Experiments of Distributional Semantics in Stylometry.

Giulia Benotto, Emiliano Giovannetti, Simone Marchi

Institute for Computational Linguistics "A. Zampolli" – Pisa, Italy

First things first: Introduction and A little bit of background knowledge

Theoretic Background-
Stylometry

- Stylometry: the statistical analysis of literary style

- It 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 used for Authorship Attribution (AA) tasks: by measuring some textual features, we can distinguish between texts written by different authors.

Theoretic Background-
Authorship Attribution

- Simple lexical features: sentence length counts and word length counts, that can be applied to any language and corpus with no additional requirements

- Character measures: a text can be viewed as a mere sequence of characters, so that measures such as alphabetic, digit, uppercase and lowercase characters count can be defined

- Syntactic information: authors tend to use similar syntactic patterns unconsciously, so syntactic information is considered a reliable authorial fingerprint

- Application independent features: they can be extracted from any textual data.

Theoretic Background-
Authorship Attribution

- Full syntactic parsing, semantic analysis, or pragmatic analysis cannot yet be handled adequately by current NLP technologies for unrestricted text: few attempts have been made to exploit high-level features for stylometric purposes

- The most important method of exploiting semantic information is based on the theory of Systemic Functional Grammar (SFG) [Halliday, 1994]. It consists on defining a set of functional features that associate words with semantic information

Theoretic Background-
Distributional Semantics

- Distributional Semantics: approaches to semantics that assume that the statistical distribution of words in contexts plays a significant role in defining their semantic behavior

- Based on the Distributional Hypothesis: The degree of semantic similarity between two linguistic expressions *a* and *b* depends on the similarity of the linguistic contexts in which *a* and *b* can appear

- On the cognitive level, this corresponds to a model of the mental lexicon in which meanings are defined by contextual representations

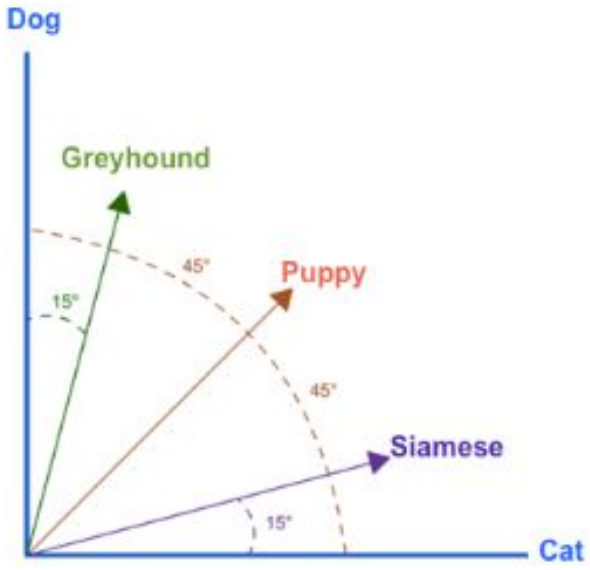
Theoretic Background-
Distributional Semantics

- Distributional Semantics Models (DSMs): evaluate the semantic similarity between words, by assessing their proximity in the distributional space

- Each word is represented as a n-dimensions vector, each dimension recording the number of times that the word under examination appears in a certain context

- When the number of dimensions in which the vectors have similar values increase, they tend to get closer in the distributional space. Based on the assumption underlying the distributional hypothesis [Harris, 1954], the semantic similarity of the corresponding words increases as the vectors get closer in the distributional space

Theoretic Background-
Distributional Semantics



- When the vectors are geometrically aligned on the same line, in the same direction, the angle they form measures 0, and their cosine measures 1, which indicates maximum similarity. When the vectors are independent, the angle they form is 90. The cosine of 90 is equal to 0, which indicates absence of similarity

Theoretic Background-

Is there a bond between an author's style and its semantics?

- One of the less investigated stylistic feature is the way in which authors use words from a semantic point of view:
  a. if they tend to use more, when dealing when polysemous words, a certain sense over the others
  b. If they tend to use senses that differ from the one that's more commonly used

- 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 the application of distributional semantics to the stylometric analysis of literary texts

Experimental Setup: Data preparation and Experiment Description

Data Set Construction

- In order to build reference and test corpora, we started from texts pertaining to the work of six Italian writers working at the turn of the 20th century.
- The selected authors were: Luigi Capuana, Federico De Roberto, Luigi Pirandello, Italo Svevo, Federigo Tozzi and Giovanni Verga.
- We used texts freely available for download from the digital library of the Manunzio project, via the Liber-Liber website
- The texts were encoded in various formats, such as .epub, .odt and .txt, so we pre-processed them by converting them all in .txt format and getting rid of all xml tags, together with footnotes and editors' notes and comments.

Why texts from the turn of the 20th century?

