Automatic Word Association Norms (AWAN)

Jorge Reyes-Magaña[1,2] **Gerardo Sierra**[1] **Gemma Bel-Enguix**[1]
jorge.reyes@correo.uady.mx gsierram@iingen.unam.mx gbele@iingen.unam.mx
Helena Gómez-Adorno[3]
helena.gomez@iimas.unam.mx
[1]Universidad Nacional Autónoma de México, Instituto de Ingeniería,
Ciudad de México, México
[2] Universidad Autónoma de Yucatán, Facultad de Matemáticas, Mérida, México
[3] Universidad Nacional Autónoma de México, Instituto de Investigaciones
en Matemáticas Aplicadas y en Sistemas, Ciudad de México, México

Abstract

Word Association Norms (WAN) are collections that present *stimuli* words and the set of their associated responses. The corpus is widely used in diverse areas of expertise. In order to reduce the effort to have a good quality resource that can be reproduced in many languages with minimum sources, a methodology to build Automatic Word Association Norms is proposed (AWAN). The methodology has an input of two simple elements: a) dictionary, and b) pre-processed Word Embeddings. This new kind of WAN is evaluated in two ways: i) learning word embeddings based on the *node2vec* algorithm and comparing them with human annotated benchmarks, and ii) performing a lexical search for a reverse dictionary. Both evaluations are done in a weighted graph with the AWAN lexical elements. The results showed that the methodology produces good quality AWANs.

1 Introduction

Word associations is a technique that helps researchers to learn how words are connected by their meanings and the relationships among them in the human mind. Although vocabulary diversity and lexicon size depend on a variety of social elements among individuals, the final result is a kind of word distribution in the population. The method is used in psychology and linguistics to discover how the human mind structures knowledge (De Deyne et al., 2013). This type of resources reflect both semantic and episodic contents (Borge-Holthoefer and Arenas, 2009). In free association tests, a person is asked to say the first word that comes to mind in response to a given *stimulus* word. The set of lexical relations obtained with these experiments is called Word Association Norms (WAN).

The development of technological tools that will help gather these kinds of resources is starting to draw attention, mostly taking advantage of distributed technologies like the Internet. Small World of Words[1] is a clear example of that. We believe that this way of collaborative construction could bring a variety of problems, biasing the final results. On the other hand, the classic methodologies of WAN's construction are very time-consuming. Just to mention some disadvantages, many people are needed to compile the data. Furthermore, good control of the environment conditions of the experiments is important, as well as carefully selecting a set of metadata that must be annotated: age, education years, gender, etc.

In the end, the complete WAN could take years to be polished and shared with the scientific community. Nevertheless, this effort is worthwhile, as WAN could help diverse areas of study: psychologists, linguists, neuroscientists and others, to test new theories about how we represent and process language.

In this paper, a methodology to build automatic WAN is presented. We called the resource generated Automatic Word Associaton Norms (AWAN). The language used to prove our methodology is Spanish, more specifically Mexican Spanish.

The only WAN corpus for Mexican Spanish is the *Normas de Asociacion de Palabras para el Español de México* (Arias-Trejo et al., 2015) (from here, this corpus will be referred to as Mexican Spanish WAN; i.e., MSWAN), which was built using a classic methodology.

[1]https://smallworldofwords.org/en/project

Proceedings of the Workshop on Cognitive Aspects of the Lexicon, pages 142–153
Barcelona, Spain (Online), December 12, 2020

The AWAN methodology presented here pretends to infer semantic relatedness between stimuli and their responses. The main reason is that word association has been of great interest as a tool to research mechanisms related to semantic memory (Barrón-Martínez and Arias-Trejo, 2014). The main relations shown in MSWAN are: metonymy, meronymy, functionality, cohyponymy, qualification, hyponymy, 'made of' and synonymy (Mijangos et al., 2017). Our objective is to capture the semantic relatedness but not the types of relation.

Gómez-Adorno et al. (2019) presented Word Embeddings based on *node2vec* (Grover and Leskovec, 2016) and a graph constructed with the MSWAN corpus. Besides, the work presented by Reyes-Magaña et al. (2019a) used the MSWAN to develop a lexical search model for the implementation of a reverse dictionary. With these two works, we can obtain a gold standard to be compared with our AWAN running on the same tasks.

The elements we used to build the AWAN are a general dictionary and a set of pretrainned word vectors. Specifically, we used the Mexican Spanish Dictionary, *Diccionario del Espanol de México* (DEM, 2010), as the main input for our methodology and the pretrained embeddings available for Spanish [2]. The algorithms that were used to train these embeddings are: FastText(Bojanowski et al., 2017), Word2Vec (Mikolov et al., 2013), and Glove (Pennington et al., 2014).

