UNIVERSITY OF SÃO PAULO
INSTITUTE OF MATHEMATICS AND COMPUTER SCIENCE

Authorship attribution with traditional methods and complex networks

VANESSA QUEIROZ MARINHO

Advisor: Prof. Dr. Diego Raphael Amancio
Advisor of Internship Abroad: Prof. Dr. Graeme Hirst

Work Plan for an internship abroad during the Master's

Abstract. Concepts and methods of complex networks have proven useful to probe several real systems of very distinct nature. The discovery that methods from complex networks can be used to analyse texts in their different complexity levels has allowed the study of natural language processing tasks from a new perspective. Examples of tasks studied via topological analysis of networks are keyword identification, automatic extractive summarization and authorship attribution. The latter task, which is the focus of this Master's project, has been studied with some success by representing texts as words adjacency networks. Even though networked representations have been applied to study the authorship recognition problem, such approaches have not outperformed other traditional models relying upon statistical paradigms. Because network models are able to grasp textual patterns that can not be with traditional statistical models, we intend to devise hybrid systems that combine both traditional NLP techniques with properties provided by the topological analysis of complex networks. By combining such distinct paradigms in a complementary way, we aim to improve the performance of textual stylistic characterization and authorship attribution systems. We are bold to predict that such combination shall probably improve the performance of related applications, such as the analysis of stylistic inconsistencies and plagiarism.

1 Introduction

The representation and characterization of real systems with complex networks has been useful to describe a large variety of systems found in nature and in society [1]. Some examples include the cell, which can be described as a network of substances connected by chemical reactions, and the Internet, a network of routers and computers connected by physical links [3]. Many scientific works that use complex networks are interdisciplinary, benefiting from ideas of different areas, such as mathematics, physics, biology, computer science [2]. Traditionally, the study of networks was related to the theory of graphs undergoing random processes. In the last few years, however, the finding that many real systems could be characterized by networks with non-trivial patterns [2] allowed a fast development in the area. A fast increase of data availability and computational capacity allowed the analysis and development of efficient algorithms in different applications [5], including text analysis via topological characterization of networks. There are several representations of texts as networks, where both nodes and edges may represent distinct textual aspects. The most common models where nodes represent words and edges are established

according to syntactic [4], semantic [6] or empirical [7] relationships. Interestingly, it has been shown that small-world and scale-free properties arises in such networks [8].

A special case of syntactic networks is the co-occurrence network (or word adjacency network) [8]. In that type of network, links are created by connecting adjacent words, as most of the syntactical relations occurs in very short scales [28]. The representation of text as word co-occurrence networks has proven useful for example to identify literary movements [10] and to create automatic extractive summarizers [4]. A study carried out by Amancio et al. [11] demonstrated that most of the topological measurements of these networks capture syntactic and stylistic characteristics of the language. This means that these networks are more adequate to handle stylistic tasks, however a dependency between topology and the semantic of words has also been observed [12]. In this context, we intend to use the discriminant power of word co-occurrence networks to quantify different styles in the authorship attribution task.

Authorship attribution methods are relevant in practice because they can be applied to classify literary works, solve copyright disputes [13] or even intercept terrorist messages [14]. The first statistical authorship attribution techniques were devised by Mosteller and Wallace to probe the authorship of the so-called Federalist Papers [15]. After this seminal study, researchers have been proposing new attributes in order to characterize writing styles [16]. Some attributes traditionally used in the task include statistical properties of words (for example, the average length, the frequency, burstiness and the vocabulary richness) [5] and characters (frequencies and long-range correlations) [13]. Besides the lexical attributes, syntactic (for example, frequency of specific chunks) and semantic attributes have been also used as relevant attributes [5, 18].

