Evaluating word embeddings for computational pragmatics

Leonardo Sanna

University of Modena and Reggio Emilia, Modena, Italy

Abstract

Word embeddings are a popular set of machine learning techniques that are used in natural language processing to create a semantic model of a corpus, representing semantic relationship within a geometric space. In this space every word is mapped to a vector so that words with similar meanings are represented close to one another while words with different meanings are distant.

In this work we briefly present a novel use of word embedding for computational pragmatics in which we used word2vec Skip-Gram (henceforth SG) to explore the inferential model generated by algorithmic personalization on Facebook. We used their experiment in a recent study in which we implemented Eco's model reader as an inferential model. In a nutshell, we found that we can simulate the model reader with word embedding to create computational inferential models, that are then compared to identify a possible divergence of interpretation paths. Our experiment conducted on Facebook showed that algorithmic personalization influences model readers, differentiating inferences across different timelines.

Though for pragmatic goals, word embeddings have two significant issues that are 1) accuracy 2) variability. The standard way of evaluating the accuracy of word embeddings is to use intrinsic evaluation, using predefined analogies and similarities such as WordSim-353 (Finkelstein et al. 2002) and SimLex. However, this accuracy is quite hard to estimate while inquiring pragmatic aspects as we do not have any predefined pragmatic output.

This work aims to review these methods for a computational pragmatic approach, highlighting possible limitations and adaptations. The review will cover three aspects:

1) Reproducibility of computational pragmatics

2) Reliability assessment for computational pragmatics inferences

3) Accuracy test for computational pragmatics goals

Keywords 1

Word Embedding, model reader, evaluation, computational pragmatics

1. Introduction

Word embeddings are a popular set of techniques that are used in natural language processing to create a semantic model of a corpus, representing semantic relationship within a geometric space. In this space every word is mapped to a vector so that words with similar meanings are represented close to one another; instead, distance within the space means that words have less semantic affinity. This semantic representation originates in the distributional hypothesis [1], affirming that words with similar meanings occur in similar contexts. In mathematical terms, we can say that "The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear" [2].

In a nutshell, there are two main ways of doing word embeddings. On the one hand, we have count-based models, on the other hand, predictive-probabilistic models. In the first case, embeddings are built using a co-occurrence matrix while in the letter case machine learning is used to build a stochastic model of the language. A renowned empirical evaluation [3] showed that the performance of predictive models was far superior compared to models based on counting co-occurrences.

Proceedings of the Linguistic Forum 2020: Language and Artificial Intelligence, November 12-14, 2020, Moscow, Russia
EMAIL: leonardo.sanna@unimore.it
ORCID: 0000-0003-3021-6606


In this paper we are using word2vec, perhaps the most famous and used word embedding predictive algorithm. The goal of this work is to discuss some issues in the evaluation of probabilistic word embeddings for computational pragmatics. In the following sections we will briefly illustrate the functioning of word2vec and how could they be used for studying pragmatics. Then, we will introduce the evaluation problem, proposing a first outline of a quali-quantitative framework to evaluate our embeddings while inquiring pragmatics.

2. Related work

Predictive word embeddings might be performed with different algorithms such as word2vec [4], fastText [5], Star Space [6] and RAND-WALK [7]. More recently, other approaches such as ELMO and BERT [8,9] have gained popularity in the field of NLP. In this work we used word2vec because it is the most used and renowned version of probabilistic word embeddings.

Word2vec works on probabilistic neural networks and it might be realized in two different ways [10]: continuous-bag-of-words models (CBOW) and skip-gram (SG). A CBOW trains the neural network to predict a word given a context, while the SG model learns how to predict a context starting from a target word (Figure 1).

