



Towards Pattern Recognition with Network Science and Natural Language Processing for Information Retrieval

Muskan Garg¹^a, Mukesh Kumar²^b and Debabrata Samanta³^c

¹Artificial Intelligence & Informatics, Mayo Clinic, Rochester, MN, U.S.A.

²University Institute of Engineering & Technology, Panjab University, Chandigarh, India

³Department of Computational Information Technology, Rochester Institute of Technology, Kosovo

Keywords: Co-Word Analysis, Graph Theory, Information Retrieval, Linguistic Analysis, Pattern Recognition.


Abstract: A surge in text-based information retrieval such as topic detection and tracking has increasingly shown growth from static to dynamism in the last decade. We posit the need of investigating an interdisciplinary approach of network science and natural language processing for graph-based information extraction. Post-lockdown era, it makes sense to consider Graph of Words (GoW) evolved from user-generated text from social media platforms amid increase in the internet traffic. The idea is to unfold the latent patterns in graph-based text representation with limited resource availability resulting in effective models, in comparison of computationally expensive pre-trained models, limited to a certain type of information extraction. As a solution towards advancing statistical approach for language independent models, we plot three different information retrieval applications: (i) Structural analysis: find unique patterns in domain/ language/ genre-specific GoW for keyword extraction, (ii) Language independence: design objective function for language-independent information retrieval, (iii) Dynamism: mathematical modeling for concept-drift and evolving trends/ events in dynamic GoW evolved from streaming data. We associate recent developments and open challenges with our position as potential research direction.


1 INTRODUCTION


Evolution of text mining from rule-based analysis to learning based mechanism has enriched this domain with NLP-centered applications. To investigate latent associations among words, we pose the integration of network science and natural language processing for syntactic pattern analysis. Such patterns facilitate information retrieval and suggest the viability of NLP-centered information retrieval applications with cost and time-effective solutions for NLP-centered applications with streaming data. The development in the field of natural language processing points to responsible AI information extraction methods. The pre-trained model inherit theoretical limitations of its ability to model certain *type of information*, to transformer-based models (Chernyavskiy et al., 2021). The sizeable improvements in existing models are expensive hindering the environment with design

of the next generation deep NLP architectures. A major challenge with use of learning-based mechanisms and transformer-based models for NLP-centered applications is the *language dependency*. To handle this, we need a Layer of Language-Independence (LLI). We post this LLI as a potential encoding of any given language in a high-dimensional vector. We further discuss its applicability to real-time streaming data. The structural analysis of complex networks enables mathematical modeling of its dynamics through *preferential attachment* model. Thus, formalizing different languages through network science facilitates many NLP-centered applications such as inferring mental states, topic detection and tracking, concept drift, NLP-based recommendation systems and many more.

Motivation. Past studies perform spectral analysis for GoW through adjacency matrix to glean eigenvalues. The proportion of difference between subsequent eigenvalues is inversely proportional to the size of GoW (Liang, 2017). To this end, a set of all eigenvalues defines *spectrum*. Eigenvalues depicts semantics of GoW which is a key component in ex-

^a <https://orcid.org/0000-0003-0075-9802>

^b <https://orcid.org/0000-0002-4920-670X>

^c <https://orcid.org/0000-0003-4118-2480>

isting random-walk based keyword extraction mechanisms (Campos et al., 2020). (Garg and Kumar, 2018a) illustrates the significance of edge weights over node degree distributions in path-based networks such as GoW superseding the traditional random walk mechanism on hub-authority based core-periphery structure. To this end, we examine structural difference in GoW evolved for documents of various predefined languages.

2 OUR POSITION

Although, the tools of Natural Language Processing (NLP) experienced a major shift during 1990's in transitioning from rule-based methods to statistical approaches, most of today's NLP research focuses on 20 of the 7000 languages of the world, leaving the vast majority of languages under-explored (Magueresse et al., 2020). We posit the integrated studies of network science and natural language processing in three categories:

1. **Structural Analysis:** find unique patterns in domain/ language/ genre-specific GoW for keyword extraction.
2. **Language Independence:** design objective function for language-independent information retrieval.
3. **Dynamism:** mathematical modeling for concept drift and evolving trends/ events in dynamic GoW evolved from streaming data.

