Using a Variety of n-Grams for the Detection of Different Kinds of Plagiarism

THE UNIVERSITY OF ALABAMA AT BIRMINGHAM
Knowledge that will change your world

Prasha Shrestha and Thamar Solorio
Department of Computer and Information Sciences
University of Alabama at Birmingham

CORAL
Computational Representation and Analysis of Language

Motivation

Text Alignment Task

<u>House appraisal reports</u> *for* FHA loans *are not* considered standard *and* tend *to have* higher <u>fees</u> due *to the* <u>extra</u> amount *of* work *and* time.

<u>Internal evaluations</u> *for* FHA loans *are not* considered standard *and* tend *to have* higher <u>costs</u> due *to the* <u>extraordinary</u> amount *of* work *and* time.

Years later, a sword called "Souunga" reappears and recalls Takemaru from the grave, and Takemaru decides he wants to play dogcatcher

Years, sword called "Souunga" reappears and grave, Takemaru decides he wants to play dogcatcher

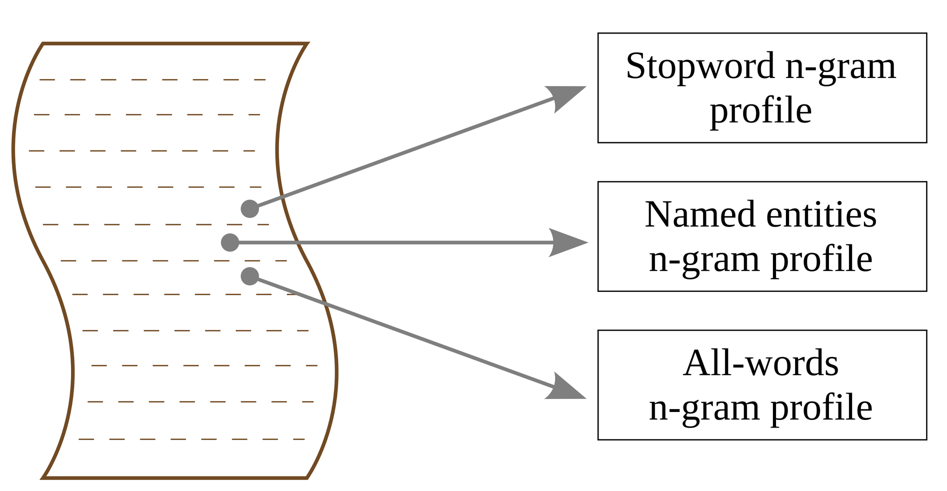
Discovery Channel Stores worked closely together to identify the scope of the project, conduct a needs assessment, and select vendors. Marketmax will work with Deloitte Consulting on software implementation.

Discovery Channel Stores worked together closely to the scope across the cinch, conduct a assessment, and select vendors. Marketmax will work with Deloitte of software.

Certain Methods are Better Suited to Detect Certain Kinds of Obfuscations.

Methodology

Seeding



Matching

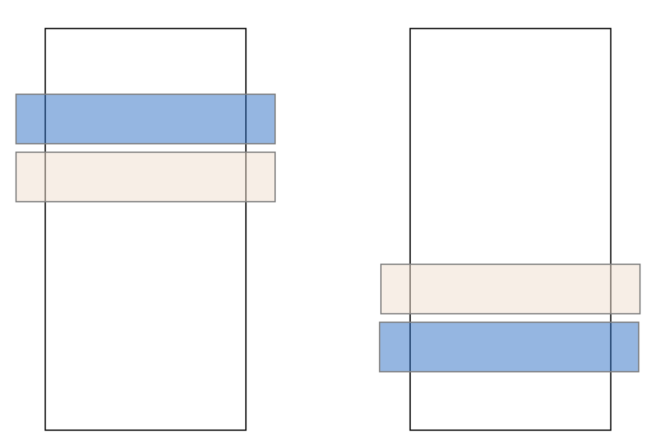
We had a very dry summer
We had a very dry season

Exact Matches
(stopword n-grams)

