Comparison between the Structures of Word Co-occurrence and Word Similarity Networks for Ill-formed and Well-formed Texts in Taiwan Mandarin

Huang, Po-Hsuan[1, 2][0000-0002-9044-8196] and Shao, Hsuan-Lei[2][1111-2222-3333-4444]

[1] Graduate Institute of Linguistics, National Taiwan University, 10617 Taipei, Taiwan
[2] Department of East Asian Studies, National Taiwan Normal University, 10646 Taipei, Taiwan
hlshao@ntnu.edu.tw

Abstract. The study of word co-occurrence networks has attracted the attention of researchers due to their potential significance and applications. Understanding the structure of word co-occurrence networks is therefore important to fully realize their significance and usages. In past studies, word co-occurrence networks built on well-formed texts have been found to possess certain characteristics, including being small-world and disassortative, and following a two-regime power law distribution. On the flip side, word co-occurrence networks built from ill-formed texts have been found to behave differently. While both kinds of word co-occurrence networks are small-world and disassortative, the latter are scale-free and follow the power law instead of the two-regime power law distribution. However, since past studies only investigated English, the cross-linguistic universality of such characteristics remains unknown. In addition, it is yet to be investigated whether there could be possible similitude/differences between word co-occurrence networks and other comparable networks. This study therefore investigates and compares the structure of word co-occurrence networks and word similarity networks based on Taiwan Mandarin ill-formed/well-formed texts, and seeks to explore whether the three aforementioned properties (scale-free, small-world, and disassortative) for the two types of texts are universal among languages and between word co-occurrence and word similarity networks.

Keywords: Word Co-occurrence Network, Word Similarity Network, Taiwan Mandarin, Structure.

1 Introduction

Word co-occurrence networks (WCN) have attracted the attention of researchers due to both their potential significance (e.g., semantic similarity [1]) as well as their applications (e.g., keyword extraction, text summarization, and author affiliation, cf. [2]). The understanding of the structure of WCN is therefore crucial if one is to grasp a holistic picture of their significance and applications. Moreover, it is equally important to explore the possible similitudes/differences between WCN and other potentially comparable networks. In this study, we investigate and compare WCN and word similarity networks (WSN) for a Taiwan Mandarin internet forum, PTT and for judicial

judgments made by Taiwanese courts from the years 2004 and 2008. In past studies, WCN based on well-formed documents have been found to share certain properties, including being small-world [3, 4], following a two-regime power law distribution [5], and being generally disassortative [6]. On the other hand, WCN built with less well-formed microblog data in English has been found to behave differently than WCN based on well-formed documents. While both kinds of WCN are small-world and disassortative, word co-occurrence networks built from ill-formed texts are scale-free and follow the power law distribution instead of the two-regime power law distribution [2]. However, whether such likeness and discrepancies between WCN for well-formed and ill-formed texts can be universally found across languages requires further investigation. In addition, it remains to be seen whether such similarities and differences are reserved for WCN or are in fact shared among other networks such as networks based on word similarity. As such, the current study seeks to examine 1) the universality of the similarities and differences for the three parameters (degree distribution, small-worldness, and disassortativity) among different languages and 2) the universality of the three properties between WCN and WSN.

2 Methods

2.1 Data Collection

For the PTT data, 139,578 posts from the *Gossiping*, *Food*, and *HatePolitics* forums on PTT were collected between January 1st to July 31st, 2023. With the comments included, the dataset contained a total of 4,148,879 texts. For the judicial judgment data, 53,272 judgments during the years 2004 and 2008 were collected from the Judicial Yuan, R.O.C. Sentences were segmented with punctuation, leading to a total of 4,017,811 texts.

2.2 Data Preprocessing

The texts were first preprocessed, with the numbers converted to *0*, the alphabets converted into lower cases, and all other non-Mandarin characters removed. The preprocessed texts were then segmented with the CKIP segmentation system [7].

