, Aitor Soroa, Oier de Lacalle, and Eneko Agirre. Two birds with one stone: learning semantic models for text categorization and word sense disambiguation. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2317–2320. ACM, 2011.

[13] Julien Tissier, Christophe Gravier, and Amaury Habrard. Dict2vec: Learning word embeddings using lexical dictionaries. 2017.