The rest of the paper is organized as follows. In Section 2, the related work is discussed. In Section 3, a description of the methodological framework for the construction of the Automatic Word Association Norms is presented. Section 4 shows the evaluation of the generated norms, using a word similarity dataset in Spanish and the lexical search model. Finally, in Section 5, we establish some conclusions and discuss possible directions of future work.

2 Related Work

In linguistics and psycholinguistics, semantic networks (Sowa, 1992) are defined as graphs relating words (Aitchison, 2012). Their use is not exclusive to learn the organization of the vocabulary, but also to draw the structure of knowledge.

WAN are a special kind of semantic networks, and they are available in many languages. The creation of WAN is not new. The first example is Roget (1911), and two very well-known resources are the *Edinburgh Associative Thesaurus*[3] (EAT) (Kiss et al., 1973) and the collection of the University of South Florida (USF) (Nelson et al., 1998)[4]. Thanks to the Internet and new technologies, WAN lists have been more efficiently compiled in the last years, with the help of a large number of volunteers. Some examples are: *Jeux de Mots*[5], in French (Lafourcade, 2007) and the multilingual dataset *Small World of Words* [6].

For Spanish, there exists several corpora of word associations. Algarabel et al. (1998) integrate 16,000 words, including statistical analyses of the results. Macizo et al. (2000) build norms for 58 words in children, and Fernández et al. (2004) work with 247 lexical items that correspond to Spanish (Sanfeliu and Fernández, 1996).

As stated above, the only resource designed and compiled for Mexican Spanish is the MSWAN. Reyes-Magaña et al. (2019a) introduced a method for lexical search based on that compilation that worked from clue words or definitions to the concept, i.e., from the responses to the *stimuli*.

In some cases, authors create this type of corpus from scratch and in other cases, they extend the available WAN to learn more responses to the *stimuli*. In recent years, Bel-Enguix et al. (2014) used techniques of graph analysis to calculate associations from large collections of texts. Additionally, Garimella et al. (2017) published a model of word associations that was sensitive to the demographic context. This was based on a neural network architecture with n-skip-grams and improved the performance of the generic techniques, which do not take into account the demography of the participant.

Sinopalnikova and Smrz (2004) showed that Word Association Thesaurus (WAT) is comparable to balanced text corpora and can replace them in case of absence of a corpus. The authors presented a

[2]https://github.com/dccuchile/spanish-word-embeddings
[3]http://www.eat.rl.ac.uk/
[4]http://web.usf.edu/FreeAssociation
[5]http://www.jeuxdemots.org/
[6]https://smallworldofwords.org/

methodological framework for building and extending semantic networks with WAT, including a comparison of quality and information provided by WAT vs. other language resources.

Borge-Holthoefer and Arenas (2009) used free association information for extracting semantic similarity relations with a Random Inheritance Model (RIM). The obtained vectors were compared with LSA-based vector representations and the WAS (word association space) model. Their results indicate that RIM can successfully extract word feature vectors from a free association network.

In the work by De Deyne et al. (2016), the authors introduced a method for learning word vectors from WANs using a spreading activation approach in order to encode a semantic structure from WAN. The authors used part of the *Small World of Words* network. The word-association-based model was compared with a word embeddings model (Word2Vec) using relatedness and similarity judgments from humans, obtaining an average of 13% of improvement over the Word2Vec model.

In the recent work by Bel-Enguix et al. (2019), the authors used two WAN in English, EAT and USF to produce word embeddings that were tested against human-annotated benchmarks and some external tasks, Showing that this kind of learning method produces good quality vectors without a training corpus based on billions of words.

WANs are proved to be good in a reverse dictionary task since they suitably represent the connections between words and the way concepts are linked in the human mind. The whole scenario of onomasiological searches changed with the universalization of the Internet and language technologies that allowed to build online resources powered by the huge corpus the World Wide Web provides. In the last two decades, several online dictionaries have been designed that allow natural language searches. Users enter their own definition in natural language and the engine looks for the words that match such definition.

One of the first online dictionaries allowing this type of search was the one created for French by Dutoit and Nugues (2002). Bilac et al. (2004) designed a dictionary for Japanese where the users can freely enter their definitions. It has an algorithm that calculates the similarity between concepts comparing the words.