2.3 Word Embedding

Word similarities were obtained with word embedding. A skip-gram word2vec model was first trained for the PTT data and judicial judgment data respectively, with a window size of 10. The vector sizes were 500. Only words with an occurrence more than 3 were taken into the vocabularies.

2.4 Network Building

Four networks were built. For the PTT and judicial judgment data respectively, one unweighted and undirected network was built based on word co-occurrence and word similarity respectively. These four networks were thus: a WCN and WSN for the PTT data (WCN-P and WSN-P) and a WCN and WSN for the judicial judgment data (WCN-J and WSN-J). To reduce the computational load, only 10% of the vocabulary was randomly selected as the database used for network building. For WCN, in the current study, the word co-occurrence was determined as two words that both occurred in the same text. For WSN, the similarity threshold was determined at the 99th centile, and two words were determined as similar if the similarity was above the threshold. The numbers of the nodes and edges in these four networks are listed in Table 1.

Table 1. Numbers of nodes and edges in the four networks.

Network	Number of nodes	Number of edges
WCN-P	24,035	208,759
WSN-P	228,163	43,106,445
WCN-J	16,847	1,469,475
WSN-J	86,262	43,248,161

2.5 Calculation of the Parameters

Degree Distribution
The degree distributions of the four networks were accessed by fitting the degree distributions of the networks to power law vs. two-regime power law models. The models were then assessed with goodness of fit using the sums of squared residuals (SSR) and AIC.

Small-world Property
The small-world property was determined by comparing the average clustering coefficients (CC) of the target network and the Erdos-Renyi (ER) random network [8]. A network is said to possess the small-world property if the CC in the target network is far larger than the CC in the ER random network ($\cong 0$; cf. [2]).

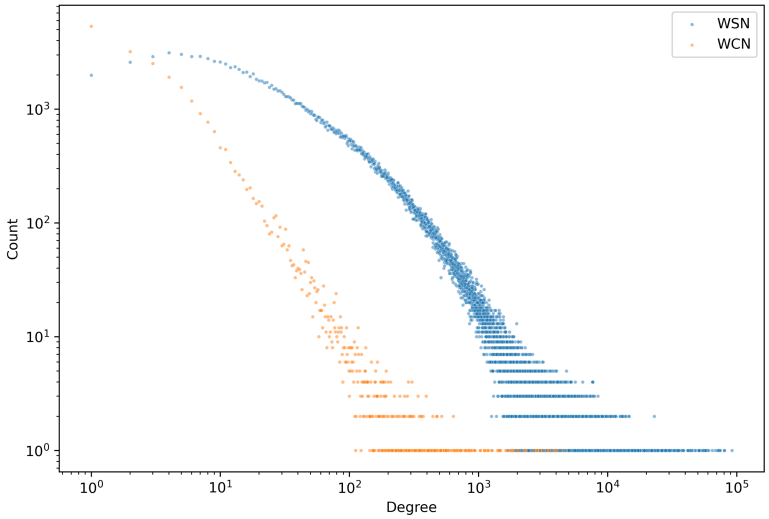
Assortativity
Lastly, the assortativity was decided with the degree assortativity coefficient (DAC), which indicated the tendency for a node to be connected with nodes with higher/lower degrees. If DAC is positive, the network is said to be assortative; if DAC is negative, the network is said to be disassortative. If the coefficient is close to zero, the network is said to be more randomly distributed in terms of its assortativity [2].