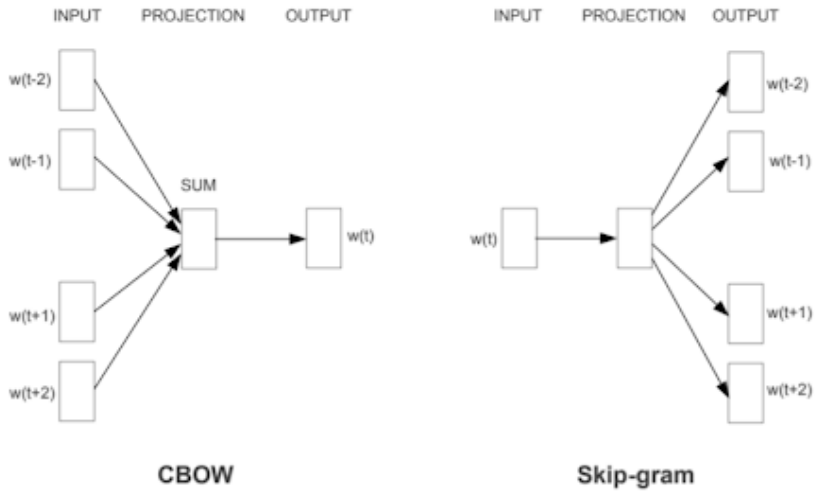

Figure 1: The functioning of word2vec, from Mikolov et al. (2013b)

Word2Vec can work either with negative sampling or with hierarchical softmax; in the case study presented we used SG with negative sampling because of its better performance on analogical reasoning as explained in the paper by Mikolov et al. [4].

We can say that word2vec creates a model computing the probability of words to co-occur, meaning that it can also capture semantic relations between words that rarely or never co-occur in the corpus whilst being semantically related. Moreover, word2vec models can perform actual inferences based on words' semantics[2]. For this reason, we used word2vec to evaluate pragmatic inferences [11], simulating Eco's model reader [12] with word embeddings.

We can think of the model reader as the pragmatic competence needed to interpret a text with reference to its producer's intention, meaning that, in Umberto Eco's semiotics, text interpretation is always guided and limited by the text itself. On the other hand, what we call "producer intention", although partially embedded in the text instance, has to be guessed by the empirical reader who has to formulate a hypothesis on the producer intention. To sum up, in Eco's theory the interpretation process is successful only when the interpretation matches the author intention by following the instructions of the model reader, meaning that the reader understood what the producer wanted to communicate. A crucial component of the model reader is what Eco called encyclopedic competence, namely the shared

[2] The renowned example is that given the words "king" and "man", the word2vec model can infer the word "queen", given the word woman,

cultural context between the reader and the author. For instance, to understand the sentence "word embeddings are a machine learning technique used in NLP", the reader has to know what is machine learning and what is NLP and, most important, how do these two things are related. In Eco's theory, encyclopedic competence is essential to fulfil the interpretation process. This competence is acquired in previous textual and verbal exchanges. In a nutshell, the encyclopedia is the shared information that we must know to communicate and the model reader is the set of instructions on how to combine this shared knowledge to understand a text.

Hence, we approached the model reader theory empirically using word embedding arguing that it could be as an inferential model, precisely a function that takes as input a target text and that produces interpretation paths as an output. In other words, we can use word embeddings in qualitative analysis, creating a computational version of the inferential model.

We used this theory to investigate algorithmic personalization [13] on Facebook. A study on the Italian elections made in 2018 [14] highlighted that Facebook algorithm treated politically sources unevenly, fostering the political ideology of each profile. From a semio-linguistic perspective, it is interesting to inquire whether also the model readers were different across each political ideology. The study was made creating six different profiles, all following the same 40 political pages but interacting only with one political party. We used these six profiles to generate six different word embeddings, investigating differences within the inferential models like in the example showed in Figure 2.

