# MMG at SemEval-2022 Task 1: A Reverse Dictionary approach based on a review of the dataset from a lexicographic perspective

**Alfonso Ardoiz**[1,2]
alfonso.ardoiz@dezzai.com
aardoiz@ucm.es

**Óscar García-Sierra**[1,2]
oscar.garcia@dezzai.com
oscarg02@ucm.es

**Ignacio Arranz**[1]
ignacio.arranz@dezzai.com

**Miguel Ortega-Martín**[1,2]
m.ortega@dezzai.com
m.ortega@ucm.es

**Jorge Álvarez**[1]
jorge.alvarez@dezzai.com

**Adrián Alonso**[1,3,4]
a.alonso@dezzai.com
adrian.barriuso@urjc.es

**1** dezzai by MMG **2** Universidad Complutense de Madrid
**3** Universidad Rey Juan Carlos **4** Data Science Laboratory - Universidad Rey Juan Carlos

## Abstract

This paper presents a novel and linguistic-driven system for the Spanish Reverse Dictionary task of SemEval-2022 Task 1. The aim of this task is the automatic generation of a word using its gloss. The conclusion is that this task results could improve if the quality of the dataset did as well by incorporating high-quality lexicographic data. Therefore, in this paper we analyze the main gaps in the proposed dataset and describe how these limitations could be tackled.

## 1 Introduction

The CODWOE (Comparison of Word Glosses and Word Embeddings) task at SemEval-2022 (Mickus et al., 2022) encouraged participants to analyze the relation between two types of semantic descriptions, word embeddings and dictionary glosses, by proposing two sub-tasks: Reverse Dictionary (RD) (Hill et al., 2016), in which participants must generate vectors from glosses, and Definition Modeling (DM) (Noraset et al., 2017), in which participants must generate glosses from vectors. These subtasks aim to be useful for explainable Artificial Intelligence (AI) by including human-readable and machine-readable data.

Given the didactic nature of these tasks, the output generated by these models should be as accurate as the most prestigious dictionaries. Hence, the process of selecting a quality dataset is a critical phase, as Garg et al. (2020) state: “a small number of data examples prevents an effective convergence to the task, while noisy data leads to incorrect convergence”. In this case, tasks require that the glosses used in the training represent the exact meaning of the word being defined in the context that the embeddings were extracted. However, as per our understanding, coherence, rigour and lexicographical prestige of the provided dataset should be improved; although accessing a prestigious dictionary is not an easy task.

A Reverse Dictionary takes a description in natural language and generates a list of words satisfying it (Siddique and Sufyan Beg, 2018). First RD were Information Retrieval systems for Turkish (El-Kahlout and Oflazer, 2004) and Japanese (Bilac et al., 2004). Other approaches used lexical graphs (Thorat and Choudhari, 2016; Ortega-Martín, 2021) that capture the relationships between the words of the definition itself and between these and others similar to them at different levels (synonymy, hyperonymy, etc.). Other systems create a vector space from these lexical resources, such as Wordnet (Dutoit and Nugues, 2002; Calvo et al., 2016; Méndez

*Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 68 - 74
July 14-15, 2022 ©2022 Association for Computational Linguistics

| ID | Definition |
|---|---|
| es.train.212 | “Biología.— Se dice de los microorganismos que no aceptan los colorantes habituales.” |
| es.train.250 | “Zoología.— Cualquiera de los colibríes del género Chlorostilbon .” |
| es.train.119 | “Servirse , darse ayuda mutuamente .” |
| es.train.120 | “Trabajar ( uso pronominal de ... )” |

**Table 1:** Examples of glosses

et al., 2013). As in many other NLP fields, more recently have appeared approaches based on Neural Networks (NNs) such as Long Short-Term Memory (LSTMs) (Malekzadeh et al., 2021; Zhang et al., 2020b) or Transformers (Qi et al., 2020; Yan et al., 2020). Language Models (LMs) based on Recurrent Neural Networks (RNNs) (Hill et al., 2016) have also been used. Finally, from a linguistic point of view, Shaw et al. (2011) add syntactic knowledge, Zock and Schwab (2008) try to replicate the model of the mental dictionary and Zhang et al. (2020b) use morphological knowledge in their system.

