COMSATS Institute of Information Technology, Lahore, Pakistan

Note book for PAN at CLEF 2016

Abdul Sittar, Hafiz Rizwan Iqbal and Rao Muhammad Adeel Nawab

{ abdulsittar72@yahoo.com, rizwan.iqbal@ciitlahore.edu.pk, adeelnawab@ciitlahore.edu.pk }

Author Diarization (Task A, Task B, Task C)

Abstract

Author Diarization is a new task introduced in PAN'16, toidentify portion(s) of text with in a document written by multiple authors. This paper presents, our proposed approach for author diarization task. Various types of stylistic features which include lexical features, used to uniquely identify an author. Furthermore, to find anomalous text within a single document, ClustDist method used. Finally, clusters were generated by using simple k-means clustering algorithm. Experiments were performed both on training and testing data sets. It has been observed that by changing the text fragments length, promising results can be achieved.

Lexical Features

1.Characters Count
2. Digits Count
3. Uppercase Letters Count
4. Spaces Count
5. Letters Count
6. Tabs Count
7. Ratio of Interrogative Sentence
8. Words Count
9. Average Word Length
10. Ratio of Digits
11. Average Sentence Length
12. Ratio of Spaces
13. Ratio of Upper case letters
14. Ratio of Letters
15. Ratio of Tabs

ClustDist Function

Technique to compute the average distance from one portion of text to all other pieces of text, and then calculate the average of all resulted distances.

- ClustDist Formula
- Document with n sentences
- d = distance
- V = Feature Vector

$$ClustDist(\vec{x}, V) = \frac{\sum_{k} d(\vec{x}, \vec{v})}{n}$$

Step-by-Step Author Diarization by ClustDist Approach

Step 1: Read Raw Input Text

Step 2: Break Down Text into Sentences

Step 3: Lexical Features Computation

Step 4: Distance Calculation

Step 5: ClustDist Computation

Step 6: Generating Clusters

Training Dataset : Best Results

Task A	Micro-recall	Micro-Precision	Micro-F	Macro-Recall	Macro-Precision	Macro-F
	0.1493	0.2655	0.1911	0.1648	0.2664	0.2036

Task B	Bcubed-recall	Bcubed-Precision	Bcubed-F
	0.4823	0.2861	0.3591

Task C	Bcubed-recall	Bcubed-Precision	Bcubed-F
	0.5464	0.2822	0.3722

Testing Dataset : All Tasks Results

Results of Task A

Table 2: Task A Results

Sentences	Micro-recall	micro-precision	micro-f	macro-recall	macro-precision	macro-f
2	0.1338	0.2006	0.1605	0.1216	0.2006	0.1514
3	0.1045	0.1828	0.1330	0.1109	0.1823	0.1379
4	0.1291	0.2492	0.1701	0.1178	0.2492	0.1600
5	0.1392	0.2599	0.1813	0.1337	0.2599	0.1766
6	0.1461	0.2572	0.1864	0.1421	0.2572	0.1830
7	0.1493	0.2655	0.1911	0.1648	0.2664	0.2036
8	0.1130	0.1998	0.1444	0.1129	0.1995	0.1442
9	0.1280	0.2323	0.1651	0.1304	0.2322	0.1670
10	0.1045	0.1828	0.1330	0.1109	0.1823	0.1379
11	0.1379	0.2547	0.1875	0.1481	0.2523	0.1866
12	0.1165	0.2315	0.1550	0.1103	0.2307	0.1492
15	0.1301	0.2573	0.1728	0.1242	0.2565	0.1674

Results of Task B

Table 3: Task B Results

Sentence Length	bcubed-recall	bcubed-precision	bcubed-f
5	0.4823	0.2861	0.3591
10	0.5951	0.1315	0.2154
12	0.6143	0.1080	0.1838
13	0.6260	0.1051	0.1800
14	0.6376	0.0887	0.1558

Results of Task C

Table 4: Task C Results

Sentence Length	bcubed-recall	bcubed-precision	bcubed-f
5	0.5464	0.2822	0.3722
10	0.6253	0.1339	0.2206
12	0.6386	0.1076	0.1842

Conclusion

- Different size of fragments to get better results.
- Calculation of 17 lexical features.
- Calculation of ClustDist.

Future Work

- content based, topic based and stylistic features in combination with the ClustDist method will be explored.