Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.11, No.3, 2020

Building WordNet for Afaan Oromoo

Biru Abera Bacha
College of Electrical Engineering and Computing, Ambo University

Abstract
WordNet is a lexical database which has many relations to disambiguate the sense of words for natural languages. From the WordNet relations synonyms and hyponym has major role for natural language processing and artificial intelligence applications. In this paper, word embedding (Word2Vec) and lexico-syntactic pattern (LSP) are developed to extract automatically synonyms and hyponyms respectively. For this study, the word embedding is evaluated on two specialized domain algorithms such as a continuous bag of words and Skip Gram algorithms and show superior results. Applying word embedding (Word2Vec) algorithms for Afaan Oromo texts has been registered 80.09% and 85.04% for the continuous bag of words and Skip Gram respectively. According to the result achieved in this study, the skip-gram algorithm does a better job for frequent pairs of words than a continuous bag of words. But, a continuous bag of words algorithm is faster while skip-gram is slower. A lexical syntactic pattern with the combination of Word2Vec and without Word2Vec is also evaluated using information retrieval evaluation metrics such as precision, recall and F-measure to extract hyponym relation from Afaan Oromoo texts. The precision, recall and F-measure have been registered by lexical syntactic patterns without the combination of Word2Vec is 66.73%, 72%, and 69.26% respectively and with the combination of Word2Vec 81.14%, 80.8%, and 81.1% have been registered for precision, recall and F-measure respectively. There are factors that could affect the accuracy of results: 1) the style of writer of Afaan Oromoo i.e. they write a noun phrase with many adjective to express the noun for the reader; and, 2) it is possible that some instances of the LSP are missed due to misspellings and other typographical errors.
Keywords: Afaan Oromoo WordNet, Word embedding, Lexico syntactic patterns, Extraction of WordNet relations.
DOI: 10.7176/CEIS/11-3-01
Publication date: May 31st 2020

1. Introduction

1.1. Background

WordNet is a lexical database for natural languages (Wales & Sanger, 2017). WordNet is also an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory (Miller et al., August 1993). It groups words of specific language into sets of synonyms called ***synsets***, provides short, general definitions, and records the various semantic relations between these synonym sets. It has two objectives namely (1) to combine dictionary with a thesaurus to make more automatically usable, and (2) to help automatic text analysis and AI applications. WordNet, a manually constructed electronic lexical database for English, was conceived in 1986 at Princeton University, where it continues to be developed (Fellbaum, 2006). Experiments by researchers in artificial intelligence (Collins & Quillian, 1968) probing human semantic memory inspired the psycholinguist George a. Miller to test the underlying theories on a large scale (Miller et al., August 1993).

WordNet has been used as a complete semantic lexicon in a module for full-text message retrieval in a communication aid, in which queries are expanded through keyword design. WordNet has started to be used as a language knowledge tool to symbolize and understand the meaning of, and offer the user with well-organized and integrated access to, information; integration, indeed, has become an increasingly necessary feature with the development of multiple database access systems and one in which WordNet's identification and interpretation of semantic equivalents is extraordinarily useful (Morato et al., 2003).

A collection of words exist in the WordNet has been used by major search engines, IR research projects (Fishkin, 2005), many natural language processing (NLP) application and also usable for many Artificial intelligence (AI) fields for many years. But, in many cases, an application doesn't understand the meaning of terms used by them. WordNet is powerful to get information about the following for a given word or given phrase: **synonyms** - words that have the similar meaning (same = similar), **hypernyms** - the generic term used to designate a class of specifics (i.e. Soil is a kind of land), **hyponyms** - a member of a class of terms (i.e. Clay is a kind of soil), **homonyms** - name of a whole of which other words are a part (i.e. Nutrients are a part of the soil), **meronyms** - parts of the holonym (i.e. Soil is part of the ground) (Wales & Sanger, 2017).

Generally, WordNet is the basic and relevant component for the development of most of NLP related applications. But, Afaan Oromoo doesn't have a lexical database called WordNet so far. Thus, the development of such a system for Afaan Oromoo is essential for easing the development of Afaan Oromoo related NLP applications and that is what this study is intended for. As far as the fact of the researcher is concerned no same study was conducted so far for the language. But in this study, approaches used to extract the WordNet relations

such as synonyms and hyponyms from Afaan Oromoo texts are developed to solve the problem of Afaan Oromoo IR.