El-Kahlou & Oflazer (El-Kahlout and Oflazer, 2004) built a similar resource for Turkish. They took into account some synonymy relations between words, as well as the similarity of definitions by means of a counter of similar words in the same order and in subsets of such words. For English, there exists an online onomasiological dictionary, OneLook Reverse Dictionary,[7] that retrieves acceptable results.

One of the main works in Spanish is the one by Sierra and McNaught (2000). DEBO is an onomasiological dictionary that works with user queries given in natural language and a search engine, which was later improved; the database structure was also optimized (Sierra and Hernández, 2011).

Finally, the use of WANs to build a reverse dictionary in Spanish is presented by Reyes-Magaña et al. (2019a). The authors used the corpus MSWAN and graph-based techniques, specifically a measure of betweenness centrality, to perform searches in the knowledge graph. The results of the search model overcome the information retrieval systems it was compared to. The same methodology is successfully applied to English in Reyes-Magaña et al. (2019b). In the latter work, another graph algorithm was presented additionally to perform the search, the PageRank. Nevertheless, the results show that betweenneess centrality is more suitable for the reverse dictionary task.

3 Methodology of Automatic Word Association Norms

The aim of this work is to present a general methodology that could serve as a model to build WAN for any language. The main process consists in parsing the entire dictionary, working with the entries and their definitions. We consider that all the entries become the stimuli words, and each one of the words that define the entries become the associate responses to them. The process also involves the inference of a numeric value that measures the relationship between words, allowing us to obtain the weight the classic WANs have.

Algorithm 1 presents the overall schema of our model. The dictionary, *Diccionario del Español de México* (DEM, 2010), is the result of a set of investigations of the vocabulary used in Mexico since 1921. The investigations have been carried out since 1973 at the Center for Linguistic and Literary Studies of *El Colegio de México*. The Mexican Dictionary of Spanish is a comprehensive dictionary of Spanish in

[7] https://www.onelook.com/reverse-dictionary.shtml

Algorithm 1: Automatic Word Association Norms

```
Data: Dictionary, Word embeddings
Result: AWAN
pre-process(Dictionary)
 for each entry do
    for each word in definition do
        similarity = cosine_similarity(entry,word);
        weight = similarity * tf_idf(words);
    ordering(words)
```

its Mexican variety, prepared on the basis of an extensive study of the Corpus of contemporary Mexican Spanish (1921-1974) and a set of data after that last date to the present.

Sometimes, the definitions of each one of the entries bring examples of use. All of this additional information was removed because we consider that this kind of data could contaminate the final WANs. Then, in order the prepare the definitions, we performed some preprocessing steps, as described.

- All the words are lemmatized using Freeling (Padró and Stanilovsky, 2012) for the Spanish language.
- All the functional words were removed using the Spanish *stop words* list available in the *NLTK* package (Bird and Loper, 2004).
- Some specific words were added to the *stop list* in order to be removed as well. These words are very common in dictionaries but do not provide meaningful data for our purpose of building AWAN. Some of them are: 'etc.', 'approximately', 'generally', 'specifically', 'type', among others.

Later, with the remaining words, we calculate the cosine similarity between the entry and each word corresponding to its definition. For this purpose, we use pretrainned word embeddings[8] for Spanish language. Table 1 presents the main characteristics of each embeddings model. The corpora used to train these embeddings are the following: FastText, Glove and Word2Vec with a Spanish Billion Word Corpus, and FastWiki with Wikipedia Spanish Dump.

Short name	**Model file**	**Dimensions**	**# vectors**	**Algorithm**
FastText	FastText from SBWC	300	855,380	FastText with Skipgram
Glove	GloVe from SBWC	300	855,380	GloVe
Word2Vec	Word2Vec from SBWC	300	1,000,653	Word2Vec with Skipgram
FastWiki	FastText from Spanish Wikipedia	300	985,667	FastText with Skipgram

Table 1: Description of the pretrained vectors in Spanish used to measure similarities.

With the remaining lexical elements, the *tf-idf* of each word is calculated; every definition is considered as a different document. The value will be used as adjustment factor of the cosine similarity between words. The weight is calculated as follows:

$$W_{as}(stimulus, response) = tf\_idf(response) * cosine\_similarity(stimulus, response) \quad (1)$$

We called this weight Approximation Strength (W_{as}). The final step is to order from high to low the weights of all the associated responses (words in a definition) to the entries (*stimuli*).

[8]https://github.com/dccuchile/spanish-word-embeddings

The corpus of the AWANs is available in github[9]. We generate four different collections, one for each embedding used to calculate the cosine similarity.

