A Lexical Search Model based on word association norms

Article *in* Journal of Intelligent and Fuzzy Systems · April 2019

DOI: 10.3233/JIFS-179010

4 authors:

Jorge Reyes-Magaña
Universidad Autónoma de Yucatán
3 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Gemma Bel-Enguix
Universidad Nacional Autónoma de México
93 PUBLICATIONS **209** CITATIONS
SEE PROFILE

Helena Gomez Adorno
Universidad Nacional Autónoma de México
47 PUBLICATIONS **531** CITATIONS
SEE PROFILE

Gerardo Sierra
Universidad Nacional Autónoma de México
115 PUBLICATIONS **513** CITATIONS
SEE PROFILE

Journal of Intelligent Fuzzy Systems xx (0) 1
IOS Press

A Lexical Search Model Based on Word Association Norms [1]

Jorge Reyes-Magaña [a,c], Gemma Bel-Enguix [a], Helena Gómez-Adorno [b,*] and Gerardo Sierra [a]
[a] *Instituto de Ingeniería (II), Universidad Nacional Autónoma de México (UNAM), Mexico City, Mexico*
[b] *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México (UNAM), Mexico City, Mexico*
[c] *Facultad de Matemáticas (FM), Universidad Autónoma de Yucatán (UADY), Merida-Yucatan, Mexico*

Abstract. This work introduces a lexical search model based on a type of knowledge graphs, namely word association norms. The aim of the search is to retrieve a target word, given the description of a concept, i.e., the query. This differs from traditional information retrieval models were complete documents related to the query are retrieved. Our algorithm looks for the keywords of the definition in a graph, built over a corpus of word association norms for Mexican Spanish, and computes the centrality in order to find the relevant concept. We performed experiments over a corpus of human-definitions in order to evaluate our model. The results are compared with a Boolean information retrieval (IR) model, the BM25 text-retrieval algorithm, an algorithm based on word vectors and an online onomasiological dictionary–OneLook Reverse Dictionary. The experiments show that our lexical search method outperforms the IR models in our study case.

Keywords: Information retrieval, Word association norms, Natural language graphs, Lexical search

1. Introduction

Two types of dictionaries can be distinguished in order to link a concept with its meaning: semasiological and onomasiological. The former provides meanings, ie. given a word, the user obtains the meaning of such word. The latter works in the opposite way, given the description of a word, the user obtains the related concept [4]. This kind of dictionary can be seen as an information retrieval system because it satisfies a user's information need.

In printed onomasiological dictionaries the words are not isolated, but usually arranged by shared semantic or associated features grouped under headwords [46,47]. The main disadvantage in this type of search is that a really specific idea of the concept is needed in order to search in the right place of the index or headwords. Currently, we have a lot of information accessible through different digital resources such as the internet. It is easy to search for almost any kind of topic in the most common search engines, i.e. Google, Bing, Yahoo, etc. Unfortunately the outcome of the search tends to be even more confusing or it simply shows other results that do not correspond to the concept.

For psychology, and especially psycholinguistics, the problem, formulated as lexical access, is also relevant. The most important modern contributions come from Bonin [12], Levelt [31], Aitchinson [37] and Jarema et al. [27]. The main discussion in the area is how to deal with the latencies and errors in the access, the tip-of-the-tongue phenomenon, and the distinction lexeme/lemma. In recent years, it has been clear the need of interaction between linguistics, psychology and language technologies in order to tackle some disease-related dysnomia.

The line of research related to cognition produced the inclusion of a shared task on lexical access in

[1]The final publication is available at IOS Press through http://dx.doi.org/10.3233/JIFS-179010
[*]Corresponding author. Helena Gómez-Adorno. Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), Universidad Nacional Autónoma de México, Mexico City, Mexico. E-mail: helena.gomez@iimas.unam.mx

language production in CogaLex Workshop at Coling 2014.

The present paper presents an algorithm that performs a lexical search over a knowledge graph in a similar way onomasiological dictionaries help to find a concept starting from a definition or a set of clue words. We developed a model based on a graph-based technique, the betweenness centrality, to perform the search of a given concept on a small corpus of word association norms for Mexican Spanish.

We developed an evaluation corpus composed of 56 concepts. For each concept we collected 5 definitions from human sources. We used the 280 definitions as queries in our searching model and compared the results with two information retrieval (IR) models (Boolean and Vector Space). Our model consistently achieved higher results than the baseline IR model for this case of search scenario.

The rest of the paper is organized as follows: in Section 2 we present some related work. Section 3 explains how word association norms are structured and the criteria for designing our graph. Section 4 describes the onomasiological search model. Section 5 presents the experimental settings and the results. The evaluation of the system is reported in Section 5.3. The paper ends with some conclusions and future work in Section 6.