Definition Modeling is a relatively new task based on using distributed word representations to generate its definition. Noraset et al. (2017) use an RNN to compute the probability of a word being part of the definition. Different approaches have been proposed to this Natural Language Generation (NLG) task. Usually these new methods are heavily focused on the importance of the context of the word being defined, like fine-tuning a BART model to define groups of words (Bevilacqua et al., 2020), using attention and a Skip-gram model to smooth the problems of word selection in the generation step (Gadetsky et al., 2018), or exploring new ways to understand the embeddings and their capabilities, resulting these in a new task named “usage modeling” (Zhang et al., 2020a). There have been few attempts to use pure linguistics traits to improve definition generation, like accounting polysemy as a generative target using multi-sense word embeddings (Kabiri and Cook, 2020) or using sememes to condense the semantic core of generated sentences (Yang et al., 2020).

This paper has the following structure. Chapter 2 contains a review of the data along with some linguistic knowledge we consider relevant. In chapter 3 the RD approach and results are presented. Finally, chapter 4 contains the conclusions and future work. Our contributions are the following:

- We point out the main problem of this task, the lack of high-quality lexicographic data.
- We present the third best model for the Spanish “sgns” embeddings Reverse Dictionary task, which due to the use of external resources is not valid for the challenge.
- We compare various approaches for the previous task, analysing different preprocessing strategies, model architectures, loss functions and embedding initialization tactics.

# 2 Data analysis

This section contains a review of the Spanish dataset structure, an introduction about relevant lexicography concepts and the RD task preprocessing techniques.

## 2.1 The data

The dataset can be used for both subtasks. It is stored in a JSON file where each element contains four or five keys: its ”ID”, its ”gloss” or definition, the character-based embeddings (”char”), the Word2Vec Skip-gram Negative Sampling embeddings (”sgns”) and, just for some languages, the ”ELECTRA” (Clark et al., 2020) embeddings. All of these embeddings have a dimensionality of 256. For the development of the Spanish RD model “sgns” embeddings were used, since it was considered that using a more static approach such as “char” would lower the performance of the model and ”ELECTRA” embeddings were not available. However, as it will be explained later, the model was found out to be scalable to other embedding types and languages. Table 1[1] contains some words from the Spanish dataset that will be useful in the subsequent analysis.

## 2.2 Linguistic analysis

Even though this is not a linguistic paper, there are lexicographic concepts that should be explained in order to reach a deeper understanding of the dataset flaws. One of the most common approaches to

[1] Appendix A contains the translation of these examples

classify lexical dictionaries distinguishes between general dictionaries (also known as usage dictionaries) and specific dictionaries (this classification includes, to name a few, encyclopedic, synonym and scientific dictionaries). Therefore, given the significant number of different possible uses of the CODWOE tasks, the dataset should only include generic term definitions found in usage dictionaries, not specific ones.

In second place, there is no consensual standard structure for definitions. However, two principles must be followed (del Teso Martín, 1987).

1. The definitions must specify the hyperonym of the word being defined.
2. The definition should explain the main distinction between the class and its instance.

In other words, the definition has to include a hyperonym which clusters the word being defined into a category (for instance, "person who...", "a block of rock that...") and the definition should specify the specific traits of its meaning (following the last examples, "...plays badminton", "...shines even at night").

Lastly, a given definition is not the only way that a word could be defined, but just a context related meaning. This is why, from our perception, static embeddings should be avoided in modern Natural Language Processsing (NLP) tasks. For that reason, "ELECTRA" embeddings, used in other languages but not available for Spanish, could be more representative than "sgns" or "char" embeddings. Furthermore, these contextual embeddings should have been extracted from solid examples of use which represent the exact meaning of the gloss.

### 2.3 Dataset review

Dataset review revealed that glosses did not only go against the previously explained notions, but also lack coherence and exactitude. As seen in table 1, many definitions include the category which they belong (for instance, "Zoology"), or some grammatical information ("pronominal use"). Although dictionaries usually include this kind of information, it should not be in the definition (García and José, 2017).

Another drawback is that the dataset combines generic and specific definitions. Generic definitions usually can be found in usage dictionaries like "DLE" for Spanish or "Oxford Dictionary" for English, meanwhile specific definitions include terms from a certain domain, like zoology or linguistics. In our opinion, using specific definitions in this phase of the task just add noise to the training and evaluation.

