External Plagiarism Detection using Information Retrieval and Sequence Alignment

Rao Muhammad Adeel Nawab, Mark Stevenson and Paul Clough

Natural Language Processing Group
Department of Computer Science
University of Sheffield, UK

22 September 2011

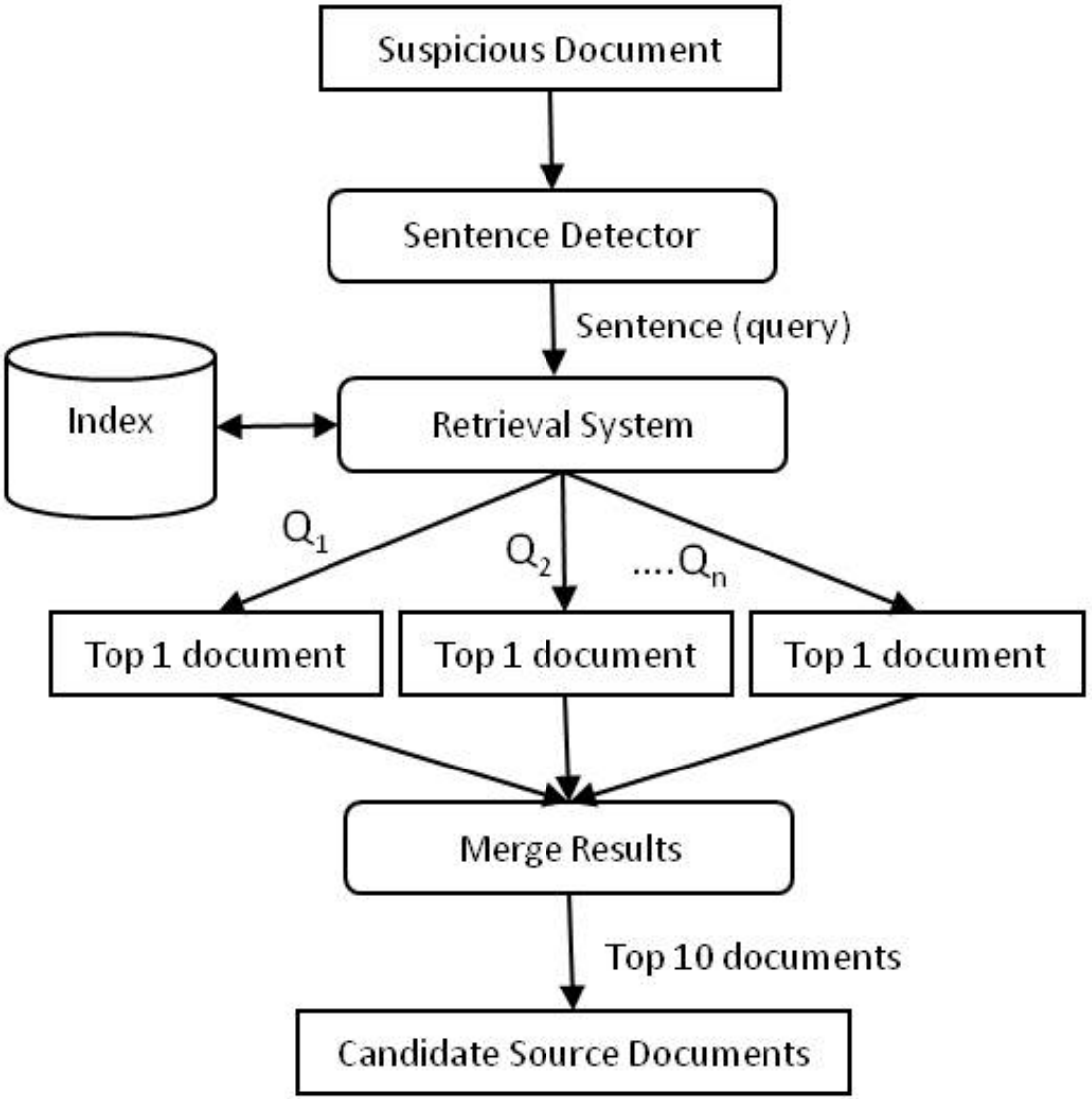
Outline

Three Stage Framework for Monolingual External Plagiarism Detection

1. Preprocessing and Indexing

- Each document split into sentences
- Lower cased and non-alphanumeric characters removed
- Source collection indexed using Terrier IR system

2. Candidate Document Selection



2. Candidate Document Selection cont...

Retrieval

- TF.IDF

Result Merging using CombSUM method

- $S_{finalscore}$ is obtained by adding the scores obtained against each query q:

$$S_{finalscore} = \sum_{q=1}^{N_q} S_q(d) \tag{1}$$

where N_q is the total number of queries to be combined and $S_q(d)$ is the similarity score of a document d for a query q.

3. Detailed Analysis Stage

Greedy String Tiling (GST)

- A string matching algorithm called Running Karp-Rabin Greedy String Tiling (RKR-GST) was used in combination with heuristics to identify suspicious-source section pairs.

An Example of GST

Source a dog$_{[1]}$ bit the postman$_{[2]}$.
Rewrite the postman$_{[2]}$ was bitten by a dog$_{[1]}$.

- [1] and [2] indicate aligned matches between the two texts.

3. Detailed Analysis Stage

Parameters

1. *length of longest match* (α_{length})
   - filters candidate documents for further analysis
   - Best value: $\alpha_{length} > 5$
2. *minimum match length* (*mml*)
   - minimum length of a match in aligning two sequences of tokens
   - Best value: $mml = 3$
3. *length of gap* (α_{merge})
   - distance between pairs of aligned passages which are merged into a single passage
   - Best value: $\alpha_{merge} \leq 35$ characters
4. *discard length* ($\alpha_{discard}$)
   - minimum length for a merged section, any shorter than this are discarded
   - Best value: $\alpha_{discard} \leq 230$ characters

Evaluation

System Performance

Overall System Performance

Precision	Recall	Granularity	PlagDect
0.28	0.09	2.18	0.08

System Performance

Candidate Document Selection Stage

Obfuscation	Precision	Recall	F1
Entire corpus	0.1313	**0.5596**	0.1950
None	0.1807	0.7280	0.2895
Low	0.1642	0.6890	0.2652
High	0.1091	**0.5223**	0.1805
Simulated	0.2648	**0.1675**	0.2052

Detailed Analysis Stage

Obfuscation	Precision	Recall	F1
Entire corpus	0.3316	**0.2827**	0.3052
None	0.6808	0.7280	0.7036
Low	0.6547	0.5803	0.6153
High	0.0643	**0.0422**	0.0510
Simulated	0.5361	**0.0859**	0.1481

Sources of Error

Candidate Document Selection Stage

- Only 10 candidate documents
- Computationally expensive to process more than 10 documents

Detailed Analysis Stage

- GST parameter setting using small dataset due to computational reasons.
- GST can only detect exact copy and fails to detect rewritten text.

Future Work

Future Work

Future Work

- Adapt GST to identify correspondences between paraphrased texts, for example, synonym replacement, morphological changes etc.
- Use automatic machine learning approach for parameter setting.
- For my PhD, incorporate NLP techniques into candidate retrieval framework to identify highly obfuscated text.

Thank you

Questions?

References I