2. Related Work

2.1. Onomasiological searching

There are some specialized texts that aim to help writers who need to go from a meaning or concept to a corresponding word. These resources are gathered according to their behaviour in the following three features: a) the type of information they contain, b) the structure of the wordbook, and c) the type of search undertaken. We distinguish four different groups:

Thesauri are an example of the most conceptual reference books. They have a systematic table of subjects. To find a target word, the user typically has to start from an approximation to the concept or from a clue word. This process is similar to what we do when searching for a concept in a search engine, so that we choose the words that we think are closer to it and can retrieve the desired result. There are two examples of thesauri that are very different, but share the same theoretical bases: the Roget's Thesaurus of English Words and Phrases [42] and WordNet [35].

Reverse dictionaries allow the users to search from a clue rather than from an index or conceptual tree. To find a target word in either dictionary, users think of a concept and a clue word referring to it, and then go to the main body of the dictionary, the "reverse dictionary". The macro-structure is alphabetical, which allows the users to go from the clue word to the concept without an index. Every clue word has a reduced list of related words following a brief definition for each concept. However, these resources have several difficulties. First, there could not be a suitable clue or word, and secondly, it can happen that such clue exists in the dictionary, but it is not linked to the target.

Synonymy and antonymy dictionaries. These have a key difference from thesauri, which is the fact that they operate with words rather than concepts, as thesauri do [2]. If the tool is oriented to synonyms, the user must think of words that have similar meaning to the target. If it is oriented to antonyms, the clues will be words with a meaning that is the contrary to the target. Usually, when a person uses a dictionary of this type, he or she knows the target, and looks for words that can be used in the same context, with the same or exactly the opposite meaning. Unfortunately, it seems they are not the most appropriate tools to find a target concept.

Pictorial dictionaries. They present concepts through pictures rather than words. The way these tools work is very simple: the user, that does not remember a word, is able to look for the image of the concept. Of course, no abstract ideas can be represented, and only several types of physical objects seem to be suitable for this type of visual representation.

The whole scenario of onomasiological searches changed with the universalization of 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. The users enter their own definition in natural language and the engine looks for the words that match the definition.

One of the first online dictionaries allowing this type of search was the one created for French by Dutoit & Nugues [16]. The resource takes into account the differences between a regular user and formal dictionaries when describing a term. To look for the smaller difference, it uses a database hierarchically organized, so that hypernyms and hyponyms are automatically identified. One of the disadvantages of this system is that synonymy is not considered.

Bilac et al. [9] 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. Such measure clearly decreases when the definition contains words that are not exactly the ones they have in their database, being this one of the main problems of this web application.

Both resources have not been tested to have an adequate evaluation, provided that only a small set of selected words has been used, or the definitions have been taken from dictionaries, which is a very different purpose than the one claimed by the authors. Nevertheless, those techniques imply a qualitative advance in the topic.

El-Kahlou & Oflazer [17] 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. The results are mostly positive: 66% of the times the term that is searched is among the 50 first candidates. When using the definitions of other dictionaries as input, the score reaches 92% among the 50 first. However, this implementation does not take into account the use of colloquialism; the number of candidate terms, 50, is very high, and it does not take the average position of the targets on the list.

For English, there exists an online onomaisological dictionary, OneLook Reverse Dictionary,[1] that retrieves acceptable results. It does not only allow queries in natural language, but it also deals with regular expressions.

One of the main works in Spanish is the one by Sierra [45]. DEBO is an onomasiological dictionary that works with user queries given in natural language and a search engine improved by Hernández [26], who also optimized the database structure.

For evaluation, definitions from non-system users were collected, and the average of the target was computed. Hernández's algorithm improved by 15% the initial Sierra's results. Compared with other works, like El-Kahlou & Oflazer [17], this search engine improved the outcome by 5%.

2.2. Lexical Access

Zock et al. [51] have defined the problem of lexical access as a problem of search, encouraging the development of new interdisciplinary approaches to the problem.

Ferret [20,21] and Zock et al. [50] suggested a matrix-based model to deal with topical detection and collocation links, i.e., syntactic and semantic contexts in which a single word appeared. Zock's proposal needs complex double processing matrices.

A very intuitive model to tackle the problem is the use of networks. A simplest solution to the need of having balanced syntagmatic-paradigmatic relations between words can be collocation networks [19]. The authors used the BNC corpus to build two graphs: G1 and G2. First, a so-called co-occurrence graph G1 in which words are linked if they co-occur in at least one sentence within a span of maximum three tokens. Then, a collocation graph G2 is extracted in which only those links of G1 whose end vertices co-occur more frequently than expected by chance are retained.

Widdows & Dorow [49] suggested the possibility of constraining the corpus with PoS annotations. The graph must be annotated according to the criteria that have been followed to build the network. Zock et al. [51] use the tags AKO (a kind of), ISA (subtype), TIORA (Typically involved object, relation or actor). Since the suggested network also involves syntagmatic links, more labels should be introduced to describe syntactic relationships.

Some computational resources have been built to help the users to find what they are looking for. An example can be the application designed by Lafourcade [29] to assist the writers when they experience the problem of the Tip-of-the-tongue problem.

