Towards a classification system for Arabic texts

Rami AYADI

ISIM-Sfax institute & UTIC laboratory
Ayadi.rami@planet.tn

Maraoui MOHSEN

UTIC laboratory – Monastir TUNISIA
maraoui_m7@yahoo.fr

Mounir ZRIGUI

UTIC laboratory – Monastir TUNISIA
Mounir.zrigui@fsm.rnu.tn

Abstract

Our researches works are interested on the application of the intertextual distance theory on the Arabic language as a tool for the classification of texts. This theory assumes the classification of texts according to criteria of lexical statistics, and it is based on the lexical connection approach. Our objective is to integrate this theory as a tool of classification of texts in Arabic language. It requires the integration of a metrics for the classification of texts using a database of lemmatized and identified corpus which can be considered as a literature reference for times, kinds, literary themes and authors and this in order to permit the classification of anonymous texts.

1. Introduction

The abundance of information, due to the development of the Internet, storage media (containing large text corpora) and encyclopedias scanned, makes it difficult if not impossible, its exploration and analysis. Hence arises the need to explore new approaches to using automatic text analysis.

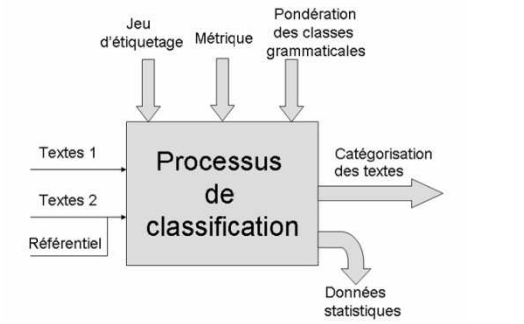
The classification of texts is defined as an operation which identifies classes of equivalence between segments of texts reflecting their information content (words, n-gram, etc.).. Therefore we have to define the degree of similarity or dissimilarity between the segments. Us within the framework of a mathematical and statistical approach, we have the opportunity to explain this level of digital evidence. Several approaches are used to define each method to calculate the degree of similarity or dissimilarity, especially the theory of intertextual distance which was introduced by Charles Muller (Muller, 1977) in the late sixties under the name of connection lexicale to meet the need to define a metric for measuring the degree of similarity between texts. This theory is based on statistics relating to vocabulary and expressions used in the texts, creating a need for decomposition according to a set of grammatical classes and texts by means of lemmatiszeurs.

The classification does not represent an end in itself, classifying the treatment must be a precise moment in a learning process much more complex. That is why we propose in our research work to establish a platform for text classification in Arabic based on the theory of intertextual distance. This requires the definition of a process that encompasses all the steps needed for classification. There is also a need to establish a repository for the classification of texts in the form of a structured corpus. This repository contains defined and categorized corpus used as a benchmark for classification. This definition of the process has enabled us to achieve a platform classification of texts into Arabic. This platform provides a generic lemmatiseur based learning that allows text to label giving the user the choice of game to use labels (Jaouadi, 2006). It also offers a classifier using an enhanced version of the following formula of the theory of intertextual distance to associate the game of labeling the concept of weight. Finally we have come to associate with this platform a structure to house the repository of classification as a categorized and selected corpus.

2. Architecture of the classification system

The system we realize a fundamental aim of allowing the classification of texts into Arabic for purposes of categorization and indexing. To do this we propose to define a process to have as input a text and present the output of the categorization. This categorization can be compared to an existing reference or relative to another text input. Our use of the theory of intertextual distance for the establishment of a metric classification us to the integration process at a stage lemmatization texts

(Pretreatment). This step is necessary to prepare the texts decomposing which allows the use of grammatical structures in detection of classes of equivalence between segments of text. We have exploited the wealth of the grammar of the Arabic language to incorporate the concept of grammatical classes in the metric in this way lemmatiseur game operates independently of the structure adopted and we introduce the concept of weight associated with the grammatical classes level metrics.



2.1 A formula based on the weighting of grammatical classes

Intertextual distance is a measure of degree of similarity or dissimilarity between texts. To be able to say whether the texts are "closer" or "fairly distant" relative to the use of a common vocabulary. The distance between two texts is measured by the number of words (tokens) that contain different (formulas in Labbé & Labbé 2001). This measure is a distance - and not a simple measure of dissimilarity - because it has three characteristic properties:

- Positivity: d (a, b) $\geq$ 0 and d (a, a) = 0 (the distance of a text itself is zero if d (a, b) = 0, then A and B contain the same words with the same frequencies in the same grammatical class);
- Symmetry: d (a, b) = d (b, a) (the result is the same as the measurement is carried out first with A or B)
- Triangular inequality: d (a, b) $\leq$ d (a, c) + d (c, b) (equality in a single case where the C is a subset of A and B)

The question of intertextual distance was first debated in the late sixties by the pioneer of statistical lexicology Charles Muller as lexical connection (Muller 77) which defines the intersection between the language of both texts. The indices used were not taking into account the difference between the sizes of the texts. This had the effect of forcing Muller use sliced pieces of equal size. Then, Etienne Burnet took up this problem by applying the theory on vocabulary Giraudoux (Burnet 88) where he proposes to calculate the distance between two texts is based solely on the number of common words regardless of their frequency and on text size without taking into account the difference in sizes. In 90 years, Dominique Labbé (Labbé 2001) was also interested in the intertextual distance, based on its predecessors. He proposed that the frequency of use of each of the terms, ie the overall extent of texts compared, also taking over the difference between sizes. The term "intertextual distance", the word text indicates that the calculations cover the whole text (N) and not on their own vocabulary (V).