While much study has been devoted to single techniques based either on statistical or networked representation and characterizations, no comprehensive study has probed the benefits of combining such distinct paradigms. For this reason, the main objective of this Master's project is to combine traditional and networked representations in order to improve the representation of texts for the purpose of grasping styles in a more adequate and accurate manner. In our preliminary experiments, we have extended the traditional co-occurrence model taking into account features borrowed from traditional textual representations. More specifically, we have devised a technique to establish relevant links between words in text networks, which has allowed a significant improvement in the performance of classification systems. This is evidence that a much deeper linguistic analysis is able to improve current networked models.

In this proposal, specifically, the main goal is to develop hybrid classifiers, in which we will combine network techniques with the traditional ones used in the authorship attribution task. In these classifiers, both components will be analysed, the topological component (obtained by complex networks measurements) and the component resultant from traditional methods. Besides, we intend to analyse the robustness of these classifiers according to the text length, as there exists a clear dependency between classification performance and text length [21]. The main hypothesis of this work is an improvement of performance in the authorship attribution task, since the disadvantages of each technique should be overcome by the hybrid

classification. Besides improving the performance of the task, we expect to create classifiers more robust than the actual ones, that can be vulnerable to attacks [17].

This research proposal is organized as follows. In Section 2, we present a literature review of authorship attribution and complex networks. In Section 3, the research proposal of this internship abroad is presented. Finally, in Section 4, we present some final remarks.

2 Literature Review

In the first part of this section, some authorship attribution techniques are presented. In Subsection 2.2, we explain how texts are modelled as complex networks. Finally, we present some works that use co-occurrence models for the authorship attribution task.

2.1 Authorship Attribution

In a typical authorship attribution problem, a text whose authorship is unknown is assigned to an author, from a set of possible ones. The first authorship attribution activities supported by statistical methods are from the XIX century. The goal of these methods is to maximize the probability $P(x|a)$ of a text x to belong to a possible author a. In [15], Mosteller and Wallace analysed the authorship of a collection of political essays, known as The Federalist Papers. With that study, they made a huge contribution to the area showing that the frequency of common words (as 'and' and 'to') can distinguish different authors.

Differently from the authorship recognition techniques used by linguistics, the research carried out by Mosteller and Wallace [15] was based mostly on simple textual statistics. Since then, many works tried to define characteristics to quantify the writing style of authors, known as stylometry [26]. According to Stamatatos [5], current stylometric features can be classified in a three-fold way: character, lexical, syntactic and semantic features.

Lexical and character features consider the text as a sequence of words or characters, respectively. Examples of lexical features commonly used in the authorship attribution task are word and sentence lengths, frequency of words, part-of-speech tags, vocabulary richness and others [32]. Considered as one of the first authorship recognition approaches using lexical features, Mendenhall [27] used the length of sentences and words to identify the authorship of documents. Currently, most of the studies use, at least partially, lexical features to characterize writing styles [5]. An important finding related to lexical features is the fact that common words (articles, prepositions, pronouns), referred to as function words, are one of the best characteristics to distinguish a set of authors [29]. Once function words are topic-independent and used in a largely unconscious manner, they may capture stylistic choices of each author [5]. Besides that, frequent words account for the successful approach proposed by physics in [38]. In that work, the rank distances of the word frequencies are used to calculate the similarity between texts. Another possible approach is to extract the most frequent words in a text [29, 30]. In this approach, each text is represented as a vector of word frequencies. Then, suitable machine learning techniques can be applied to distinguish these vectors.

Character-based features analyse the text as a simple sequence of characters. Some examples of these measurements are alphabetic characters count, digit characters count, frequency of letters and punctuation marks and others [5]. Another approach is to extract frequencies of specific n-grams. Because this representation has proven useful to grasp authors' styles, one may infer that it is possible to create bigram networks to characterize texts. More specifically, in the context of authorship attribution, Jankowska et al. [31] observed that the frequency of common characters could distinguish different authors, showing better results than the ones with lexical features. One of the advantages of using character-based features is that they can be extracted from any natural language. Therefore, this is a language independent method.