Lexical access was the topic of a shared task in the workshop Cogalex at Coling 2014. In this gathering several new strategies were presented on how to approach the problem. Ghosh et al. [24] introduced a new two-stage model that has proven to be very efficient.

2.3. Free Word Associations

Free word associations (WA) are commonly collected by presenting a stimulus word (SW) to the participant and asking him to produce in a verbal or written form the first word that comes to his mind. The answer generated by the participant is called response word (RW).

Compilations of WA are called Word Association Norms. Many languages have this type of resources, which are time-consuming and need many volunteers. Among the available compilations, the best-known

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

in English are the *Edinburgh Associative Thesaurus*[2] (EAT) [28] and the collection of the University of South Florida [36][3].

In recent years, the web has become the natural way to get data to build such resources. *Jeux de Mots*[4] provides an example in French [30], whereas the *Small World of Words*[5] contained datasets in 14 languages at the time of writing. Such repositories have the problem of being collected without control over who is actually playing, the linguistic proficiency of the users, and their age, gender or level of studies.

For Spanish, there exist several corpora of word associations. Algarabel et al. [1] integrate 16,000 words, including statistical analyses of the results. Macizo et al. [32] build norms for 58 words in children, and Fernández et al. [18] work with 247 lexical items that correspond to Spanish [43].

The use of free word associations to compute relationality between words is not new. Borge-Holthoefer & Arenas [13], describe a model (RIM) to extract semantic similarity relations from free association information. The authors applied a network methodology to discover feature vectors on a free association network. The obtained vectors were compared with LSA-based vector representations and WAS (word association space) model.

In recent years, Bel-Enguix et al. [8] used techniques of graph analysis to calculate associations from large collections of texts. Additionally, Garimella et al. [23] 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. The method improved the performance of the generic methods to calculated associations that do not take into account the demography of the writer.

The only resource designed and compiled for Mexican Spanish is the *Corpus de Normas de Asociación de Palabras para el Español de México*[6] (NAP, from here) [3]. This work proposes the use of this corpus to be the basis of the design of a lexical search system that works from the clues or definitions to the concept, i.e., from the responses to the stimuli.

[2]http://www.eat.rl.ac.uk/
[3]http://web.usf.edu/FreeAssociation
[4]http://www.jeuxdemots.org/
[5]https://smallworldofwords.org/.
[6]http://www.labpsicolinguistica.psicol.unam.mx/Base/php/general.php

2.4. Related tasks in NLP

The work we are presenting is also related to other NLP tasks, like entity search and entity retrieval. Both tasks are more focused on web-based searches.

Entity search [5,6] aims at finding words, instead of documents, as a result of a query. The outcome is an entity or a list of entities. When asked for questions like 'Countries that border the Baltic Sea', the system is supposed to retrieve a list of entities. Balog et al. [6] suggest three different types: term-based, category-based and example-based.

Entity retrieval is based on the assumption that a user has a need of information that is well defined and can be expressed using a set of keywords that are submitting to an entity ranking system [44].

These tasks typically need external sources of information, i.e., Knowledge Bases, to locate the entities and retrieve similar ones. The main difference between these tasks and the lexical search we are introducing here is that the former works with entities while the latter with terms, avoiding named entities.

3. NAP Corpus and Graph

The NAP corpus consists of 234 stimulus words. There were 578 informants - 239 men and 339 women. Stimuli were divided into two lists of 117 words. All the participants were young university students whose mother tongue is Mexican Spanish, aged 18 to 28, and at least 11 years of education. The total number of words of the resource is 65,731, with 4,704 different words.

So far, the associations in NAP do not cover compound terms, except in very rare cases (ie. New_York), which are treated as an only word. An additional limitation of the resource could be the fact that, every *stimulus* in the NAP is a concrete noun. This unbalance the category of the lexical items of the resource, because of the tendency to retrieve a noun to another noun.

The graph representing the NAP has been elaborated with lemmatized 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.
- $\phi : E \rightarrow \mathbb{R}$, is a function over the weight of the edges.

The graph is undirected so that every stimulus is connected to every associated word without any precedence order.

For the weight of the edges there are three different functions:

Time (T) Measures how many seconds it takes the participant to retrieve a response for every *stimulus*.

Frequency (F) Counts the number of occurrences of every associated to its *stimulus* in the whole corpus. For the system to work in the shortest paths, we need to calculate the IF, inverse frequency, that is defined in the following way: being F the frequency of a given associated word, and ΣF the sum of the frequencies of the words connected to the same *stimulus*, $IF = \Sigma F - F$

Association Strength (AS) Establishes a relation between the frequency (F) and the number of associations for every stimulus. It is calculated as follows: being F the frequency of a given associated word, and ΣF the sum of the frequencies of the words connected to the same *stimulus* (the total number of responses), the association strength (AS) of the word W to such *stimulus* is given by the formula:

$$AS_W = \frac{F * 100}{\Sigma F}$$

For our experiments, we need to calculate the inverse association strength, IAS, in order to prepare the system to work with graph-based algorithms:

$$IAS_W = 100 - \frac{F * 100}{\Sigma F}$$

Figure 3 depicts a subgraph of the NAP corpus, containing only four stimuli with their corresponding associates. It can be observed that there are some associate words that are common to different stimuli, even for this small subgraph. We can also find relationships between two stimuli; for example, *flor* (flower) and *abeja* (bee).