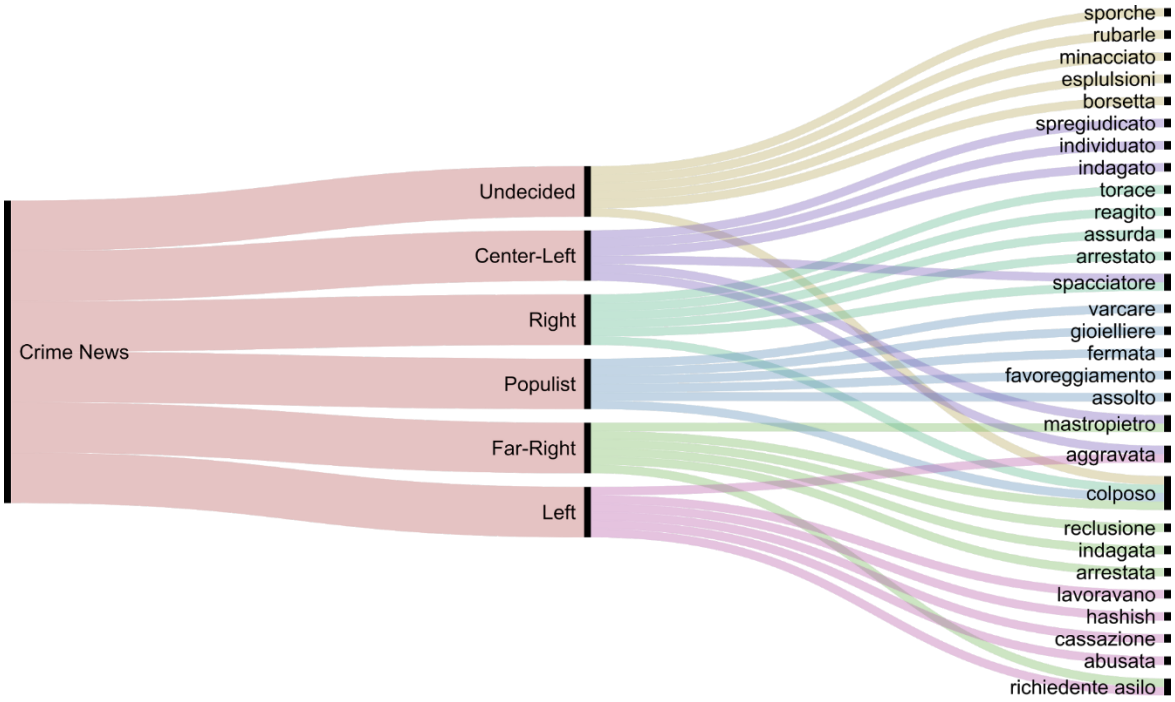

Figure 2: An example of different inferences made by the six different profiles regarding Crime News

3. Data and methods

The dataset of the experiment is the same used for the aforementioned experiment and it is composed by two different parts: sources and impressions. In the sources dataset we had the entire collection of each Facebook post made by the whole set of pages followed by the bots, while in the impressions dataset we had the actual posts that were shown to our profiles.

We created an embedding model both for the sources dataset and then one embedding for each alleged model reader, as shown in the example in Fig.2. We noticed that there were some important issues regarding the explicability and the evaluation of our experiment.

Explicability and evaluation of word embeddings still an issue in NLP. Predictive models are not fully reproducible, since each time we start a model from scratch, even with the same hyperparameters, the model would be slightly different. Research on these themes, to best of or knowledge, is very recent [15].

To sum up, we had three types of problems related to our research.