The extraction of syntactic and semantic characteristics requires a more detailed text analysis. According to Stamatatos [5], syntactic information is considered more representative than lexical characteristics. An example of the usefulness of this information is the good performance of function words at characterizing the writing style, since these words are found in some syntactic structures. Some syntactic approaches include the extraction of rewrite rule frequencies. Rewrite or production rules indicate how a sentence may be decomposed into subparts. A sentence is considered syntactically valid, if it is possible to generate all its terminals (words) according to some rewrite rules [39]. An example of a rewrite rule is $S \rightarrow NP \;\; VP$, in which S means Sentence, NP means Noun Phrase and VP means Verb Phrase. Besides, the analysis of sentences or chunks and part-of-speech (POS) taggers [5] are also employed in syntactic approaches. One of the first works that used syntactic information applied to the authorship attribution task was carried out by Baayen et al. [33]. They extracted rewrite rule frequencies and those measurements provided better results than the ones with lexical features. Hirst and Feiguina [18] combined the idea of bigram frequencies with syntactic analysis. Their syntactic label bigrams were found useful to distinguish different authors, even when applied in short texts.

2.2 Complex Networks

Modelling text as complex networks To model text as complex networks, some pre-processing steps are required. In the first step, usually employed in many studies [10], a list of function words (or stopwords) is removed. This list usually contains articles, prepositions and adverbs, mainly words conveying little semantic content. The second part is the lemmatization step. This is applied to the remaining words using a part-of-speech tagger. After the lemmatization, words related to the same concept will be associated with the same vertex, despite the different inflections. Table 1 illustrates the application of the two pre-processing steps in an extract from the book "The Adventures of Sherlock Holmes", by Arthur Conan Doyle.

After the pre-processing steps, the texts are modelled as complex networks. Once the written language is formed by linear chains of words, the easiest way to represent a text as a network is through the connection of adjacent words. This type of network, called word co-occurrence (or adjacency) network, has been used in many studies [8, 11, 22]. In a traditional co-occurrence network, each node represents a distinct word in the text and the edges are established between two adjacent words. Figure 1 illustrates two possible co-occurrence networks for the sentence used in Ta-

Table 1. Example of the two pre-processing steps applied to the text. Some sentences from the book "The Adventures of Sherlock Holmes" are shown after the removal of the stopwords and then after lemmatization.

Original	**Without stopwords**	**After lemmatization**
"There are three men waiting for	*three men waiting*	*three men wait*
him at the door", said Holmes.	*door said holmes*	*door say holmes*
"Oh, indeed! You seem to have	*oh indeed seem*	*oh indeed seem*
done the thing very completely.	*done thing completely*	*do thing completely*
I must compliment you."	*must compliment*	*must compliment*
"And I you", Holmes answered.	*holmes answered*	*holmes answer*

ble 1. The network on the left uses the traditional co-occurrence model. This graph was created connecting each word to its first nearest neighbor. On the other hand, in the network on the right, each word was connected to its first and second nearest neighbors.

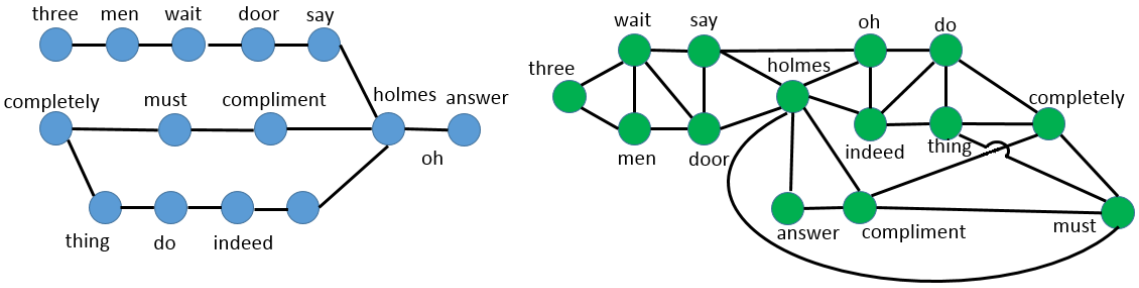

Fig. 1. Graphs obtained for the sentence "three men wait door say holmes oh indeed seem do thing completely must compliment holmes answer". The graph on the left represents the traditional co-occurrence network. Another possible co-occurrence network is shown on the right. In this network, each word is connected to its first and second nearest neighbors.

