Mixed Style Feature Representation and B_0-maximal Clustering for Style Change Detection

Notebook for PAN at CLEF 2020

Daniel Castro-Castro[1], Carlos Alberto Rodríguez-Losada[1], and Rafael Muñoz[2]

[1] Oriente University, Santiago de Cuba, Cuba
{danielbaldauf, carlosarl1999}@gmail.com
[2] Department of Software and Computing systems, Alicante University, Spain
rafael@dlsi.ua.es
http://www.dlsi.ua.es

Abstract The goal of Style Change Detection task in a document is to determine if it was written by more than one author and in such case, to delimit which paragraph (or more generally a portion of text) corresponds to each one of them. The objective of our proposal is to build a paragraph representation based on general Style Feature computed considering characters, lexical and syntactic features, without the use of semantic words. The paragraphs were grouped employing a non overlapped variant of the B_0-maximal clustering algorithm, where the overlapping was eliminated considering the order of paragraphs in the document.

1 Introduction

Authorship detection is important for determining which author or group of authors should get credit for writing a given document. In particular, in our digital modern society, it is a complex task when the objective is to determine who wrote a piece of digital text or if a document could be written by more than one author.

Thanks to the research community and in particular to the organizers of PAN [3] evaluation forum [4], in recent years there is a growing interest in sharing methods and algorithms to solve many of the tasks involved in Authorship Attribution (AA). One of these tasks is the Style Change Detection in a document, with the purpose of detection if a document was written by only one author or more than one, and in the last scenario, what piece of text corresponds to each one of the authors [1].

Overviews of the past style change detection task [5][3][6] resumed the description of the task, approaches presented by participants and the results obtained. It is important to highlight, that a priori, there is no information about authors or the numbers of them involved in a problem, that's why, the tasks are mainly solved considering text clustering solutions.

[3] http://pan.webis.de

One of the key aspects tackled to solve the task, corresponds to, the representation of textual contents, and in the majority of proposals, it was used the Bag of Word model considering lexical and syntactical linguistic features. For clustering algorithms have been used hierarchical and non-hierarchical traditional methods. Also, due to the nature of the task, the clusters of documents may not be overlapped, because a document or a paragraph (depending of the practical problem proposed) belongs to an unique author, so then, a document or paragraph must be part of just one cluster (or group).

In section 2 our proposal is described with emphasis in paragraph representation based on the construction of a Mixed Style Set of Features and the clustering algorithm employed. Section 3 presents the results obtained and the conclusions of the work, and in Section 4 a brief discussion of the main problems.

2 Proposal for Style Change Detection - PAN 2020

Our main goal is to determine clusters of paragraphs in which is considered that all paragraphs in a cluster are written by the same author. If the algorithm obtains more than one cluster, then the document was written by more than one author and the number of distinct authors corresponds to the number of clusters.

In the next two sections are described the representation of paragraphs and the clustering algorithm. The computational representation is based on the formulations proposed in Logical Combinatorial Pattern Recognition [4] and the clustering algorithm is explained in [2].

2.1 Paragraph Mixed Style Feature Representation and Similarity

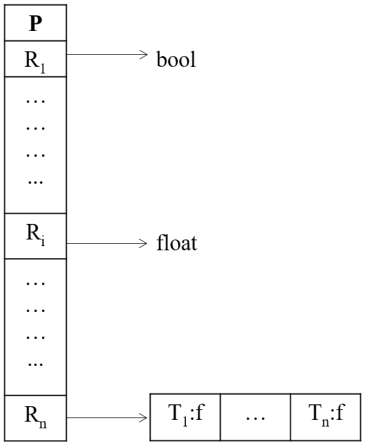
The objective of our proposal was to build a paragraph representation based on general Style Feature computed considering characters, lexical and syntactic features, without the use of content words. In Figure 1, are illustrated the paragraph representation and similarity functions implemented to compare two of them.

The paragraph representation is build considering a finite mixed set of 185 features from three types of data values, Boolean, Float or n-gram Vector. At the left corner in Figure 1 there is a section "Features examples" with three examples of the type of features analyzed.

Features were structured in six subsets considering different textual layers on the text. These layers are boolean, character, sentence, paragraph, syntactic and the text.