4. Lexical Search Model (LSM)

Given a definition, we search in the graph the word that better matches with it. For this purpose, we considered several graph centrality measures such as degree centrality, closeness centrality, betweenness centrality, load centrality, page Rank, Katz centrality, percolation centrality. Centrality measures identify the most important nodes in a graph; for example, the degree centrality assumes that important nodes have many connections. The degree centrality is not suitable for our purposes because we need to find the most important nodes for a specific subset of nodes (the nodes that represent the words in a definition).

We choose a variation of the *betweenness centrality* (BT) algorithm [22] which instead of computing BT of all pairs of nodes in a graph, calculates the centrality based on a sample (subset) of nodes [14]. The traditional betweenness algorithm assumes that important nodes connect other nodes. For a given node (v) in a graph (G), the BT is calculated as the relation between the number of shortest paths between nodes i and j that pass through node v and the number of shortest paths between nodes i and j. It is formally described as follows:

$$C_{btw}(v) = \sum_{i,j \in V} \frac{\sigma_{i,j}(v)}{\sigma_{i,j}} \quad (1)$$

where:
$V =$ is the set of nodes, $\sigma_{i,j}$ is the number of shortest paths between i and j, and $\sigma_{i,j}(v)$ is the number of those paths that pass through some node v that is not i or j.

In a non-weighted graph, the algorithm looks for the shortest path. In a weighted graph, like the one we have built, it finds the path that minimizes the sum of the weights of the edges.

BT algorithm was introduced having as a basis the general idea that when a particular person in a group is strategically located on the shortest communication path connecting pairs of others, that person is in a central position [7]. Remarking the importance of the shortest paths, we adapted the information available in NAP, letting the most important nodes and their relations were represented as minimal values as explained before. This is why we have adopted the weighting function based on inverse frequency and inverse association strength.

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. Our hypothesis is that if we use a subset, the nodes of the NAP graph (NG) that represent the words of a definition as initial and final nodes in the BT algorithm, and calculate the

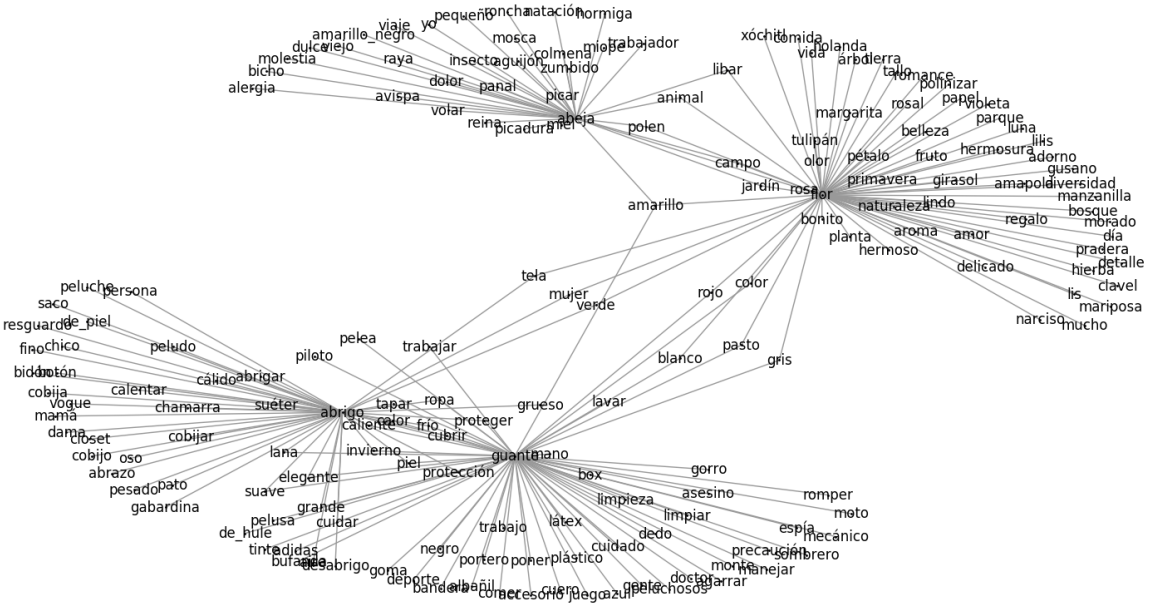

Fig. 1. Subgraph with the stimuli *flor* (flower), *abeja* (bee), *abrigo* (coat), and *guante* (glove) with their corresponding associates.

centrality of the other nodes in NG taking these nodes as pairs, then the more central nodes will be the concept of such definition. This approximation is formally described as follows:

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

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.

