Assessing Gene Regulatory Network Inference Algorithms Using Word Embeddings: A Novel Approach for NLP and Systems Biology Integration

Sergio Peignier[1]   Patricia Zapata[2]

[1]Univ Lyon, INSA Lyon, INRAE, BF2I UMR 203, 69621, Villeurbanne, France
[2]Universidad Mayor de San Andrés, La Paz, Bolivia

1 Introduction

Since their conceptualization, vectors have been extensively used to describe objects in different scientific domains [5]. In natural language processing, representing words as numerical vectors has been a popular prerequisite to apply state-of-the-art machine learning techniques for text analysis. A common way to represent words, called bag-of-words [13], is through sparse vectors of counts in different documents. Another popular representation is based on dense vectors, also known as word embedding (e.g. [14]). These vectors are designed to capture the meaning and context of words, and each dimension is considered to model specific semantic or syntactic features of words. Similarly, but in a different domain, i.e., biology, advancements in high-throughput sequencing have enabled the representation of genes as numerical vectors of expression levels in different experimental condition and tissues. Such vector representations have also been used to apply statistical and machine leaning techniques to untangle biological problems. In this context, a very important research domain in biology consists in using gene expression to reverse engineer the architecture of regulatory interactions between genes [19]. Indeed, the inference of gene regulatory networks (GRN), from gene expression, is a challenging problem for the systems biology community, and a several techniques have been proposed to tackle it [19]. Moreover, evaluating these algorithms is not a trivial task, their results are often compared against a benchmark dataset [12, 16]. Two major kinds of GRN inference benchmark datasets have been proposed so far: i) real organisms gold-standard datasets consisting of experimentally verified regulatory interactions between genes. ii) Synthetic gold-standard based on underlying in-silico GRN models. Nevertheless, biological benchmark are rather rare, expensive to produce and often incomplete, and synthetic datasets may include biases related to their underlying in-silico model [12]. In both cases, it reveals to be difficult to evaluated and understand the behavior of the algorithms using such benchmark datasets.

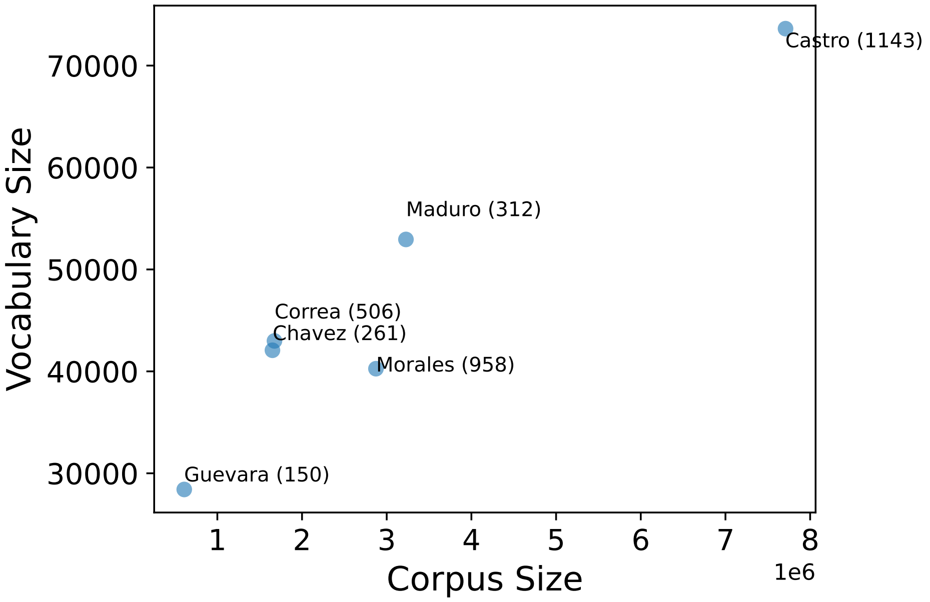


Figure 1: Vocabulary size as a function of the corpus size, in words for each politician (number of speeches in parenthesis).

Given the similarity between words embeddings and gene expression matrices, and given the intuitive interpretability of natural language, a natural question is whether word embeddings can be used to evaluate and leverage the understanding of GRN inference tools. In this work, we propose to consider words as genes, and their word embeddings as gene expressions, and then analyze these datasets using GRN inference tools, with a two-folded objective: i) Answer common questions related to the inference task, studying thus the potential of such datasets to assess GRN inference methods. ii) Investigate the possibility to use GRN inference tools, as new NLP techniques, to build knowledge graphs and study the relationships between words.