3.1 AWAN Corpus and Graph

The corpus AWAN has a total of 17,330 *stimuli*. The vocabulary size of each AWAN is : a) FastText with 22,699 b) Glove with 21,867 c) Word2Vec with 22,045 and d) FastWiki with 22,517. The discrepancy in the vocabulary sizes is due to the embeddings corpus, not all the words that appear in the definitions are in the vector resources. The richest AWAN, in terms of amount of lexical items, is the FastText version.

The graph representing the AWAN is elaborated with all the lexical items. It is formally defined as: $G = \{V, E, \phi\}$ where:

- $V = \{v_i | i = 1, ..., n\}$ is a finite set of nodes of length n, $V \neq \emptyset$, that corresponds to the *stimuli* and their *associates*.
- $E = \{(v_i, v_j) | v_i, v_j \in V, 1 \leq i, j \leq n\}$, is the set of edges.
- $\phi : E \rightarrow R$, is a function over the weight of the edges.

The graph is undirected so that every stimulus is connected to their associated words without any precedence order. For the weight of the edges we use the Approximation Strength measure. Table 2 presents a brief snapshot of the AWAN corpus, in specific for the stimulus *stimulus* Bee *(Abeja)* and its responses, using the FastText and Glove embeddings. It can be observed that they share the same responses, but they are located in different positions. The cosine similarity obtained on each embedding corpus produces the arrangement adjustment.

4 AWAN evaluation

To measure the quality of the AWANs, two types of experiments were performed. The first one allows us to know the representativeness of the embeddings that were trained using AWAN and node2vec, trying to describe similarity against human-annotated benchmarks. The second experiment is about the lexical search model used in the reverse dictionary. This evaluation is done because the WANs prove to be well-performing lexical searches using this kind of corpus as input (Reyes-Magaña et al., 2019a). Each one of the experiments will be compared with the results of these tasks using MSWAN. We select these outcomes as the gold standard because the WAN corpus used to perform the experiments corresponds to Mexican Spanish, same as our AWANs.

4.1 Node2vec

Node2vec (Grover and Leskovec, 2016) finds a mapping $f : V \rightarrow R^d$ that transforms the nodes of a graph into vectors of d-dimensions. It defines a neighborhood in a network $N_s(u) \subset V$ for each node $u \in V$ through a S sampling strategy. The goal of the algorithm is to maximize the probability of observing subsequent nodes on a random path of a fixed length.

The sampling strategy designed in *node2vec* allows it to explore neighborhoods with skewed random paths. In this work, we used the implementation of the project *node2vec*, which is available on the web[10] considering a dimension of 300.

With the embeddings trained on AWAN, we evaluated the ability of word vectors to capture semantic relationships through a word similarity task. Specifically, we used two widely-known corpora: a) the corpus *WordSim-353* (Finkelstein et al., 2001) composed of pairs of terms semantically related to similarity scores given by humans and b) the MC-30 (Miller and Charlees, 1991) benchmark containing 30 word pairs. Both datasets in their Spanish version (Hassan and Mihalcea, 2009).

We calculated the cosine similarity between the vectors of word pairs contained in the above mentioned datasets and compared it with the similarity given by humans using the Spearman correlation. To deal with the non-inclusion of every word of the testing datasets in our AWAN, we introduced the concept

[9]https://github.com/jocarema/AWAN

[10]http://snap.stanford.edu/node2vec/

Table 2: Responses for *stimulus* Bee, using FastText and Glove.

***Abeja* (Bee)**			
FastText		**Glove**	
Response	**Approximation Strength**	**Response**	**Approximation Strength**
mellifera	0.599	*apis*	0.521
miel	0.580	*miel*	0.442
zángano	0.552	*mellifera*	0.441
apis	0.550	*hembra*	0.393
himenóptero	0.546	*zángano*	0.353
aguijón	0.532	*reina*	0.343
insecto	0.520	*nido*	0.336
néctar	0.511	*insecto*	0.321
cera	0.506	*macho*	0.320
hembra	0.485	*cera*	0.310
polen	0.482	*aguijón*	0.304
macho	0.464	*panal*	0.303
polinizador	0.461	*néctar*	0.290
apidae	0.431	*polen*	0.286
nido	0.412	*estéril*	0.259
reina	0.403	*apidae*	0.195
panal	0.367	*himenóptero*	0.181
domesticar	0.339	*fértil*	0.179
estéril	0.318	*colonia*	0.164
amarillo	0.315	*amarillo*	0.160
rojizo	0.311	*alimentar*	0.131
colonia	0.300	*solo*	0.126
obrero	0.284	*vivir*	0.117
fértil	0.273	*domesticar*	0.107
solo	0.257	*obrero*	0.106
vello	0.240	*rojizo*	0.099
producto	0.230	*misión*	0.081
galería	0.220	*vello*	0.076
alimentar	0.214	*existir*	0.073
medir	0.195	*producto*	0.064
vivir	0.185	*construir*	0.062
existir	0.176	*galería*	0.061
constituir	0.171	*frecuencia*	0.058
frecuencia	0.169	*polinizador*	0.056
numeroso	0.163	*cubrir*	0.055
cubrir	0.162	*medir*	0.047
misión	0.155	*constituir*	0.043
aprovechar	0.146	*aprovechar*	0.035
subterráneo	0.146	*subterráneo*	0.025
proveer	0.138	*numeroso*	0.007
construir	0.107	*proveer*	0.003