1. Reproducibility: How can we reproduce an inferential model so that we can perform more experiments? Recently Hellrich [15] experimented a Singular Value Decomposition of a PPMI matrix with weighting-based downsampling to generate reliable word embeddings without losing performance. In this case, it is possible to have a fully reproducible word embedding model, meaning that with the same parameters we always get the same vectors. SVD works on three different matrixes, two orthogonal containing vectors and one diagonal containing values. These three matrices are then decomposed each time in the same way, so that the same input would always generate the same matrix. Particularly interesting the solution proposed by Hellrich that introduced a variant using a PPMI matrix populated via weighting, to enrich its embeddings.
2. Reliability; intended as the measurable stability of word embeddings, namely how much word embeddings change each time that they are initialized with the same hyperparameters. SG model is trained with a stochastic gradient descent that works finding one of the possible optimum local minima. The fact that the gradient can find slightly different local minima combined with the fact that the vectors are randomly initialized creates a model which is not fully reproducible.
3. Accuracy; in probabilistic word embeddings is usually measured using predefined word pairs and analogies [16, 17], testing then the ability of embeddings to produce semantic inferences.

The first two points are related, as a fully reproducible algorithm also generates a reliable model. Though, in pragmatics, it is hard to find stability because inferences are never determined but only suggested, the same text might produce different inferences even with the same reader. The model reader is a set of instructions that guides the empirical reader in its interpretation; the conditions of the model reader are not rigid and mandatory. We argue that, for our goals, it would be misleading to use the proposed SVD matrix to build our embeddings, because it would represent a deterministic view of the inferential model.

Yet the problem of reproducibility is real and the important work made by Hellrich brought us clear evidence that we cannot avoid dealing with robust evaluation methods. Thus, we decided to use two different NLP methods in our experiment.

The model reader is generated in each text from the combination of author's intention and encyclopedic knowledge, namely the shared information among reader and author. This shared information might be assumed as equal for each model reader and each model author, meaning that we can use a fully reproducible method to create the semantic model of our encyclopedia.

For our goals, we were also interested in extracting the main topics that emerged from our corpus. Therefore, we created a topic modelling using a hierarchical classification [18] on the sources dataset. This hierarchical classification allows us to extract the most meaningful topics and the words associated with each one of them. We chose Reinert hierarchical classification (CHD) because of how his algorithm works; it creates its classification mapping lexical forms with contexts. Although there is not any stochastic elaboration, this feature makes CHD comparable with word embeddings, since the representation of words semantics is not created among words but words and contexts, intending contexts as text chunks. The reason why we chose word embedding to represent the model reader is that they consider the relations between words and contexts, CHD does the same on a count-based framework.

We then selected the most relevant words in the most relevant topics, and we used these words to create our semantic query to perform the pragmatic move within our model readers. In fact, one of the characteristics of word2vec is that we can perform algebraic operations with our vectors, adding or subtracting semantic traits from each word.

Regarding the last point, accuracy is not easy to evaluate in word embeddings. A review of methods used to test the quality of embeddings has been made by Wang et al. [19]. These methods could be divided into two families: intrinsic evaluation and extrinsic evaluation. Extrinsic evaluation is a downstream method while the former evaluates our model internally. As said, intrinsic evaluation is probably the most common standard evaluation for word embeddings. Word similarities and analogies are computed to verify consistency within the vectors. In particular, for our goals analogy was the best option because we want to investigate inferences. The equation proposed by Mikolov et al. [20]

computes the $argmax$ of a given linear representation. For instance, if we calculate the analogy $man: king :: woman: x$, we compute $x = man - king + woman$ and then compute the $argmax$ to find the vector with the highest cosine similarity to x. In our case, analogy testing is excellent but at the same time problematic. Inquiring pragmatics in political ideology implies that we are interested in discovering inferences that have an ideological value; with the word "discovering", we mean that we are using word embedding to explore an inferential model that we do not know, hence we do not have a predefined set of analogies to test the quality of our embeddings.