For any given document D , consider a Graph of Words (GoW) G representing unique words in vocabulary as nodes $V = \{1, 2, 3, \dots, n\}$ and co-occurring words in a given document as edges $E = \{E_1 : (V_1, V_2), E_2(V_2, V_3), \dots, E_m = (V_{n-1}, V_n)\}$. The network science metrics and models on GoW glean insights for information retrieval tasks. Based on latent patterns, we further pose an *objective function* to optimize the parameters such that a given text T_i that belongs to one of the languages given in predefined set L is processed through objective function and converted in low-dimensional embedding vector. Then, pre-trained models for information retrieval tasks including keyword extraction, topic detection, and concept drift are provided an input of a low-dimensional embedding vector.

3 STRUCTURAL ANALYSIS

In graph-based representations, complete graph among all words of a sentence is often avoided in past

studies because its a common belief from linguistic studies that a gap between two words is inversely proportional to their relevancy (Hulth, 2003). A graph-based textual representation are (i) Sliding window-based GoW, (ii) Context aware graph, (iii) Multi layer network, (iv) Line graphs (Garg, 2021), (v) Knowledge graphs and (vi) Semantic graphs which we discuss as follows:

1. **Sliding Window-Based GoW.** A graph-based text representation with edges between two words co-occurring within a predefined window called sliding window n . A special case of GoW construct Word-Adjacency Graph (WAG) or Word Co-occurrence Network (WCN) for sliding window $n = 2$ delineating connections between adjacent terms.
2. **Context Aware Graph.** Entailment is defined as a phenomenon where consecutive sentences use the context set by a given sentence, imparting continuity in communication (Duari and Bhatnagar, 2019). A window slides over two consecutive sentences to (i) capture the contextual co-occurrence of words, and (ii) eliminate the need of integer-valued window-size parameter.
3. **Multilayer Network.** A potential network construction happens in multiple layers to represent words, syllable and grapheme in different layers for natural language understanding for different languages such as Croatian and English (Martinčić-Ipšić et al., 2016).
4. **Line Graphs.** Line graphs represents co-occurrence as a node in and word as an edge of in the graph (Harary and Norman, 1960).
5. **Knowledge Graph.** Knowledge graphs use a graph-based data model to capture knowledge in NLP-centered applications employing benefits of graph-based abstraction over relational models (Hogan et al., 2021).
6. **Semantic Graph.** Semantic languages and models are increasingly used in order to describe, represent and exchange data in multiple domains and forms through RDF knowledge bases (Čebirić et al., 2019).

Inferences. We observe network science metrics for the structural aspects of GoW evolved for different genres: (i) Microblogs, (ii) essays, (iii) novels, (iv) science articles, and (v) news reports in Table 1. We study the chief characteristic feature of WCN to examine the behavioural aspects of GoW. The nature of graph-based text representation is un-directed and un-weighted. We use *First Story Detection (FSD)* dataset (Petrović et al., 2010) to compare structural aspects

Table 1: Network metrics in GoW for different genres.

Network Metrics	Microblogs	Essays	Novels	Science Articles	News Reports
Length	183466	1142	5224	961	748
#Nodes	34925	440	1314	426	343
#Edges	116477	826	3199	734	581
Average degree	2.166	3.26	4.69	3.32	3.35
ASPL/ASPL _r	1.258/ 3.366	3.61/ 4.70	3.29/ 4.60	3.92/ 5.04	3.91/ 4.83
CC%/CC _r %	0.493/ 0.00039	10.61/ 0.97	17.62/ 0.45	8.15/ 0.87	7.16/ 1.02

of the Microblog WCN with different genres: essays, novels, popular science articles, and news reports. We use first 25000 tweets in FSD for experiments and evaluation. With decrease in the number of nodes in GoW, the value of CC for WCN to Erdos-Renyi model decreases suggesting scalability for all genres. The GoW for all genres follow the small-world property. Furthermore, we enlist following open challenges and research gaps for handling different genres:

1. Variations in the average degree of nodes in GoW for different genre illustrates the disparity in vocabulary and thus, patterns among words.
2. The extent to which observed metrics differs plays a pivotal role in tracking behavioral and structural aspect for different genres.

