A Hybrid Architecture for Plagiarism Detection
Demetrios Glinos
University of Central Florida, Orlando, Florida USA
glinos@eecs.ucf.edu

Hybrid Design
- Combine different techniques for different types of text plagiarism
- Use text alignment method (extended Smith-Waterman dynamic programming algorithm) for order-based plagiarism
- Use concept clustering method (several variations) for non-order based plagiarism

Processing Flow

Order-Based Plagiarism
- Key feature:
  Concepts in both documents appear in substantially the same order, possibly with some additions, deletions and differences.

Source sentence:
This essay discusses Hamlet's famous soliloquy in relation to the major themes of the play.

Suspect sentence:
This article discusses the famous Hamlet monologue of the main themes of the game.

This essay discusses Hamlet's famous soliloquy in relation to the major themes of the play.

Text Alignment

Recursive Descent

Matrix Splicing

Non-order Based Plagiarism

Passage Detection

Summary Detection

Clustering Algorithms

Basic Clustering:
- Form trigrams centered on 30 most frequent tokens at least 5 characters long that start with letter in the source document.
- Find occurrences of bigrams in suspect document and merge into clusters if within 15 tokens.
- Keep clusters at least 40 tokens long that contain at least 8 source terms and choose maximal cluster that has Jaccard coefficient at least 0.65 computed on source concept words and content words in the suspect cluster, excluding stop words.
- Attempt to find a passage in the source document with Jaccard coefficient at least 0.50 for concept words in maximal suspect cluster.

Recursive Descent:
- Search for occurrences in suspect document.

Test Data

Test Results