Examples for each of the layers subset of features:
1- Boolean layer: Uses the same word to finish a sentence and to begin the next sentence.
2- Character layer: Average length of words.
3- Sentence layer: Average number of words. Average number of distinct prepositions.
4- Paragraph layer: Average number of sentences. Average number of words.
5- Syntactic layer: Proportion of nouns over adjective.
6- Text layer: Average length of sentence. Bag of Words of conjunctions.
In order to compare two representations, one comparison criteria (CC) for each type of

[4] https://www.uci.cu/reconocimiento-logico-combinatorio-de-patrones



$$\mathcal{F}(P_i, P_j) = \begin{cases} 1\ if\ \left(\sum_{s=1}^{n} CC(R_{sPi}, R_{sPj})\right) > \beta_0 \\ 0\ in\ other\ case \end{cases}$$

$$CC_{bool}(R_{sPi}, R_{sPj}) = \begin{cases} 1\ if\ R_{sPi} == R_{sPj} \\ 0\ in\ other\ case \end{cases}$$

$$CC_{float}(R_{sPi}, R_{sPj}) = \begin{cases} 1\ if\ \left|R_{sPi} - R_{sPj}\right| < \gamma \\ 0\ in\ other\ case \end{cases}$$

$$CC_{vector}(R_{sPi}, R_{sPj}) = \begin{cases} 1\ if\ sim(R_{sPi}, R_{sPj}) > \delta \\ 0\ in\ other\ case \end{cases}$$

Features examples

R_1: if two or more sentence begin with the same word.
R_i: proportion of Personal Pronouns over all pronoun used.
R_n: 1-gram of auxiliary words.

Figure 1. Description of paragraph representation and similarity function

data feature value is introduced. The three CC formulas are exposed at the right section in Figure 1. For features of Bool type, two features are similar, if they have the same value (true or false), see CC_{bool}. For features of Float type, two features are similar, if the difference between values are less than a predefined threshold, see CC_{float}. For features of Vector type, two features are similar if the similarity between them is greater than a predefined threshold, see CC_{vector}. We used $MinMax$[5] similarity to compare two vector. Finally, the two paragraphs are similar, if the number of features, in which they are similar, are greater than a percentage defined, see $F(P_i, P_j)$.

2.2 B_0-maximal Clustering Method

The clustering proposal generates all the subsets of paragraphs in order to achieve that the similarity between each of the paragraphs in a cluster should be larger than a predefined B_0 parameter. The B_0-maximal clustering algorithm obtains compact groups of paragraphs and some overlapped groups.

For the task, this overlapping needs to be eliminated and we used an approach based on the order of paragraphs in the document. If a paragraph could be part of two or more clusters, it will be considered only in the cluster where a paragraph with the lower index in the order of appearance in the text exists. This decision is based on the assumption that the style in a document are characterized by the style reflected in the firsts paragraphs and in general the main author tend to write the majority of the paragraphs and the firsts one. To accomplish that, the overlapped paragraphs were sorted by their index of appearance in the document. When the cluster assignment was defined, then all edges from these paragraph to other clusters were eliminated.

In Figure 2 and Figure 3 are presented an example based on a graph construction, where the vertices are the paragraphs and the number of the vertices, the order of the paragraphs in the document. The edges that connect two vertices represent that the similarity of vertices is greater or equal than a B_0 parameter and in our proposal we use a percentage of similar features.

The Figure 2 corresponds to the output of the clustering algorithm, and it can be seen that paragraph 4 and 5 could be part of two clusters. Considering the heuristic explained to eliminate the overlapping, the final clusters will correspond to the two illustrated at Figure 3.

3 Evaluation

The data-set distributed contains documents for two problems of Style Change Detection, a narrow data-set and a wide data-set [1]. The description of the data and evaluation measures are discussed on the overview published for the task.

In Table 1, are resumed the average results for task1 and task2 considering results in both data-set. For task1 the objective was to answer if a document was written by one author or more than one. In task2 had to be answered, in which paragraph (could be more than one) of the text there was a style change. As an additional data, the organizers informed, that a maximum of three authors could be involved in a document,