Therefore, we define a subgraph composed by the words (nodes) of the definition. This subgraph is used as both initial and final nodes, for calculating 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.

Algorithm 1 presents the overall schema of our model. First, we perform some pre-processing steps. All the *stimuli* and the *responses* are lemmatized, leaving each word as the most representative of the flexed forms. The same pre-processing is applied to the definitions to be searched by the model. This process provides us with more matches in the case when the definition contain *table*, *tables*, etc. because it will be transformed into its lemma, *table*. For this purpose, we use the lemmatization process available in *FreeLing* [38] for the Spanish language.

Algorithm 1: Lexical search

```
Data: NAP Corpus, definitions to search
Result: list of ranked concepts
pre-process(NAP Corpus);
pre-process(definitions to search);
GraphNAP = build-graph(NAP Corpus);
for each definition do
    definition = remove-StopWords(definition);
    definition = filter-WordsInNAP(definition);
    build_subgraph(definition);
    ranking_nodes = BT(GraphNAP,subgraph);
    ascending_order(ranking_nodes);
```

Later, we built the GraphNAP with the Python package Networkx [25]. Then, for each definition to be searched we removed all the functional words using the Spanish *stop words* list available in the *NLTK* package [10]. Next, with the list of words with lexical meaning, we kept only the ones that belong to the vocabulary in NAP. With this we built a subgraph to be the input in the Betweenness Centrality algorithm. Fi-

nally, the nodes were sorted out according to the highest centrality measure, which corresponds to the words that are closer to the ones of the definition.

5. Experiments and Results

5.1. Evaluation Corpus

For the experiments, a small corpus containing 5 definitions for 56 concepts corresponding to *stimuli* of the NAP was collected, with a total number of 280 definitions. The corpus was gathered with the collaboration of students who gave their own description of the word. Note that this task is almost equivalent to the one of providing clue words. In fact, the task was not restricted, and some of the contributions of the students were lists of words. Moreover, when dealing with definitions to implement the sentence.

All of the words defined by the participants are nouns grouped into six semantic domains: animals, transportation, body parts, household appliances, clothes, and food. We do not use definitions taken from dictionaries, because they tend to be more precise and understandable, but they are not the type of clues a human looking for a word tends to give. However, the using canonical definitions from regular dictionaries is also a possible task.

Table 1, presents an example of 5 definitions of the same concept given by different students.

5.2. Results with the Lexical Search Model

The experiments were performed taking into account weighted graphs with the 3 previously mentioned functions: Time (T), Inverse Frequency (IF) and Inverse Association Strength (IAS).

For the evaluation of the inference process, we used the technique of precision at k ($p@k$) [33], 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 2. It is clear that when the model searches over the graphs weighted with IF and IAS the results are higher than when searching on the graph weighted with T. Furthermore, the search on the IAS weighted graph achieves the higher precision in all the evaluation measures. This was the expected outcome. According to psycholinguistics, reaction time does not necessarily indicate relatedness between *stimulus* and *response*, although the intuition says that there could be some connection. However, so far, no study has been able to establish a systematic relation.

Psychologists agree that Association Strength is the measure that implies a cognitive relationship between two terms, and this idea is reflected in our results. Frequency is closely related to AS, but it lacks the generalization of the latter function.

5.3. Results

In order to evaluate the relevance of our method we have performed experiments with other well-known IR methods.

First, we compared the performance of our method with the results of a reverse dictionary. To do that, we have used the OneLook Thesaurus, that allows you to describe a concept, and returns a list of words and phrases related to that concept. Although there is a Spanish version of the resource[7], it is is clearly outperformed by the English one, so that the use of the Spanish one has been dismissed. To use OneLook, we have translated each of the definitions of the corpus literally as well as the target concepts, using Google Translator. The definitions have been manually checked using the OneLook web application[8].

Secondly, we compared our results with those obtained by a baseline IR model using a Boolean search. The experiments were performed on two different reference corpora: a) *Diccionario de la Real Academia Española* [39], and b) *Corpus de Normas de Asociación de Palabras para el Español de México* [3]. The Boolean Retrieval Engine[9] takes each definition of the corpus and generates a query joining the words with logical connectors AND to obtain the most relevant documents containing all the items in the search. For this experiment, the engine first looked for a file containing every word of the definition. In case it did not find it, the Quorum function [15] is applied, i.e., another search was carried out with all the words except one and every possible combination. The process continued until finding a combination retrieving a document.

As the LSM shown in algorithm 1, the Boolean search requires several corpora and definitions preprocessing tasks. Moreover, it is important to mention that

[7] http://www.rimar.io/

[8] https://www.onelook.com/thesaurus/

[9] https://github.com/jin-zhe/boolean-retrieval-engine

Table 1
Definitions of 'león' (*lion*) and 'queso' (*cheese*) given for the students. Google translations are provided in order to keep the literalness of the expression and the precise words.

