Investigating the Application of Distributional Semantics to Stylometry

Giulia Benotto, Emiliano Giovannetti, Simone Marchi
Istituto di Linguistica Computazionale "A. Zampolli"
Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1, 56124, Pisa - Italy
{name.surname}@ilc.cnr.it

Abstract

English. The inclusion of semantic features in the stylometric analysis of literary texts appears to be poorly investigated. In this work, we experiment with the application of Distributional Semantics to a corpus of Italian literature to test if words distribution can convey stylistic cues. To verify our hypothesis, we have set up an Authorship Attribution experiment. Indeed, the results we have obtained suggest that the style of an author can reveal itself through words distribution too.

Italiano. *L'inclusione di caratteristiche semantiche nell'analisi stilometrica di testi letterari appare poco studiata. In questo lavoro, sperimentiamo l'applicazione della Semantica Distribuzionale ad un corpus di letteratura italiana per verificare se la distribuzione delle parole possa fornire indizi stilistici. Per verificare la nostra ipotesi, abbiamo imbastito un esperimento di Authorship Attribution. I risultati ottenuti suggeriscono che, effettivamente, lo stile di un autore pu rivelarsi anche attraverso la distribuzione delle parole.*

1 Introduction

Stylometry, that is the application of the study of linguistic style, offers a means of capturing the elusive character of an author's style by quantifying some of its features. The basic stylometric assumption is that each writer has certain stylistic idiosyncrasies (a "human stylome" (Van Halteren et al., 2005)) that define their style. Analysis based on stylometry are often used for Authorship Attribution (AA) tasks, since the main idea behind computationally supported AA is that by measuring some textual features, we can distinguish between texts written by different authors (Stamatatos, 2009).

One of the less investigated stylistic feature is the way in which authors use words from a semantic point of view, e.g. if they tend to use more, when dealing with polysemous words, a certain sense over the others, or senses that differ (even slightly) from the one that's more commonly used (as it happens, typically, in poetry).

A possible approach to the analysis of this characteristic is to consider the textual contexts in which certain words appear. According to Distributional Semantics (DS), certain aspects of the meaning of lexical expressions depend on the distributional properties of such expressions, or better, on the contexts in which they are observed (Lenci, 2008; Miller and Charles, 1991). The semantic properties of a word can then be defined by inspecting a significant number of linguistic contexts, representative of the distributional behavior of such word.

In this work we would like to investigate if the analysis of the distribution of words in a text can be exploited to provide a stylistic cue. In order to inspect that, we have experimented with the application of DS to the stylometric analysis of literary texts belonging to a corpus constituted by texts pertaining to the work of six Italian writers of the late nineteenth century.

In the following, Section 2 gives a short insight on the state of the art of computational stylistic analysis, Section 3 describes the approach together with the corpus used to conduct our investigation and Section 4 discuss about results. Finally, Section 5 draws some conclusions and outlines some possible future works.

2 State of the Art

The very first attempts to analyze the style of an author were based on simple lexical features such as sentence length counts and word length counts, since they can be applied to any language and any corpus with no additional requirements (Koppel and Schler, 2004; Stamatatos, 2006; Zhao and Zobel, 2005; Argamon et al., 2007). Similarly, character measures have been proven to be quite useful to quantify the writing style (Grieve, 2007; De Vel et al., 2001; Zheng et al., 2006). Basically, a text can be viewed as a mere sequence of characters, so that various measures can be defined (including alphabetic, digit, uppercase and lowercase characters count, etc.). A more elaborate text representation method is to employ syntactic information (Gamon, 2004; Stamatatos et al., 2000; Stamatatos et al., 2001; Hirst and Feiguina, 2007; Uzuner and Katz, 2005). The idea is that authors tend to use similar syntactic patterns unconsciously. Therefore, syntactic information is considered a more reliable authorial fingerprint in comparison to lexical information.

More complicated tasks such as full syntactic parsing, semantic analysis, or pragmatic analysis cannot yet be handled adequately by current NLP technologies for unrestricted text. As a result, very few attempts have been made to exploit high-level features for stylometric purposes. Perhaps the most important method of exploiting semantic information so far was described in (Argamon et al., 2007). This work was based on the theory of Systemic Functional Grammar (SFG) (Halliday, 1994) and consisted on the definition of a set of functional features that associate certain words or phrases with semantic information.

The previously described features are application independent since they can be extracted from any textual data. Beyond that, one can define application-specific measures in order to better represent the nuances of style in a given text domain (such as e-mail messages, or online forum messages) (Li et al., 2006; Teng et al., 2004).