of overlap in the experiments, and calculated the total number of common words between the lists that are being compared. The others are excluded from the evaluation. In principle, having large overlaps is a positive feature of this approach. Tables 3 and 4 present the Spearman correlation of the similarity

given by human taggers with the similarity obtained with word vectors (learned from MSWAN and AWAN separately). We also report the overlap, which is the number of words that can be found in both, the given WAN corpus (MSWAN or AWAN) and the evaluation dataset (ES-WS-353 or MC-30).

Table 3: Spearman rank order correlations between Mexican Spanish WAN embeddings (300 dimension) and the ES-WS-353 dataset.

WAN	Weighting function	Overlap	Correlation
MSWAN (Gómez-Adorno et al., 2019)	Inv. Frequency	140	0.489
	Inv. Association		0.463
	Time		0.461
AWAN FastText	Inv. Approximation	291	**0.595**
AWAN Glove			0.555
AWAN Word2Vec			0.550
AWAN FastWiki			0.572

Table 4: Spearman rank order correlations between Spanish WAN embeddings (300 dimension) and the MC-30 dataset

WAN	Weighting function	Overlap	Correlation
MSWAN (Gómez-Adorno et al., 2019)	Inv. Frequency	11	0.305
	Inv. Association		0.563
	Time		0.545
AWAN FastText	Inv. Approximation	22	0.747
AWAN Glove			0.698
AWAN Word2Vec			0.706
AWAN FastWiki			**0.771**

It can be observed that the word embeddings obtained from the AWAN corpus achieved better correlation with the human similarities than the embeddings obtained from the MSWAN corpus in both datasets, ES-WS-53 and MC-30.

4.2 Lexical Search Model

Given a definition, the search in the graph is done considering the word that better matches with it. For this purpose, centrality measures identify the most important nodes in a graph; the variation of the *betweenness centrality* (BT) algorithm (Freeman, 1977) which instead of computing BT of all pairs of nodes in a graph, calculates the centrality based on a sample (subset) of nodes (Brandes, 2008). This approximation is formally described as follows:

$$C_{btw\_aprox}(v) = \sum_{i \in I, f \in F} \frac{\sigma_{i,f}(v)}{\sigma_{i,f}} \tag{2}$$

where: I is the set of initial nodes, F is the set of final nodes, $\sigma_{i,f}$ is the number of shortest paths between i and f, and $\sigma_{i,f}(v)$ is the number of those paths that passes through some node v that is not i or f.

In a non-weighted-graph, the algorithm looks for the shortest path. In a weighted graph, such algorithm finds the path that minimizes the sum of the weight of the edges. When using WAN as the input corpus, we obtain the weighted one.

We employ the approximation of the BT algorithm in order to search for the concept related to a given definition because it only uses a subset of nodes to find the most central nodes in the graph. Therefore, we define a subgraph composed by the words (nodes) of the definition. This subgraph is used as both initial and final nodes to calculate the shortest paths from each of the nodes of the initial nodes set to each one of the nodes of the final nodes set. Finally, the nodes are ranked taking the measure of BT as a parameter for the comparison of the most important nodes found by the algorithm.

We constructed the AWAN graph considering only the 234 *stimuli* of MSWAN but having the response associated and the weights, using the algorithm 1 previously described.

For the experiments, we use the small corpus available in github[11]. It is reported that this corpus contains 5 definitions for 56 concepts corresponding to *stimuli* of the MSWAN, with a total number of 280 definitions. The corpus was gathered with the collaboration of students who gave their own description of the word. For the evaluation of the inference process, we used the technique of precision at k ($p@k$) (Manning et al., 2009), for example, $p@1$ shows that the concept associated to a given definition was ranked correctly in the first place; in $p@3$ the concept was in the first three results, and the same applies to $p@5$.