Nonetheless, there are for sure some "wrong inferences". We do not perform inferences in a completely free environment but instead following some textual constraints that are posed by the text itself and by the encyclopedic context. Suppose I read the word "flat tax". In that case, I expect to infer words related to this term, namely the political party which is proposing it, words associated with the economic lexicon and words that are related to the evaluation of the political proposal. It would be wrong for sure to infer words that are completely unrelated to the semantic field of "flat tax" such as "dog" or "Travis Scott". We propose to use this method to evaluate the quality of our embeddings, looking for words that are somehow “intruders” in our semantic field. In this case, we are assuming Fillmore’s theories on semantic fields [21]. To use the words of Violi [22] words select their contexts, hence each word would select an appropriate semantic field. We called words that do not belong to the given semantic field “semantic noise” and we used this concept to test the accuracy of our embeddings. We took into account the top-50 most similar words, looking for semantic noise. At this stage of the research, we have not defined how to quantify semantic noise yet. We assumed that up to 10% of semantic noise in the top-50 would have been fine, however semantic noise within the top-20 or top-10 should be weighted differently. We found significant semantic noise only in the centre-left model reader, that was also the smallest.

To sum up, using word embeddings for computational pragmatics we would need two have at least three corpora: the first corpus representing the encyclopedic knowledge and then at least two to represent the model readers so that we can compare their inferences. This methodology allows a fully reproducible experiment as long as we share criteria and methods used to select both corpora. In the case presented, the experiment is limited to model readers guessed and proposed by Facebook algorithm, thus only Facebook posts were used to generate the embeddings.

Our experimented concluded that there were some differences among the model reader, but that further analysis was necessary to understand the differentiation and their semantic reasons better.

4. Discussion

In this paper we are not discussing the findings of the experiment but instead possible flaws and implications of our evaluation method.

To have a fully reproducible and reliable model we used two different algorithms to represent the shared cultural knowledge and the actual model reader. One possible limitation of this approach is that we are relying on the topic modelling that is built in the first part of the experiment. If this representation is biased, then also our inferential model would be biased. Although Reinert classification is well-trusted in literature [23, 24, 25], using different algorithms for topic modelling would be the best solution so that at least we can compare the most relevant topics and inquiry their consistency. Besides this, there is also a theoretical question.

We said that we avoided using an SVD matrix because it was too deterministic in representing pragmatic inferences. On the other hand, we assumed that we could use a fully reproducible topic modelling to represent our shared knowledge. This might be questioned. The encyclopedia is the whole cultural context that each one of us must have and share with other people to which he intends to communicate. Since each one of us has different cognitive capacities, we might argue that even the encyclopedic knowledge should be represented with a probabilistic algorithm. However, according to Eco’s theory, the shared cultural context is something stable and objective. What changes is the capacity to use and combine this knowledge for each empirical reader and, in the case of the model reader, the suggestions that might guide the empirical reader in combining and interpreting his encyclopedia.

We might argue that shared knowledge must be stable and equal, while the ability to combine this shared knowledge might differ in each interpretation process, meaning that even the same reader might accomplish different interpretation of the same text, without adding new information to its encyclopedic competence. Even though it might sound too simplistic, this is the way our communication works. For instance, the sentence "Boris Johnson has finally made Brexit" requires shared knowledge on Boris Johnson and Brexit in each model reader. What changes it is the interpretation of these words, that would reveal pragmatic moves of each model reader. In Umberto Eco semiotics, this is called *intentio operis* [26], meaning that the text has a stable value of signification that exists regardless of the author's intention and no matter of the reader's interpretation. The *intentio operis* could be seen as the objective semiotic dimension of the textual instance, namely what the text actually means because of its semantic structures. Thus, we conclude that using a fully reproducible method to represent the encyclopedic knowledge is necessary both to account for reliability and to correctly represent the semiotic dimension of *intentio operis*.

