Authorial studies using ranked lexical features

Liviu P. DINU, Sergiu NISIOI
University of Bucharest, Faculty of Mathematics and Computer Science,
Centre for Computational Linguistics, Bucharest
ldinu@fmi.unibuc.ro, sergiu.nisioi@gmail.com

ABSTRACT

The purpose of this article is to propose a tool for measuring distances between different styles of one or more authors. The main study is focused on measuring and visualizing distances in a space induced by ranked lexical features. We investigate the case of Vladimir Nabokov, a bilingual Russian - English language author.

KEYWORDS: stylometry, L1 distance, translationes, rankings, Nabokov.

Proceedings of COLING 2012: Demonstration Papers, pages 125–130,
COLING 2012, Mumbai, December 2012.

1 Introduction. Selecting the right features

The features generally considered to characterize the style of an author are function words - conjunctions, prepositions, pronouns, determiners, particles, etc.. These words consist of non-content words, mostly words without a semantic referent. Also they have a crucial role in the construction of a phrase, holding syntactic features and tying semantics together. These words form the closed class set of a language and can easily be extracted from Wiktionary (url), a database which is constantly being improved and provides a great source of linguistic knowledge for languages that do not usually have tools and resources. The function words can sometimes be a compound token composed from two function words, for example some Russian declensions requires two words (e.g. for masculine, singular, prepositional case of ves' is the token obo vsem). We have treated this case by analysing the occurrences of an expression. Selecting the right lexical features is difficult, on one hand, using the entire list of function words from a language to designate the style of author has the disadvantage of containing words that are hapax legomena or do not exist in the analysed corpus. On the other hand, this disadvantage can provide a spectrum of used and unused words of an author, this being a mark of style. Also, it is a fixed feature that belongs to the language and does not depend on additional decisions regarding the corpus. This type of procedure is also discussed by (Argamon and Levitan, 2005). In order to obtain features that are strictly related to the corpus, one can concatenate all the texts to be analysed and extract the first n (function) words (Burrows, 2002),(Argamon, 2008). The drawback of this procedure is that we can not always know if the value chosen for n is optimal with regard to the expected hypothesis. Cases appear when completely different values of n increase the accuracy of the result depending on the type of language and other factors discussed by (Rybicki et al., 2011). Our tools can handle both of these situations, for the first we can input a list of tokens, one on each line and for the second we are developing a special semi-automatic process. This, second list is constructed from the first n most frequent words which, agreeing with the study of (Jockers and Witten, 2012), have a *mean relative frequency of at least 0.05%*. We want to implement a special procedure that, given n_1 and n_2, two integers, computes the adjusted Rand index (Hubert and Arabie, 1985) between the cluster i and the cluster $i-1$ with $n_1 < i \leq n_2$. The label for two clusters A and B to be joined will be given by $\min(A,B)$. This way we can label all the remaining clusters recursively. We were looking for a sequence of i where the clustering becomes stable and the adjusted Rand index remains close to 1. Meaning that when adding new words to the list we obtain the same result. The sequence obtained can be trusted to offer one literary hypothesis from the entire set that is the most stable to the way an author uses his most common words.

All the token - function words retrieved from a document are stored in a trie structure, with small caps. In a normal trie in each node there will be an extra field to indicate that the respective node is an end of word. We have used this field to indicate the frequency of each token. After filling the trie structure we can traverse it to retrieve the entire list of tokens with the frequencies. Because frequencies are positive integers we have used Radix sort, a non-comparison, sorting method by least significant digit in order to obtain the rank-ordered list of words in linear complexity. Computing distances or measurements between documents is more efficient this way.

Our algorithms are based on the use of rankings induced by the frequencies of function words, e.g the most frequent word has rank one, the second most frequent rank two and so on. We call a tie the case when two or more frequencies are equal. In order to solve ties we

apply the standard Spearman's rank correlation method. This means that if k objects claim the same rank (i.e. have the same frequency) and the first x ranks are already used by other objects then they will share the ranks and will receive the same rank number (the median from the k objects) which is in this case:

$$\frac{(x+1)+(x+2)+\cdots+(x+k)}{k} = x + \frac{(k+1)}{2} \tag{1}$$

Using rankings instead of raw frequencies has proved to offer better hypothesis regarding similarities between documents (Dinu et al., 2008).