3 Results

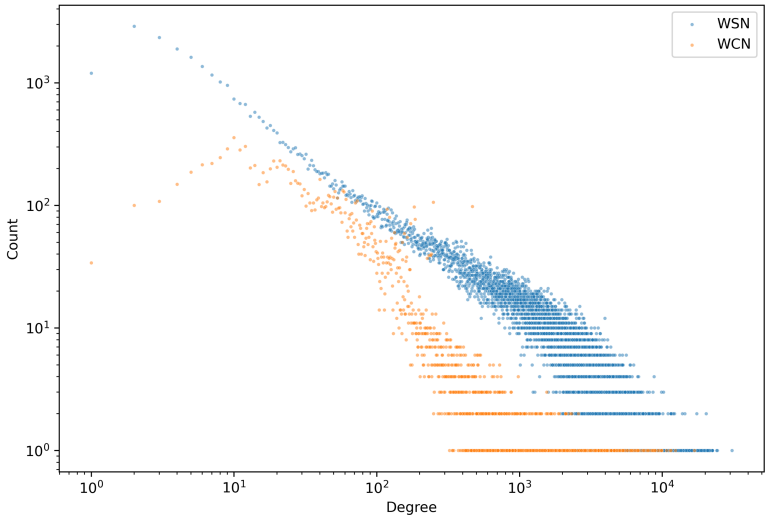
3.1 Degree Distribution

The degree distributions of the four networks are illustrated in Fig. 1.

Fig. 1. Degree distributions of word co-occurrence and similarity networks for the PTT and judicial judgment data.



(a) PTT data



(b) Judicial judgment data

It can be observed that while WCN-P seems to be quite straight-forwardly scale-free, as is found in past studies [2] for ill-formed texts, the other three networks seem to be rather ambiguous between being more similar to the power law distribution or to the two-regime distribution. However, upon examination, our goodness-of-fit results showed that all four networks were generally scale-free. As seen in Table 2, for all four networks, the SSR were lower for the fitted power law models than the fitted two-regime power law models, and the AIC were also all lower for the fitted power law models than the fitted two-regime power law models.

Table 2. Sums of squared residuals and AIC for fitted power-law and two-regime-power-law models for the four networks' degree distributions.

Network	Model	SSR	AIC
WCN-P	Power law	201.09	-284.67
	Two-regime power law	263.91	-169.74
WSN-P	Power law	4,199.72	-2,760.59
	Two-regime power law	4,910.08	-1,748.96
WCN-J	Power law	558.69	-1,000.83
	Two-regime power law	667.51	-774.55
WSN-J	Power law	3,660.20	-4,321.66
	Two-regime power law	4,395.24	-3,060.43

3.2 Small-worldness

The clustering coefficients for the four networks are listed in Table 3.

Table 3. Clustering coefficients for the four networks.

Network	CC
WCN-P	0.60
WSN-P	0.32
WCN-J	0.95
WSN-J	0.58

It can be seen that all four networks had clustering coefficients far larger than that of the ER random network, suggesting that all four models were small-world in nature.

3.3 Assortativity

The degree assortativity coefficients for the four networks are listed in Table 4.

Table 4. Degree assortativity coefficients for the four networks.

Network	DAC
WCN-P	-0.27
WSN-P	-0.18
WCN-J	-0.33
WSN-J	-0.04

The negative DAC of the four networks suggest general disassortativity. However, the near zero value of the DAC for WSN-J suggests that WSN-J is rather neutral in terms of assortativity compared with the other three networks. In addition, it can also be observed that in general, WCN networks are more disassortative than WSN networks, for both well-formed and ill-formed data.

The three investigated parameters are summarized in Table 5.

Table 4. Summary of the three parameters for the four networks.

Network	Degree distribution	Small-worldness	Assortativity
WCN-P	Scale-free	Small-world	Disassortative
WSN-P			
WCN-J			
WSN-J			Neutral

4 Discussion and Conclusion

4.1 Degree Distribution for Networks Based on Ill-formed and Well-formed Data

As mentioned in [2], past studies of WCN built with well-formed texts generally fit two-regime power law distributions. On the other hand, their WCN built with Twitter microblog data is scale-free, and follows the power law distribution. However, our results suggest that such scale-free property may not be reserved to networks for ill-formed texts. Our networks for judicial judgment data also showed scale-free properties. Since judicial judgments are undoubtedly well-formed, such properties might not be directly related to the well-formedness of the texts, but rather may be attributed to the specificity of the texts. Since similar power law distribution has been found for other specific texts such as bioinformatics literature [9], it is likely that in both ill-formed microblog texts such as Twitter and PTT data and specific texts such judicial judgment and academic literature data, a handful of specific words are connected to especially large numbers of words. In microblogs, such words may be acronyms or community-specific pronouns reserved for the community only. In specific texts, such words may be professional jargon. This might suggest that the determinant for degree distribution is the specificity of the texts, instead of well-formedness. This, however, would require further investigation.

