WINS: Web Interface for Network Science via Natural Language Distributed Representations

Dario Borrelli(✉), Razieh Saremi, Sri Vallabhaneni, Antonio Pugliese, Rohit Shankar, Denisse Martinez-Mejorado, Luca Iandoli, Jose Emmanuel Ramirez-Marquez, and Carlo Lipizzi

School of Systems and Enterprises, Stevens Institute of Technology, 525 River Street, Hoboken, NJ 07030, USA
{dborrell,clipizzi}@stevens.edu

Abstract. This work proposes a novel approach to visually interact with semantic networks constructed via natural language processing techniques. The proposed web interface, WINS, allows the user to select a textual document to be analyzed, choose the algorithm to construct the semantic network, and visualize the network with its metrics. Unlike previous works, which are typically based on co-occurrence matrix for constructing the text network, the proposed interface embeds an additional approach based on the combination of network science with distributed representations of words and phrases.

Keywords: Distributional hypothesis · Networks · Natural language

1 Introduction

The large-scale amount of digital information produced nowadays, especially textual information, can be an important resource for studying how words, idioms and their semantic meanings vary depending on different variables. These variables may be time, domain knowledge, socio-cultural bias, political bias, the author of the text, the audience, just to name few examples. Therefore, meaning is a relative concept that may assume a different connotation in different contexts.

Books, newspapers articles, scientific articles, patents, unstructured text from social media, web search engines, medical reports, contracts, government forms, all are examples of textual information that is available in digital format or can be converted into digital format. Having such resource potentially available naturally fosters research on methodological, computational, and visual ways to find relationships among meaning and concepts in these documents. Detecting these relationships would enable an analytic approach that could (i) reduce the effort that results from a manual analysis, (ii) will provide a scalable, computational way to compare different documents that could make latent knowledge emerge.

C. Stephanidis and M. Antona (Eds.): HCII 2020, CCIS 1224, pp. 614–621, 2020.
https://doi.org/10.1007/978-3-030-50726-8_80

An application of this type can not substitute humans in related task but can augment their capability of analysis and decision making.

To address such challenge, this paper presents a work-in-progress Web Interface for Network Science (WINS) via Natural Language Processing techniques. The proposed approach is inspired by previous works (i.e. Wordij [3]) for creating semantic networks starting from text documents of any typology. Unlike traditional approaches based on word-pair occurrences [15], this work introduces a layer of novelty by applying computational distributed representations of words generated with the artificial-neural-network model proposed by Mikolov et al. [16]; these representations are based on the distributional hypothesis [8], which states that words occurring in similar contexts have linked meanings. This feature embedded in WINS enables the detection of latent semantic proximity between textual units leveraging spatial distance among vectors, and use it to construct the semantic network. The identified textual units can be words, n-grams, idioms, sentences, or paragraphs depending on how the user performs the text pre-processing phase. Each unitary element will form a node in the network generated by WINS, while the edges of the network are created using proximity measures among respective distributed representations of words.

The proposed interface is currently at an early development stage. The present paper aims at showing the conceptual idea behind WINS and a preliminary view of the functionalities of the interface.

The paper is structured as follows: the section "Related Works" examines previous works and related approaches. Then, the "Methods" section introduces the reader to the proposed interface giving an overview of the approach and features embedded in the user interface at the current stage. The "Results" section shows examples of outputs generated by WINS. Finally, "Conclusion and Future Works" will provide the reader with information on possible applications, limitations, and future work.

2 Related Works

The increasing amount of digital textual information potentially available is making research interests grow with respect to methods and tools to analyze such large-scale data. Potential applications of such methods could benefit many research fields due to the fact that natural language is the main vehicle for communicating domain-specific concepts and expressions [6].

To analyze such expressions, semantic network analysis[1] and Natural Language Processing (NLP) are both techniques that are typically used to visualize and quantify semantic links among different concepts expressed in a textual corpus [7]. Previous works [3–5] propose user interfaces capable of analyzing textual documents via semantic networks analysis; more recently, they integrate semantic networks and NLP techniques such as topic modeling [17]. Most of

[1] Also called Network-Text Analysis (NTA) [19] when referred to networks created with measures of proximity between concept, or Socio-Semantic Networks when referred to social media text data [11].

the previous approaches construct the semantic network using the co-occurrence matrix. This matrix can capture links between concepts when they appear close to each other or within a user-defined window size of surrounding words.

