Rank Distance as a Stylistic Similarity

Marius Popescu
University of Bucharest
Department of Computer Science
Academiei 14, Bucharest, Romania
mpopescu@phobos.cs.unibuc.ro

Liviu P. Dinu
University of Bucharest
Department of Computer Science
Academiei 14, Bucharest, Romania
ldinu@funinf.cs.unibuc.ro

Abstract

In this paper we propose a new distance function (rank distance) designed to reflect stylistic similarity between texts. To assess the ability of this distance measure to capture stylistic similarity between texts, we tested it in two different machine learning settings: clustering and binary classification.

1 Introduction

Computational stylistics investigates texts from the standpoint of individual style (author identification) or functional style (genres, registers). Because in all computational stylistic studies / approaches, a process of comparison of two or more texts is involved, in a way or another, there was always a need for a distance function to measure similarity (more precisely dissimilarity) of texts from the stylistic point of view. Such distance measures were proposed and used for example in authorship identification (Labbé and Labbé, 2001; Burrows, 2002) or clustering texts by genre (Luyckx et al., 2006).

In this paper we propose a new distance measure designed to reflect stylistic similarity between texts. As style markers we used the function word frequencies. Function words are generally considered good indicators of style because their use is very unlikely to be under the conscious control of the author and because of their psychological and cognitive role (Chung and Pennebaker, 2007). Also function words prove to be very effective in many author attribution studies. The novelty of our approach resides in the way we use information given by the function word frequencies. Given a fixed set of function words (usually the most frequent ones), a ranking of these function words according to their frequencies is built for each text; the obtained ranked lists are subsequently used to compute the distance between two texts. To calculate the distance between two rankings we used *Rank distance* (Dinu, 2003), an ordinal distance tightly related to the so-called *Spearman's footrule* (Diaconis and Graham, 1977).

Usage of the ranking of function words in the calculation of the distance instead of the actual values of the frequencies may seem as a loss of information, but we consider that the process of ranking makes the distance measure more robust acting as a filter, eliminating the *noise* contained in the values of the frequencies. The fact that a specific function word has the rank 2 (is the second most frequent word) in one text and has the rank 4 (is the fourth most frequent word) in another text can be more relevant than the fact that the respective word appears 349 times in the first text and only 299 times in the second.

To assess the ability of this distance function to capture stylistic similarity between texts, we tested it in two different machine learning settings: clustering and binary classification.

Compared with other machine learning and statistical approaches, clustering was relatively rarely used in stylistic investigations. However, few researchers (Labbé and Labbé, 2001; Luyckx et al., 2006) have recently proved that clustering can be a useful tool in computational stylistic studies. Apart of this, clustering is a very good test bed for a distance measure behavior. We plugged our distance function into a standard hierarchical clustering algorithm and test it on a collection of 21 nine-

Coling 2008: Companion volume – Posters and Demonstrations, pages 91–94
Manchester, August 2008

teenth century English books (Koppel et al., 2007). The results are very encouraging. The family trees produced grouped together texts according to their author, genre, even gender.

Also a distance measure can be used to solve classification problems if it is coupled with proper learning algorithm. One of the simplest such algorithms is nearest neighbor classification algorithm. We chose nearest neighbor algorithm because its performance is entirely based on the appropriateness to the data of the distance function on which it relies. In this way the accuracy of the classification will reflect the adequacy of the distance measure to data and domain on which the method was applied. We used the new distance function in conjunction with nearest neighbor classification algorithm and tested it on the well known case of authorship of disputed Federalist papers. The method attributed all disputed papers to Madison, the result being consistent with that of Mosteller and Wallace.

To check if the usage of ranks of function words is better suited for capturing stylistic differences than the usage of actual frequencies of the function words, we repeated the above experiments on clustering and binary classification with the standard euclidean distance between the vectors of frequencies of the same function words that were used in computing the rank distance. The comparison is in favor of rank distance.

2 Rank Distance and Its Use as a Stylistic Distance Between Texts

Rank distance (Dinu, 2003) is an ordinal metric able to compare different rankings of a set of objects. It is tightly related to the Spearman's footrule (Diaconis and Graham, 1977), and it had already been successfully used in computational linguistics, in such problems as the similarity of Romance languages (Dinu and Dinu, 2005).

A ranking of a set of n objects can be represented as a permutation of the integers $1, 2, \ldots, n$, $\sigma \in S_n$. $\sigma(i)$ will represent the place (rank) of the object i in the ranking. The Rank distance in this case is simply the distance induced by L_1 norm:

$$D(\sigma_1, \sigma_2) = \sum_{i=1}^{n} |\sigma_1(i) - \sigma_2(i)| \qquad (1)$$

This is a distance between what is called full rankings. However, in real situations, the problem of *tying* arises, when two or more objects claim the same rank (are ranked equally). For example, two or more function words can have the same frequency in a text and any ordering of them would be arbitrary.