1.2. Afaan Oromoo language

The Afaan Oromoo language is an Afro-Asiatic macro language which is primarily composed of four distinct languages: Southern Oromo, which includes the Gabra and Eastern Oromo, Orma, Sakuye varieties, which includes the Orma, Munyo, Waata/Sanye varieties, and West Central Oromo. Like with the varieties of Arabic, Oromo is a dialect continuum, so language varieties spoken in neighboring regions differ only slightly, but the differences accumulate over distance so widely-separated varieties are not mutually intelligible (Eggi, 2012)

The Afaan Oromoo language is also known as Oromo language. It is a Cushitic language spoken by above 50 million people in Somalia, Kenya, Ethiopia, and Egypt and is the third-largest language in Africa. There are additional Oromo speakers out of the Ethiopian country than the resident population in Ethiopia. In the United States, Australia, Canada, and different Europe cities people are speaking and communities are teaching their kids and foreigners those interested in communications in Afaan Oromoo also taking the Oromo class. In Oromia, it has a high rank and it is an official language. It has its own writing and it can be written with Latin script. The verbalized tradition is very rich and nowadays there are sufficient literary works written in Afaan Oromoo; modern arts like music and folk arts. Oromo people speak Afaan Oromoo, as well as Amharic, Tigrinya, Guragegna and Omotic languages. They are mainly Muslim and Christian, while only 3% still follow the customary religion based on the worshipping of the God Waaqa. Oromo are mainly farmers and cattle herders. They have distinguished themselves throughout history for their strong military organization (Erena, 2017).

2. Materials and methods

Among the different WordNet, building approaches proposed by different researchers this study have adopted merged and the distributional semantics approaches. These approaches are automatic approach. To achieve this, the study starts by creating the word embedding model to generate WordNet terms by collecting the terms and their relations from the Afaan Oromoo documents.

The system architecture is depicted in Figure 1. The architecture is modularized into sub-components including preprocessing tasks like tokenization, stop words and number removal, applying a lexical syntactic pattern, co-occurrence manipulation, and similarity. In addition to this, the system architecture has text operation (i.e., lexical analysis, and stop word elimination), extracting synonyms relations by word embedding, and extracting hyponym relations by lexical syntactic patterns with and without combination of word embedding components. The system accepts document collections and applies text operation on the documents to generate the nearest neighbors of the term. The system also accepts document collections and applies text operation on the documents to retrieve the hyponym relations exist in texts which can match the created patterns These nearest neighbors have automatically generated WordNet terms for that particular term.

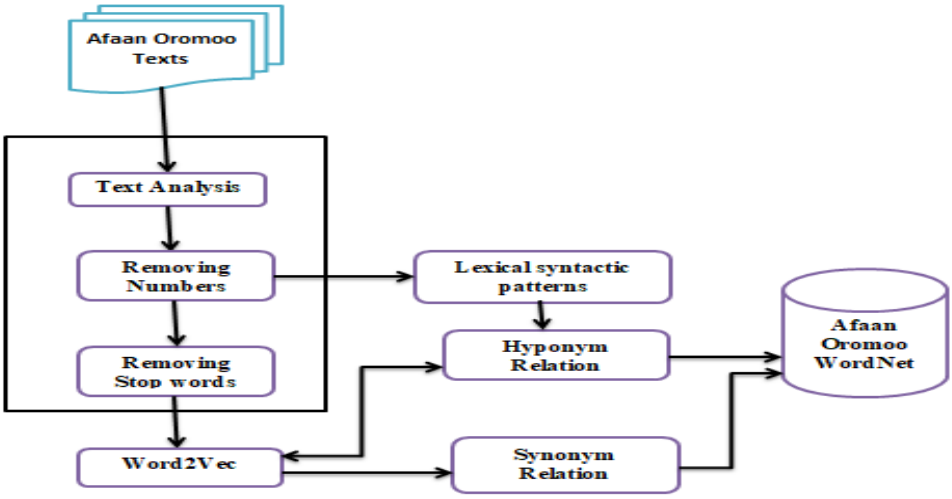

Figure 1 the system architecture of the automatic Afaan Oromoo WordNet generation system.