However, the co-occurrence matrix fails in capturing semantic links that may exist among concepts that do not appear close or inside a window size. For instance, considering the document *"Rome is the capital of Italy ... Paris is the capital of France,"* one might assume that there is a semantic link between *"Rome"* and *"Paris"* because both are capital cities, even if they do not appear close to each other but both appear close to the word *"capital"*. One practical way to address this challenge without any pre-defined taxonomy is to represent words of a document using numerical distributed representation obtained with artificial neural networks models, so-called, word embeddings [16]. Leveraging these techniques, WINS aims at combining the traditional co-occurrence approach with the word embeddings approach to generate semantic networks representative of an input textual corpus for exploring and interacting with semantic information.

3 Methods

Semantic networks can be constructed with different numerical approaches. A typical approach is the one that uses the co-occurrence of words in a document to create links between nodes, where nodes are elementary textual units such as words or any idiomatic expression[2]. More formally, a semantic network is a graph $G(N, E)$ composed by a set of nodes N and a set of edges E. The generic set of edges E can be generated according to different estimates of semantic proximity. The traditional approach uses co-occurrence matrix as the adjacency matrix of an undirected graph $G(N, E)$. With this approach, a link is added among two different words w_i and w_j, if these words appear together or within a user-defined window (Fig. 1).

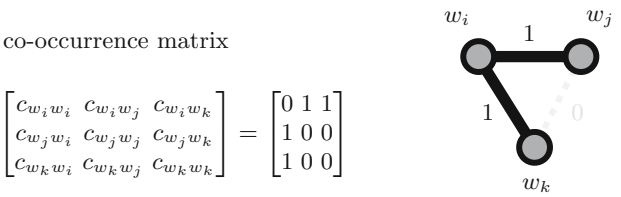

Fig. 1. On the left: the co-occurrence matrix where a generic element $c_{w_iw_j}$ is equal to the number of times w_i and w_j appear close to each other or within a user-defined window size. On the right: the resulting semantic network of co-occurrences.

[2] Chunking, n-gramming are text pre-processing phases to segment raw text into these units [2].

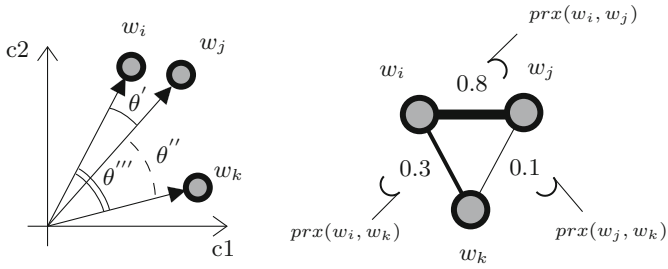

Fig. 2. On the left: vector representations of words in $\mathbb{R}^2$. On the right: semantic network generated with vector proximity measures as weights for the edges.

On the other hand, the second approach that is proposed to generate a semantic network is based on the following steps: 1) the textual units in the documents are vectorized and distributed in a euclidean space $\mathbb{R}^n$ via the artificial neural network method by Mikolov et al. [16]; then, 2) a measure of proximity among vectors of $\mathbb{R}^n$ such as ones' complement of a spatial distance measure is calculated and used as a weight for each of the edges[3]. Eventually, the user can vary a threshold to add a cut-off on the number of edges (Fig. 2).

WINS includes both the described approaches in its architecture. When the semantic network is created, the user can extract additional metric based on theory. Leveraging both Network Science and NLP, the design of WINS intends to enable the user to an interactive exploration of concepts' relationships contained in text data in order to support analysis for research and academic purposes.

3.1 Interface Design

The user selects the textual document to be analyzed using a dedicated card[4]. This card contains three different options to select a text document: 1) upload of a file, which can be a text file or a PDF file; 2) application of a filter on a database of patents, papers, and news via keywords' queries; 3) selection a Wikipedia page article via an embedded search bar. Then, two options for the construction of the semantic network are provided.

One approach consists in using the word-pair co-occurrence [3]. The other approach uses a distributed vectorization technique, i.e. Word2Vec [16], to transform words into vectors and construct the semantic network with vectors' spatial proximity. Afterward, the interface processes the previous inputs and returns a visual representation of the semantic network, and three different files available for download: (ii) a spreadsheet with network analysis metrics (e.g. *centrality measures*), (iii) a Graph Modeling Language file that can be opened with external network software, (iv) an interactive HTML file for visually exploring the network within the web browser.

[3] Proximity can be quantified with any spatial proximity measurement; for example, using *euclidean* distance, or *cosine* distance.

[4] The term "card" refers to the HTML division class used for the aesthetic layout design.

4 Results

The proposed interface embeds two main tabs. Figure 3 shows the first tab with its respective features for gathering/uploading text documents, constructing the network, visualizing the network, and saving results to the local environment of the user. The second tab, on the other hand, is populated when the network is constructed using a vectorization algorithm (Word2Vec [16]). This tab allows the user to interact visually with the vectors and apply additional transformation and measures (see Fig. 4) such as measuring the spatial proximity.