2 Materials and Methods

2.1 Corpus

In this work we relied on a corpora of discourses of six latin-american socialist politicians from the XX-th and XXI-st century, namely, Fidel Castro, Ernesto Guevara, The number of speeches, corpora sizes and vocabulary sizes are represented in Figure 1.. These corpora are suitable candidates for benchmark datasets for this study, since they were analyzed using a word-embedding-based technique to represent words as numerical vectors, and then they were connect similar word vectors to form networks of words that were studied in the context of discourse analysis [20], and thus the relationships between words are well-characterized.

2.2 Embedding Algorithm

Among the word embedding techniques, Word2Vec [14] is a state-of-the-art technique that relies on a neural network architecture, termed Skip-Gram, trained on a textual corpus. The network is trained to take each word at a time and predict its surrounding context words, within a window of fixed size. To do so, the network adapts the word vector representations, such that similar words (semantically and/or syntactically) tend to have similar representations in the word vector space. Word2vec has been applied successfully to tackle a wide range of NLP tasks such as classification (e.g., [11]), and language modeling including Discourse Analysis (e.g. [14]). In practice, we used the gensim library [17] implementation, to form word embeddings of 300 dimensions, running 10 iterations over each corpus, and considering a word context window size of 10 words.

In this work, we analyzed the word embeddings from each politician independently, as well as combinations of embeddings: we concatenated the word embeddings from the two politicians from XX-th century (Guevara and Castro), and from the remaining four politicians from XXI-st century. This allowed us to study the impact of combining different datasets to infer GRNs, a common question relative to the GRN inference task.

2.3 Discours analysis

The corpora described in Section 2.1 have been analyzed in [20] by first building word-embeddings, then clustering them, representing each cluster' words using a graph-based technique to form so called prototypical-discourses, and then applying the well-known discourse analysis methodology described in [7, 6], to study these prototypical-discourses. This discourse analysis methodology aims at grouping discursive strategies following the Aristotelian categorization of rhetoric art [3]: i) Ethos: increases the speaker's trustworthiness, and defines speaker's discursive identity: often as an expert or a charismatic leader (for example, the systematic use of the pronoun "us" aims at building a bringing closer the relationship between the politician and the audience). ii) Pathos: appeals the audience emotionally using two major strategies: a) the triadic scenario: the speech is organized as a story where the audience is the victim of an enemy (which may be internal or external and can depend on the historical context, and which is often associated to negative terms), and a hero (the speaker and his political party) fights the enemy and protects the audience (in this context, the politician tends to use words with an strong emotional load such as "war", "misery", "danger"). b) Recruitment process: the speaker refers to positive values (social welfare, economic growth and technological development ...) in order to make the audience accept the speaker's project willingly. iii) Logos: appeal to rational arguments, this strategy is under-represented in political speeches, and mostly involves citing historical characters and events of reference and giving many details to increase the veracity of the speech. The major discursive strategies from these families are schematized in Figure 2.

The word relationships identified in this work, were analyzed in terms of grammatical and discourse analysis-based consideration