Nevertheless, we used word2vec to generate our model readers. In this case we still have the reliability problem and also the accuracy problem. As we said before, accuracy is challenging to estimate for pragmatic inferences and we used a novel method, defining "semantic noise" words that do not belong to the inquired semantic field. However, evaluating semantic noise is not always possible, even relying on databases such as WordNet. In fact, we might have what look like a semantic intruder that is, instead, a valid inference coming from our corpus. For this reason, we used a quali-quantitative approach to evaluate accuracy, using corpus linguistic tools to evaluate co-occurrences and concordances [27]. Besides, we also evaluated the density of our vectors. According to our findings, similar words usually have a cosine similarity between 0.5 and 0.8 in the top-50 most similar words, with the most two-three similar words much closer compared to the other. Instead, words that have a poor representation might either have a sparse cluster of most similar words (<0.50) or a really dense one (>0.90 for more than three words in the cluster).

Hellrich also introduced a quantitative evaluation for the reliability of word embedding comparing cosine similarity and Jaccard coefficient among different word embedding models. However, these metrics have not yet been explored on large corpora and we decided that it was not suitable for our experiment since we had relatively small models (< 50k words). Anyhow, should we evaluate pragmatics embeddings simply calculating a coefficient of variability? Perhaps the answer is yes, since an extremely high variability would mean that we have a problem in our corpus or our data. On the other hand, it is difficult to quantify how much variability should we accept while studying computational pragmatics. This part would probably need dedicated research with a detailed analysis of the variability of semantic fields and their related pragmatic instructions.

What about readers embeddings then? We should accept as a compromise a probabilistic model, to account for variability and complexity of human interpretation.

Summing up, the encyclopedic model must be computed with methods that allow for complete reproducibility, so that we have an accurate representation of the *intentio operis*. On the other hand, we use probabilistic embeddings to represent the variability of the interpretation proposed by the model reader. Regarding accuracy evaluation, the first step is to analyze the semantic field that has been computed for the target word in our model, looking for possible semantic noise. The second step is to verify the hypothesis of semantic noise using standard corpus linguistics metrics to evaluate the density of our vectors.

Concluding, we would like to pinpoint that extrinsic evaluation for computational pragmatics is not possible. Evaluating word embeddings for tasks such as translation or classification is relatively easy as we know what we are trying to accomplish with our models. In the case of computing a model reader, we want to use our computational model to discover something that we do not (perhaps cannot) know using other methods. What is fascinating of word embeddings models is not merely the fact that we can compute the model readers but the fact that, for the first time in semiotic analysis, we can study them data-driven with computational tools. We do not know in advance whether an inference is bad or good, nor we know if our inferences are deep enough or if they could be improved. This is a new level of complication in evaluating a process that is already hard to keep transparent and reproducible. The problem then shifts from computational linguistics to epistemology.

5. Conclusions

Computational pragmatics with word embeddings is still unexplored in linguistics and semiotics, while evaluation methods for word embeddings are vastly understudied also in other applications of word embedding techniques.

This work was meant as a first attempt to outline possible methods of evaluation for embeddings specifically meant to study pragmatics. The evaluation is possible creating a fully reproducible semantic model of an encyclopedic context using non-probabilistic algorithms; in our case study, we used Reinert hierarchical classification (CHD). On the other, the variability of the model reader should be computed with probabilistic embeddings to avoid a deterministic view of its functioning. We should then accept, as a compromise, to compute our inferences with non-reproducible word embeddings.

Further research is needed to find reliable quantitative metrics to evaluate word embeddings for computational pragmatics as, by the time we are writing this paper, for pragmatics goals there are still no studies on large corpora.

6. References

[1] Z.S. Harris. "Distributional structure". *WORD*, 10:2-3 (1954), 146-162.

[2] A. Lenci. "Distributional semantics in linguistic and cognitive research". Italian journal of linguistics, (2008) 20(1):1-31.

[3] M. Baroni, G. Dinu, and G. Kruszewski. "Don't Count, Predict! A Systematic Comparison of Context–counting vs. Context–Predicting Semantic Vectors", Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1) (2014): 238–47.

[4] T. Mikolov, I. Sutskever., K. Chen., G. S Corrado, and J. Dean. "Distributed representations of words and phrases and their compositionality" in: Advances in neural information processing systems, (December 2013), Leake Tahoe, 3111-3119.