The results are shown in Table 5. It is clear that when the model searches over MSWAN graphs weighted with any function, the results are higher than when searching on the AWAN graph. We consider that the precision obtained with the AWAN corpus is still competitive. We can affirm this because in the work of Reyes-Magaña et al. (2019a), the authors describe and implement other retrieval information systems applied to the reverse dictionary, being all outperformed by our AWAN graphs. These methods were: Boolean IR, *Onelook reverse dictionary* [12], BM-25 (Robertson and Zaragoza, 2009) and CAS (Ghosh et al., 2014).

Table 5: Lexical Search Results in terms of precision.

WAN	**Weighting function**	**p@1**	**p@3**	**p@5**
MSWAN (Reyes-Magaña et al., 2019a)	Inv. Frequency	0.616	0.741	0.774
	Inv. Association	**0.655**	**0.804**	**0.829**
	Time	0.362	0.550	0.652
AWAN FastText	Inv. Approximation	0.329	0.526	0.584
AWAN Glove		0.333	0.544	0.587
AWAN Word2Vec		0.340	0.537	0.584
AWAN FastWiki		0.326	0.526	0.580

We did some additional experiments to prune the graph. For this purpose, on each AWAN we vary the weight with incremental intervals of .05. Figure 1 shows the precision of the lexical search; this value is seen on the vertical axis. The horizontal axis represents from left to right the reductions of responses that satisfy the filter, meaning that, if we have the value of .1, the responses to be considered will be those whose weights vary from 1 to .1. In the case of .55, we only select responses with weight from 1 to .55, and so on. With this technique, we could see if there is an improvement of precision as we vary the values in weights. The reason to perform this experiment is that in some cases, a more compact graph yields more efficient searches. When the reduction reaches a value of .60, the filtered responses are bigger, having fewer words to work with and making the precision of lexical search, turns almost to 0. We can see that in the first intervals, reducing the graph does not make a significant difference in the precision outcomes. A slight peak can be reached before the precision starts to decrease. For this reason, we provide full AWANs without any reduction.

5 Conclusions and Future Work

We introduced a method for learning Word Association Norms in Mexican Spanish from a dictionary. Although we could use a general Spanish dictionary like the *Real Academia de la Lengua (RAE)*, the experiments did not yield good results, mainly because the test corpus is based on definitions made by people that use Mexican Spanish as their mother tongue. Nevertheless, the methodology we provide in this paper can be applied to any kind of dictionary. To evaluate the AWANs, we used two types of test. The first one is the intrinsic test which uses the *node2vec* algorithm to learn word vectors on the graph built with the AWAN corpus. The results determine that these vectors overcome the Spearman correlation presented with the MSWAN corpus. The second one is the extrinsic test which presents a more realistic