With the co-occurrence network of each book, a set of measurements is extracted from them. Examples of these measurements are the accessibility [23], betweenness [24], selectivity of words [25] and others. The extracted values are used as features of each book and evaluated with different machine learning algorithms.

Authorship Attribution with Complex Networks The authorship attribution task through the use of word co-occurrence networks is described in [19]. The measurements extracted from the network were able to distinguish different authors, emphasizing its importance to this task. After the combination of some measurements, it was also possible to cluster different authors, characterizing a common writing style.

Dependencies between the network topology of co-occurrence networks and the writing style of authors were observed in the work "Comparing intermittency and network measurements of words and their dependence on authorship" [11]. In that work, after removing the function words from the text, network measurements were extracted in order to evaluate how they relate with the writing style. As one of their results, the mean shortest path between words was found to be dependent on

the writing style. Other works also use complex networks to perform authorship attribution, even in short texts [20, 21]. These successful application of network concepts for identifying styles is the basis for the research we intend to conduct during the internship.

3 Work Proposal

The main research question of this internship project reside precisely in the investigation of a way to combine traditional approaches used in the authorship attribution task (for example, the ones listed in Section 2.1) with techniques from complex networks.

The Master's project, to which this proposal is related, aims to develop new models for authorship recognition using complex networks. These new models include the extension of the word co-occurrence network and the development of models that combine different techniques for the authorship attribution task. The first proposed task of this Master's project was to extend the traditional co-occurrence model. We developed it using the strategy of further hierarchies. In this strategy, to account for possible relevant links between non-adjacent words, we connect every pair of words whenever they are separated by less than $W-1$ intermediary words, for $W = 1, 2, 3$. Some of these values are explained below:

- $W = 1$: This represents the traditional co-occurrence model. The left network in Figure 1 was created with this W.
- $W = 2$: In this network, we connect each word to its adjacent word (called X) and to the word adjacent to X. It represents the right network in Figure 1.

In our preliminary results, we evaluated 40 novels (5 from each different author) and our success rate (number of books correctly classified) was improved when we used the further hierarchies strategy with $W = 2$ and $W = 3$. In a different experiment, the only features extracted from the networks were the number of motifs [34] of size 3 and 4. Motifs are interconnections patterns that occur in real networks at numbers significantly higher than in randomized networks. With only those characteristics, we achieved a 55% of performance (the guessing baseline is 12.5%) using four machine learning methods.

In this proposal, specifically, the main goal is to develop hybrid authorship attribution classifiers. One of the components of these classifiers will be the traditional approaches, successfully applied in many works. As we described in Section 2.1, some traditional techniques include lexical, character, syntactic and semantic features. We also expect to use some techniques based on patterns of local coherence. Feng and Hirst [35] demonstrated that these patterns can be considered a powerful stylometric feature. The other component of the classifier will use complex networks methods, specially the models we already created during this project, which showed good performance in our preliminary results. We intend to analyse the robustness of these classifiers according to different text sizes.

During the development of the hybrid classifier, we intend to propose new strategies that can be applied to connect words in co-occurrence networks. In order to improve the topological representation, we can use traditional techniques to extract

information from texts, as the output of a part-of-speech tagger. Then, besides connecting adjacent or intermediary words, this extracted information can be used to establish more relevant links among words.

As an extension of our main goal, we can use the strategies we developed to characterize and identify different writing styles. We aim to apply them to analyse the database used in [37]. Hirst and Feng analysed books from three different authors in order to investigate whether the Alzheimer's disease affects an authors' writing style. Since co-occurrence networks can capture syntactic and stylistic characteristics of the language, we expect they can extract meaningful information from these books.