[5] https://rdrr.io/cran/stylo/man/dist.minmax.html

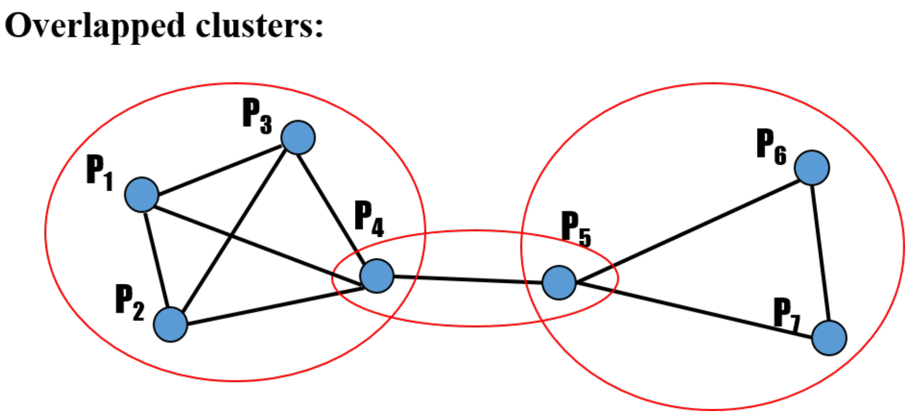


Figure 2. Example of B_0-maximal compact cluster graph representation

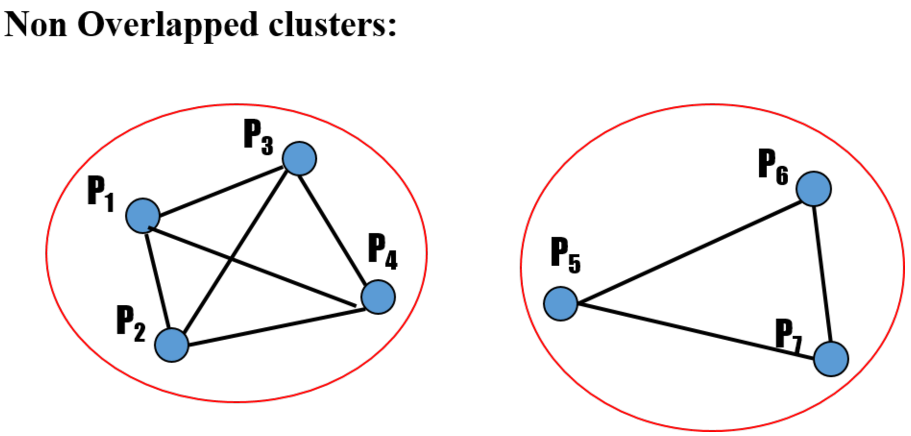


Figure 3. Example of B_0-maximal compact cluster graph representation, with non overlapped clusters

but our proposal is not restricted by a predefined number of clusters. Task1 was evaluated by F_1 measure and task2 using $micro - F$ measure. Using train and validation data-set distributed for the task, it was selected the values for parameters B_0, γ and δ, considering all combinations of these three.

Table 1. Style Change Detection results.

Team	task1	task2
iyer20	**0.640**	**0.856**
castro20	0.539	0.757
nath20	0.520	0.752

As a baseline we considered for task1 that the answer was always multi-authored, and as the data-set are balanced in the number of problems for multi-authored and single-authored documents, the result is 0.5. Similar result is obtained if the answer were single-authored for all documents. For task 2 we could not compare the results with a baseline.

4 Discussion

Using training and validation data-set, we got better results processing the wide data-set than the narrow one, and this is interesting, considering that we did not use content (topic related) words, concluding that the syntactic and structural style features are used differently when the topics change. At the contrary, we got no significant difference of style between authors when they wrote about the same topic.

Several of the features get duplicated values, because they capture the same values, considering that two or more structural layers are fused, for example, when the unit to be analyzed as a document is a paragraph, then the paragraph layer and text layer are considered distinct but they are the same.

5 Conclusion and Future Work

It was presented a proposal based on a paragraph representation, considering general Style Features at character, lexical and syntactic layers of analysis, without the use of topic or content words. The paragraphs were grouped employing a non overlapped variant of the B_0-maximal clustering algorithm, where the overlapping was eliminated considering the order of the paragraph in the document.

As future work could be interesting to combine semantic and topic vector representation as features of the mixed model in order to distinguish between paragraph of different topics. Also, the heuristic employed to eliminate the overlapping scenarios can be improved, if some characteristics of the groups are considered, for example: the size, strength of similarity or the adjacency of paragraphs.

References

1. Eva Zangerle, Maximilian Mayerl, G.S.M.P.B.S.: Overview of the Style Change Detection Task at PAN 2020. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
2. Gil-García, R., Badía-Contelles, J.M., Pons-Porrata, A.: A parallel algorithm for incremental compact clustering. In: Euro-Par. Lecture Notes in Computer Science, vol. 2790, pp. 310–317. Springer (2003)
3. Kestemont, M., Tschuggnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection. In: Cappellato, L., Ferro, N., Nie, J., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2125/invited_paper_2.pdf
4. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World. Springer (Sep 2019)
5. Tschuggnall, M., Stamatatos, E., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Overview of the author identification task at PAN-2017: style breach detection and author clustering. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, vol. 1866. CEUR-WS.org (2017), http://ceur-ws.org/Vol-1866/invited_paper_3.pdf
6. Zangerle, E., Tschuggnall, M., Specht, G., Stein, B., Potthast, M.: Overview of the style change detection task at PAN 2019. In: Cappellato, L., Ferro, N., Losada, D.E., Müller, H. (eds.) Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019. CEUR Workshop Proceedings, vol. 2380. CEUR-WS.org (2019), http://ceur-ws.org/Vol-2380/paper_243.pdf