Intrinsic Plagiarism Detection using N-gram Classes

Imene Bensalem[1], Paolo Rosso[2], Salim Chikhi[1]

[1] MISC Lab., Constantine 2 University , Algeria
[2] NLE Lab.,PRHLT Research Center, Universitat Politècnica de València, Spain

What is intrinsic plagiarism detection ?

It is to identify, in the given suspicious document, the fragments that are not consistent with the rest of the text in terms of writing style.

Why using character n-grams ?

Character n-grams allowed for characterizing the writing style for authorship attribution and plagiarism direction.

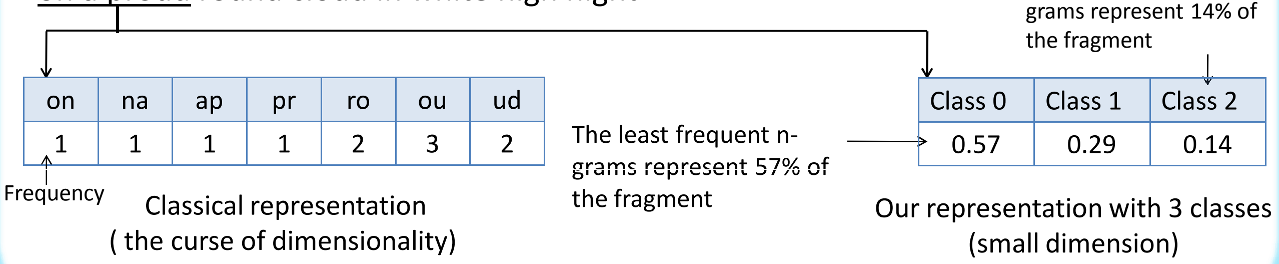
Character n-grams have been used in one of the best intrinsic plagiarism detection methods [1].

What is different in our method ?

We introduced a new text representation using character n-grams.

Example

"on a proud round cloud in white high night"



on	na	ap	pr	ro	ou	ud
1	1	1	1	2	3	2

Frequency

Classical representation
(the curse of dimensionality)

The least frequent n-grams represent 57% of the fragment

The most frequent n-grams represent 14% of the fragment

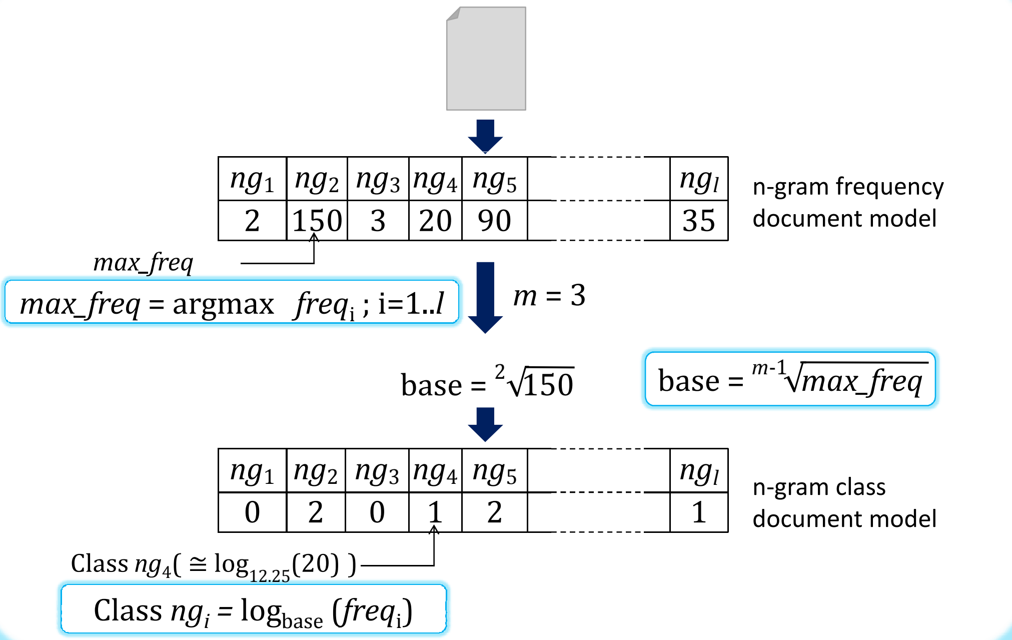
Class 0	Class 1	Class 2
0.57	0.29	0.14

Our representation with 3 classes
(small dimension)

What is a n-gram class ?