2.1. Development Tools and Techniques

Finding a large size and standard corpus for the Afaan Oromo language is one of the challenges faced in this study. The state of the art in the area of text processing indicates that there is no developed standard corpus for the Afaan Oromo language and also as a result of time factor it takes a lot of time to prepare a large size of the corpus with

many documents. Thus, in this study, small size corpus (2746KB) which contains **680** pages with **387224** tokens were collected from different news and media, such as VOA, Bariisaa, Holy Bible and online educational resources that are written in Afaan Oromo and available on the web are used and to implement the lexical syntactic patterns for Afaan Oromoo texts for the purpose of extracting the hyponym relation of Afaan Oromoo WordNet java program language is used and the python programming language is used to implement word embedding as shown in the figure 2.

```
Python 3.7.4 (tags/v3.7.4:e09359112e, Jul  8 2019, 19:29:22) [MSC v.1916 32 bit
(Intel)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
==== RESTART: C:\Users\Abdii-isaa\Desktop\data data\nlp\preprocessing.py ====
 Jecha duraa  galchaa!
gadaa
 Jecha itti aanu  galchaa!
gadaa
Cosine similarity between gadaa and gadaa by CBOW :  1.0
Cosine similarity between gadaa and gadaa by skip Gram :  1.0
```

Figure 2 Python program to calculate the similarity of two Afaan Oromoo words

2.2. Word Embedding (Word2Vec)

Word embedding is the combined name for a number of language modeling and feature learning techniques in NLP where words or phrases from the vocabulary are mapped to vectors of real numbers called Word2Vec (Sahlgren, 2015). These models are thin, two-layer neural networks that are trained to rebuild linguistic contexts of words. Word2Vec takes large texts as its input and output a vector space, typically of several dimensions, with each distinctive word in the corpus being assigned a corresponding vector in the space. Word vectors are located in the vector space such that words that share mutual contexts in the corpus are located in close nearness to one another in the space (Mikolov et al., 2013).

Word2Vec can use either of two model constructions to create a distributed representation of words: continuous skip-gram or continuous bag-of-words (CBOW). In the CBOW algorithm, the model forecasts the current word from a window of nearby context words. The sequence of context words does not affect prediction (bag-of-words assumption). In the skip-gram algorithm, the model uses the current word to forecast the surrounding window of context words. The skip-gram architecture weighs nearby context words more heavily than more distant context words (Mikolov et al., 2013) (Mikolov et al., 2013). According to the authors' note (Archive, 2013) CBOW is faster while skip-gram is slower but does a better job for infrequent words.

```
Begin
    Input: target words
    Load Afaan Oromo texts
    Tokenizing the text contents
    Removing the stop words and numbers
    Normalizing the text contents
    Calculating the similarity between the words exist in the texts
    Output: Word2Vec diagram of the words exist in texts
    If the target word is in texts
        Calculating the most similar to the target word
      Output: list of the most similar word of the target word
End
```

Algorithm 1 Word Embedding

2.3. Lexical syntactic patterns (LSP)

Lexicon-syntactic patterns (LSP, cf. LSPs at ontologydesignpatterns.org) are generalized linguistic structures or schemas that indicate semantic relationships between words and can be useful for the identification of official concepts and conceptual relations in natural language text. It is also a string of words paired with syntactic structures; they depend only on the syntactic categories of the component words, with no reference to their meaning. Lexicon semantic patterns are strings of words paired with semantic categories. The lexicon-ontology must allow representing such patterns though not necessarily as lexical entries and also allow representing the lexicon-syntactic structure of the patterns as well as the semantic relation it expresses.

For instance, $NP1 be (the same as/synonym of)/know as/call/(refer to as) $NP2. This pattern might be said to explain equivalence related between an OWL class C1 considered with NP1 and an OWL class C2 labeled with NP2.

Lexico-syntactic patterns are appropriate for automatic ontology building since they model semantic relations. These extract exactly the kind of relationship between their parts that makes them easily changeable into an ontology structure. The lexico-syntactic pattern in (Hearst, 1992) (Klaussner & Zhekova, 2011) corresponds to the classic hyponymy relation:

(1) If (NP0 such as NP1, NP2..., (and | or) NPn)
for all NPi, $1 \leq i \leq n$, hyponym(NPi,NP0)

This study creates seven (**7**) lexico-syntactic and semantic patterns automatically from Afaan Oromoo texts for extracting conceptual knowledge from Afaan Oromoo texts. The created patterns are general and domain/ application-independent and work at the sentence level. They are used to get taxonomic and non-taxonomic relations and axioms from sentences and phrases. Among the created patterns and templates, semantic patterns are language independent and although linguistic (lexico-syntactic) patterns are created for Afaan Oromoo language. To identify the lexical syntactic patterns from the documents, this study used the following algorithm.