It should also be noted that the terms glosses have not global coherence, and most of them do not follow the hyperonym and main distinction principles. There are plenty of synonym definitions (not optimal for these tasks as the definition length is too low) and encyclopedic definitions (which add a lot of noise as they have to fully describe the word being defined). Data could be improved if basic lexicographic notions were applied, but it is understood that being a multilingual dataset and given the available resources, CODWOE team has done a great work.

### 2.4 Data preprocessing

Data was preprocessed by deleting stopwords and category words (a term at the beginning of the definition that indicates its semantic category). In the second RD approach, the lexical graph, which is explained in section 3, words from the definitions were also lemmatized using the Spacy Spanish models lemmatizer [2]. Evaluation showed that the definitions preprocessing has been the most useful factor in the RD task, which indicates that the quality of the original definitions is what has penalized the model the most.

## 3 Our approach

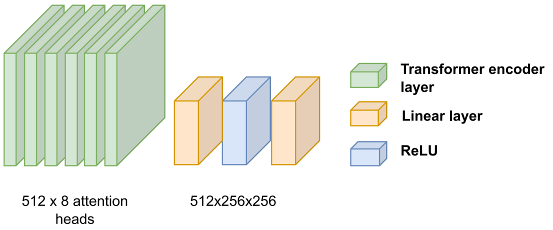

**Figure 1:** Model architecture

For the RD task, model was trained trying to make the definition embeddings as similar as possible as the defined word ones, focusing just on Spanish "sgns" (Word2Vec Skip-gram Negative Sampling embeddings of 256 dims) embeddings, al-

[2] `https://spacy.io/api/lemmatizer`

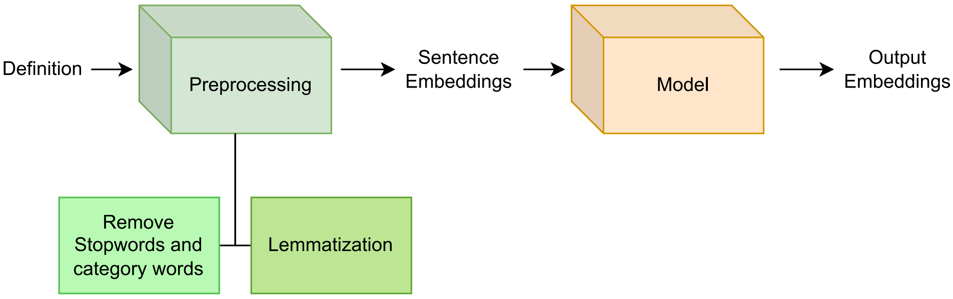

**Figure 2:** Full Reverse Dictionary pipeline

though system architecture was found scalable to other languages and embedding types. After test phase the model was applied also to Spanish "char" and to English "sgns" embeddings, achieving proportionally similar results which will be mentioned later.

Different tactics to initialize definition embeddings were also used. In our first approach we use Sentence-Transformers (Reimers and Gurevych, 2019). The "distiluse-base-multilingual-cased-v2"[3] was found to be the most appropriate model for Spanish tasks. Secondly, a lexical graph was built with the training dataset, in such a way that each defined word is related to the words in its definition. Then a SAGE Graph Neural Network (GNN) (Hamilton et al., 2017) with 2 SAGE layers with dimensions of sizes 256, 512 and 512 was used to perform message passing from every defined word to the words in its definition. To train this GNN we used 1 negative example for every positive edge, and trained the model for 50 epochs. Adam with 0.001 as learning rate was used as optimizer.

More than one model architecture were compared. As seen in figure 1, the final model was built with a Transformer encoder and an additional Multi-Layer Perceptron (MLP) with two linear layers with dimensions of sizes 512, 256, 256 and a ReLU layer between them, which during evaluation in the development set achieved better results than models based just on Transformer encoders or MLPs. Adam was used as optimizer. During training two loss functions from PyTorch were compared: Cosine Embedding Loss and Mean Square Error (MSE) Loss, which correspond to two of the task evaluation metrics. Regarding the hyperparameters, the following optimized the evaluation on the development set: 8 attention heads, 6 encoder layers, batch size=2048, learning rate=0.001. Model converges after around 10 epochs. Models were trained using 4 Nvidia Tesla v100 32GB.