Our approach to the theory of intertextual distance is essentially motivated by a need to define a metric for the classification of texts into Arabic. This is our starting point, we considered the problem from this point of view to reformulate the problem so that it more responsive to our needs. We identified the following points:

- The formula Labbé is our starting point
- The intertextual distance is essentially based on the respective vocabularies of texts to compare.
- A vocabulary is the set of words used to generate text. These words belong to different grammatical classes
- The principle of the scheme of Labbé is to compare the frequency of vocabulary in both texts regardless of the nature of vocabulary.

We then ask a very important issue for us to know: is that all grammatical classes have the same relevance in the vocabulary?

We can reformulate in:

(i) Is that all grammatical classes have the same relevance in the writing of text?

(ii) Can we define a number of classes as a parasite vocabulary?

We found, through some experiments by eliminating a number of classes such as numbers and names, as the comparison index improved more and more. Hence the idea of parasite in the vocabulary. And through our reading, we've noticed that the level of other methods of classifying and indexing the concept already exists for example in the work on n-grams of Jalam Radwan and Jean-Hugues Chauchat (Jalam, 2002) or term is often to define words that by chance one of the n-grams features of the class, but the word itself is interesting. In our case we will call any parasite component that could distort the result by its irrelevance or the fact that it is not representative of the style vocabulary of the author.

Note for example that the names and numbers can be regarded as not representative of lexical style of an author. More specifically, comparing the same text written by the same person describing two characters will not give a zero which is the logical result because the names and dates differ thus distorting the result.

To remedy this deficiency that we met, we propose to develop the concept of vocabulary in a text. We will consider a vocabulary V as the sum of Vi vocabularies where i refers to the various classes grammar.

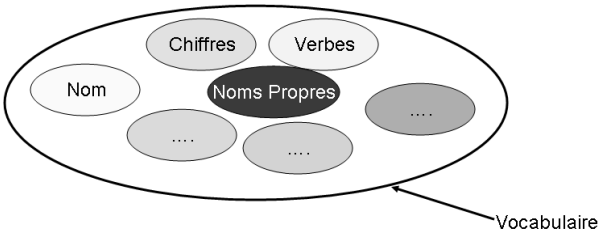


Figure 2 : les classes grammaticales dans le vocabulaire
V=∑Vi ; i ♠ {Classes grammaticales}

To answer the other question, namely the variation in degree of importance of grammar classes, we have a concept of relative weights to each class grammar. Thus we will define a variable representing the weight in the vocabulary, this variable will be related to grammatical classes. Ai is therefore the weight of the grammatical class i. Note that the value of αi is between 0 and 1 and this take us back to where it is equal to 1 at the approach of D. Labbé, which is our starting point.

2.2 Mathematical Formulation

Two texts A and B have respectively the vocabularies Va and Vb. We use the method Labbé to put the same size and get the vocabulary Va and V'b. Then we integrate the weights in the frequency of vocabulary according to grammatical class membership.
The new formulation: αCi is the weight of class C to which the term i
This gives the following distance:

- absolute distance between the vocabulary of text A and B '(B' = B reducing the size of A)

$$D_{V_{a,b(u)}} = \sum_{V_a, V'_{b(E)}} \alpha_{Ci} \mid F_{ia} - E_{ia(u)} \mid$$

- Relative distance:

$$D_{(a,b)} = \frac{\sum_{V_a, V'_{b(E)}} \alpha_{Ci} \left| F_{ia} - E_{ia(u)} \right|}{\sum F_{ia} - \sum E_{ia(u)}} = \frac{\sum_{V_a, V'_{b(E)}} \alpha_{Ci} \left| F_{ia} - E_{ia(u)} \right|}{N_a - N'_b}$$

$$E_{ia(u)} = F_{ib} * U_{(a,b)} \quad et \quad U(a,b) = \frac{N_a}{N_b}$$
$$N'_b = \sum_{V_b} E_{ia(u)}$$

2.3 Advantages of the integration weights

The new formula supports the reduction of noise level vocabularies by cancellation or mitigation of the impact of certain grammatical classes at frequencies of vocabularies and this is set to zero by weight or by its decrease. Also taking care of other criteria in the calculation of the intertextual distance as the weighting of grammatical classes at a number of authors, literary genres ... In fact by assigning a number of weighting each reference text it is possible to make the index more specific distance to a reference and thus strengthen its representation of the projection of the text to compare the reference retenue.et finally the generalization of the formula written by Labbé. Indeed, this new formulation allows for a configuration αi = 1 to comply with the approach of Labbé.