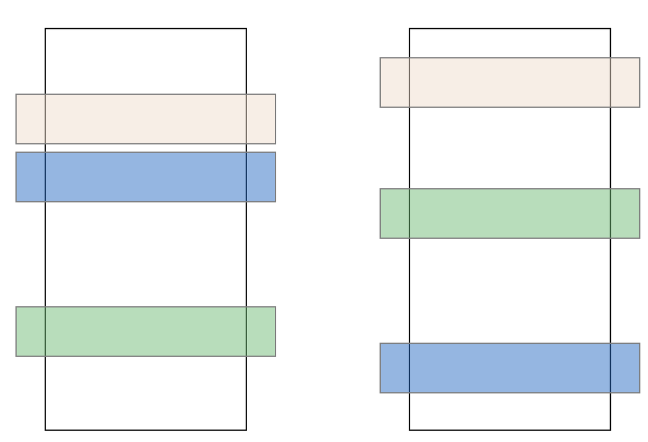
We had a very dry summer
The summer was very dry

Unordered Matches
(named entity and all-words n-grams)

Merging



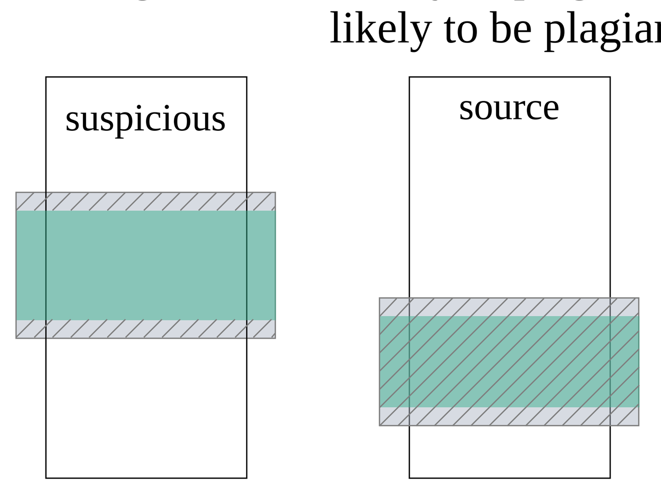
Merged



Not Merged

Passages in the vicinity of plagiarised passages are more likely to be plagiarised.

Postprocessing



Unordered matching of shaded portions of text using all word n-grams with more relaxed parameters

Remove passages that are too short.

Results

Table 1. Evaluation Results for the Training Corpus

Plagiarism Type	Precision	Recall	Granularity	Pladget
No Obfuscation	0.99723	0.80425	1.00000	0.89040
Random Obfuscation	0.90482	0.71842	1.27195	0.67649
Translation Obfuscation	0.87069	0.61710	1.23666	0.62194
Summary Obfuscation	0.91405	0.10747	1.98930	0.12174

Table 2. Evaluation Results for the Test Corpus

Plagiarism Type	Precision	Recall	Granularity	Pladget
No Obfuscation	0.99902	0.80933	1.00083	0.89369
Random Obfuscation	0.92335	0.71461	1.30962	0.66714
Translation Obfuscation	0.88008	0.63618	1.26184	0.62719
Summary Obfuscation	0.90455	0.09897	1.83696	0.11860
Overall	0.87461	0.73814	1.22084	0.69551
Best System	0.89484	0.76190	1.00141	0.82220
Baseline	0.92939	0.34223	1.27473	0.42191

Conclusion

- Three different types of n-grams, each with a different characteristic, collectively can catch passages obfuscated differently. These methods can be combined in such a way that they do not hurt the overall quality of detection of the system.
- Main area that needs improvement is granularity. Named entity n-gram matching inherently produces sparse matches. Although we removed too short passages, removing any more would cost us precision and recall.
- Our postprocessing approach helps to increase detection without compromising the precision. Making our postprocessing approach lenient will help us reduce granularity but will decrease the precision.
- Our approach produces comparatively balanced results across different forms of obfuscations.

Acknowledgement

This research is partially funded by
The Office of Naval Research under grant N00014-12-1-0217.