- An Italian literary movement peaked between approximately 1875 and the early 1900s: Verismo (meaning "realism", from Italian vero, meaning "true")

- Literary verismo did not constitute a formal school, but it was still based on specific principles. Its birth was influenced by a positivist climate which put absolute faith in science, empiricism and research

- It was based on Naturalism, a literary movement which spread in France in the mid-19th century. For naturalist writers, literature should objectively portray society and humanity like a photograph, strictly representing even the humblest social class in even its most unpleasant aspects, with the authors analysing real modern life like scientists

Why texts from the turn of the 20th century?

- Giovanni Verga and Luigi Capuana were main exponents of the literary verismo. Unlike French naturalism, which was based on positivistic ideals, Verga and Capuana rejected claims of the scientific nature and social usefulness of the movement

- Literary verismo developed in the urban cultural life of Milan, which brought together intellectuals from different areas, but tended to portray central and southern Italian life – especially Sicily (described in the works of Verga, Capuana and Federico de Roberto).

- All the authors we selected pertain to the temporal span in which the literary verismo developed, but not all of the them are proponents of such genre

Verist Authors vs. Non-Verist Authors

Verist Authors: a short biography

- Federico De Roberto (January 16, 1861 Naples – July 26, 1927 Catania) began his writing career as a journalist. Among all his works (novels and short stories), he is best known for his novel I Vicerè, published in 1894

- Giovanni Carmelo Verga (September 2, 1840 Vizzini– January 27, 1922 Catania) was best known for his depictions of life in Sicily, and especially for the short story Cavalleria Rusticana and the novel I Malavoglia (The House by the Medlar Tree)

- Luigi Capuana (May 28, 1839 Catania– November 29, 1915 Catania) was one of the first authors influenced by the works of Émile Zola. He was the author of plays, stories, novels, theatrical works and poetry in Sicilian

Verist Authors vs. Non-Verist Authors

Non-Verist Authors: a short biography

- Luigi Pirandello (June 28, 1867 Agrigento – December 10, 1936 Rome) was a dramatist, novelist, poet and short story writer. He was awarded the 1934 Nobel Prize for "his almost magical power to turn psychological analysis into theatre". His farces are seen as forerunners of the Theatre of the Absurd.

- Federigo Tozzi (January 1, 1883 Siena – March 21, 1920 Rome). He wrote poems and novels (the most known of which, Con gli occhi chiusi, is a highly autobiographical text). He is considered a classic of Italian modernism

- Italo Svevo, born Aron Ettore Schmitz (19 December 1861 – 13 September 1928), was a novelist, playwright, and short story writer. He is best known for his classic Modernist novel La Coscienza di Zeno (1923).

Modernism

- The three non-verist authors are, indeed, italian representative of another literary movement: Modernism

- Literary modernism, originated in the late 19th and early 20th centuries, mainly in Europe and North America, and is characterized by a self-conscious break with traditional ways of writing, in both poetry and prose fiction

- Modernists experimented with literary form and expression, adhering to Ezra Pound's maxim to "Make it new".

- The horrors of the First World War saw the assumptions about society reassessed, and modernist writers raised questions about the rationality of the human mind

Corpus Construction

- 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, (covering all the authors' different stylistic and thematic phases), we built a reference corpus as the composition of the 70% of each single work. The same technique was used to create the test corpus by using the remaining 30% of each work.