4.2 Universality of the Three Parameters

The results of the analysis for the three parameters for both word co-occurrence networks and word similarity networks based on Taiwan Mandarin texts showed that the two networks had similar structures as the word co-occurrence networks built with ill-formed data as well as academic literature data in English in previous studies [2, 9]. This indicates that these characteristics of networks for ill-formed/specific texts are universal across languages and are potentially shared among different networks.

4.3 Potential different tendencies of word similarity networks

While both the word co-occurrence networks and the word similarity networks investigated in the current study demonstrated similar properties, different tendencies seemed to exist between the two kinds of networks. It can be noticed that, for the measured parameters, WSN seemed to behave differently from WCN investigated in previous studies. The WSN in the current study had a smaller CC and also a DAC closer to 0, suggesting that their small-worldness was less significant than that of WCN, and that they were also less disassortative than WCN. However, it remains to be seen whether such discrepancies between WCN and WSN can be extended to a more balanced corpus or are present only when based on microblogs or specific texts.

4.4 Conclusion

In this study, word network properties (i.e., degree distribution, small-worldness, and assortativity) were explored based on ill- vs. well-formed texts and co-occurrence vs. similarity in Taiwan Mandarin. Both similar and discrepant findings were attested compared to past studies on English data. Similar to past findings, the networks investigated in this study demonstrated general disassortativity and small-worldness. On the flip side, unlike what has been found in past studies, networks built on both the ill- and well-formed texts demonstrated scale-free properties, whether they were based on co-occurrence or similarity. Our study therefore suggests text specificity as a more revealing determinant for degree distribution as compared with well-formedness. In addition, our study also supports the cross-linguistic universality of the investigated network properties. By comparing domain-specific texts and ill-formed texts and exploring a language other than English, we hope to further shed light on the different structures of word networks, and provide potential research directions for future studies.

References

1. Lancia, F.: Word co-occurrence and similarity in meaning, http://www.mytlab.com/wcsmeaning.pdf, last accessed 2023/11/15
2. Garg, M. and Kumar, M.: The structure of word co-occurrence network for microblogs. Physica A: Statistical Mechanics and its Applications **512**, 698–720 (2018)
3. Masucci, A. P. and Rodgers, G. J.: Network properties of written human language. Physical review E: Statistical, nonlinear, and soft matter physics **74**(2), 026102 2006.

4. Kramer, A.: Dependency lengths in speech and writing: A cross-linguistic comparison via Youdepp, a pipeline for scraping and parsing Youtube captions. In: Proceedings of the Society for Computation in Linguistics 2021, pp. 359–365. Association for Computational Linguistics, online (2021)
5. Kapustin, V. and Jamsen, A.: Vertex degree distribution for the graph of word co-occurrences in Russian. In: Proceedings of TextGraphs-2: Graph-Based Algorithms for Natural Language Processing, Association for Computational Linguistics, Rochester, NY (2007)
6. Millington, T. and Luz, S.: Analysis and classification of word co-occurrence networks from Alzheimer's patients and controls. Frontiers in Computer Science **3**, 649508 (2021)
7. Tsai, Y.-F. and Chen, K.-J.: Reliable and cost-effective PoS-Tagging. Computational Linguistics and Chinese Language Processing **9**(1), 83–96 (2004)
8. Erdös, P. and Rényi, A.: On the evolution of random graphs. In: Newman, M., Barabási, A.-L., Watts, D.J. (eds.) The Structure and Dynamics of Networks, pp. 38–82. Princeton University Press, Princeton (2006)
9. Li, T., Bai, J., Yang, X., Liu, Q., Chen, Y. Co-occurrence network of high-frequency words in the bioinformatics literature: Structural characteristics and evolution. Applied Sciences 8(10), 1994 (2018)