Plagiarism Detection: Techniques, Tools and Policies

Tomáš Foltýnek
Mendel University in Brno, CZ :: University of Konstanz, DE
tomas.foltyn@academicintegrity.eu
Introduction

• Literature review on Plagiarism detection
  • Overview of methods
  • Strengths, weaknesses
• Do PDS vendors learn from scientists?
• Plagiarism policies
Methodology

• Keyword-based automated search
• Google Scholar
  • restricted time 2013 – 2018
  • manual review of titles
  • manual review of abstracts
  • exclusion of papers in journals on the Beall’s list
• Topically related conferences
  • Plagiarism across Europe and Beyond 2013, 2015, 2017
  • PAN competitions
• Sort to categories, identify research gaps
Typology of plagiarism

• Copy & Paste
  • Sources possibly mentioned (pawn sacrifice, cut & slide)

• Weak obfuscation
  • Technical disguise
  • Synonym substitution

• Translation

• Paraphrase
  • Mosaic, clause quilts

• Structural plagiarism
  • Boilerplate, template
  • Idea plagiarism

• Shake & Paste
  • Patch-writing, compilation, remix, mosaic, mash-up
Number of papers (Google Scholar)
Three Layers of Plagiarism Detection

Policies

Tools

Techniques
Plagiarism Detection Tasks

• What “Plagiarism detection” means?

• Traditional task
  • Document Level Detection
    • Given suspicious document D and set of source documents
    • Select source document(s) similar to D

• Up-to-date tasks
  • Candidate retrieval
  • Text alignment
  • Paraphrase identification
  • Intrinsic PD
Extrinsic Plagiarism Detection

• Formerly Document Level Detection

• Candidate Retrieval
  • Given a suspicious document D and a search engine / database
  • Retrieve all documents from which the text was reused in D
    • with minimum costs

• Text alignment
  • Given a suspicious document D and a set of candidate documents S
  • Identify passages
    • from documents in S
    • which were re-used in D

• Paraphrase Identification
  • Given two passages, decide whether they have same meaning

Source: Potthast et al. (2014)
Intrinsic Plagiarism Detection

• Given suspicious document
• Identify borders between parts with different style
• Group these parts according to authorship

• Stylometry
• Machine learning

• Related tasks
  • Author identification
  • Author profiling

• Applications: Marketing, law enforcement
Howe to Solve these Tasks?

• Typology of Plagiarism Detection Techniques
  • Lexical-based
  • Syntax-based
  • Semantics-based

• Applications
  • Recommender systems
  • Detection of similar documents
    • related news
    • duplicated work
Lexical-based Methods

• N-Grams
  • N consecutive characters/words
  • Former approach: Fingerprinting

• Vector Space Models
  • Bag-of-words approach
  • Each word is a dimension in a vector space
  • Values weighted (tf-idf)
  • Documents/passages treated as vectors
    • cosine similarity measure
Syntax-based Methods

• Part-of-speech tagging
  • marks each word with its class (noun, verb, ...)
  • pre-processing for semantic analysis
  • comparison within classes

• Syntactic graphs
  • sentence represented as a graph
  • syntactic relations marked
  • allows structural plagiarism detection
Semantics-based Methods

• Latent Semantic Analysis
  • Singular Value Decomposition
  • Grouping documents by meaning

• Explicit Semantic Analysis
  • Vector of concepts

• Knowledge Graph Analysis
  • Passage represented as a graph
  • Graph similarity metrics

• Bag of words – meaning problem
  • A loves B \approx B loves A

Knowledge graph built from the sentence “I opened a new bank account” (Franco-Salvador, 2016)
Tools for Semantic Analysis

- WordNet (Princeton)
  - English dictionary + thesaurus
- EuroVoc (EU)
  - Multilingual thesaurus
- Wikipedia
- Wikidata
- BabelNet
Evaluation of Techniques

- PAN corpuses, Microsoft Research Paraphrase Corpus, German Paraphrase Corpus, P4P...

- **Recall**
  - What portion of plagiarism was revealed?
  - \( R = \frac{TP}{TP+FN} \)

- **Precision**
  - What portion of revealed cases are plagiarism?
  - \( P = \frac{TP}{TP+FP} \)

- **F-measure**
  - Harmonic mean of Recall and Precision
  - \( F = \frac{2*P*R}{P+R} \)
Accuracy of Extrinsic Plagiarism Detection

• Copy-paste detection ≈ 100 %
• Synonym replacement detection ≈ 90 %
• Paraphrase identification ≈ 80 %
• Summary identification ≈ 75%
• Cross-language plagiarism detection ≈ 70 %
• Structure, idea,... ???
Author Identification & Profiling