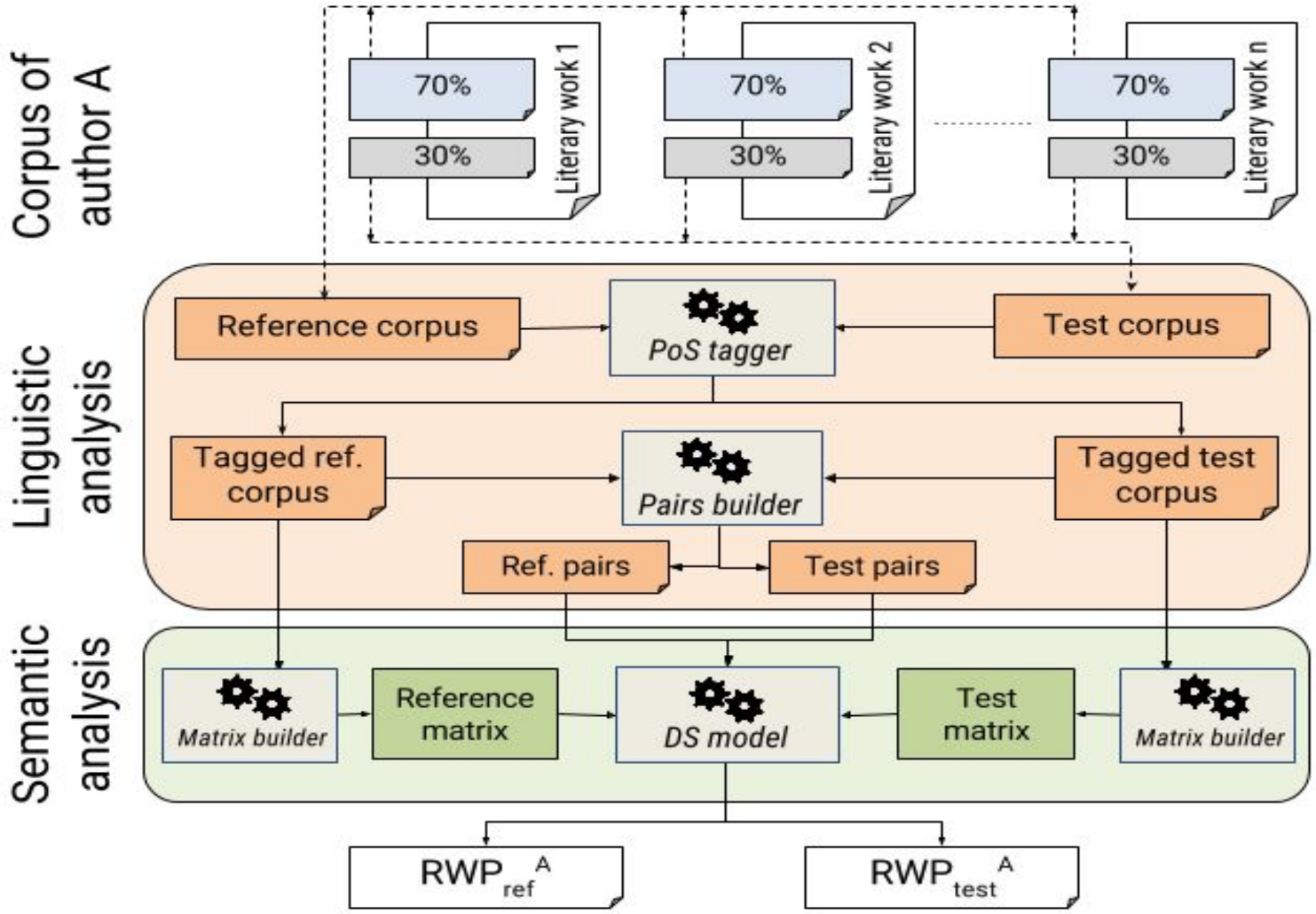
Corpus Construction

- Typical authorship attribution 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

- In order 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 train portion and a test portion.

Data Construction



Data Construction

- Each reference and test corpora was analyzed 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 co-occurrence matrixes for each author, using a context window of 4

Data Construction

Matrix construction

- 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 (thus not losing any information).

- We performed different empiric setup of the window's size and chose the one that showed more suitable results, according to Kruszewski and Baroni (2014).

- 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 to evaluate the semantic relatedness between the words by assessing their proximity in the distributional space

- We obtained two related word pair lists for each author A: RWP_{ref}^{A} and RWP_{test}^{A}

Data Construction

Removing lexical bias

- As already said, this is not a canonical AA task: we wanted to focus on the analysis of the semantic distribution of words. So we decided to exclude any possible “lexical bias”.

- We restricted the analysis on a common vocabulary, i.e. a vocabulary constituted by the intersection of the six authors’ vocabularies: by doing so, we wanted to prevent our classifier to exploit the presence of words used by some (but not all) of the authors.

- 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 could be pertaining with the signature lexicon of each author.

Data Construction

Removing lexical bias

- We computed the number of times words appeared together in the context window, as well as their total number of occurrences (#occ_a and #occ_b)

- 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.

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

Data Construction

Removing lexical bias

- 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 pairs with the highest cosine, selected using a percentage parameter θ

- We chose to introduce the parameter θ for two reasons:

  a. to avoid the classification algorithm to be disturbed by noisy (i.e. not significative) pairs which would not hold any relevant stylistic cue,
  b. 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.

Data Classification

- 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.

Classification Results

	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 showing classification results reported for each author and for each θ value.

Classification Results

	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 showing classification results obtained via the nearest-cosine method for θ = 5%

Classification Results

- To help in interpreting the failure of the algorithm in classifying Tozzi's test list for θ values lower than 5% we calculated the cardinality of the RWP test lists for each author with the change in θ value

- 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 (RWP_{test}^{Tozzi}).

- Increasing the value of θ and consequently the number of significant RW pairs that are analysed, the system is able to correctly classify RWP_{test}^{Tozzi}

Classification Results

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

- Cardinality of RWP_{test} for each author and for each θ value.

Discussion and Next Steps

Further works

- The here reported results seem to suggest that the way words are distributed across a text, can actually represent a stylometric characteristic of an author.

- 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.

Further works

- Another thing we plan to investigate in the future is the ability of our method to recognize the genre of a literary work

- For this reason we selected works pertaining to verism writer as well as modernist writer. Using the same data we already have, we can try to understand which works and authors are pertaining to verism and which are pertaining to modernism

Thank you for your attention

- Contacts:
  - giulia.benotto@ilc.cnr.it

- Literary Computing Group @ Institute for Computational Linguistics:
  - http://licolab.ilc.cnr.it

Short Biography

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

- 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 AK Halliday. 1994. Functional grammar. London: Edward Arnold.

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

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

- Hans Van Halteren et al. 2005. New machine learning methods demonstrate the existence of a human stylome. Journal of Quantitative Linguistics, 12(1):65–77