Concept	**León**	***Lion***
Definition 1	Ruge y vive en la selva	*Roars and lives in the jungle*
Definition 2	Rey	*King*
Definition 3	Animal carnívoro, de cuatro patas, grande melena, pelaje amarillo. Es el rey de la selva	*Carnivorous animal, with four legs, big mane, yellow fur. He is the king of the jungle*
Definition 4	El animal del escudo de Gryffindor	*The animal of the Gryffindor shield*
Definition 5	Animal conocido como el rey de la selva	*Animal known as the king of the jungle*
Concept	**Queso**	***Cheese***
Definition 1	Alimento elaborado con leche. Existen diferentes tipos: manchego, cotija, panela entre otros	*Food made with milk. There are different types: manchego, cotija, panela among others*
Definition 2	El producto que se saca de la leche de la vaca	*The product that is taken out of the milk of the cow*
Definition 3	Amarillo y con agujeros	*Yellow and with holes*
Definition 4	Derivado lácteo que ponen en trampas para ratones	*Milk derivative that they put in traps for mice*
Definition 5	Como la crema pero sólido	*Like the cream but solid*

Table 2
Results in terms of precision of our model with three weighting functions.

Weighting function	p@1	p@2	p@5
Time (T)	0.3623	0.5507	0.6522
Inverse Frequency (IF)	0.6165	0.7419	0.7742
Inverse Association Strength (IAS)	0.6558	0.8043	**0.8297**

a stop condition in the loop is a query containing a minimum of two words in the definition because it will retrieve too many concepts that will match any word.

We have performed additional experiments with one of the most successful text-retrieval algorithms, Okapi BM25, based on probabilistic models and developed in the seventies by Stephen E. Robertson and Karen Spärck Jones [41]. The algorithm, implemented following Robertson & Zaragoza [40] is based on the bag-of-words method. Given a query, it ranks a list of documents according to their relevance for such query. We have applied it, considering as a document every definition and every set of responses to a stimulus.

Finally, we have performed an experiment using pretrained vectors. We took as basis the first stage of the work Computing Associative Responses (CAS from here) [24]. This work involves generating a ranked list of responses to a set of stimulus words. In our case the stimuli were the words in a definition and the inferred response is the concept we were trying to find in the onomasiological dictionary.

The vectors of our implementation of CAS were in Spanish using *FastText* [11][10]. *FastText* is a method that has been designed to improve the performance of *word2vec*, based on the skip-gram model, with the difference that in this method every word is represented as a bag of character n-grams. *Mikolov et al.* [34] claim that the models trained with *FastText* exhibit the best degree of accuracy compared to other systems, becoming the new state-of-the-art in distributed representations of words.

For this experiment, we used the vector representation of each word in the definition and calculated the similarity of a target concept by measuring the cosine distance between the words in the definition and the ones in the FastText resource. It is formally described as follows:

Let $S = \{x_i, \ldots, x_j\}$ definition words

$$sim(r, S) = \frac{1}{|S|} \times \sum_{i=1}^{|S|} \frac{x_i \cdot r}{|x_i| \cdot |r|} \tag{3}$$

It is important to mention that the definitions were tested without functional words and in the lemmatization form. With this experiment we only retrieved 88

[10] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

target concepts of a total of 280 definitions in an average position of 166.

Table 3
Comparative precision results

Method	P@1	P@3	P@5
OneLook	0.1151	0.2050	0.2554
LSM (IAS)	**0.6558**	**0.8043**	**0.8297**
Boolean NAP IR	0.3776	0.3776	0.3776
Boolean RAE IR	0.0359	0.0359	0.0359
BM25	.330	.490	.570
CAS	0	.007	.014

The results achieved using the four mentioned baselines: OneLook and Boolean IR, BM25 and the two-stage system with pretrained versions, are reported in Table 3, where they are compared with the best result obtained by the Lexical Search Model.

In order to perform a Boolean search over the NAP corpus, each *stimuli* was considered as a document and its associate words were the content of the documents. For searching over the RAE, each definition in the dictionary was considered as a document. The Boolean search showed better performance than the Onelook reverse dictionary when the search is performed over the corpus NAP. However, when the search is performed over the RAE dictionary, the results are very low. We believe that this behavior is due to the short nature and the technical vocabulary of a dictionary definition.

Regarding BM25, the algorithm shows a quite good performance, although it is far form the LSM. On the contrary CAS, implemented with FastText vectors, has the worst results. Although word embeddings are very useful in other tasks related to semantics, the model fails in lexical search. This can be caused, among other reasons, by the big diversity of vocabulary and the presence of all the inflected forms of a word in this kind of resources. We didn't performed the second stage of the experiment with this method (CAS) because it only performs an arrangement of the results obtained in the first stage in order to have the target concepts in the first positions. By this procedure, in the very best scenario, the improvement of the performance would reach 30% and would not be significant enough to outperform LSM.

