

External and Intrinsic Plagiarism Detection using a Cross-Lingual Retrieval and Segmentation System

Markus Muhr, Roman Kern, Mario Zechner, Michael Granitzer
{mmuhr, rkern, mzechner, mgrani}@know-center.at

CLEF 2010 / PAN / 2010-09-22

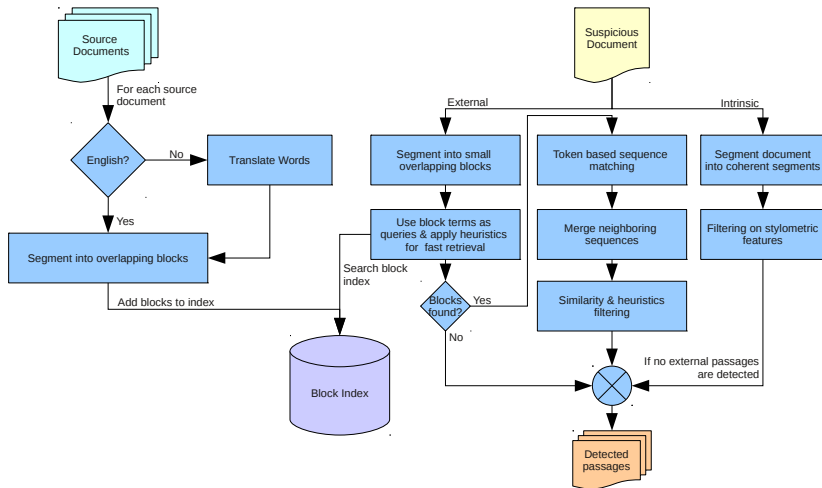
Hybrid System

- ▶ External
 - ▶ Based on information retrieval techniques
 - ▶ Post-processing based on sequence analysis
- ▶ Intrinsic
 - ▶ Detect style change
- ▶ Cross-lingual plagiarism detection
- ▶ No heuristics for high obfuscation
 - ▶ No word reordering
 - ▶ No synonym resolution

Focus

- ▶ Simulate a production system
- ▶ Scalable architecture

Flowchart



Overview

- ▶ Two step approach
 - ▶ Search for potentially matching suspicious document blocks
 - ▶ Apply heuristic post-processing on the potential matches

Work-Flow

- ▶ Build index out of source documents
 - ▶ Build overlapping blocks (40 terms)
- ▶ Split suspicious documents into blocks (16 terms)
 - ▶ Transform blocks into queries
 - ▶ Search source index for matching source blocks

Query Construction

- ▶ For each block in the suspicious document build a query
- ▶ Sort query terms by document frequency
- ▶ Join the low frequent terms by AND
- ▶ Join the remaining terms by OR
- ▶ Additional heuristics to keep number of queries low

Post-Processing

- ▶ Starting with query-block pairs
 - ▶ Expand the text around the query and the block
 - ▶ Build token by token matrix
 - ▶ Match for 3 consecutive tokens (and at least 10 characters) - other thresholds for translated documents
- ▶ Process the sequences
 - ▶ Merged by a neighborhood criterion
 - ▶ Finally a similarity between merged sequences is calculated

Overview

- ▶ Approach: Normalize all documents to English
- ▶ Multiple alternative translations
 - ▶ Not the single-best translation, but multiple candidates
- ▶ Word translations
 - ▶ First step of a complete statistical machine translation system

Word translations

- ▶ Sentence aligned multi-lingual corpus
 - ▶ Europarl v5 Koehn [2005]
- ▶ Apply word alignment algorithm
 - ▶ BerkeleyAligner Liang et al. [2006]
- ▶ Number of translation candidates sorted by probability
- ▶ Replace each non-English word by up to 5 translation candidates

| task | <i>time</i> |
|----------------|-------------|
| no translation | 7 ms |
| translation | 9.38 ms |

Overview

- ▶ Style change detection
- ▶ Focus on features without semantics

Work-Flow

- ▶ Identify regions within a document
- ▶ Build feature centroid vector
- ▶ Compare regions with centroid

Region Detection

- ▶ First idea: Split document in blocks of equal size
- ▶ Approach: Linear text-segmentation algorithm
 - ▶ Build blocks of coherent topics
 - ▶ Stop-word filtered stems as features
- ▶ TextSegFault Kern and Granitzer [2009]
 - ▶ Efficient $O(n)$
 - ▶ Open-source

Candidate Retrieval Step

- ▶ How many false positives are retrieved by the block candidate selection?
- ▶ Left: Based on 500 suspicious document in the development corpus
- ▶ Right: Based on the evaluation corpus

| task | <i>hit</i> | <i>all</i> | <i>ratio</i> |
|------------|------------|------------|---------------|
| high | 2543 | 3676 | 0.6918 |
| low | 6614 | 6988 | 0.9465 |
| none | 9381 | 9592 | 0.9780 |
| translated | 2349 | 2543 | 0.9237 |

| task | <i>hit</i> | <i>all</i> | <i>ratio</i> |
|------------|------------|------------|---------------|
| high | 13348 | 14756 | 0.9046 |
| low | 14832 | 14883 | 0.9966 |
| none | 16784 | 16784 | 1.0 |
| translated | 5462 | 6314 | 0.8651 |

Overall System Performance

- Performance results of detected plagiarism separated by different sub-tasks for the hybrid evaluation corpus

| task | <i>Precision</i> | <i>Recall</i> | <i>Granularity</i> | <i>Score</i> |
|---------------------|------------------|---------------|--------------------|--------------|
| non-translated all | 0.9299 | 0.8967 | 1.0553 | 0.8785 |
| non-translated none | - | 0.9497 | 1.0025 | - |
| non-translated low | - | 0.9207 | 1.0968 | - |
| non-translated high | - | 0.8122 | 1.0771 | - |
| translated | 0.8036 | 0.61616 | 2.1655 | 0.4195 |
| external | 0.9053 | 0.8631 | 1.1611 | 0.7949 |
| intrinsic | 0.212 | 0.1566 | 1.0 | 0.1802 |
| Overall | 0.8417 | 0.7057 | 1.1508 | 0.6948 |

- ▶ Hybrid system
 - ▶ External plagiarism detection
 - ▶ Support for cross-lingual plagiarism detection
 - ▶ Intrinsic (style-based) plagiarism detection
- ▶ Issues
 - ▶ Scalable (but slow implementation)
- ▶ Outlook
 - ▶ We plan to build a web service initialized with the Wikipedia as source

The End

Thank you!

References

- R. Kern and M. Granitzer. Efficient linear text segmentation based on information retrieval techniques. In *MEDES '09*, pages 167–171. ACM, 2009. ISBN 978-1-60558-829-2.
- P. Koehn. Europarl: A parallel corpus for statistical machine translation. *MT summit*, 5:12–16, 2005.
- P. Liang, B. Taskar, and D. Klein. Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 104–111, June 2006.