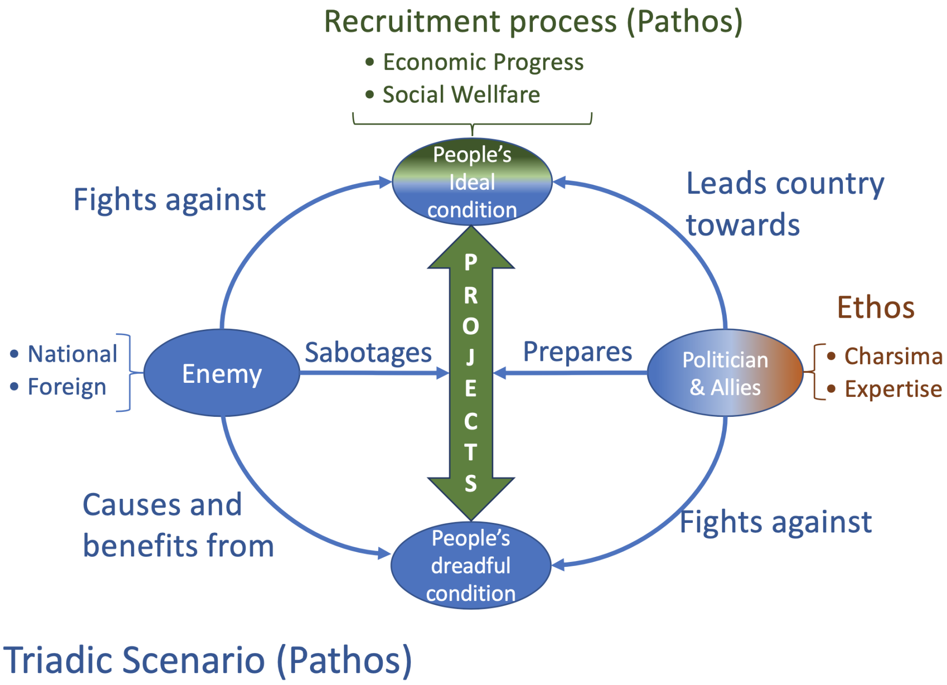


Figure 2: Schematic representation of major discursive strategies

2.4 GRN inference

The different methods have been proposed so far to infer GRNs from gene expression matrices only, can been categorized in three major families [19]: i) Probabilistic Model-Based Methods that aim at using the gene expression data to fit the parameters of a pre-established probabilistic model (e.g., Gaussian graphical models, Bayesian networks), that corresponds to the GRN model. ii) Dynamical Model-Based Methods aim at relying on temporal gene expression data to fitting a dynamical model (e.g., Dynamic Bayesian Networks and Ordinal Differential Equations) to model the temporal changes in the expression of genes. iii) Finally the so-called Data-Driven Methods analyze gene expressions to score all possible links between regulatory genes and target genes, and then select the best links [19]. These methods are generally recognized as simple, computationally efficient and accurate techniques [19]. Among these tools, early approaches (e.g., [8, 21]), termed co-expression networks, have used correlation statistics to infer undirected links between pairs of genes. More recent tools (e.g., [10, 9]), aim at training regression or classification algorithms to predict the expressions of each target gene from those of regulatory genes, and score the contribution of each regulator using a feature importance procedure. The selection of a limited list of potential regulatory genes is an important step, and with this aim GRN inference methods often use prior functional gene annotation information. In this work, we used the GXN•OMP inference method from the GXN open-source Python library [15].

2.5 Applying GRN inference to word embeddings

Given the large vocabularies present in this corpora, we selected the most important words from each corpus, as target words. To do so, we computed the word frequencies, in each corpus, and compared them to the word frequencies of reference in Spanish distributed by the Royal Academy of Spanish Language CREA[1], in order to score each word using a weighted logit function. More formally, let $p^c_w \in [0,1]$ be the frequency of word w in corpus c, and let $p^r_w \in [0,1]$ be the frequency of word w in the CREA reference. The weighted logit score associated to word w in the corpus c is $wlod^c_w = p^c_w \times log(p^c_w/p^r_w)$. When both frequencies are very close, then $lod^c_w \simeq 0$, and when w is more (respectively less) present in corpus c than expected in the CREA reference, $lod^c_w > 0$ (respectively $lod^c_w < 0$). In this work, we selected the top 500 words for each dataset to be considered as target words.

Regarding regulatory words, we have tested two strategies to study the impact of the list of regulatory elements in the inferred GRN: i) We considered all target words as potential regulators. ii) We only selected verbs among the target words as potential regulators. The first strategy corresponds to a null model, while the second strategy aims at mimicking a list of regulatory genes such as Transcription Factors, that control the expression of other genes. Indeed, the verb acts as a central word that regulates the meaning of a sentence, structuring the relationship between subject and complement. It would be interesting in a future work to investigate other types of regulating words, for example, we could select the words that appear more frequently with respect to their standard use in Spanish, since from these words, with high semantic load, we can infer the central message of the text.

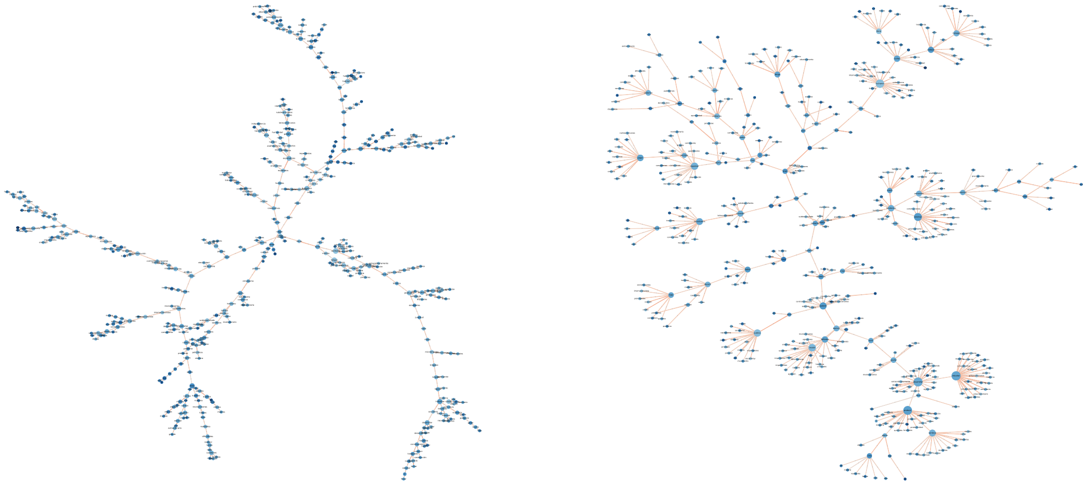
Finally, in order to enhance the visual analysis of the inferred network, we used the Edmonds' algorithm to represent the network as its maximum spanning arborescences.