To the best of our knowledge, the application of DS to the analysis of literary texts has been documented in a rather small number of works (Buitelaar et al., 2014; Herbelot, 2015). In both these works, DS is used as a theoretical basis in order to verify some hypotheses on specific semantic characteristics of poetic works. In more details, in (Buitelaar et al., 2014) the authors investigated through DS the influence of Lord Byron's work on Thomas Moore trying to find a shared vocabulary or specific formal textual characteristics. In (Herbelot, 2015) it is argued how distributionalism can support the notion that the meaning of poetry comes from the meaning of ordinary language and how distributional representations can model the link between ordinary and poetic language. However, the role of DS in the study of a style of an author was not the aim of these works.

3 Experimental Setup

First, we want to specify that it is not our purpose to propose new ways to improve state-of-the-art AA algorithms. Indeed, our aim is just to verify the hypothesis that the distribution of words can provide an indication of a distributional stylistic fingerprint of an author. To do this, we have set up a simple classification task. Subsection 3.1 briefly depicts the data set we used, while Section 3.2 describes the steps implemented in our experiment.

3.1 Data Set Construction

In order to build the reference and test corpora, we started from texts pertaining to the work of six Italian writers working at the turn of the 20th century, namely, Luigi Capuana, Federico De Roberto, Luigi Pirandello, Italo Svevo, Federigo Tozzi and Giovanni Verga. We chose contiguous authors in chronological sense, whose texts are available in digital format (in fact we could not do a similar survey on the narrative of the 90s because it is still under copyrights). Indeed, we used texts freely available for download from the digital library of the Manunzio project, via the LiberLiber website[1]. Since they were encoded in various formats, such as .epub, .odt and .txt, our pre-processing consisted in converting them all in .txt format and getting rid of all xml tags, together with footnotes and editors' notes and comments.

3.2 Experiment Description

According to Rudman (1997), a striking problem in stylometry is due to the lack of homogeneity of the examined corpora, in particular to the improper selection or fragmentation of the texts, that might cause alterations in the writers' style. In order to create balanced reference corpora, i.e. covering all the authors' different stylistic and thematic phases, for each author, as shown in Figure 1, we built a reference corpus as the composition of the 70% of each single work (usually a novel). The same technique was used to create the

[1] http://www.liberliber.it/

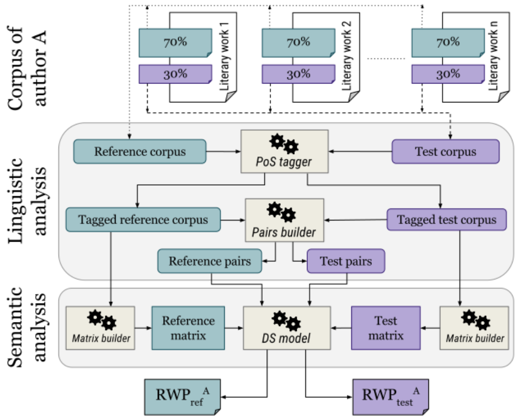

Figure 1: RWP$_{ref}$ and RWP$_{test}$ creation process for an author.

test corpus by using the remaining 30% of each work. Typical AA approaches consist in analyzing known authors and assigning authorship to previously unseen text on the basis of various features. Train and test sets should then contain different texts. Contrary to the classical AA task, our train and test sets contain different parts of the same texts. Indeed, with this experiment, we wanted to understand if the semantics that an author bestows to a word, is peculiar to his writing. To prove this, we wanted to cover all the different stylistic and thematic phases an author can go through during his activity, hence the partition of all his texts in a reference and a test portion.

We then analyzed each reference and test corpora with a Part-of-Speech (PoS) tagger and a lemmatizer for Italian (Dell'Orletta et al., 2014). For every author, we built two lists of word pairs (with their lemma and PoS), one relative to the tagged reference corpus (*reference pairs*) and the other to the tagged test set (*test pairs*), where each word was paired with all the other words with the same PoS. We also filtered the pairs to leave only nouns, adjectives and verbs. Starting from the tagged corpora, we built two words-by-words matrixes[2] of co-occurrence counts (co-occurrence matrixes) for each author, using a context window of 4[3]. The chosen DS model (Baroni and Lenci, 2010) was applied to each matrix to calculate the cosine between the vectors representing the two words of each pair. This allowed us to evaluate the semantic relatedness between the words by assessing their proximity in the distributional space as represented by the cosine value: the more this value tends to 1, the more the two words of the pair are considered to be related. We then obtained two related word pair (RWP) lists for each author *A*: $\text{RWP}_{\text{ref}}{}^{\text{A}}$ and $\text{RWP}_{\text{test}}{}^{\text{A}}$. Figure 1 shows the process described above.