2 Data visualisation

In order to inspect our results we have opted for a hierarchical clustering method based on the extension provided by (Szekely and Rizzo, 2005) for Ward's method. Their approach is concerned with increasing inner cluster homogeneity and inter-cluster heterogeneity. We have taken advantage of this joint-between within clustering method and we have adapted it for our $l1$ space to suit our purpose:

$$e_{l1}(A,B) = \frac{n_1 n_2}{n_1+n_2}\Big(\frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\| a_i - b_j \|_1 - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\| a_i - a_j \|_1 - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\| b_i - b_j \|_1\Big) \tag{2}$$

Where $A = \{a_1, \cdots, a_n\}$ and $B = \{b_1, \cdots, b_n\}$ are sets of size n_1 and n_2 respectively of m-dimensional vectors, $\| \cdot \|_1$ is the $l1$ norm. The Lance-Williams (Lance and Williams, 1967) parameters are exactly the same as for Ward's method (see (Szekely and Rizzo, 2005), Appendix):

$$d(C_i \cup C_j, C_k) = \frac{n_1+n_3}{n_1+n_2+n_3}\, d(C_i, C_k) + \frac{n_2+n_3}{n_1+n_2+n_3}\, d(C_j, C_k) - \frac{n_3}{n_1+n_2+n_3}\, d(C_i, C_j). \tag{3}$$

where n_1, n_2, n_3 are the sizes of cluster C_i, C_j, C_k and becomes:

$$d_{(ij)k} := d(C_i \cup C_j, C_k) = \alpha_i\, d(C_i, C_k) + \alpha_j\, d(C_j, C_k) + \beta\, d(C_i, C_j) \tag{4}$$

where

$$\alpha_i = \frac{n_i+n_3}{n_1+n_2+n_3}, \qquad \beta = \frac{-n_3}{n_1+n_2+n_3}, \qquad \gamma = 0.$$

Many of the valuable e properties proved only by coefficient handling like ultrametric property (Milligan, 1979) (i.e. $d_{ij} < \max\{d_{ik}, d_{jk}\}$) or space-dilatation (Everitt et al., 2009) (i.e $d_{k,(ij)} \geq \max\{d_{ki}, d_{kj}\}$) of the algorithm are inherited with this shift to e_{l1}. If A and B would be singletons, the e_{l1} distance is proportional to Manhattan distance and is recommended to be used with it and not with an Euclidean distance. Such an algorithm is best suited for our ranked data.

3 Measurements

3.1 Manhattan Distance

The most natural measure to be applied on an $l1$ space is Manhattan distance. When used on rankings it is also called Spearman's foot-rule or Rank distance by (Dinu and Popescu, 2008). Given two tied ranked vectors $X = \{x_1, \cdots, x_n\}$ and $Y = \{y_1, \cdots, y_n\}$ the equation for Manhattan distance is:

$$D(X,Y) = \sum_{i=1}^{n} |x_i - y_i| \tag{5}$$

Notice that the distance remains the same if our tied ranked vectors are obtained by an ascending ordering relation (e.g. assign rank one to the most frequent function word, rank two to the second most frequent and so on) or by a descending ordering relation . This is simple to prove once we observe that for some frequencies $\{f_1 > f_2 > \cdots > f_n\}$, that generated an ascending tied rank $X_> = \{x_1, \cdots, x_n\}$, its descending tied rank can be obtained by the next equation from $X_>$:

$$X_< = (n - X_>) + 1 \tag{6}$$

This suggests that ranking the frequencies does not imply just a simple change of the weights, but rather a change of space in which distances between documents become more measurable and more stable.

4 Application

Considering the previous studies (Dinu et al., 2008) regarding the use of lexical ranked features we have used our tools to investigate further the case of Vladimir Nabokov, a bilingual Russian - English language author. The works after 1941 are written in English. Those before, are in Russian.

We have gathered a corpus consisting of his original works in English: *The Real Life of Sebastian Knight* (1941), *Bend Sinister* (1947), *Lolita* (1955), *Pnin* (1957), *Pale Fire* (1962) , *Ada or Ardor: A Family Chronicle* (1969), *Transparent Things* (1972), *Look at the Harlequins!* (1974) together with the Russian translation of each. And his original works in Russian: *Mashenka* (1926), *Korol' Dama Valet* (1928), *Zashchita Luzhina* (1930), *Podvig* (1932), *Kamera Obskura* (1933), *Otchayanie* (1934), *Priglasheniye na kazn* (1936), *Dar* (1938) together with the English translation of *Mary* (translation year: 1970) , *The (Luzhin) Defence* (translation year: 1964), *Laughter in the Dark* (translation year: 1938), *Invitation to a Beheading* (translation year: 1959).