4 LANGUAGE INDEPENDENCE

With evolving digitization and surge in globalization in multilingual countries having more than one official languages points to a dire need of constructing the language-independent platforms for NLP-centered applications. N-gram a language-independent method paves a path to exploit statistical measures and graph theory to find interesting patterns among words of different languages (Majumder et al., 2002). On the other hand, pre-trained models for NLP-centered applications have proved their effectiveness for different languages, genres and problem domains. To resolve language-dependency, we pose a Layer for Language Independence (LLI) which converts a given text in of any language from predefined set of languages into a common representation. Consider a given text T

$$T = \{T_{1(L_1)}, T_{2(L_2)}, T_{3(L_3)}, T_{4(L_2)}, T_{5(L_1)}, \dots, T_{x(L_2)}\} \quad (1)$$

where a text T_i for $i = \{1, 2, 3, \dots, x\}$ belongs to a predefined set of given languages $L = \{L_1 : \text{English}, L_2 : \text{Spanish}, L_3 : \text{Hindi}, \dots\}$. Consider GoW (G_i) for a given text T_i of language L_j , we encode the graph-based representation of T_i by employing vector embedding such as graph2vec or node2vec such that

the final embedding results into alike outputs for every text having similar information in any predefined language.

4.1 Related Work

Centrality-based network metrics derive many language-independent methods for NLP-centered applications (Beliga et al., 2015) deducing the need of natural language understanding with latent patterns among words in GoW. The graph-based information extraction moves from graph theory to soft computing by deploying well-known traditional methods such as Random-walk (Vega-Oliveros et al., 2019; Biswas et al., 2018) and Markov decision process (Garg and Kumar, 2022). Recent advancements with semantic graphs, knowledge graphs and pre-trained models, enrich this domain even more. We unify different languages by structuring given texts in graph-based representation enabling vector embeddings such as node2vec, graph2vec and x2vec (Grohe, 2020).

4.2 Inferences

The structural variations in GoW for different languages is illustrated in Table 2 through average degree, Average Shortest Path Length (ASPL), and Clustering Coefficient (CC). The GoW construction for parallel texts of 12 Slavic languages and 2 non-Slavic languages compare and contrast the behaviour of GoW for formal texts in different language (Liu and Cong, 2013). To test the robustness of informal texts, (Garg and Kumar, 2018b) examines the structure of Microblog WCN thereby introducing BARank, a keyphrase extraction mechanism. We compare the structure of WCN in 14 different languages as shown in Table 2.

As per literature, the GoW evolved from informal (social media Twitter data) is defined as Microblog WCN for edges among words in sliding window of 2. A consistent value of ASPL validates the average length of words in the range of 3 and 4 for various languages, both formal and informal. The value of average degree and CC in Microblog WCN and English

Table 2: The structural aspect for GoW evolved for various languages, both formal and informal.

Language	Average Degree	ASPL	CC
Microblog	2.166	3.144	0.076
Belarusian	4.819	3.797	0.100
Bulgarian	5.690	3.354	0.186
Chinese	8.684	2.944	0.283
Croatian	5.353	3.479	0.151
Czech	4.945	3.627	0.199
English	9.043	2.964	0.299
Macedonian	6.206	3.225	0.220
Polish	4.983	3.628	0.118
Russian	4.504	3.891	0.091
Serbian	5.348	3.485	0.147
Slovak	5.166	3.592	0.128
Slovenian	5.367	3.406	0.164
Ukrainian	4.865	3.814	0.096
Upper-Sorbian	5.347	3.550	0.131

WCN is approximately in the ratio of 1 : 4. Such variations illustrate difference in semantics of GoW for different languages. We thus, pose the need to study the structural aspects of texts for different language, domain and genres. Furthermore, we enlist following open challenges and research gaps for handling different languages to construct LLI:

1. **Language.** The words and grammar which are used in a language are based on the script which it follows. Every script may have latent patterns of connectivity among words.
2. **Script.** There are variations in the number of basic units (Alphabets) and the number of dialects.
3. **Vocabulary.** The active vocabulary may vary for each language may vary such as only 33% of the total words in English dictionary is actively used by human beings. This active vocabulary limits the perplexity facilitating strong connections among words with latent patterns and thus, better understanding.
4. **Case Variations:** Challenging special features such as Hungarian words have 18 to 35 cases that presume meaning of the a specific through the context of its use.
5. **Word Order:** Most languages depend on word order to convey meaning such as Warlpiri, a language of indigenous Australians. Semantics of GoW may illustrate the latent patterns in word orders.