[11] https://github.com/jocarema/Natural-Language-definitions

[12] https://www.onelook.com/reverse-dictionary.shtml

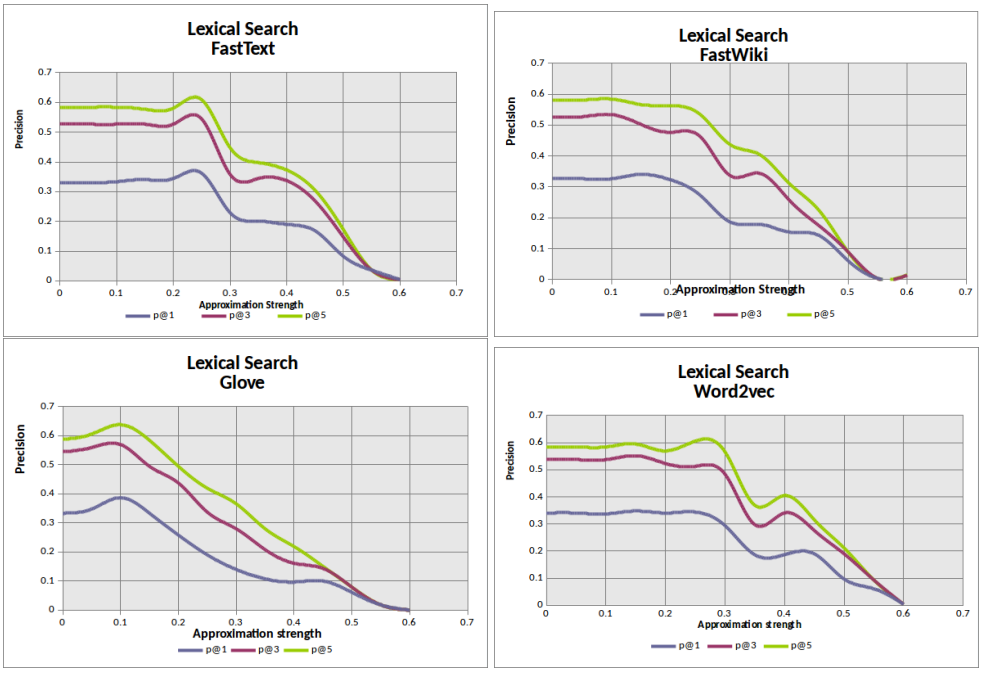

Figure 1: Lexical Search precision based on AWAN

use of this type of corpus; the lexical search model shows that even if we did not outperformed the results of the MSWAN, it is competitive enough to outperform classic information retrieval systems. We employ a weighting function on the graph edges considering the inverse approximation strength because all the tests use the shortest paths.

We consider that the methodology proposed is a helpful tool for the construction of Word Association Norms. The input elements to produce AWANs are somehow easy to get, and consist mainly of a dictionary, and the pretrained word embeddings. We also believe that the MSWAN collected and processed by humans, will bring more accurate results depending on the task that will be used. Nevertheless, in some cases where time and availability of WAN is urgent or simply impossible to collect in the classic way, the creation of AWAN is a reliable and fast solution. In a more advanced stage, the success of the technique can make unnecessary the effort and resources that are currently dedicated to collect WANs.

Besides, as a parallel result, we provide the Word Embeddings [13] that we trained using the *node2vec* algorithm, having as the most important feature that these vectors are based on Mexican Spanish. We claim that this methodology can be used to produce embeddings for specific variants of a language without a huge amount of data.

As future work, we plan to do some additional experiments to increase the precision in lexical search in order to apply some additional filters in the response words, like having only nouns, verbs, and/or adjectives, with all the possible combinations a POS tagging can produce. Roth and im Walde (2008) showed that the WANs can be enriched using diverse types of corpora in addition to a dictionary. Hence, as future work, we plan to add some encyclopedic and co-ocurrence corpora resources in order to improve the tasks performance on WANs. Also, it is possible to have the incorporation of multi-terms is possible to have, adding the vector representation of each word in the multi-term. This could be done by applying the same methodology.

Acknowledgments

This work was supported by UNAM-PAPIIT AG400119, IA401219, TA100520 and CONACyT A1-S-27780. We thank *El Colegio de México* for the availability of the *Diccionario del Español de México*.

[13] https://drive.google.com/drive/folders/1nmApEvi4ywQl1CDjK5umiSQE79MuGP9C

References

Jean Aitchison. 2012. *Words in the mind: An introduction to the mental lexicon*. John Wiley & Sons.

Salvador Algarabel, Juan Carlos Ruíz, and Jaime Sanmartín. 1998. *The University of Valencia's computerized Word pool*. Behavior Research Methods, Instruments & Computers.

Natalia Arias-Trejo, Julia B. Barrón-Martínez, Ruth H. López Alderete, and Francisco A. Robles Aguirre. 2015. *Corpus de normas de asociación de palabras para el español de México [NAP]*. Universidad Nacional Autónoma de México.

Julia B Barrón-Martínez and Natalia Arias-Trejo. 2014. Word association norms in mexican spanish. *The Spanish journal of psychology*, 17.

Gemma Bel-Enguix, Reinahrd Rapp, and Michael Zock. 2014. A graph-based approach for computing free word associations. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference*, pages 221–230.

Gemma Bel-Enguix, Helena Gómez-Adorno, Jorge Reyes-Magaña, and Gerardo Sierra. 2019. Wan2vec: Embeddings learned on word association norms. *Semantic Web*.

S Bilac, W Watanabe, T Hashimoto, T Tokunaga, and H Tanaka. 2004. Dictionary search based on the target word description. In *Proceedings of the Tenth Anual Meeting of the Assocviation for Natural Language Processing*, pages 556–559.

Steven Bird and Edward Loper. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Javier Borge-Holthoefer and Alex Arenas. 2009. Navigating word association norms to extract semantic information. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Ulrik Brandes. 2008. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145.