In the first image (Figure 1) we observe that the translated Russian period novels *Mary*, *Luzhin Defence*, *Camera Obscura* and *Invitation to a Beheading* are clustered separately from the novels written after 1940 in English. In the second image (Figure 2) we are looking at the Russian translations of the novels. A similar result with the previous one is presented: two clusteres, one with the original works in Russian and one with the translated ones. Nabokov's last Russian novel *Dar* is clustered near the English period.

Conclusion and perspectives

We have presented a reliable quantitative method with which we can measure distances between different styles of one or more authors. We have proved that, by using rankings,

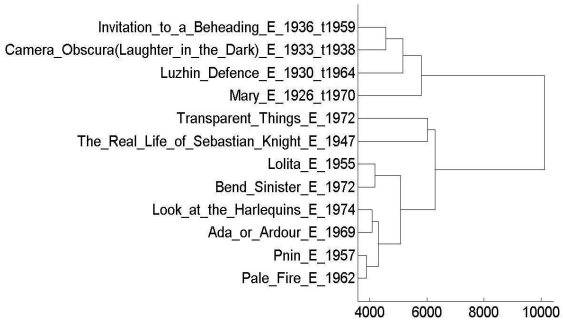

Figure 1: L1 distance applied with ranked lexical features of English.

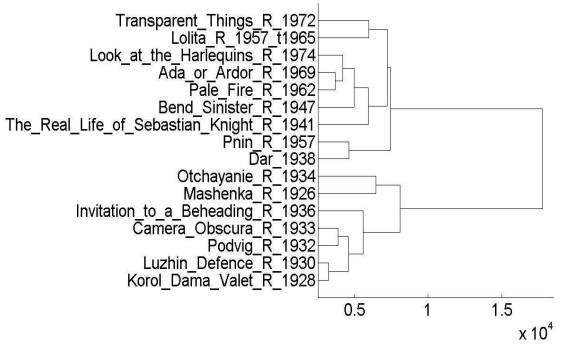

Figure 2: L1 distance applied with ranked lexical features of Russian.

Manhattan distance (or rank distance) was effective in distinguishing the style of an author. In future works we want to see if rankings can improve the accuracy of Burrows Delta (Burrows, 2002). Our $l1$ adapted clustering algorithm was able to distinguish between Nabokov's early Russian novels and his later English ones in both translation and original. Furthermore our results proved that on one hand Nabokov's style had changed significantly during his two literary periods and on the other hand that a translation affects a text in such a measure that stylistically it does not preserve the pattern of the original author. Therefore, although a method is oriented to simplicity, we can obtain significant results about an author's style if we rank an adequate set of lexical features. For further investigation we take into consideration the importance of pronouns in depicting an author's style. Moreover we intend to apply the methods on Samuel Beckett's and Milan Kundera's works and implicitly

to draw comparisons between English and French, but also Czech and French. During the course of our study we have found an interesting coincidence. Manhattan distance, which is identical with the distance a rook makes between two squares on a chess table, proved to be suitable for the literary works of Nabokov, who was a chess composer.

Acknowledgments

The research of Liviu P. Dinu was supported by the CNCS, IDEI - PCE project 311/2011, "The Structure and Interpretation of the Romanian Nominal Phrase in Discourse Representation Theory: the Determiners." We want to thank to Anca Bucur from the University of Bucharest for the helpful discussions. Note that the contribution of the authors to this paper is equal.

References

Wiktionary. `ru.wiktionary.org/`.

Argamon, S. and Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. In *In Proceedings of the 2005 ACH/ALLC Conference*.

Argamon, S. E. (2008). Interpreting burrows' delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2):131–147.

Burrows, J. F. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(1):267–287.

Dinu, A., Dinu, L. P., and Popescu, M. (2008). Authorship identification of romanian texts with controversial paternity. pages 2871–2874, Marrakech, Maroc. LREC, ELRA.

Dinu, L. P. and Popescu, M. (2008). Rank distance as a stylistic similarity. pages 91–94, Manchester. Coling, ELRA.

Everitt, B. S., Landau, S., and Leese, M. (2009). *Cluster Analysis*. John Wiley & Sons.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Jockers, M. L. and Witten, D. M. (2012). A comparative study of machine learning methods for authorship attribution. *Literary and Linguist Computing*, pages 215–223.

Lance, G. N. and Williams, W. T. (1967). A general theory of classificatory sorting strategies. *The Computer Journal*, 9(4):373–380.

Milligan, G. W. (1979). Ultrametric hierarchical clustering algorithms. *PSYCHOMETRIKA*, 44(3):343–346.

Rybicki, J., Eder, M., and Eder, M. (2011). Deeper delta across genres and languages: do we really need the most frequent words? pages 315–321.

Szekely, G. J. and Rizzo, M. L. (2005). Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of Classification*, (22):151 – 183.