Authorship Recognition with Short-Text using Graph-based Techniques

Laura Vanessa Cruz Quispe
Interinstitutional Center for Computational Linguistics
Institute of Mathematics and Computer Sciences
University of São Paulo / São Carlos
lcruzq@usp.br

Abstract

In recent years, studies of authorship recognition has aroused great interest in graph-based analysis. Modeling the writing style of each author using a network of co-occurrence words. However, short texts can generate some changes in the topology of network that cause impact on techniques of feature extraction based on graph topology. In this work, we evaluate the robustness of global-strategy and local-strategy based on complex network measurements comparing with graph2vec a graph embedding technique based on skip-gram model. The experiment consists of evaluating how each modification in the length of text affects the accuracy of authorship recognition on both techniques using cross-validation and machine learning techniques.

1 Introduction

Recent studies in the literature (Adamovic et al., 2019; dos Santos et al., 2017; Akimushkin et al., 2017; Martincic-Ipsic et al., 2017; Amancio et al., 2012; Stamatatos, 2009) show that human language is an adaptive complex system that can be modeled as a word co-occurrence network with the same characteristics of scale-free networks and small-world networks (Cancho and Solé, 2001). In text classification, measurements of complex networks are successful for the ability to capture semantic and syntactic information; for instance, betweenness measure can describe if a word is used in a broad or restricted context, identifying keywords in text (Amancio et al., 2011). However, despite the success of complex networks in text classification, it is still necessary improve some factors to allow the model to be able to deal with short-text issue since it could results that measurements of complex networks applied to linear networks become to less discriminant values (Amancio, 2014). We use a special case of text classification, authorship recognition that focuses on writing style of texts (Markowitz, 2019). We evaluate the robustness of two approaches complex network measurements and graph-embedding technique varying the length of text since it will modify the topology of graph that represents each text. We choose graph2vec technique over node2vec (Grover and Leskovec, 2016) since it can represent the entire graph on a single vector embedding. Moreover, it is also based on skip-gram model which works better than n-gram models (Sari et al., 2017; Escalante et al., 2011). Our evaluation is over co-occurrence words network for feature extraction to apply machine learning techniques to classify each book with their respectively author. Finally, it is evaluated using cross-validation technique.

2 Methodology

Our purpose is to model the writing style of each text using word co-occurrence networks to authorship recognition. Each word on books becomes a node and words that are adjacent in the text become an edge. We apply two approaches, complex network measurements: global strategy and local strategy and graph embedding: graph2vec. Our methodology is following described:

- **Complex Network Measurements**: (a) Global Strategy characterizes each co-occurrence word networks on a feature vector using the following value measures (X): betweenness (B), pagerank (PR), average shortest path (L), clustering coefficient (C), backbone symmetry ($Sb^{(h)}, h = 2, 3$), merged symmetry ($Sm^{(h)}, h = 2, 3$) and accessibility ($\alpha^{(h)}, h = 2, 3$), applied over each node i.

We calculate the average $\mu(X)$, standard deviation $\sigma(X)$ and skewness $\gamma(X)$ for each measurement to get a global representation of a network. (b) Local Strategy obtains the following measurements values (X): backbone symmetry ($Sb^{(h)}, h = 2, 3$), merged symmetry ($Sm^{(h)}, h = 2, 3$) and accessibility ($\alpha^{(h)}, h = 2, 3$) from nodes (words) that are common between each networks (books). Additionally, we apply linear discriminant analysis (LDA) on both strategies with dimension $d = 2$, over each feature vector to extract the most relevant information.

- **Graph Embedding Technique**: Graph2vec is based in skip-gram model, capture structural/syntactic information which is relevant on task of authorship recognition. It uses the common words between each documents to represent each graph with a single vector embedding (Narayanan et al., 2017). Each graph represent a single book. On the other hand, we evaluate performance of graph2vec varying the number of dimensions of each node embedding vector.