Since we wanted to focus on the analysis of the semantic distribution of words, we decided to exclude any possible "lexical bias". For this reason, we restricted the analysis on a common vocabulary, i.e. a vocabulary constituted by the intersection of the six authors' vocabularies. In this way, we prevent our classifier to exploit, as a feature, the presence of words used by some (but not all) of the authors. Moreover, we removed from the RWP_{test} lists all those pairs of words occurring frequently together in the same context, since they might constitute a multiword expression that, once again, could be pertaining with the signature lexicon of each author. To remove them, we computed the number of times (*#co-occ* in Table 1) they appeared together in the context window, as well as their total number of occurrences (*#occ$_a$* and *#occ$_b$*) and we excluded from the analysis those pairs for which the ratio between the number of co-occurrences and the total occurrences of the less frequent word was higher than the empirically set threshold of 0.5. The first two pairs of Table 1 would be removed as probable multiword (PM column in Table 1): "*scoppio*" (burst) and "*risa*" (laughter) could mostly co-occur in "*scoppio di risa*" (meaning "burst of laughter") and the words "*man*" and "*mano*" (both meaning "hand") could mostly co-occur in "*man mano*" (meaning "little by little", or "progressively").

W_a	W_b	#occ$_a$	#occ$_b$	#co-occ	ratio	PM
scoppio–s	risa–s	19	9	7	0.78	yes
man–n	mano–n	50	1325	47	0.94	yes
nausea–n	disgusto–n	27	26	0	0	no
piccolo–a	grande–a	248	237	14	0.06	no

Table 1: An example of co-occurring RWs from Pirandello's test list: the first two pairs would be removed.

Finally, we reduced the size of the six RWP_{ref} and RWP_{test} lists by sorting them in decreasing order of the cosine value and then by keeping the

[2] Being the corpus relatively small and not having particular computability issues, we chose not to apply decomposition techniques to reduce the size of the matrixes (and thus not losing any information).

[3] We performed different empiric setup of the window's size and chose the one that showed more suitable results, according to what is stated by Kruszewski and Baroni (2014).

pairs with the highest cosine, selected using a percentage parameter θ as threshold[4]. We chose to introduce the parameter θ for two reasons: i) to avoid the classification algorithm to be disturbed by noisy (i.e. not significative) pairs which would not hold any relevant stylistic cue, and ii) to ease a literary scholar in the interpretation of the results by having to analyze just a limited selection of (potentially) semantically related word pairs.

For the last phase of our experiment we defined a classification algorithm to test the effective presence of stylistic cues inside the obtained RWP_{test} lists. We defined a classifier using a nearest-cosine method to attribute each test list to an author. The method consisted in searching for a pair of words contained in the test list inside each reference list and incrementing by 1 the score of the author whose reference list included the pair with the more similar cosine value (i.e. having the minimum difference): the chosen author was the one with the highest score. Table 2 shows the classification results for $\theta = 5\%$.

	Capuana	De Roberto	Pirandello	Svevo	Tozzi	Verga
Capuana	1884	1269	1321	797	755	1054
De Roberto	729	1041	712	498	451	579
Pirandello	1387	1278	2114	937	747	1056
Svevo	353	371	341	593	372	356
Tozzi	199	219	183	242	281	244
Verga	650	671	656	473	430	851

Table 2: Classification results, obtained via the nearest-cosine method for $\theta = 5\%$.

4 Interpreting the Results

As summarized in Table 3, a correct classification of all RWPs in RWP_{test} lists has been obtained with a θ value of 5%.

To help in interpreting the failure of the algorithm in classifying Tozzi's test list for θ values lower than 5% (as shown in Table 3) we calculated the cardinality of the RWP_{test} lists for each author with the change in θ value (Tables 4).

It is possible to observe how the choice of θ influences the correct classification of Tozzi's test list. Indeed, the use of a θ value below 5% has the effect of remarkably reducing an already small test list ($\text{RWP}_{\text{text}}{}^{\text{Tozzi}}$) as shown in Table 4. It is apparent that increasing the value of θ and consequently the number of significant RW pairs that are analysed, the system is able to correctly classify $\text{RWP}_{\text{test}}{}^{\text{Tozzi}}$ (see the values in Tozzi's row of Table 3).