3. Classification trees

For this experiment, we used a set of 10 texts formed from a textbook, all texts are subject to the same treatment: spelling correction, standardization of spellings and lemmatization
Calculates the distance intertextual applied to 10 texts of our corpus - taken two to two - generates a table of 100 cells - 10 columns and 10 lines - whose size prohibits reproduction.

	Nbr de mots	Nbr de vocable		Nbr de mots	Nbr de vocable
Text1	3443	3122	Text6	3225	3079
Text2	3943	3655	Text7	3738	3403
Text3	4255	3935	Text8	5431	5057
Text4	2949	2720	Text9	4388	4121
Text5	3193	2902	Text10	3849	3599

By virtue of ownership of identity, the diagonal of this table is zero (90 cases non-zero) and the fact the property of symmetry, there are 45 different distances (90 / 2).
As indicated in Labbé 2007, the shortest distances bring together correctly almost all the texts. However, before such large populations, the use of classifications is a need to better visualize the results.
The usual method is to represent all of the texts by points whose coordinates in space are determined by their position relative to all others. Here the 10 texts form a "cloud" of points, including 45 different distances.
The analysis tree based on the assumption that if all individuals studied is separated by distances (with all three properties listed above), there is a "tree" that represents exactly the positions of these individuals by some Compared to other (for more details or Luong 1988). However, the construction of such a tree "perfect" would require that all possible combinations are examined, while their number is increasing exponentially due to the size of the series (the conclusion returns to this point). Various algorithms have been devised to build the tree without having to consider all these combinations. We use the algorithm developed by

X. Luong (the principles and formulas are also presented in Luong 1994).

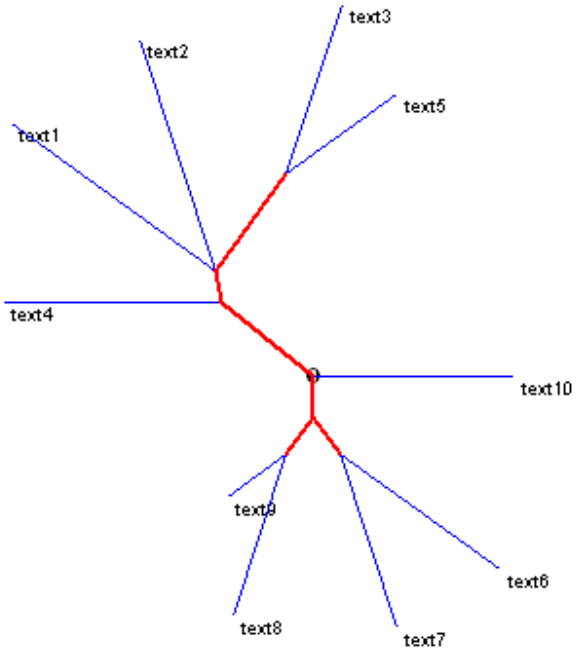


Figure1: A tree classification for 10 text

4. Conclusion

Our work focused on the classification of texts in Arabic based on the theory of intertextual distance. To do so we made a system to from a number of components to achieve this tâche.Une both reached our initial goal, we propose a third component of our work to establish the repository. This is a detailed study of various books and the selection, with the assistance of an expert in the field of Arabic language texts representing a set of criteria to genres, types, registration, themes, periods, authors. We propose to conduct an experimental study that will allow us to associate each text reference to a set of labeling and weighting of grammatical classes can enhance its features. This experimental study will finalize the system of automatic classification of text in Arabic based on the theory of intertextual distance.

Références

MULLER CHARLES (1997), PRINCIPES ET METHODES DE STATISTIQUE LEXICALE, PARIS. HACHETTE UNIVERSITE.

BRUNET E.(1988) Une mesure de la distance intertextuelle : la connexion lexicale, Le nombre et le texte. Revue informatique et statistique dans les sciences humaines, Université de Liège.

DEERWESTER S., DUMAIS. S. T., FURNAS, G. LANDAUER. T. K.HARSHMAN (1990). Indexing by latent semantic analysis. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 391-407.

JALAM R., CHAUCHAT J. H.(2002), Pourquoi les n-grammes permettent de classer des textes ? recherche de mots-cléfs pertinents à l'aide des n-grammes caractéristiques. JADT 2002 : 6ES JOURNEES INTERNATIONALES D'ANALYSE STATISTIQUE DES DONNEES TEXTUELLES.

JAOUADI W., ZRIGUI M.(2006), Application de la théorie de la distance intertextuelle pour la classification de texte en langue arabe. CONFERENCE REALITER 2006, RIO DE JANEIRO.

LABBE D.(2002), L'attribution d'auteur et la distance intertextuelle. Revue Corpus.

LABBE D. ET LABBE C.(2003), La distance intertextuelle. REVUE CORPUS 2, LABORATOIRE "BASES, CORPUS ET LANGAGE", UMR 6039 DU CNRS.

LABBE D.(2003), Réponses à M. J.-M. VIPREY Corneille et Molière.

LABBE D.(2004), CORNEILLE ET MOLIERE. 7E JOURNEES D'ANALYSE DES DONNEES TEXTUELLES.