Simon De Deyne, Daniel J. Navarro, and Gert Storms. 2013. Associative strength and semantic activation in the mental lexicon: Evidence from continued word associations. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Simon De Deyne, Amy Perfors, and Daniel J Navarro. 2016. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1861–1870.

2010. *Diccionario del Español de México (DEM)*. El Colegio de México, A.C.

M Dutoit and P Nugues. 2002. A lexical database and an algorithm to find words from definitions. In *Proceedings of the 15th European Conference on Artificial Intelligence*, pages 450–454.

I.D El-Kahlout and K Oflazer. 2004. Use of wordnet for retrieving words from their meanings. In *2nd Global WordNet Conference.*

Ángel Fernández, Emilio Díez, M. Ángeles Alonso, and M. Soledad Beato. 2004. Free-association norms form the spanish names of the snodgrass and vanderwart pictures. *Behavior Research Methods, Instruments & Computers*, 36:577–583.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10$_{th}$ International Conference on World Wide Web*, pages 406–414. ACM.

Linton C Freeman. 1977. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

Aparna Garimella, Carmen Banea, and Rada Mihalcea. 2017. Demographic-aware word associations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2285–2295.

Urmi Ghosh, Sambhav Jain, and Paul Soma. 2014. A two-stage approach for computing associative responses to a set of stimulus words. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex-IV). COLING 2014 25th International Conference on Computational Linguistics*, pages 15–21.

Helena Gómez-Adorno, Jorge Reyes-Magaña, Gemma Bel-Enguix, and Gerardo E Sierra. 2019. Spanish word embeddings learned on word association norms. In *13th Alberto Mendelzon International Worshop on Fundations of Data Managment*.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining*, pages 855–864. ACM.

Samer Hassan and Rada Mihalcea. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1192–1201. Association for Computational Linguistics.

G.R. Kiss, Ch. Armstrong, R. Milroy, and J. Piper. 1973. *An associative thesaurus of English and its computer analysis*. Edinburgh University Press, Edinburgh.

Mathieu Lafourcade. 2007. Making people play for lexical acquisition. *In Proceedings of the 7th SNLP 2007, Pattaya, Thaïland*, 7:13–15, December.

Pedro Macizo, Carlos J. Gómez-Ariza, and M. Teresa Bajo. 2000. Associative norms of 58 spanish for children from 8 to 13 years old. *Psicológica*, 21:287–300.

C.D. Manning, P. Raghavan, and H. Schutze. 2009. *Introduction to Information Retrieval*. Cambridge University Press.

Victor Mijangos, Julia B Barrón-Martınez, Natalia Arias-Trejo, and Gemma Bel-Enguix. 2017. A graph-based analysis of the corpus of word association norms for mexican spanish.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Computing Research Repository*, arXiv:1301.3781.

G.A. Miller and W.G. Charlees. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. 1998. *Word association rhyme and word fragment norms*. The University of South Florida.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Jorge Reyes-Magaña, Gemma Bel-Enguix, Helena Gómez-Adorno, and Gerardo Sierra. 2019a. A lexical search model based on word association norms. *Journal of Intelligent & Fuzzy Systems*, 36(5):4587–4597.

Jorge Reyes-Magaña, Gemma Bel-Enguix, Gerardo Sierra, and Helena Gómez-Adorno. 2019b. Designing an electronic reverse dictionary based on two word association norms of english language. *Electronic lexicography in the 21st century (eLex 2019): Smart lexicography*, page 142.

S. Robertson and H. Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

R. Roget. 1911. *Roget's Thesaurus of English Words and Phrases (TY Crowell co*.

Michael Roth and Sabine Schulte im Walde. 2008. Corpus co-occurrence, dictionary and wikipedia entries as resources for semantic relatedness information. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Carmen Sanfeliu and Ángel Fernández. 1996. A set of 254 snodgrass' vanderwart pictures standardized for spanish: Norms for name agreement, image agreement, familiarity, and visual complexity. *Behavior Research Methods, Instruments, & Computers*, 28:537–555.

Gerardo Sierra and Laura Hernández. 2011. A proposal for building the knowledge base of onomasiological dictionaries. *Journal of Cognitive Science*, 12(3):215–232.

Gerardo Sierra and John McNaught. 2000. Design of an onomasiological search sistem: A concept-oriented tool for terminology. *Terminology*, 6(1):1–34.

Anna Sinopalnikova and Pavel Smrz. 2004. Word association thesaurus as a resource for extending semantic networks. pages 267–273.

John F Sowa. 1992. Conceptual graphs as a universal knowledge representation. *Computers & Mathematics with Applications*, 23(2):75–93.