Table 4 shows what each system retrieves with two different definitions of the words *Queso* [Cheese] and *Abeja* [Bee]. The outcomes indicate that the Onelook reverse dictionary is not adequate to solve this problem, retrieving the correct concept in the first 5 results only 25% of the times. Moreover, checking carefully the outcomes of every one of the systems, a limitation of LSM can be observed. Our results are the best when using the $p@k$ evaluation model; however, every method that works over NAP has some 'non-consistent' results. For instance, when comparing the tests of BM25 with the word 'queso' in NAP and RAE, the model retrieves several words that are difficult to explain with NAP: 'pelo' [hair], 'colores' [colours], while RAE only shows a non-coherent result, 'champús' [shampoos]. Although the system is fast, efficient and demonstrates a high performance, the structure of resource we built favors the fact that two words that are nor really related by association have a short path between them because they share a connected word. This is is expected to be a feature of LSM, that can be minimized by performing some kind of lexical filter in the future.

6. Conclusions and Future Work

This paper introduces a model for onomasiological searches that has some novelties; among them the simplicity, the use of graph-based techniques and the small corpus the method is based on.

We have shown how descriptions of concepts that are made by common people with non-scientific specifications can retrieve accurate results using our method. This is possible thanks to the nature of the corpus. For the word association norms group words that are closely related in a cognitive way, and taking advantage of the weighted edges the original resource provides.

In the near future, we plan to design an online application for the users to be able to 'play' with the resource and test the results with different definitions.

The success of the system with non-scientific input can drive new lines of applied research, and the implementation of different assistant writing systems especially oriented to people with a range of aphasias, like dysnomia and Alzheimer's disease.

Our algorithm has shown a very good performance compared with other baseline systems. Its main problem is the restricted number of words that can participate in the search. The model could retrieve better results with larger Word Association Norms. For the future, we plan to extend the model with other corpora, like the Edinburgh Associations Thesaurus [28], a database with 8,000 *stimuli*. Moreover, and follow-

Table 4
Results for Cheese and Bee.

	Queso		**Abeja**	
Definition Method	Alimento elaborado con leche. Existen diferentes tipos: manchego, cotija, panela entre otros.	El producto que se saca de la leche de la vaca	Insecto volador rayado que produce miel	Insecto volador amarillo y negro
LSM IF	1. queso 2. torta 3. colores 4. pelo 5. dulce	1. queso 2. torta 3. calabaza 4.colores 5.pelo	1. cuchara 2. circo 3. luz 4. grande 5. feo	1. cuchara 2. circo 3. palo 4. martillo 5. manzana
LSM T	1. queso 2. colores 3. vaca 4. comer 5. blanco	1. queso 2. calabaza 3. colores 4. alimento 5. vaca	1. abeja 2. mariposa 3. ardilla 4. cacahuate 5.conocimiento	1. abeja 2. martillo 3. ardilla 4. agua 5. cacahuate
LSM IAS	1. queso 2. pelo 3. ratón 4. pastel 5. colores	1. queso 2. calabaza 3. leche 4. pelo 5. mercado	1. abeja 2. mariposa 3. conocimiento 4. minifalda 5. 2.0	1. abeja 2. mariposa 3. arańa 4. tractor 5. plastilina
BM25 NAP	1. queso 2. torta 3. pelo 4. colores 5. mamila	1. a leche de la vaca 2. tienda 3. calabaza 4. queso 5. pala	1. palomita 2. crayola 3. circo 4. cebra 5. pluma	1. mariposa 2. helicoptero 3. ardilla 4. palomita 5. circo
BM25 RAE	1. queso 2. chéster 3. gorgonzola 4. brandy 5. champús	1. mantequilla 2. cuajada 3. lacteado, da 4. lácteo, a 5. natilla	1. fatula 2. lapizar 3. eraje 4. meloja 5. bresca	1. mapanare 2. fatula 3. doral 4. agüío 5. cacuy
OneLook	1. atom 2. expenses 3. aphid 4. rounds 5. meal	1. strip 2. ghee 3. buttermilk 4. stroking 5. mess	1. bee 2. manna 3. moth 4. virgin 5. dor	1. wasp 2. gnat 3. bee 4. whippoorwill 5. slug
CAS	1. piloncillo 2. nixtamalizado 3. nixtamalización 4. saborización 5. quesillo	1. leche 2. producto 3. pasteurizar 4. trío 5. sacar	1. rayar 2. insecto 3. rayadura 4. miel 5. espesarla	1 amarillo 2. negro 3. amarillo/naranja 4. amarillo/blanco 5. anaranjado
Boolean NAP IR	Query: alimento AND leche AND manchego AND panela	Query: leche AND vaca	Query: insecto AND rayar AND miel	No combination of terms retrieved abeja

ing the work of [48] we plan to design a method able to learn the word associations from NAP and extend them to larger corpora.