A n-gram class is a number from 0 to *m*−1 *such that the class labeled* 0 involves the *least frequent n-grams and the* class labeled *m−1 contains the most frequent n-grams* in a document.

How are the n-gram classes computed ?



ng_1	ng_2	ng_3	ng_4	ng_5		ng_l
2	150	3	20	90		35

n-gram frequency document model

max_freq

$max\_freq = \text{argmax } freq_i$; i=1..l

$m = 3$

base = $\sqrt[2]{150}$

base = $\sqrt[m-1]{max\_freq}$

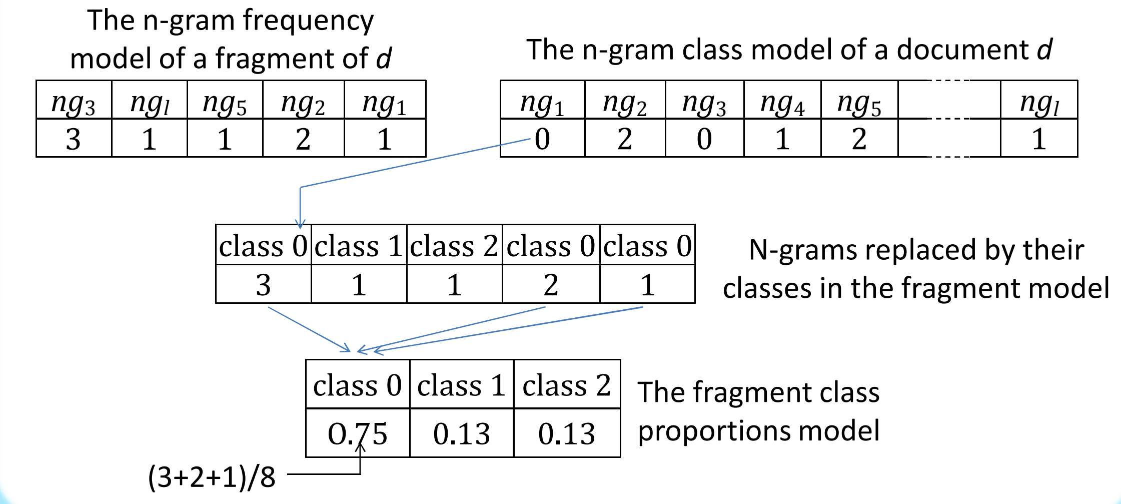
ng_1	ng_2	ng_3	ng_4	ng_5		ng_l
0	2	0	1	2		1

n-gram class document model

Class ng_4($\cong \log_{12.25}(20)$)

Class $ng_i = \log_{base} (freq_i)$

How do we use n-gram classes to represent a text ?