a	been	had	its	one	that	was
all	but	has	may	only	the	were
also	by	have	more	or	their	what
an	can	her	must	our	then	when
and	do	his	my	shall	there	which
any	down	if	no	should	things	who
are	even	in	not	so	this	will
as	every	into	now	some	to	with
at	for	is	of	such	up	would
be	from	it	on	than	upon	your

Table 1: Function words used in computing the distance

The Rank distance allocates to tied objects a number which is the average of the ranks the tied objects share. For instance, if two objects claim the rank 2, then they will share the ranks 2 and 3 and both will receive the rank number $(2+3)/2 = 2.5$. In general, if k objects will claim the same rank and the first x ranks are already used by other objects, then they will share the ranks $x+1, x+2, \ldots, x+k$ and all of them will receive as rank the number: $\frac{(x+1)+(x+2)+\ldots+(x+k)}{k} = x + \frac{k+1}{2}$. In this case, a ranking will be no longer a permutation ($\sigma(i)$ can be a non integer value), but the formula (1) will remain a distance (Dinu, 2003).

Rank distance can be used as a stylistic distance between texts in the following way:

First a set of function word must be fixed. The most frequent function words may be selected or other criteria may be used for selection. In all our experiments we used the set of 70 function words identified by Mosteller and Wallace (Mosteller and Wallace, 1964) as good candidates for author-attribution studies. The set is given in Table 1.

Once the set of function words is established, for each text a ranking of these function words is computed. The ranking is done according to the function word frequencies in the text. Rank 1 will be assigned to the most frequent function word, rank 2 will be assigned to the second most frequent function word, and so on. The ties are resolved as we discussed above. If some function words from the set don't appear in the text, they will share the last places (ranks) of the ranking.

The distance between two texts will be the Rank distance between the two rankings of the function words corresponding to the respective texts.

3 Clustering Experiments

One good way to test the virtues of a distance measure is to use it as a base for a hierarchical cluster-

Group	Author	Book
American Novelists	Hawthorne	Dr. Grimshawe's Secret
		House of Seven Gables
	Melville	Redburn
		Moby Dick
	Cooper	The Last of the Mohicans
		The Spy
		Water Witch
American Essayists	Thoreau	Walden
		A Week on Concord
	Emerson	Conduct Of Life
		English Traits
British Playwrights	Shaw	Pygmalion
		Misalliance
		Getting Married
	Wilde	An Ideal Husband
		Woman of No Importance
Bronte Sisters	Anne	Agnes Grey
		Tenant Of Wildfell Hall
	Charlotte	The Professor
		Jane Eyre
	Emily	Wuthering Heights

Table 2: The list of books used in the experiment

ing algorithm. The family trees (dendogram) thus obtained can reveal a lot about the distance measure behavior.

In our experiments we used an agglomerative hierarchical clustering algorithm (Duda et al., 2001) with average linkage.

In the first experiment we cluster a collection of 21 nineteenth century English books written by 10 different authors and spanning a variety of genres (Table 2). The books were used by Koppel et al. (Koppel et al., 2007) in their authorship verification experiments.

The resulted dendogram is shown in Figure 1. As can be seen, the family tree produced is a very good one, accurately reflecting the stylistic relations between books. The books were grouped in three big clusters (the first three branches of the tree) corresponding to the three genre: dramas (lower branch), essays (middle branch) and novels (upper branch). Inside each branch the works were first clustered according to their author. The only exceptions are the two essays of Emerson which instead of being first cluster together and after that merged in the cluster of essays, they were added one by one to this cluster. Apart of this, the family tree is perfect. Even more, in the cluster of novels one may distinguished two branches clearly separated that can correspond to the gender or nationality of the authors: female English (lower part) and male American (upper part).

For comparison, the dendogram in Figure 2 show the same books clustered with the same algorithm, but using the standard euclidean distance instead of the rank distance as measure of stylistic similarity. The same set of function words as in the case of rank distance was used. This time though, each text was represented as a vector of

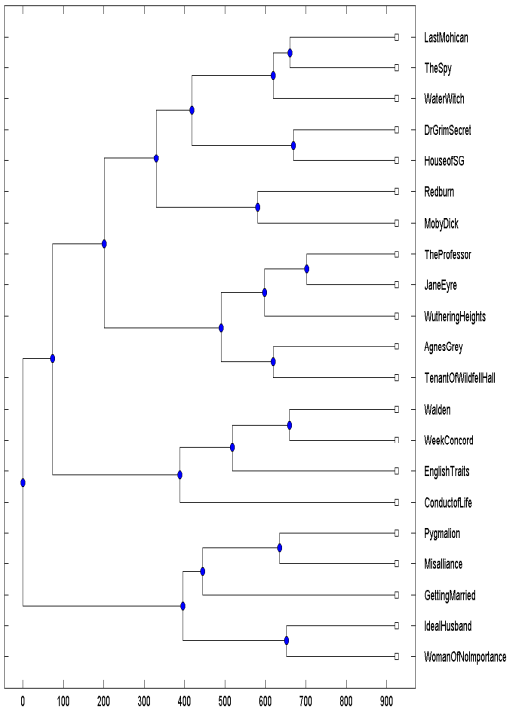

Figure 1: Dendogram of 21 nineteenth century English books (Rank Distance)