Moreover, in order to build a fast and efficient method for information retrieval focused in definitions, the algorithm could still be optimized with the use of

other search algorithms besides betweenness centrality.

Aknowledgements

This work has been supported by PAPIIT projects IA400117 and IN403016, and CONACYT Fronteras de la ciencia 002225.

References

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

[2] M Alvar Esquerra. Diccionarios ideológicos. *Libros*, 24:14–18, 1984.

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

[4] Kurt. Baldinger. *Teoría semántica: hacia una semántica moderna*, volume 12. Alcalá., 1970.

[5] Krisztian Balog, Marc Bron, and Maarten de Rijke. Category-based query modeling for entity search. In *Proceedings of the 32nd European Conference on Information Retrieval (ECIR)*, pages 319–331. Springer, 2010.

[6] Krisztian Balog, Marc Bron, and Maarten de Rijke. Query modeling for entity search based on terms, categories, and examples. *ACM Transactions on Information Systems*, 29(4, Article 22), 2011.

[7] Alex Bavelas. A mathematical model for group structure. *Social networks: critical concepts in sociology, New York: Routledge*, 1:161–88, 2002.

[8] Gemma Bel-Enguix, Reinahrd Rapp, and Michael Zock. 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., 2014.

[9] S Bilac, W Watanabe, T Hashimoto, T Tokunaga, and H Tanaka. 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., 2004.

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

[11] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Computing Research Repository*, arXiv:1607.04606, 2016.

[12] P. Bonin. *Mental lexicon: some words to talk about words*. Nova Science Publishers., 2004.

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

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

[15] Cyril Cleverdon. Optimizing convenient online access to bibliographic databases. *Information services and Use*, 4:37–47, 1984.

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

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

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

[19] R. Ferrer i Cancho and R.V. Solé. The small-world of human language. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 268(1482):2261–2265, 2001.

[20] Olivier Ferret. Using collocations for topic segmentation and link detection. In *COLING 2002*, pages 260–266, 2002.

[21] Olivier Ferret. Building a network of topical relations from a corpus. In *LREC 2006*, 2006.

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

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

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

[25] A Hagberg, D Schult, and P Swart. Networkx: Python software for the analysis of networks. *Mathematical Modeling and Analysis, Los Alamos National Laboratory*, 2005.

[26] L. Hernández. *Creación semi-automática de la base de datos y mejora del motor de búsqueda de un diccionario onomasiológico*. Universidad Nacional Autónoma de México., 2012.

[27] G. Jarema, G. Libben, and E. Kehayia. The mental lexicon. *Brain and Language*, 81, 2002.

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

[29] M. Lafourcade and A. Joubert. *TOTAKI: A Help for Lexical Access on the TOT Problem*. Text, Speech and Language Technology XI. 2015.

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

[31] W.J.M. Levelt. Spoken word production: A theory of lexical access. *PNAS*, 98(23):13464âĂŞ13471, 2005.

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

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

[34] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh*

International Conference on Language Resources and Evaluation, LREC'18, 2018.

[35] G. A. et al. Miller. Introduction to wordnet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990.

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

[37] Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. *Words in the mind: an introduction to the mental lexicon*. Blackwell., 2003.

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

[39] RAE. *Diccionario de la Lengua Española.* Real Academia Española de la Lengua, 2013.

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

[41] S.E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3), 1976.

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

[43] Carmen Sanfeliu and Ángel Fernández. 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., 1996.

[44] Sheikh Muhammad Sarwar, John Foley, and James Allan. Term relevance feedback for contextual named entity retrieval. *arXiv:1801.02687v1 [cs.IR] 8 Jan 2018*, 2018.

[45] G. Sierra. Design of an onomasiological search sistem: A concept-oriented tool for terminology. *Terminology*, 6(1):1–34, 2000.

[46] Gerardo Sierra. The onomasiological dictionary: a gap in lexicography. In *Proceedings of the Ninth Euralex International Congress*, pages 223–235, 2000.

[47] Gerardo Sierra and John. McNaught. Natural language system for terminological information retrieval. In Alexander. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 541–552., Berlin, Heidelberg, 2003. Springer.

[48] Anna Sinopalnikova and Pavel Smrz. Word association thesaurus as a resource for extending semantic networks. In *Communications in Computing.*, pages 267–273., 2004.

[49] D. Widdows and B. Dorow. A graph model for unsupervised lexical acquisition. In *19th International Conference on Computational Linguistics*, 2002.

[50] Michael Zock, Olivier Ferret, and Didier Schwab. Deliberate word access: an intuition, a roadmap and some preliminary empirical results. *International Journal of Speech Technology*, 13(4):201–218, 2010.

[51] Michael Zock, Didier Schwab, and Nirina Rakotonanahary. Lexical access, a search-problem. In *Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010)*, pages 75–84, 2010.