The n-gram frequency model of a fragment of *d*

ng_3	ng_l	ng_5	ng_2	ng_1
3	1	1	2	1

The n-gram class model of a document *d*

ng_1	ng_2	ng_3	ng_4	ng_5		ng_l
0	2	0	1	2		1

class 0	class 1	class 2	class 0	class 1
3	1	1	2	1

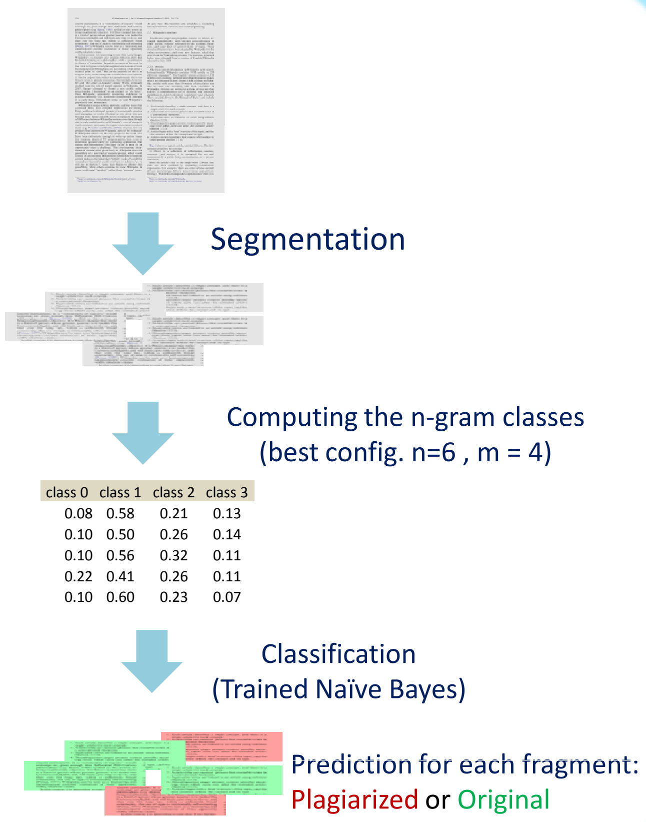
N-grams replaced by their classes in the fragment model

class 0	class 1	class 2
0.75	0.13	0.13

The fragment class proportions model

(3+2+1)/8

Our method



Segmentation

Computing the n-gram classes
(best config. n=6 , m = 4)

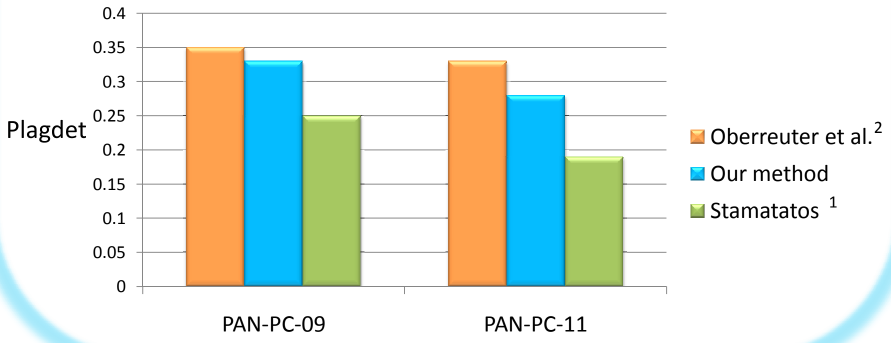
class 0	class 1	class 2	class 3
0.08	0.58	0.21	0.13
0.10	0.50	0.26	0.14
0.10	0.56	0.32	0.11
0.22	0.41	0.26	0.11
0.10	0.60	0.23	0.07

Classification
(Trained Naïve Bayes)

Prediction for each fragment:
Plagiarized or Original

Results

		Our method	Oberreuter et al. [2]
PAN-PC-09	Precision	0.31	**0.39**
	Recall	**0.49**	0.31
	F-measure	**0.38**	0.35
	Granularity	1.21	**1.00**
PAN-PC-11	Precision	0.22	**0.34**
	Recall	**0.50**	0.31
	F-measure	0.30	**0.33**
	Granularity	1.13	**1.00**
InAra [3]	Precision	0.24	**0.29**
	Recall	**0.69**	0.25
	F-measure	**0.35**	0.27
	Granularity	**1.27**	1.44



Conclusion & future work

•Representing the fragments of a given suspicious document by the proportion of character n-gram classes is a promising way for detecting plagiarism intrinsically.

Future work:

•Parameter tuning to improve the results.

•Combine n-gram classes with other stylistic features.

References

[1] E. Stamatatos. 2009. Intrinsic Plagiarism Detection Using Character n-gram Profiles. PAN at SEPLN 2009, CEUR-WS.org, vol. 502, pages 38–46.

[2] G. Oberreuter, G. L'Huillier, SA. Ríos, and JD. Velásquez. 2011. Approaches for Intrinsic and External Plagiarism Detection. PAN at CLEF 2011, pages 1–10.

[3] I. Bensalem, P. Rosso, and S. Chikhi. 2013. A New Corpus for the Evaluation of Arabic Intrinsic Plagiarism Detection. CLEF 2013, LNCS, vol. 8138, pages 53–58. Springer.