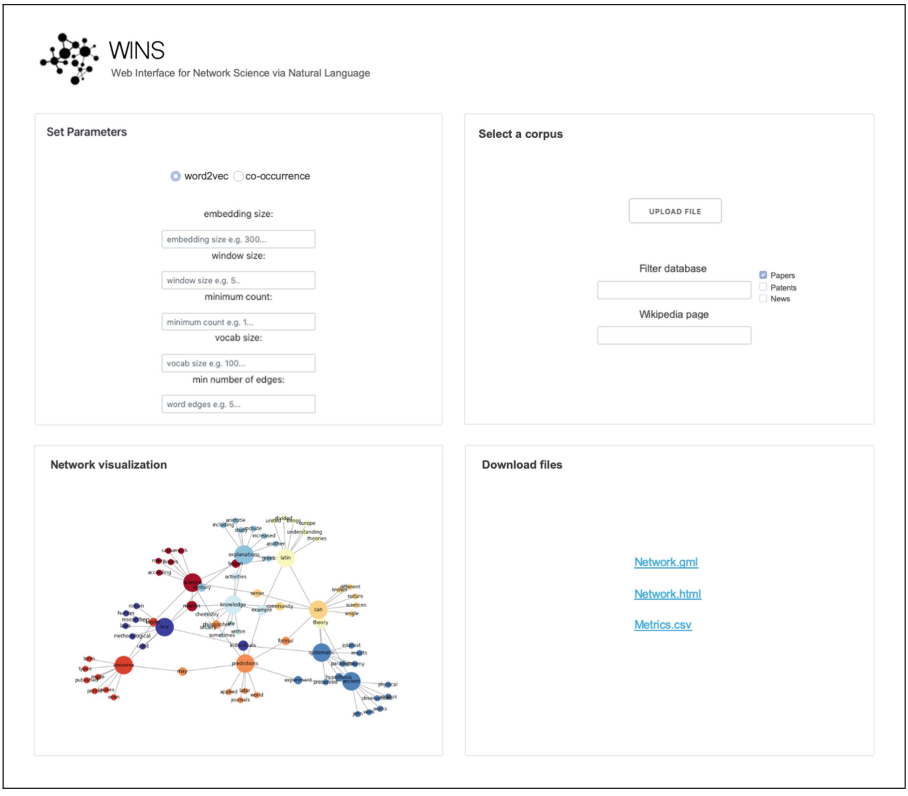

Fig. 3. Semantic network generation

4.1 Network Construction and Visualization

The upper-left card of Fig. 4 shows two options to select the approach for creating a semantic network. As introduced in the Methods section, these approaches are *word2vec* and *co-occurrence*. Then, inside the upper-left card of Fig. 3, the parameters that the user can enter for customizing the algorithm can be inserted: *embedding size* is the number of components of each generated vector, *window*

size is the number of contextual words to consider as surrounding words to a given word, *vocab size* is a filter on the maximum number of nodes in the network, *min number of edges* is a filter on the number of arcs desired by the user. The upper-right card, *select a corpus*, takes the user input for the selection of a text document. Finally, in the lower-left card, a visualization of the semantic network is shown and, in the lower-right card, hyperlinks for downloading output files are provided. In the example in Fig. 3, a *Wikipedia page* query for "science" and the default parameters for *word2vec* have been used. Colors of nodes represent clusters detected using community detection algorithms [1,9].

Fig. 4. Vectors visual exploration

4.2 Semantic Space Exploration

The tab shown in Fig. 4 enables the exploration of vectors generated with *word2vec*. The example shows the output of a word embedding model where the user added a temporal tag to each word in a pre-processing phase. When applying *word2vec* to this document, the algorithm generates vectors for words tagged with a temporal dimension. This allows the exploration of these vectors using a time filter. On the upper left, the *Embedding Matrix* card summarizes the vectors in a table that can be filtered by the user. On the upper right side of

the interface, a 2-D[5] plot is displayed for the time-interval: $[t_0, t_1] = [2015; 2016]$ is reported. The time interval $[t_0, t_1]$ can be selected by the user through the slider placed under the chart. In the lower side of this tab, the user can interact with additional tabs to perform a fitting of the vectors, and to observe which are the most similar vectors to those predicted by the fitting model.

5 Conclusions and Future Works

Previous works have documented the effectiveness of using semantic networks and NLP to analyze large textual document to extract insights about the structure of the content and its meaning [17]. These techniques have been used in a variety of research fields ranging from technological forecasts [12], national security [13], social media analysis [10,18].