As seen in figure 2, during training and prediction the process was the following. For a given sentence, stopwords and category tokens were removed. In the lexical graph model every remaining word was also lemmatized. After that the sentence embeddings are initialized, either with the Sentence-Transformers model, either by the mean of the lexical graph embeddings of every word in the definition. These initial embeddings are fed into the model along with some negative examples for the Cosine Embedding Loss model, receiving these a target label of 0. In the case of MSE Loss, negative sampling was not performed.

As stated in the CODWOE task guidelines, Reverse Dictionary submissions were evaluated using three metrics:

- mean squared error between the predicted embedding and the word embedding.
- cosine similarity between the predicted embedding and the word embedding.
- cosine-based ranking between the predicted embedding and the word embedding, which means how many other predicted embeddings have a higher cosine similarity with the target word than the right predicted one.

Since the Sentence-Transformer model was faster at generating the initial definition embeddings, it was used to initialize the definition embeddings in the final training and predictions. We understand that because of this our results in the task are not valid. However, the lexical graph approach can achieve almost similar results without the use of external data.

As seen in table 2, two different loss functions were used separately during training. Therefore,

[3] `https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased`

| | MSE | Cosine Similarity | Cosine-based ranking |
|---|---|---|---|
| Cosine Embedding Loss | 2.0157 | 0.4029 | 0.1665 |
| Mean Square Error Loss | 0.9106 | 0.2274 | 0.5003 |

**Table 2:** Reverse Dictionary results

two models were eventually presented. During evaluation, the first model trained with Cosine Embedding Loss reached more than 0.4 in cosine score and 0.16 in cosine ranking, which we consider remarkable and, according to the rankings[4], better than the top result for these particular Spanish embeddings. However, this model reached more than 2 in MSE, which considerably worsens the baseline (0.92) and the top results (0.85). On the other hand, the MSE Loss trained model slightly improves the baseline test MSE (0.91) and cosine score (0.22) but worsens the cosine ranking score. Other attempts to combine both loss functions did not success and achieved worse results in each of the evaluation metrics.

In the end, these results were found to be scalable to another languages and embedding types by using this same model architecture, and encountering in the way the same issues as for the Spanish “sgns” embeddings, that is, trouble combining MSE and cosine metrics. A Cosine Embedding Loss model for English ”sgns” embeddings achieves 0.34 cosine and 1.58 MSE, and using Spanish ”char” embeddings it reached 0.84 cosine and 1.66 MSE. As for the case of Spanish ”sgns”, compared to the top-ranked participants, a better cosine was achieved in exchange of a worse MSE. This leads to the opinion that the system architecture is easily scalable to other inputs, but it suffers from the same issues that with other languages and embeddings: a higher cosine similarity score can be achieved by using Cosine Embedding Loss, but in exchange of a worse MSE score.

## 4 Conclusions

For these tasks a combination of Machine Learning techniques and linguistic knowledge was proposed, in order to achieve good results and to understand the problems and the future challenges of these tasks.

In this paper, the main gaps in the dataset from a lexicographic perspective and its lack of coherence and exactitude were explained, and then a preprocessing solution was proposed, which was finally used in the RD system to avoid the problems that the dataset could carry in the model. Eventually, MMG team has presented a novel approach with an architecture that is easily scalable to other languages and embedding types. We understand that due to the use of external resources our results in the task are not valid for the challenge. However, we would like to remark that the lexical graph approach, which in the end we did not submit due to speed issues, achieved almost similar results.

These tasks are considered to represent an excellent starting point for research on the relationships between dictionaries and word embeddings. Both subtasks in general and our research on them in particular open up many options for further investigation. In our case, our intention is to use more linguistic knowledge at different levels, further exploring the power of linguistic graphs and putting into practice what we have learned in the Reverse Dictionary task to create quality Definition Modeling systems.

## References

Bevilacqua, M., Maru, M., and Navigli, R. (2020). Generationary or “how we went beyond word sense inventories and learned to gloss”. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221.

Bilac, S., Watanabe, W., Hashimoto, T., Tokunaga, T., and Tanaka, H. (2004). Dictionary search based on the target word description. In *Proc. of the Tenth Annual Meeting of The Association for Natural Language Processing (NLP2004)*, pages 556–559.

Calvo, H., Méndez, O., and Moreno-Armendáriz, M. A. (2016). Integrated concept blending with vector space models. *Computer Speech & Language*, 40:79–96. ISBN: 0885-2308 Publisher: Elsevier.