3 Experiments and Results

Our database contains books which are available on Gutenberg site[1]. Table 1 shows the list of books used in our experiments, the list of authors is based on (Stanisz et al., 2018; Amancio et al., 2011). To evaluate

Table 1: Gutenberg Books

Author	Books
Hector Hugh	The Toys of Peace, The Unbearable Bassington, Beasts and Super Beasts, When William Came, The Rise of the Russian Empire, The Chronicles of Clovis.
Thomas Hardy	A Pair of Blue Eyes, A Changed Man and other Tales, far from the Madding Crowd, The Return of the Native, The Hand of Ethelberta, Jude the Obscure.
Daniel Defoe	Memoirs of a Cavalier, Colonel Jack, Roxana the Fortunate Mistress, Captain Singleton, Moll Flanders, Robinson Crusoe.
Alan Poe	The Works of Edgar Allan Poe (1,2,3,4,5), The Narrative of Arthur Gordon Pym of Nantucket.
Bram Stoker	The Lady of Seven Stars, The Mystery of the Sea, The Jewel of Seven Stars, The Lair of the White Worm, The Man, Dracula Guest.
Mark Twain	Following the Equator, Life on the Mississippi, The Prince and the Pauper, The Innocents Abroad, The Adventures of Huckleberry Finn, The Adventures of Tom Sawyer.
Charles Dickens	Oliver Twist, David Copperfield, The Mystery of Edwin Drood, Barnaby Rudge, The Pickwick Papers, A Tale of Two Cities.
Pelham Grenville	Right Ho Jeeves, My Man Jeeves, The Clicking of Cuthbert, The Man with Two Left Feet, The Adventures of Sally, Tales of St Austins.
Charles Darwin	Geological Observations on South America, Volcanic Islands, The Structure and Distribution of Coral Reefs, The Different Forms of Flowers on Plants of the Same Species, The Expression of the Emotions in Man and Animals, On the Origin of Species.
Arthur Doyle	The Adventures of Sherlock Holmes, The Refugees, The Lost World, The Exploits of Brigadier Gerard, The Valley of Fear, Micah Clarke.
George Eliot	The Mill on the Floss, Adam Bede, Romola, Daniel Deronda, Middlemarch, Felix Holt the Radical.
Jane Austen	Mansfield Park, Sense and Sensibility, Northanger Abbey, Persuasion, Emma, Pride and Prejudice.
Joseph Conrad	Victory an Island Tale, Lord Jim, Chance a Tale in Two Parts, Nostromo a Tale of the Seaboard, Under Western Eyes, An Outcast of the Islands.

effects of short-texts, each book was limited to a set of their first tokens t (words), after doing the pre-processing step. We describe following the principal steps of our experiments: (1) Pre-processing step:

[1] http://www.gutenberg.org

(a) removing signal punctuation, in contrast of (Darmon et al., 2019) that uses only signal punctuation; (b) removing stop-words defined in [2]NLTK to focus on words with semantic information, since syntactic information are implicit in the graph structure; (c) lemmatization, it was applied using NLTK. (2) Feature extraction step, in this step was applied the two approaches complex network measurements and graph embedding techniques. (3) Classification step, in order to quantify the robustness of the features obtained from the two approaches to distinguish between authors, we employ machine learning algorithms such as naive bayes (NB), support vector machines (SVM) and k-nearest neighbor (KNN). Robustness of each strategies is evaluated using accuracy results of cross-validation technique with $k = 4$ folds.

Table 2: Classification accuracy of global strategy and local strategy applying linear discriminant analysis (LDA) over feature vector to reduce its dimensionality ($d = 2$), where $l - text$ is the k first token of each books.

Global Strategy										
		LDA		LDA		LDA		LDA		LDA
KNN	16.0	41.0	23.0	**57.0**	34.0	27.0	38.0	38.0	30.0	45.0
NB	29.0	44.0	31.0	46.0	25.0	38.0	33.0	41.0	32.0	46.0
SVM	23.0	**52.0**	34.0	54.0	**47.0**	29.0	46.0	**48.0**	44.0	**56.0**
Local Strategy										
KNN	14.0	91.0	15.0	43.0	11.0	30.0	21.0	**37.0**	30.0	20.0
NB	26.0	86.0	18.0	38.0	12.0	29.0	15.0	27.0	9.0	29.0
SVM	26.0	**97.0**	28.0	**51.0**	26.0	**32.0**	24.0	33.0	**37.0**	26.0
$l - text$	2500		5000		10000		15000		20000	

Table 3: Classification accuracy of k-nearest neighbor (k = 3) using graph2vec with dimension vector d over different sizes of texts ($l - text$).