3 Results

3.1 Impact of regulators choice

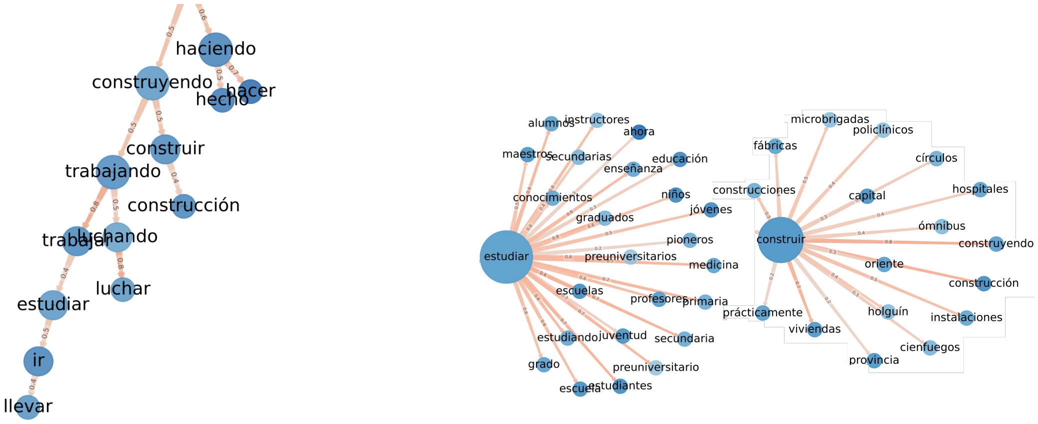
The choice of regulatory words revealed to be crucial in the network inference. Indeed, when all word can act as regulators, spanning arborescences tend to form deep branches with connections between words sharing the same grammatical category (verbs, adverbs, nouns), and inner smaller regions of inter-connected semantically similar terms, as shown in Figures 3a and 4a. When only verbs were chosen as regulators, spanning arborescences tend to be structured as inter-connected star-like verbal-ego-graphs, as shown in Figure 3b and 4b. In this case, the links connect verbs and their potential subjects and complements. The interpretation of the inter-connected verbal-ego-graphs networks was easier than the deep branches, since their structures (verb surrounded by subjects and complements) can be easily interpreted as prototypic-sentences. Notice, that edges are always oriented from the regulatory word towards the target word, nevertheless this order does not imply a linguistic or discursive order.

[1] `https://corpus.rae.es/lfrecuencias.html`



(a) (b)

Figure 3: Arborescences for Castros' network, considering all words (3a) or only verbs (3b) as regulators.



(a) (b)

Figure 4: Neighborhood for words ”estudiar [to study]” and ”construir [to build]” for Castros' network, when all words (4a) and only verbs (4b) are taken as regulators.