[5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. "Enriching word vectors with subword information". Transactions of the Association for Computational Linguistics, 5:135-146.

[6] L. Y Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes and J. Weston. "Starspace: Embed all the things!". arXiv preprint arXiv:1709.03856 (2017).

[7] S. Arora, Y. Li, Y. Liang, T Ma,., and A, Risteski. "Rand-walk: A latent variable model approach to word embeddings". (2015) *arXiv preprint*, arXiv:1502.03520

[8] M. E., Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. "Deep contextualized word representations." arXiv preprint arXiv:1802.05365 (2018)

[9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." (2018) arXiv preprint arXiv:1810.04805

[10] T. Mikolov, I. Sutskever., K. Chen., G. S Corrado, and J. Dean. "Efficient estimation of word representations in vector space. arXiv preprint, (2013) arXiv:1301.3781.

[11] L. Sanna, D. Compagno. "Implementing Eco's Model Reader with Word Embeddings. An Experiment on Facebook Ideological Bots". In: JADT 2020 proceedings (post-review pre-print) (2020) https://iris.unimore.it/retrieve/handle/11380/1220856/300738/Paper_JADT_final-3.pdf

[12] U. Eco, The Role of the Reader, Bloomington : Indiana UP, 1979.

[13] E. Parisier, The Filter Bubble: What the Internet is hiding from you. The Penguin Press - New York 2011

[14] E. Hargreaves, C. Agosti, D. Menasché, G. Neglia, A. Reiffers-Masson, E. Altman. "Biases in the Facebook News Feed: a Case Study on the Italian Elections." International Conference on Advances in Social Networks Analysis and Mining, August 2018, Barcelona

[15] J. Hellrich,. Word embeddings: reliability & semantic change (Vol. 347). IOS Press 2019.

[16] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan,, Wolfman, G., & Ruppin, E. "Placing search in context: The concept revisited". In: Proceedings of the 10th international conference on World Wide Web (pp. 406-414) (2001).

[17] F. Hill, R. Reichart & A. Korhonen. "SimLex-999: Evaluating semantic models with (Genuine) similarity estimation. In: Computational Linguistics", 41(4): 665–695 (2014).

[18] A. Reinert. "Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte." Cahiers de l'Analyse des Données 8.2 (1983) : 187-198.
[19] B. Wang, A. Wang, F. Chen, Y. Wang, and C. J. Kuo. "Evaluating word embedding models: Methods and experimental results." APSIPA transactions on signal and information processing 8 (2019).
[20] M.A. Bouchard, and S. Kasparian. "La classification hiérarchique descendante pour l'analyse des représentations sociales dans une pétition antibilinguisme au Nouveau-Brunswick, Canada." JADT'18: 142 (2018).
[21] C. J. Fillmore. "Frame semantics and the nature of language." Annals of the New York Academy of Sciences, 280(1), 20-32. (1976)
[22] Violi, P. Significato ed esperienza. Studi Bompiani, 1997
[23] T. Mikolov, Y. Wen-tau and Geoffrey Zweig. "Linguistic regularities in continuous space word representations." In Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies, pp. 746-751. (2013).
[24] I. Tzankova and E. Cicognani. "Youth Participation in Psychological Literature: A Semantic Analysis of Scholarly Publications in the PsycInfo Database." Europe's Journal of Psychology 15, no. 2 (2019): 276-291.
[25] P. Ratinaud, and P. Marchand. "Quelques méthodes pour l'étude des relations entre classifications lexicales de corpus hétérogènes: application aux débats à l'assemblée nationale et aux sites Web de partis politiques." Statistical Analysis of Textual Data (2016): 193-202.
[26] U. Eco, The limits of Interpretation. Bloomington, Indiana UP, 1990
[27] C. Manning and H. Schutze. Foundations of statistical natural language processing. MIT press, 1999.