Meta Text Aligner:
Text Alignment Based on Predicted Plagiarism Relation

Samira Abnar[1], Mostafa Dehghani[2], Azadeh Shakery[1]
University of Tehran[1], University of Amsterdam[2]

Summary Obfuscation

Text alignment is one of the main steps of plagiarism detection in textual environments. Considering the pattern in distribution of the common semantic elements of the two given documents, different strategies may be suitable for this task. In this paper we assume that the obfuscation level, i.e the plagiarism type, is a function of the distribution of the common elements in the two documents. Based on this assumption we propose the META TEXT ALIGNER that uses the predicted plagiarism type to select the best method or tune the parameters of a particular method for each document pair. Furthermore exploiting the predictions of the classifier for choosing the proper method or the optimal configuration for each type we have been able to improve the Plagdet score of the existing methods.

No Obfuscation

The performance of text alignment can be improved if we know the type of plagiarism to choose the best algorithm for that type. Our suggested mechanism for text alignment begins with predicting the type of plagiarism between a given document pair. The type of plagiarism is a function of the distribution of common elements. Depending on that different methods may be suitable for text alignment. We propose META TEXT ALIGNER which predicts plagiarism relation of two given documents and employs the prediction results to select the best text alignment strategy. Thus, it will potentially perform better than the existing methods which use a same strategy for all cases. As indicated by the experiments, we have been able to classify document pairs based on

In this work We have two concrete research questions:

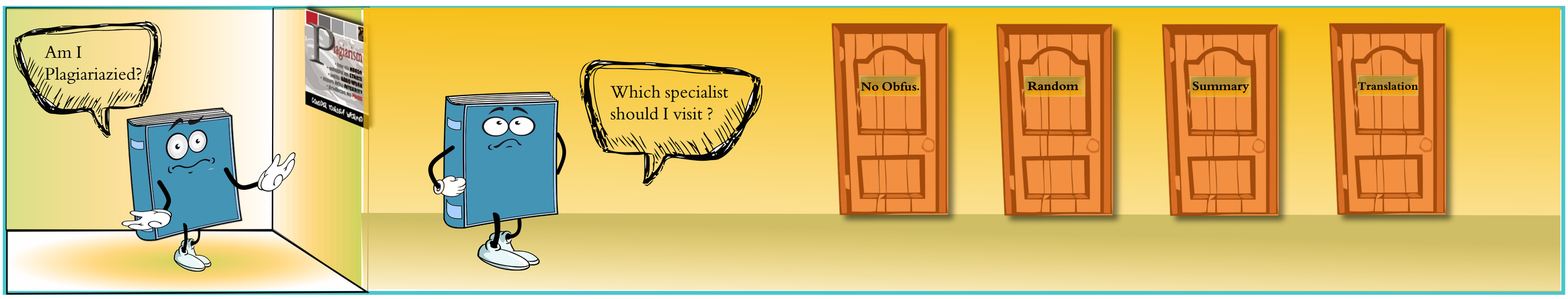
RQ1 How can we determine the type of plagiarism relation between two documents before aligning their texts?

RQ2 How can we improve text alignment performance knowing the type of plagiarism?

Regarding the first research question, the main difficulty in determining the plagiarism type is that we do not know which parts are related to the plagiarism cases. The second research question is inspired from the fact that there is no one for all optimized method for all types of plagiarism.

Our first research question is: "How can we determine the type of plagiarism relation between two documents before aligning their texts?" We solve this problem with a supervised method. Thus the problem is mapped to a classification task.

Our second research question is: "How can we improve text alignment performance knowing the type of plagiarism? ". According to the fact that there is no one for all optimized method of text alignment, an idea to improve the accuracy of text alignment is to choose the best aligning strategy based on plagiarism type.



In the first place you should visit a general practitionair!

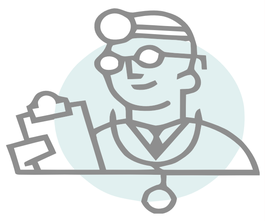
Symptoms

Given the source and the suspicious document, consider:

- Statistics of the similarity scores of common elements
- Statistics of the distances between common elements in both documents
- Frequency of the common elements in both documents
- Difference between distributions of different types of common elements

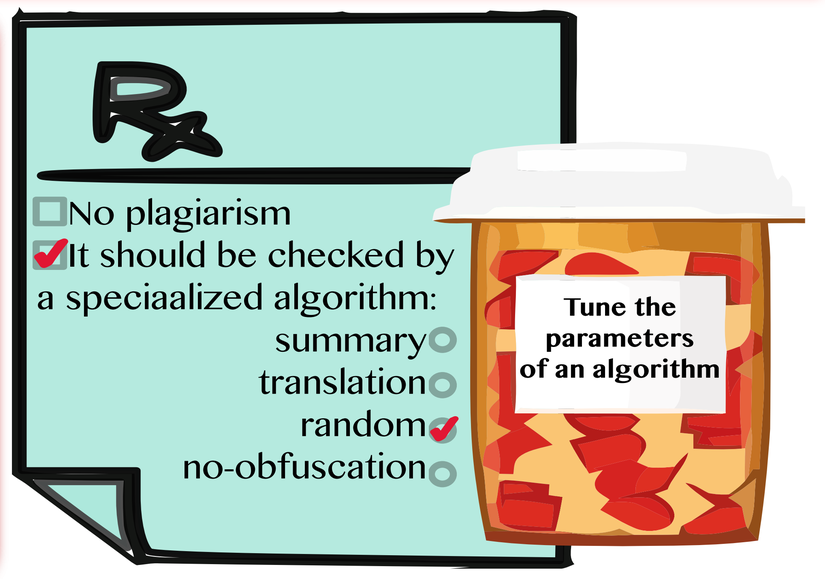
The Meta Text Aligner!

- **It is a rule-based classifier.**
- **Detects the type of plagiarism relation between a document pair.**



Feature Type	Precision	Recall	F1
distribution of frequency and intensity	0.866	0.865	0.865
distribution of frequency, intensity and positions	0.897	0.894	0.894

Performance of the classifier



Using the classifier as the core of the META TEXT ALIGNER improves the overall performance of the existing text alignment methods.

Team	Year	No-obfus.	Random-obfus.	Circular-trans.	Summary	Total
Sanchez-Perez	2014	0.9003	0.8842	0.8866	0.5607	0.8782
Glinos	2014	0.9624	0.8062	0.8472	0.6236	0.8593
META TEXT ALIGNER		0.9577	0.8698	0.8820	0.6310	0.8900

Plagdet score of the methods that have the best performance at least for one plagiarism type in PAN 2014

META TEXT ALIGNER improves the performance of a particular method by customizing its parameters per plagiarism type.

Method	No-obfus.	Random-obfus.	Circular-trans.	Summary	Total
GEN	0.8512	0.4906	0.6737	0.1715	0.6722
SEN	0.8917	0.6802	0.7008	0.5074	0.7521

Plagdet score of the general expanded n-grams based text aligner (GEN) vs the specialized expanded n-gram based text aligner (SEN)

@samiraabnar
s.abnar@ut.ac.ir