3.1 Evaluation

The evaluation of this project will be accomplished in two steps. First, the traditional techniques and the complex networks methods will be tested separately in our corpus. In the second step, our classifier will be tested. This is intended to verify the impact of each technique in the context of this research, enabling the best combination of them to obtain a robust hybrid classifier.

The corpus will be treated following some traditional approaches, such as the stratified cross validation (usually with 10 folds) and the use of different machine learning algorithms [36]. Our results will be compared with methods from the state of art and some baselines for the authorship attribution problem. We will also evaluate the performance of our classifier on short texts.

3.2 Schedule of Work

The development of hybrid classifiers for the authorship attribution task is part of the Master's project entitled "Development of new models for authorship recognition using complex networks" to which this proposal is linked. Below the tasks of the Master's are listed, as well as the schedule of it, in Table 2.

Table 2. Schedule tasks of the Master's.

	Year 1 (2015)		Year 2(2016)		Year 3(2017)
Task	1^{st} Semester	2^{nd} Semester	1^{st} Semester	2^{nd} Semester	1^{st} Semester
1					
2					
3					
4					
5					
6					
7					
8					

- **T1**: Fulfilment of disciplines;
- **T2**: Literature review of authorship attribution methods and complex networks;
- **T3**: Survey of possible target applications of our methods, besides authorship attribution problems;

- **T4**: Qualification exam;
- **T5**: Internship abroad;
- **T6**: Development of the proposed tasks;
- **T7**: Evaluation of the proposed methods;
- **T8**: Technical reports production. Publication of the results in events and journals of the area. Thesis writing and defense of the Master's.

The internship abroad is foreseen in the task T5, during which the hybrid classifiers will be designed and developed. Therefore, it foresees, during the internship, the achievement of the tasks listed below, from T1 to T6 and. In Table 3, these tasks are outlined in a schedule.

Table 3. Schedule tasks during the internship abroad

	2016					
Task	March	April	May	June	July	August
1						
2						
3						
4						
5						
6						

- **T1**: Literature review of traditional techniques applied in authorship attribution.
- **T2**: Definition of a corpus to be used and evaluated;
- **T3**: Application of the traditional techniques and complex networks methods. Development of the hybrid classifier;
- **T4**: Evaluation of the classifier in our corpus. It also will be evaluated using different text sizes;
- **T5**: Application and evaluation of the complex network methods in the database used at [37];
- **T6**: Publication of the results in events and journals of the involved areas.

4 Final Remarks

Professor Dr. Graeme Hirst has a wide experience with authorship attribution and stylistic characteristics [18, 32, 35, 37]. Because of that, we are confident that this internship will be a great learning opportunity.

The advances obtained during this project may be useful not only to improve the authorship attribution field but also to study related applications, such as the analysis of stylistic inconsistencies and plagiarism.

References

1. Albert, R., Barabási, A.-l.: Statistical mechanics of complex networks. Rev. Mod. Phys, (2002).
2. Newman, M.: Networks: An Introduction. Oxford University Press, (2010)
3. Mihalcea, R., Radev, D.: Graph-based natural language processing and information retrieval. Cambridge University Press, (2011).