Graph2vec					
d = 2	**91.0**	88.0	**94.0**	92.0	92.0
4	85.0	90.0	89.0	**93.0**	89.0
5	87.0	89.0	92.0	91.0	**95.0**
10	88.0	91.0	92.0	92.0	95.0
15	88.0	**92.0**	92.0	94.0	94.0
16	87.0	89.0	92.0	92.0	95.0
32	90.0	91.0	92.0	92.0	95.0
64	86.0	91.0	92.0	92.0	95.0
128	91.0	92.0	92.0	91.0	92.0
$l - text$	2500	5000	10000	15000	20000

4 Conclusion and Future Work

Local strategy with linear discriminant analysis is better to describe short-text than graph2vec with low dimension, it achieves 97% of accuracy. However, the performance on local and global strategies decrease while length of each text is increasing. In conclusion, the large length of our text can generate noise on complex network approach, in contrast of graph embedding technique, which have better results with large length of texts. Moreover, Graph2vec works better than complex network measurements using low dimension of vector embedding which can help to visualization tasks as future work.

[2]http://www.nltk.org

References

Sasa Adamovic, Vladislav Miskovic, Milan Milosavljevic, Marko Sarac, and Mladen Veinovic. 2019. Automated language-independent authorship verification (for indo-european languages). *Journal of the Association for Information Science and Technology*, 0(0).

Camilo Akimushkin, Diego R. Amancio, and Osvaldo N. Oliveira Jr. 2017. On the role of words in the network structure of texts: application to authorship attribution. *CoRR*, abs/1705.04187.

Diego Raphael Amancio, Eduardo G Altmann, Osvaldo N Oliveira Jr, and Luciano da Fontoura Costa. 2011. Comparing intermittency and network measurements of words and their dependence on authorship. *New Journal of Physics*, 13(12):123024.

Diego R. Amancio, Sandra M. Aluisio, Osvaldo N. Oliveira, and Luciano da F. Costa. 2012. Complex networks analysis of language complexity. *EPL (Europhysics Letters)*, 100(5):58002, dec.

Diego R. Amancio. 2014. Probing the topological properties of complex networks modeling short written texts. *CoRR*, abs/1412.8504.

Ramon Ferrer i Cancho and Richard V. Solé. 2001. The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1482):2261–2265.

Alexandra N. M. Darmon, Marya Bazzi, Sam D. Howison, and Mason A. Porter. 2019. Pull out all the stops: Textual analysis via punctuation sequences. *CoRR*, abs/1901.00519.

Leandro Borges dos Santos, Edilson Anselmo Corrêa Júnior, Osvaldo N. Oliveira Jr., Diego R. Amancio, Leticia Lessa Mansur, and Sandra M. Aluísio. 2017. Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts. *CoRR*, abs/1704.08088.

Hugo Jair Escalante, Thamar Solorio, and Manuel Montes-y Gómez. 2011. Local histograms of character n-grams for authorship attribution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 288–298, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. *KDD : proceedings. International Conference on Knowledge Discovery & Data Mining*, 2016:855–864.

David M. Markowitz. 2019. What words are worth: National science foundation grant abstracts indicate award funding. *Journal of Language and Social Psychology*, 38(3):264–282.

Sanda Martincic-Ipsic, Tanja Milicic, and Ljupco Todorovski. 2017. The influence of feature representation of text on the performance of document classification. *CoRR*, abs/1707.01321.

Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. 2017. graph2vec: Learning distributed representations of graphs. *CoRR*, abs/1707.05005.

Yunita Sari, Andreas Vlachos, and Mark Stevenson. 2017. Continuous n-gram representations for authorship attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 267–273, Valencia, Spain, April. Association for Computational Linguistics.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.*, 60(3):538–556, March.

Tomasz Stanisz, Jaroslaw Kwapien, and Stanislaw Drozdz. 2018. Linguistic data mining with complex networks: a stylometric-oriented approach. *CoRR*, abs/1808.05439.