5 DYNAMISM

With major focus on Television Rating Point (TRP) by traditional news media, they broadcast viewers' choices in lieu of current happenings (Petrović et al., 2010). After extensive academic research on retrospective event detection from Twitter (Atefeh and Khreich, 2015), NLP research community shifts the focus to streaming data (Garg and Kumar, 2016). Network science metrics such as k-core decomposition (Meladianos et al., 2015), network cliques (Rousseau, 2015) and assortativity (Garg and Kumar, 2018b) glean cohesiveness in GoW for information extraction. The cohesiveness in GoW captures dense sub-graphs and thus paves a path to construct dynamic models. The structural analysis set foundation to model dynamism of words and their connectedness through dynamic models such as preferential attachment. Recently, (Saeed et al., 2019) contribute towards dynamics of GoW through heartbeat graph approach.

For a set of i documents $D = \{d_1, d_2, d_3, \dots, d_i\}$, consider a Graph of Words (GoW) G representing unique words in vocabulary as nodes V

$$V = \{1, 2, 3, \dots, n\} \quad (2)$$

and co-occurring words in a given document as edges E :

$$E = \{E_1 : (V_1, V_2), E_2(V_2, V_3), \dots, E_m = (V_{n-1}, V_n)\} \quad (3)$$

We further introduce growth and decay mechanism for G as follows:

1. **Growth.** A graph G is updated for every time interval (t_a) where a chunk of documents is collected, every word and connections are identified to connect them with G resulting in graph at time instance (t_a) as G_{t_a} .
2. **Decay.** For computational efficiency, it becomes seemingly important to eliminate unimportant nodes and edges from G_{t_a} .

This iterative process of growth and decay function highlights significant information as we observe dense sub-graphs due to high cohesiveness among nodes in sub-graphs. We suggest employing approximation algorithms for information retrieval.

6 CONCLUSION

The past work in (i) types of GoW (Garg, 2021), (ii) examining semantics of languages-specific GoW such as Chinese and English (Liang et al., 2009),

12 Salvic languages (Liu and Cong, 2013) and Microblogs (Garg and Kumar, 2018b); (iii) gleam insights from GoW for various domains such as Text Authorship (Akimushkin et al., 2017); (iv) applicability of network science metrics such as power-law (Choudhury et al., 2010), spectral distribution (Liang, 2017), has enriched this domain to investigate information retrieval approach towards unifying language-specific text in language-independent vector representation.

Due to versatility of the project and need of native-language experts, a major challenge for introducing a generic model on GoW is the construction of dataset: samples and annotations. However, the access to in-build NLTK corpus in Python language facilitates the structural analysis of text documents in different language. Furthermore, we found new multilingual datasets for different NLP-centered tasks such as MIRACL dataset¹ for information retrieval and XCOPA² for causal commonsense. Such datasets facilitate structural analysis to find unique patterns in domain/ language/ genre-specific GoW for keyword extraction. Structural analysis act as foundation to design language-independent objective function for information retrieval. An alternative application of dynamism construct mathematical modeling for concept drift and evolving trends/events.