	0.5%	**1%**	**2%**	**5%**
Capuana	Capuana	Capuana	Capuana	Capuana
De Roberto	De Roberto	De Roberto	De Roberto	De Roberto
Pirandello	Pirandello	Pirandello	Pirandello	Pirandello
Svevo	Svevo	Svevo	Svevo	Svevo
Tozzi	Verga	Verga	Tozzi/Verga	Tozzi
Verga	Verga	Verga	Verga	Verga

Table 3: Results of the classification. Classification errors are highlighted.

	0.5%	**1%**	**2%**	**5%**
#$\text{RWP}_{\text{test}}{}^{\text{Capuana}}$	678	1357	2714	6785
#$\text{RWP}_{\text{test}}{}^{\text{De Roberto}}$	488	977	1954	4886
#$\text{RWP}_{\text{test}}{}^{\text{Pirandello}}$	692	1385	2770	6925
#$\text{RWP}_{\text{test}}{}^{\text{Svevo}}$	425	851	1702	4257
#$\text{RWP}_{\text{test}}{}^{\text{Tozzi}}$	246	493	986	2466
#$\text{RWP}_{\text{test}}{}^{\text{Verga}}$	526	1053	2106	5267

Table 4: Cardinality of RWP_{test} for each author and for each θ value.

5 Conclusion and Next Steps

In this paper we investigated the possibility that an analysis of the semantic distribution of words in a text can be potentially exploited to get cues about the style of an author. In order to validate our hypothesis, we conducted a first experiment on six different Italian authors. The results seem to suggest that the way words are distributed across a text, can provide a valid stylistic cue to distinguish an author's work. Of course, it is not our intent, with this paper, to define new methods for enhancing state-of-the-art authorship attribution algorithms. Our research will focus, in the next steps, in detecting and providing useful indications about the style of an author. This can be done by highlighting, for example, atypical distributions of words (e.g. with contrastive methods) or by analysing their distributional variability. Furthermore, it could be interesting to use a different distributional measure, than the cosine, to test our hypothesis.

[4] At the following url we have uploaded an archive containing all the data we have used and processed for our experiment: https://goo.gl/nrTqWh

References

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822, April.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Paul Buitelaar, Nitish Aggarwal, and Justin Tonra. 2014. Using distributional semantics to trace influence and imitation in romantic orientalist poetry. In *AHA!-Workshop 2014 on Information Discovery in Text*. ACL.

Olivier De Vel, Alison Anderson, Malcolm Corney, and George Mohay. 2001. Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4):55–64.

Felice Dell'Orletta, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. 2014. T2kˆ 2: a system for automatically extracting and organizing knowledge from texts. In *LREC*, pages 2062–2070.

Michael Gamon. 2004. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th international conference on Computational Linguistics*, page 611. Association for Computational Linguistics.

Jack Grieve. 2007. Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, 22(3):251–270, May.

Michael AK Halliday. 1994. Functional grammar. *London: Edward Arnold*.

Aurélie Herbelot. 2015. The semantics of poetry: A distributional reading. *Digital Scholarship in the Humanities*, 30(4):516–531.

Graeme Hirst and Olga Feiguina. 2007. Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing*, 22(4):405–417, September.

Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*, page 62. ACM.

Germán Kruszewski and Marco Baroni. 2014. Dead parrots make bad pets: Exploring modifier effects in noun phrases. *Lexical and Computational Semantics (* SEM 2014)*, page 171.

Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

Jiexun Li, Rong Zheng, and Hsinchun Chen. 2006. From fingerprint to writeprint. *Communications of the ACM*, 49(4):76–82.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Joseph Rudman. 1997. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31(4):351–365.

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 2000. Automatic text categorization in terms of genre and author. *Computational linguistics*, 26(4):471–495.

Efstathios Stamatatos, Nikos Fakotakis, and Georgios Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214.

Efstathios Stamatatos. 2006. Authorship attribution based on feature set subspacing ensembles. *International Journal on Artificial Intelligence Tools*, 15(05):823–838.

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

Gui-Fa Teng, Mao-Sheng Lai, Jian-Bin Ma, and Ying Li. 2004. E-mail authorship mining based on svm for computer forensic. In *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*, volume 2, pages 1204–1207. IEEE.

Özlem Uzuner and Boris Katz. 2005. A comparative study of language models for book and author recognition. In *Natural Language Processing–IJCNLP 2005*, pages 969–980. Springer.

Hans Van Halteren, Harald Baayen, Fiona Tweedie, Marco Haverkort, and Anneke Neijt. 2005. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1):65–77.

Ying Zhao and Justin Zobel. 2005. Effective and scalable authorship attribution using function words. In *Information Retrieval Technology*, pages 174–189. Springer.

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393, February.