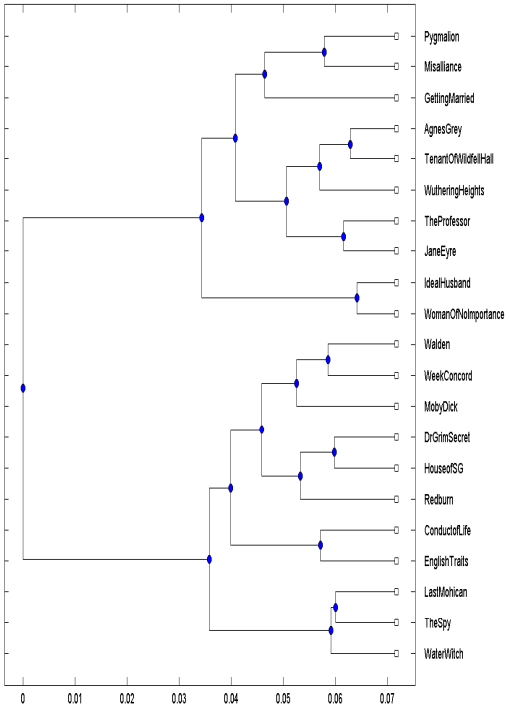

Figure 2: Dendogram of 21 nineteenth century English books (Euclidean Distance)

relative frequencies of these function words in the text. The relative frequency of a particular function word in a text is calculated as the number of appearances of the respective function word in the text divided by the length (in tokens) of the text. The distance between two texts is given by the euclidean distance between the corresponding vectors of relative frequencies of function words. In the family tree obtained using euclidean distance, most of the books are still grouped according to their author, but the distinct clusters corresponding to genre and gender disappeared and the novels of Melville were separated: one being clustered with the essays of Thoreau (Moby Dick) and the other with the novels of Hawthorne.

4 Binary Classification Experiments

When a distance measure is available, the most natural choice of a classification algorithm is the nearest neighbor algorithm (Duda et al., 2001).

We tested the nearest neighbor classification algorithm combined with both rank distance and euclidean distance on the case of the 12 disputed federalist papers (Mosteller and Wallace, 1964). In our experiments we followed the Mosteller and Wallace setting, treating the problem as a binary classification problem. Each one of the 12 disputed papers has to be classified as being written by Hamilton or Madison. For training are used the 51 papers written by Hamilton and the 14 papers written by Madison.

Tested on disputed papers, the nearest neighbor classification algorithm combined with rank distance attributed all the 12 papers to Madison. This matches the results obtained by Mosteller and Wallace and is in agreement with today accepted thesis that the disputed papers belong to Madison. When the nearest neighbor classification algorithm was combined with euclidean distance only 11 papers were attributed to Madison, the paper 56 was attributed to Hamilton.

5 Discussion

In this paper we have proposed a new distance measure based on the ranking of function words, designed to capture stylistic similarity between texts. We have tested it in two different machine learning settings: clustering and binary classification; we have compared its performance with that of standard euclidean distance on vectors of frequencies of the function words. Though testing on more data is needed, the initial experiments shown that the new distance measure is indeed a good indicator of stylistic similarity and better suited for capturing stylistic differences between texts than the standard euclidean distance.

In future work it would be useful to test this distance measure on other data sets and especially in other machine learning paradigms like one-class classification to solve authorship verification problems (Koppel et al., 2007).

Acknowledgments Research supported by MEdC-ANCS, PNII-Idei, project 228 and University of Bucharest.

References

Burrows, John. 2002. 'delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287.

Chung, Cindy K. and James W. Pennebaker. 2007. The psychological function of function words. In Fiedler, K., editor, *Social communication: Frontiers of social psychology*, pages 343–359. Psychology Press, New York.

Diaconis, P. and R.L. Graham. 1977. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(2):262–268.

Dinu, Anca and Liviu Petrisor Dinu. 2005. On the syllabic similarities of romance languages. In *CICLing-2005*, pages 785–788.

Dinu, Liviu Petrisor. 2003. On the classification and aggregation of hierarchies with different constitutive elements. *Fundamenta Informaticae*, 55(1):39–50.

Duda, R. O., P. E. Hart, and D. G. Stork. 2001. *Pattern Classification (2nd ed.)*. Wiley-Interscience Publication.

Koppel, Moshe, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8:1261–1276.

Labbé, Cyril and Dominique Labbé. 2001. Intertextual distance and authorship attribution corneille and moliere. *Journal of Quantitative Linguistics*, 8(3):213–231.

Luyckx, Kim, Walter Daelemans, and Edward Vanhoutte. 2006. Stylogenetics: Clustering-based stylistic analysis of literary corpora. In *Proceedings of LREC-2006, the fifth International Language Resources and Evaluation Conference*, pages 30–35.

Mosteller, Frederick and David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Massachusetts.