Begin
1. Select representative seed words (possibly covering a wide range of topics). To archive this 5-word pairs from all physical objects, from abstract objects, from location objects, from entity objects, from group objects, from activity objects, from psychological feature and event objects are selected by following WordNet Ontological taxonomy structure[1] relations
2. Extract sentences containing the seed word pairs from step 1
3. Extract sequence of words linking the word pairs
4. Replace the word pairs with a variable and consider the resulting sequence as a pattern
5. Clean the patterns manually to remove arbitrary words…

End

Algorithm 2 to create the Lexico syntactic patterns from the corpus

Since only a subset of the possible instance of the hyponym relation will appear in a particular form. The study need to make use of as many patterns as possible. Below is a list of created lexical syntactic patterns that indicate the hyponym relations in Afaan Oromoo texts followed by illustrative sentence fragments and the predicates that can be derived from the texts and NP is stands for noun phrase. NP can be arranged in texts with and without adjectives.

1. NP NP {keessaayyuu|keessattuu|addatti} NP (, NP)* (fi|yookin|akkasumas) NP
… Kuduraaf muduraan keessaayyuu maangoon, timaatimni, abukaadoo fi paappayyaan ...
2. NP NP {wantoota akka } NP (, NP)*, (fi|yookin|akkasumas) NP
… Meeshaalen manaa wantoota akka eelee, barcumaa, siree, wullee, hubboo, akkasumas saayinaa …
3. NP NP {kan akka |kanneen akka |warra akka } NP (, NP)*(,)? (fi|yookin|akkasumas) NP
--- magaalota (kanneen akka)|(kan akka)|(warra akka) adaamaa, finfinnee, maqalee, yookin ambo ---
4. NP {warri akka } NP (, NP)* (fi|yookin|akkasumas) NP
--- dhukkuboonni warri akka qufaa, eedisii, fanxoo, copxoo akkasumas koleeraa ---
5. NP{, NP}*(,)? (yookin|fi|akkasumas) NP NP?
Ancootee, cumboo, waaddii, yookin nyaatni ----
6. NP{, NP}*(,)? (yookin|fi|akkasumas) NP (dabalatee) NP NP?
--- boqqoolloo, garbuu, mishingaa akkasumas qamadii dabalatee Midhaan nyaataa ---
7. NP? NP (akka) (.*)(,) NP NP
--- tajaajiloota akka nyaataa, iddoo irriibaa, bishaan dhugaatii, geejjibaa kennuun ---

2.4. Lexico syntactic pattern with Word embedding

This study combines LSP with Word2Vec to improve the performance of the LSP, because, the words which have the same hyponym relation are arranged with the same neighbors. Those words are considered the same words by Word2Vec because of their neighbor's words. To do this, as illustrated in figure1, Word2Vec takes the result retrieved by LSP as input using the following algorithm.

Begin
1. Getting the result of LSP and saving it to as file
2. Reading the saved file and Splitting the hyponym from the hypernym using the connectors which used in the patterns to get the words which are hyponym of each other as word pair
3. Taking the word pair of step 2 as input for Word2Vec
4. Saving the word pairs with similarity value as a file if their similarity is greater than zero
5. Identifying the correct hyponym relation pairs from an incorrect one

End

Algorithm 3 Combing LSP with word embedding (Word2Vec)

[1] http://www.phmartin.info/CGKAT/ontologies/coWordNet.html

2.5. Extraction of synonym relations from Word2Vec and hyponym relations from LSP

The Word2Vec and LSP models are not adjust the WordNet relations which extracted by them. The Word2Vec extract the synonym relations with the weight of their similarity given for them by the system. The WordNet doesn't include the weight of similarity of the terms i.e. it needs only the more similar words. So, the system removes the given number before getting the similar words to save it in the Afaan Oromoo WordNet. To do this, the system uses the following algorithm.