REFERENCES

- Akimushkin, C., Amancio, D. R., and Oliveira Jr, O. N. (2017). Text authorship identified using the dynamics of word co-occurrence networks. *PloS one*, 12(1):e0170527.
- Atefeh, F. and Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164.
- Beliga, S., Meštrović, A., and Martinčić-Ipšić, S. (2015). An overview of graph-based keyword extraction methods and approaches. *Journal of information and organizational sciences*, 39(1):1–20.
- Biswas, S. K., Bordoloi, M., and Shreya, J. (2018). A graph based keyword extraction model using collective node weight. *Expert Systems with Applications*, 97:51–59.
- Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., and Jatowt, A. (2020). Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.
- Čebirić, Š., Goasdoué, F., Kondylakis, H., Kotzinos, D., Manolescu, I., Troullinou, G., and Zneika, M. (2019). Summarizing semantic graphs: a survey. *The VLDB journal*, 28(3):295–327.
- Chernyavskiy, A., Ilvovsky, D., and Nakov, P. (2021). Transformers: “the end of history” for natural language processing? In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 677–693. Springer.
- Choudhury, M., Chatterjee, D., and Mukherjee, A. (2010). Global topology of word co-occurrence networks: Beyond the two-regime power-law. In *Coling 2010: Posters*, pages 162–170.
- Duari, S. and Bhatnagar, V. (2019). scake: semantic connectivity aware keyword extraction. *Information Sciences*, 477:100–117.
- Garg, M. (2021). A survey on different dimensions for graphical keyword extraction techniques. *Artificial Intelligence Review*, pages 1–40.
- Garg, M. and Kumar, M. (2016). Review on event detection techniques in social multimedia. *Online Information Review*.
- Garg, M. and Kumar, M. (2018a). Identifying influential segments from word co-occurrence networks using ahp. *Cognitive Systems Research*, 47:28–41.
- Garg, M. and Kumar, M. (2018b). The structure of word co-occurrence network for microblogs. *Physica A: Statistical Mechanics and its Applications*, 512:698–720.
- Garg, M. and Kumar, M. (2022). Kest: A graph-based keyphrase extraction technique for tweets summarization using markov decision process. *Expert Systems with Applications*, 209:118110.
- Grohe, M. (2020). word2vec, node2vec, graph2vec, x2vec: Towards a theory of vector embeddings of structured data. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 1–16.
- Harary, F. and Norman, R. Z. (1960). Some properties of line digraphs. *Rendiconti del Circolo Matematico di Palermo*, 9(2):161–168.
- Hogan, A., Blomqvist, E., Cochez, M., d’Amato, C., Melo, G. d., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., et al. (2021). Knowledge graphs. *ACM Computing Surveys (CSUR)*, 54(4):1–37.
- Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 216–223.
- Liang, W. (2017). Spectra of english evolving word co-occurrence networks. *Physica A: Statistical Mechanics and its Applications*, 468:802–808.
- Liang, W., Shi, Y., Chi, K. T., Liu, J., Wang, Y., and Cui, X. (2009). Comparison of co-occurrence networks of the chinese and english languages. *Physica A: Statistical Mechanics and its Applications*, 388(23):4901–4909.
- Liu, H. and Cong, J. (2013). Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin*, 58(10):1139–1144.
- Magueresse, A., Carles, V., and Heetderks, E. (2020). Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.
- Majumder, P., Mitra, M., and Chaudhuri, B. (2002). N-gram: a language independent approach to ir and nlp.

¹<https://project-miracl.github.io/>

²<https://github.com/cambridgeltl/xcopa#cite>

- In *International conference on universal knowledge and language*.
- Martinčić-Ipšić, S., Margan, D., and Meštrović, A. (2016). Multilayer network of language: A unified framework for structural analysis of linguistic subsystems. *Physica A: Statistical Mechanics and its Applications*, 457:117–128.
- Meladianos, P., Nikolentzos, G., Rousseau, F., Stavrakas, Y., and Vazirgiannis, M. (2015). Degeneracy-based real-time sub-event detection in twitter stream. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 248–257.
- Petrović, S., Osborne, M., and Lavrenko, V. (2010). Streaming first story detection with application to twitter. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*, pages 181–189.
- Rousseau, F. M. (2015). *Graph-of-words: mining and retrieving text with networks of features*. PhD thesis.
- Saeed, Z., Abbasi, R. A., Razzak, M. I., and Xu, G. (2019). Event detection in twitter stream using weighted dynamic heartbeat graph approach [application notes]. *IEEE Computational Intelligence Magazine*, 14(3):29–38.
- Vega-Oliveros, D. A., Gomes, P. S., Milios, E. E., and Berton, L. (2019). A multi-centrality index for graph-based keyword extraction. *Information Processing & Management*, 56(6):102063.