4. Amancio, D. R., Nunes, M. G. V., Oliveira, O. N., Costa, L. F.: Extractive summarization using complex networks and syntactic dependency. Physica A: Statistical Mechanics and its Applications, (2012).
5. Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. J. Am. Soc. Inf. Sci. Technol, (2009).
6. Liu, H.: Statistical properties of Chinese semantic networks. Chinese Science Bulletin, (2009).
7. Luduena, G. A., Behzad, M. D., Gros, C.: Exploration in free word association networks: models and experiment. Cognitive Processing, (2014).
8. Cancho, R. F., Solé, R. V.: The Small World of Human Language. Proceedings of The Royal Society of London. Series B, Biological Sciences, (2001).
9. Cancho, R. F., Solé, R. V., Köhler, R.: Patterns in syntactic dependency networks. Phys. Rev. E, (2004).
10. Amancio, D. R., Oliveira, O. N., Costa, L. F.: Identification of literary movements using complex networks to represent texts. New Journal of Physics, (2012).
11. Amancio, D. R., Altmann, E. G., Oliveira, O. N., Costa, L. F.: Comparing intermittency and network measurements of words and their dependence on authorship. New Journal of Physics, (2011).
12. Silva, T. C., Amancio, D. R.: Word sense disambiguation via high order of learning in complex networks. EPL (Europhysics Letters), (2012).
13. Grant, T.: Quantifying evidence in forensic authorship analysis. International Journal of Speech Language and the Law, (2007).
14. Abbasi, A., Chen, H.: Applying authorship analysis to extremist-group web forum messages. IEEE Intelligent Systems, (2005).
15. Mosteller, F., Wallace, D. L.: Inference and Disputed Authorship: The Federalist Papers. Addison-Wesley, (1964).
16. Juola, P.: Authorship Attribution. Found. Trends Inf. Retr., (2006).
17. Brennan, M. R., Greenstadt, R.: Practical attacks against authorship recognition techniques. AAAI, 2009.
18. Hirst, G., Feiguina, O.: Bigrams of syntactic labels for authorship discrimination of short texts. Literary and Linguistic Computing, (2007).
19. Antiqueira, L., Pardo, T. A. S., Nunes, M. G. V., Oliveira, O. N., Costa, L. F.: Some issues on complex networks for author characterization. (2006).
20. Amancio, D. R.: Authorship recognition via fluctuation analysis of network topology and word intermittency, CoRR, (2015).
21. Amancio, D. R.: Probing the Topological Properties of Complex Networks Modeling Short Written Texts. PLoS ONE, (2015).
22. Roxas, R. M., Tapang, G.: Prose and Poetry Classification and Boundary Detection Using Word Adjacency Network Analysis. International Journal of Modern Physics C, (2010).
23. Travençolo, B. A. N., Costa, L. F.: Accessibility in complex networks. Physics Letters A, (2008).
24. Costa, L. F., Rodrigues, F. A., Travieso, G., Boas, P. R. V.: Characterization of complex networks: A survey of measurements. Advances in Physics, (2007).
25. Masucci, A. P., Rodgers, G. J.: Differences between normal and shuffled texts: Structural properties of weighted networks. Advances in Complex Systems (ACS), (2009).

26. Holmes, D. I.: Authorship attribution. Computers and the Humanities, (1994).
27. Mendenhall, T. C.: The characteristic curves of composition. Science, (1887).
28. Ferrer i Cancho, R.: Euclidean distance between syntactically linked words. Phys. Rev. E, (2004).
29. Burrows, J. F.: Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style. Literary and Linguistic Computing, (1987).
30. Stamatatos, E.: Authorship attribution based on feature set subspacing ensembles. International Journal on Artificial Intelligence Tools, (2006).
31. Jankowska, M., Milios, E., Keselj, V.: Author verification using common n-gram profiles of text documents. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, (2014).
32. Graham, N., Hirst, G., Marthi, B.: Segmenting documents by stylistic character. Natural Language Engineering, (2005).
33. Baayen, R., van Halteren, H., Tweedie, F.: Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing, (1996).
34. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network Motifs: Simple Building Blocks of Complex Networks. Science, (2002).
35. Feng, V.W., Hirst, G.: Patterns of local discourse coherence as a feature for authorship attribution. Literary and Linguistic Computing, (2013).
36. Duda, R. O., Hart, P. E., Stork, D. G.: Pattern Classification. Wiley-Interscience, (2000).
37. Hirst, G., Feng, V.W.: Changes in Style in Authors with Alzheimer's Disease. English Studies, (2012).
38. Havlin, S.: The distance between Zipf plots. Physica A: Statistical Mechanics and its Applications, (1995).
39. Kumar, E.: Natural language processing. I K International, (2012).