Begin
1. Taking the result retrieved by Word2Vec
2. Removing the weight of similarity of terms calculated by Word2Vec
3. Saving the more similar words to Afaan Oromoo WordNet

End

Algorithm 4 to extract synonym relations from Word2Vec

Similar to Word2Vec, LSP doesn't extract pure hyponym relations of terms. It returns the existed hyponym relations with the articles by which they are interconnected. The WordNet doesn't include the hyponym relations with those articles. So those articles must be removed from the hyponym word relations to save it to the WordNet. To do this, the system uses the following algorithm.

Begin
1. Taking the result retrieved by LSP
2. Removing the articles retrieved with terms
3. Saving the retrieved hyponym relations to Afaan Oromoo WordNet

End

Algorithm 5 to extract hyponym relations from LSP

2.6. Experimental Results

For this study, experiments are conducted to evaluate the performance of the suggested approaches. Evaluating the performance of the Word2Vec system and the lexico-syntactic pattern system is an important part of the study, which discusses the actual work of the study. An evaluation of the proposed knowledge-based WSD algorithm accuracy is performed. The experiments are conducted on **30** pairs of words to evaluate the Word2Vec system using both CBOW and skip gram algorithms. Applying word embedding (Word2Vec) algorithms for Afaan Oromo texts has been registered **80.09%** and **85.04%** for CBOW and Skip Gram respectively.

The lexical syntactic patterns that created from the texts (which listed above) are also evaluated using **relevant retrieved** results and the **irrelevant retrieved** results. Table 1 show the results of the above Afaan Oromoo lexical patterns which created automatically from the texts. The proposed system (LSP) evaluated using precision, recall, and F- measure as follows.

Approach	**Precision**	**Recall**	**F-Measure**
Lexical Syntactic pattern(LSP)	0.6673	0.72	0.6926
LSP with Word2Vec	0.814	0.808	0.811

Table 1 Summary of the results with identified patterns using LSP and LSP with Word2Vec

The ideas about the introduced patterns and tested them on simple Afaan Oromoo texts are implemented. Table 1 shows the results of testing the created patterns.

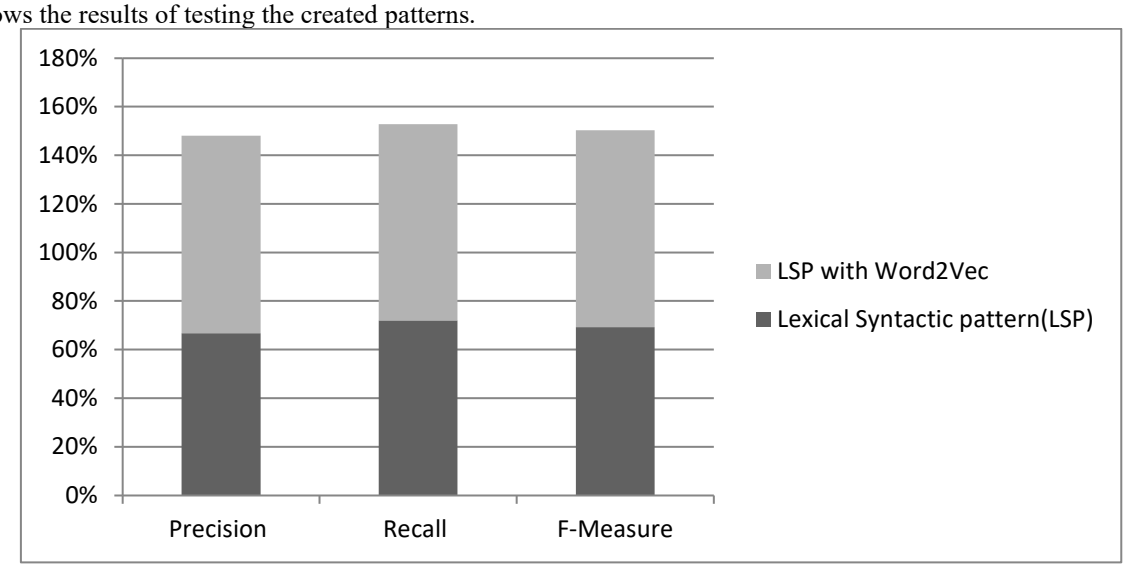

Figure 2 The graph Description for the experimental results .