[4] https://competitions.codalab.org/competitions/34022#results

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

del Teso Martín, E. (1987). En torno a la definición lexicográfica. *Contextos*, (10):29–56.

Dutoit, D. and Nugues, P. (2002). A lexical database and an algorithm to find words. In *ECAI 2002: 15th European Conference on Artificial Intelligence, July 21-26, 2002, Lyon France: Including Prestigious Applications of Intelligent Systems (PAIS 2002): Proceedings*, volume 77, page 450. IOS Press.

El-Kahlout, I. D. and Oflazer, K. (2004). Use of wordnet for retrieving words from their meanings. In *Proceedings of the global Wordnet conference (GWC2004)*, pages 118–123.

Gadetsky, A., Yakubovskiy, I., and Vetrov, D. (2018). Conditional generators of words definitions. *arXiv preprint arXiv:1806.10090*.

García, J. and José, E. (2017). Forma y función del diccionario: hacia una teoría general del ejemplo lexicográfico. *Forma y función del diccionario*, pages 1–151.

Garg, S., Vu, T., and Moschitti, A. (2020). Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7780–7788.

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Hill, F., Cho, K., Korhonen, A., and Bengio, Y. (2016). Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Kabiri, A. and Cook, P. (2020). Evaluating a multisense definition generation model for multiple languages. In *International Conference on Text, Speech, and Dialogue*, pages 153–161. Springer.

Malekzadeh, A., Gheibi, A., and Mohades, A. (2021). Predict: Persian reverse dictionary. *arXiv preprint arXiv:2105.00309*.

Méndez, O., Calvo, H., and Moreno-Armendáriz, M. A. (2013). A reverse dictionary based on semantic analysis using wordnet. In *Mexican International Conference on Artificial Intelligence*, pages 275–285. Springer.

Mickus, T., Paperno, D., Constant, M., and van Deemter, K. (2022). SemEval-2022 Task 1: Codwoe – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Noraset, T., Liang, C., Birnbaum, L., and Downey, D. (2017). Definition modeling: Learning to define word embeddings in natural language. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ortega-Martín, M. (2021). *Grafos de vinculación semántica a partir del definiens del DUE*. PhD thesis, Universidad Complutense de Madrid.

Qi, F., Zhang, L., Yang, Y., Liu, Z., and Sun, M. (2020). Wantwords: an open-source online reverse dictionary system. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–181.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Shaw, R., Datta, A., VanderMeer, D., and Dutta, K. (2011). Building a scalable database-driven reverse dictionary. *IEEE Transactions on Knowledge and Data Engineering*, 25(3):528–540.

Siddique, B. and Sufyan Beg, M. M. (2018). A review of reverse dictionary: Finding words from concept description. In *International Conference on Next Generation Computing Technologies*, pages 128–139. Springer.

Thorat, S. and Choudhari, V. (2016). Implementing a reverse dictionary, based on word definitions, using a node-graph architecture. *arXiv preprint arXiv:1606.00025*.

Yan, H., Li, X., and Qiu, X. (2020). Bert for monolingual and cross-lingual reverse dictionary. *arXiv preprint arXiv:2009.14790*.

Yang, L., Kong, C., Chen, Y., Liu, Y., Fan, Q., and Yang, E. (2020). Incorporating sememes into chinese definition modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1669–1677.

Zhang, H., Du, Y., Sun, J., and Li, Q. (2020a). Improving interpretability of word embeddings by generating definition and usage. *Expert Systems with Applications*, 160:113633.

Zhang, L., Qi, F., Liu, Z., Wang, Y., Liu, Q., and Sun, M. (2020b). Multi-channel reverse dictionary model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 312–319.

Zock, M. and Schwab, D. (2008). Lexical access based on underspecified input. In *COLING 2008: Proceedings of the Workshop on cognitive Aspects of the Lexicon (COGALEX 2008)*, pages 9–17.

## Appendix A Translation of Table 1

es.train.212: “Biology.— Said of microorganisms that do not accept the usual dyes .”
es.train.250: “Zoology.— Any of the hummingbirds of the genus Chlorostilbon .”
es.train.119: “To serve, to help each other.”
es.train.120: “Work ( pronominal use of ... )”