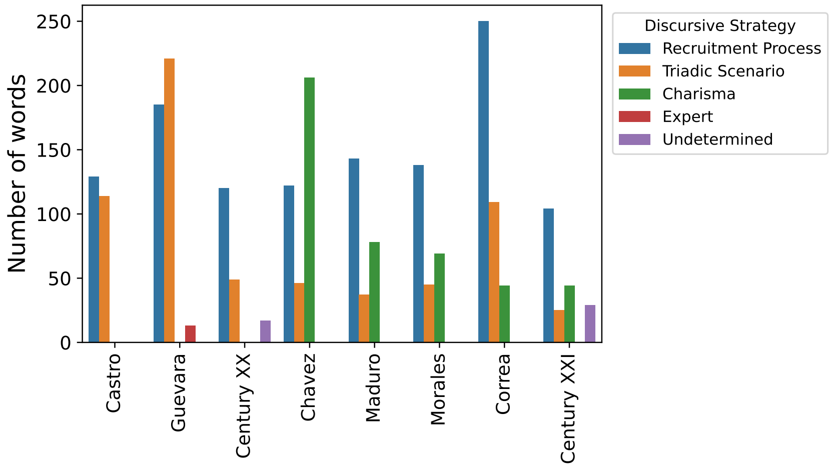


Figure 5: Number of words associated to each discursive strategy for each corpus

3.2 Impact of combining different datasets

The verbal-ego-graphs containing more than 10 words were studied using Charaudeau's methodology, and we count the number of words associated to the 4 most important discoursive strategies, namely recruitment process, triadic scenario, charisma and expert ethos. Results depicted in Figure 5 show that individual analysis reveal more details (higher number of words associated to each category) than grouped ones. Indeed, combining word embedding from different sources provided more general results, with less words associated to the different discursive strategies, including verb-ego-graphs without clear links to discursive strategies.

4 Conclusion and Discussion

This study has shown the importance of carefully choosing the list of regulatory elements to form networks, as well as the impact of combining datasets from different sources, showing that word embeddings have an important potential to evaluate and better understand GRN inference methods, including the analysis of new questions in future works. In addition, the networks that were obtained using GRN inference tools revealed valuable information that was easy to analyze, which also highlights the potential use of such techniques to infer networks of words, and extract knowledge from texts. Indeed, the word networks obtained in this work are different from state-of-the-art representations such as clusters of co-occurring words, since the latest are simply ”bags of words” that tend to appear in the same context without connections between them, while the former provide to the analyst meaningful relationships between words, that can be studied as proto-sentences.

Another promising research direction includes comparing this method with respect to dedicated Knowledge Discovery tools, based on Part-Of-Speech tagging techniques [1, 2] or based on the collection and analysis of dedicated data bases [4].

Finally, Transformer-based models [18] have been recently used to address a large variety of Natural Language Processing successfully. These methods provide contextualized embeddings for words, and therefore, it could be particularly interesting to extend this work towards the analysis of such kind of embeddings.

References

[1] Building a knowledge base from texts - practical nlp with python. `https://www.nlplanet.org/course-practical-nlp/02-practical-nlp-first-tasks/16-knowledge-graph-from-text.html`. Accessed: April 18, 2023.

[2] Knowledge graph – a powerful data science technique to mine information from text (with python code). `https://www.analyticsvidhya.com/blog/2019/10/how-to-build-knowledge-graph-text-using-spacy/`. Accessed: April 18, 2023.

[3] u. Aristotle. *Rhetoric*. Kessinger Publishing, 2004.

[4] P. Chandak, K. Huang, and M. Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023.

[5] J. M. Chappell, A. Iqbal, J. G. Hartnett, and D. Abbott. The vector algebra war: a historical perspective. *IEEE Access*, 4:1997–2004, 2016.

[6] P. Charaudeau. Le discours de manipulation entre persuasion et influence sociale. In *Acte du colloque de Lyon*, 2009.

[7] P. Charaudeau and D. Maingueneau. *Dictionnaire d'analyse du discours*. Seuil, 2002.

[8] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, 5(1):e8, 2007.

[9] A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert. Tigress: trustful inference of gene regulation using stability selection. *BMC systems biology*, 6(1):145, 2012.

[10] A. Irrthum, L. Wehenkel, and P. Geurts. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776, 2010.

[11] J. Lilleberg, Y. Zhu, and Y. Zhang. Support vector machines and word2vec for text classification with semantic features. In *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 136–140. IEEE, 2015.

[12] D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, D. Consortium, M. Kellis, J. J. Collins, and G. Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796, 2012.

[13] M. F. McTear, Z. Callejas, and D. Griol. *The conversational interface*, volume 6. Springer, 2016.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[15] S. Peignier and F. Calevro. Gene self-expressive networks as a generalization-aware tool to model gene regulatory networks. *Biomolecules*, 13(3):526, 2023.

[16] A. Pratapa, A. P. Jalihal, J. N. Law, A. Bharadwaj, and T. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 17(2):147–154, 2020.

[17] R. Rehurek and P. Sojka. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2), 2011.

[18] A. Rogers, O. Kovaleva, and A. Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2021.

[19] G. Sanguinetti and V. A. Huynh-Thu. Gene regulatory network inference: an introductory survey. In *Gene Regulatory Networks*, pages 1–23. Springer, 2019.

[20] P. Zapata and S. Peignier. *Análisis del discurso socialista latinoamericano basado en inteligencia artificial*. Instituto Internacional de Integración Convenio Andrés Bello, 2017.

[21] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.