As discussed in section 1.3.4, lexico-syntactic patterns are applied for Afaan Oromo texts without Word2Vec and it has been registered 66.73%, 72%, and 69.26% for precision, recall and F-Measure respectively and applying Word2Vec on the result of LSP has been registered 81.14%, 80.8%, and 81.1% for precision, recall and F-Measure respectively. Nevertheless, there are factors that could affect the performance of system: **1)** the POS tagging error[1] i.e. many times the users of the target language use adjective to express the noun phrase before listing the hyponym of that noun phrase and, **2)** it is possible that some examples of the LSP are missed because of misspellings and other typographical errors.

3. Conclusion

In this paper, the models used to build WordNet for Afaan Oromoo are proposed and developed. The main objective of the study is to develop the models used to extract the synonym and hyponym WordNet Relations from the Afaan Oromoo texts automatically. Word2Vec is developed to extract synonyms, LSP to extract hyponym from Afaan Oromoo texts and the Word2Vec with LSP integrated to improve the performance of LSP. In order to evaluate the Word2Vec prototype, **30** Afaan Oromo pairs of words are collected and to evaluate the lexical patterns **7** Afaan Oromoo patterns are used. In Word2Vec, the expected and the actual value are used and the correlation coefficient of the two values is calculated. In this system, system evaluation is performed in two-phase. In the first phase, the prototype system is tested using the CBOW algorithm and in the second phase, it is tested using the Skip Gram algorithm and the performance of the Word2Vec using the skip-gram algorithm is satisfactory when it is compared with the performance of Word2Vec using the CBOW algorithm. According to the achieved results, the skip-gram algorithm does a better job for frequent pairs of words than CBOW. But, the CBOW algorithm is faster while the skip-gram is slower. LSP is also evaluated with Word2Vec and without Word2Vec using IR evaluation metrics such as precision, recall, and F-measure. But, there are factors that could affect the accuracy of results: 1). the writer writes a noun phrase with many adjective to express the noun for the reader; and, 2) it is possible that some instances of the LSP are missed due to misspellings and other typographical errors.

References

Archive, G., 2013. *word2vec introduction*. [Online] Available at: https://code.google.com/archive/p/word2vec/.

Eggi, G.G., 2012. Afaan Oromo text retrieval system. *Master's Thesis, Addis Ababa University*.

Erena, B.G., 2017. *Afaan Oromo Language*. [Online] Available at: https://scholar.harvard.edu/erena/oromo-language-afaan-oromoo.

Fellbaum, 2006. wordnet(s). In: keith brown, (editor-in-chief). In *encyclopedia of language & linguistics, second edition*. Oxford: elsevier. pp.665-70.

Fishkin, R., 2005. *what can wordnet are used for?* [Online] Available at: https://moz.com/blog/what-can-wordnet-be-used-for.

Goldberg, Y. & Levy, O., 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *Computation and Language (cs.CL)*.

Hearst, M.A., 1992. Automatic Acquistion of Hyponums from Large Text Corpora. *Proceedings of the 14th conference on Computational linguistics - Volume 2, COLING 92 , Stroudsburg, PA, USA. ACL.*, pp.539–45.

Klaussner, C. & Zhekova, D., 2011. Lexico-Syntactic Patterns for Automatic Ontology Building. In *Proceedings of the Student Research Workshop associated with RANLP 2011*. Hissar, Bulgaria, 2011.

Mikolov, T., Chen, K., Corrado, G. & Dean, J., 2013. Efficient Estimation of Word Representations in Vector Space. *Computer Science Computation and Language*, pp.1-12.

Mikolov, T. et al., 2013. Distributed Representations of Words and Phrases and their Compositionality. *Computation and Language (cs.CL)*.

Mikolov, T. et al., 2015. Computing numeric representations of words in a high-dimensional space. *United States Patent*.

Miller, G.A. et al., August 1993. introduction to wordnet: an on-line lexical database. *International Journal Lexicography*.

Morato, J., Marzal, M.Á., Lloréns1, J. & Moreiro, J., 2003. WordNet Applications. *Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings,* pp.270–78.

Sahlgren, M., 2015. A brief history of word embeddings.

Wales, J. & Sanger, L., 2017. *Wikipedia*. [Online] Available at: http://www.wikipedia.com/ [Accessed 01 April 2019].

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3125434/