In this paper, a web-interface that integrates Network Science with Natural Language Processing has been introduced. This interface allows users to an augmented interaction with text files enabling the possibility to explore semantic relationships. With the proposed interface, the user can construct semantic networks and word embeddings starting from digital documents. Semantic network analysis is adopted to capture semantic relationships between the unitary components of a text, which can be words, idioms, or even symbols. Additional functionalities of word embeddings allow the user to combine the semantic network analysis with a semantic spatial analysis. It is worth noting that the approach introduced has to be tested within different domain-specific text documents, and compared with traditional approaches in order to gain a deeper understanding of the outputs that it generates. Moreover, the introduction of a feature for taking into account the dynamical nature of concepts, and the diachronic variations of language, using both networks and word embeddings, is a future challenge to address. The interface is a work-in-progress idea with the goal of developing a first prototype to grant web access to the interface for research and academic purposes.

Acknowledgements. This material is based upon work supported, in whole or in part, by the U.S. Department of Defense through the Office of the Assistant Secretary of Defense for Research and Engineering (ASD(R&E)) under Contract [HQ0034-19-D-0003, TO#0150].

References

1. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. **2008**(10), P10008 (2008)
2. Borrelli, D., Gongora Svartzman, G., Lipizzi, C.: Unsupervised acquisition of idiomatic units of symbolic natural language: an n-gram frequency-based approach for the chunking of news articles and tweets. Plos one **15**(6), e0234214 (2020)

[5] Dimensionality of vectors has been reduced to two components. This can be achieved with different techniques with some limitations as discussed by the authors of [14].

3. Danowski, J.A.: WORDij: a word-pair approach to information retrieval. NIST Special Publication, no. 500207, pp. 131–136 (1993)
4. Diesner, J.: Context: software for the integrated analysis of text data and network data. Social semantic networks in communication research (2014)
5. Diesner, J., Carley, K.M.: AutoMap 1.2: extract, analyze, represent, and compare mental models from texts. Carnegie Mellon University, School of Computer Science, Institute for ... (2004)
6. Doerfel, M.L.: What constitutes semantic network analysis? A comparison of research and methodologies. Connections **21**(2), 16–26 (1998)
7. Drieger, P.: Semantic network analysis as a method for visual text analytics. Procedia-Soc. Behav. Sci. **79**(2013), 4–17 (2013)
8. Harris, Z.: Distributional structure. Word **10**(2–3), 146–162 (1954). Reprinted in Fodor, J.A., Katz, J.J. (eds.) Readings in the Philosophy of Language
9. Khanfor, A., Ghazzai, H., Yang, Y., Massoud, Y.: Application of community detection algorithms on social internet-of-things networks. In: 2019 31st International Conference on Microelectronics (ICM), pp. 94–97. IEEE (2019)
10. Lipizzi, C., Dessavre, D.G., Iandoli, L., Marquez, J.E.R.: Towards computational discourse analysis: a methodology for mining twitter backchanneling conversations. Comput. Hum. Behav. **64**, 782–792 (2016)
11. Lipizzi, C., Iandoli, L., Marquez, J.E.R.: Extracting and evaluating conversational patterns in social media: a socio-semantic analysis of customers' reactions to the launch of new products using Twitter streams. Int. J. Inf. Manag. **35**(4), 490–503 (2015)
12. Lipizzi, C., Iandoli, L., Marquez, J.E.R.: Combining structure, content and meaning in online social networks: the analysis of public's early reaction in social media to newly launched movies. Technol. Forecast. Soc. Change **109**, 35–49 (2016)
13. Lipizzi, C., Verma, D., Korfiatis, G., Borrelli, D., Capela, F., Clifford, M., Desai, P., Giffin, R., Hespelt, S., Hoffenson, S., et al.: Meshing capability and threat-based science and technology (S and T) resource allocation. Technical report, Stevens Institute of Technology Hoboken United States (2019)
14. Liu, S., Bremer, P.T., Thiagarajan, J.J., Srikumar, V., Wang, B., Livnat, Y., Pascucci, V.: Visual exploration of semantic relationships in neural word embeddings. IEEE Trans. Vis. Comput. Graph. **24**(1), 553–562 (2017)
15. Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. Behav. Res. Methods Instrum. Comput. **28**(2), 203–208 (1996)
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
17. Paranyushkin, D.: Infranodus: generating insight using text network analysis. In: The World Wide Web Conference, pp. 3584–3589 (2019)
18. Primario, S., Borrelli, D., Iandoli, L., Zollo, G., Lipizzi, C.: Measuring polarization in Twitter enabled in online political conversation: the case of 2016 us presidential election. In: 2017 IEEE International Conference on Information Reuse and Integration (IRI), pp. 607–613. IEEE (2017)
19. Taskin, Y., Hecking, T., Hoppe, H.U.: ESA-T2N: a novel approach to network-text analysis. In: International Conference on Complex Networks and Their Applications, pp. 129–139. Springer, Cham (2019)