- Up-to-date accuracy
  - Style change identification ≈ 60 %
  - Grouping by authorship ≈ 60 %
  - Native language prediction ≈ 65 – 85 %
  - Gender identification ≈ 80 %
  - Age-group identification ≈ 50 – 55 %

- Real author: J.K. Rowling

- Related task: Author obfuscation
Three Layers of Plagiarism Detection

Policies

Tools

Techniques
Layer 2: Plagiarism Detection Tools

• Which one is the best?
  • For given language
  • For given type of plagiarism
  • Considering other criteria

• Prof. Debora Weber-Wulff Testing

• Less sound tests
  • Vani&Gupta, 2016; Chowdhury&Bhattacharyya, 2016
# Technical Disguise Test

<table>
<thead>
<tr>
<th>System</th>
<th>Indexes IJEI</th>
<th>Homoglyphs</th>
<th>White characters</th>
<th>Synonyms</th>
<th>Text as image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turnitin</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Urkund</td>
<td>NO</td>
<td>YES</td>
<td>NO (PLAINTEXT)</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Theses.cz</td>
<td>NO</td>
<td>NO (PLAINTEXT)</td>
<td>NO (PLAINTEXT)</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Slovenia</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Slovakia</td>
<td>NO</td>
<td>NO (PLAINTEXT)</td>
<td>NO (PLAINTEXT)</td>
<td>NO</td>
<td>NO</td>
</tr>
</tbody>
</table>
ENAI Standardized Testing

• Research gap: No standardized testing in place

• Develop sound methodology

• Prepare good testing set
  • Various European languages
    • cross-language plagiarism
  • Various plagiarism types
  • Various types of sources
    • Wikipedia, Internet, OA papers, paid papers

• More evaluation criteria
  • Results, speed, user friendliness,...
Three Layers of Plagiarism Detection

Policies

Tools

Techniques
Layer 3: Policies for Plagiarism

• Projects exploring policies
  • IPPHEAE project (2010-2013)
  • SEEPPAI project (2016-2017)
  • Follow-up (CoE) in 2018-2019

• Great books
  • Debora Weber-Wulff (2014): False Feathers
  • Tracey Bretag (2016): Handbook for Academic Integrity
Systemic Effort Works

Curtis & Vardanega, 2016

- 10-years study (2004-2014)
  - Introduced course on academic writing
  - Started using Turnitin
  - Introduced criterion and standards-based assessment
  - Implemented educational changes
- All forms of plagiarism DECREASED
  - except recycling and ghostwriting

Owens & White, 2013

- 5-years study (2007-2011)
  - PD SW as both deterrent and formative tool
  - Educational about academic writing and authorship
- Self-reported plagiarism DROPPED significantly
- Nature of assessment has a central role
Academic Integrity Maturity Model

- **Transparency** in academic integrity and quality assurance
- **Policies**: fair, effective and consistent
- **Sanctions**: standard range
- **Software tools**
- **Prevention** strategies and measures
- **Communication** about policies and procedures
- **Knowledge** and understanding about academic integrity
- **Training** for students and teachers
- **Research** and innovation in academic integrity

(IPPHEAE, 2013)
Institutional Policy

• Problem identification
  • Students? Teachers? Scientists? Management?
  • Plagiarism? Exam cheating? Contract cheating?
  • Inconsistent approach of teachers?
  • Risk of a scandal?

• Propose set of measures
  • Inspiration from others

• Consider differences
  • Culture: focus on motivation or penalties?
  • Society: positive/negative examples
  • Field of study: type of assignments, type of assessment
  • Institution...
General Findings

• One-fits-all strategy ineffective
• Multilayered, evidence-based, longitudinal strategy

• Central role of assessment
• Address both educational approaches
  • for “good” students
• and deterrence strategies
  • for “bad” students
  • students who cheat tend to cheat repeatedly
• Research for gathering evidence and impact evaluation
ENAI Tools Coming Soon

• Glossary of terms
  • already published
• Guidelines for academic integrity
  • almost ready
• Self-evaluation tools
  • this autumn
• Manual for cross-sector cooperation
  • this summer
• Collection of educational materials
  • this summer
• Announcements in ENAI Newsletter
Sources

- https://en.wikipedia.org/wiki/The_Cuckoo%27s_Callin
- https://cdn-images-1.medium.com/max/1600/1*AV35no2cjqhMr19_14c9A.png
- https://c1.staticflickr.com/4/3738/12221192503_c3eff8db68_b.jpg
- https://upload.wikimedia.org/wikipedia/commons/2/26/Precisionrecall.svg
- http://www.alglib.net/matrixops/general/lsvd1.gif
Thank you for your attention

Tomáš Foltýnek

Mendel University in Brno, Czechia
University of Konstanz, Germany

tomas.foltynek@academicintegrity.eu