Intrinsic Plagiarism Detection

Register & Genre

Oren Halvani
Outline

• Motivation

• External / Internal Plagiarism Detection

• Text layers

• Features / Feature Extraction

• Similarity Functions

• Putting It All Together

• Take-Home-Messages / References
Motivation

• Thanks to the internet the amount of information nowadays has grown drastically during the last two decades


• A huge part of these information is represented through text, e.g. e-Mails, websites, office documents, e-Books, etc.
Motivation

• Due to [8] current world population estimate: 6,852,472,823

• From this it follows: many different authors have been involved in the generation of these textual information 😊

• Sometimes an author A “borrows” some information (text) from an author B

• Furthermore A may not mention this action...

• This act refers to the term of *plagiarism*
Motivation

• What is plagiarism exactly?

"Definition: Plagiarism is the act, intentional or otherwise, of copying or borrowing words or ideas without properly acknowledging the original source."

[3]

• Since manual detection is very time-consuming, research focuses on automatic plagiarism detection

• Majority of existing approaches refer to External Plagiarism Detection (EPD)

• Rough example...
External Plagiarism Detection

Reference corpora (sample documents)

External Plagiarism Detector

X is / isn’t plagiarized
External Plagiarism Detection

• Limitation: in Real-World-scenarios reference material isn’t always available

• In such cases we wish at least to figure out if a document has been written by only ONE author

• How can we make it possible, if a given document is simultaneously

  → the reference corpus itself?

• The discipline which occupied with this non-trivial challenge is the so-called:

  **Intrinsic Plagiarism Detection (IPD)**
Intrinsic Plagiarism Detection

- IPD is a relatively young discipline, in comparison to EPD

- Massively influenced by the following pioneer in this field

- Prof. Dr. Benno Stein,
  Chair of the Web-Technology & Information Systems Group
  → Bauhaus-Universität Weimar

- Stein and his team published many fundamental methods & theories for IPD

- Furthermore they’ve developed a unique corpus for text plagiarism cases (PAN)
Intrinsic Plagiarism Detection

• What does the term intrinsic stand for? One possible definition:

“Metaphysically, an intrinsic property of an object is a property that the object has by virtue of itself, depending on no other thing...” [9]

• In the context of IPD intrinsic stands for “the analysis from inside”

• Main purpose: assists to judge if a document has been plagiarized

• Core concept: feature extraction → outlier analysis → decision!
Intrinsic Plagiarism Detection

• For which plagiarism delicts can we use IPD?
**Intrinsic Plagiarism Detection**

- IPD can only be achieved if features are suitable enough to detect breach of style within the text

- Key question: which features should be used?

- In order to answer this question we should first understand one important fact

- Text typically consists of several layers...
Text layers (small overview)

Grapheme Layer: This is a text

Symbol Layer: This is a text

Constituents Layer: (This (is (a (text))))

Token Layer: [This] [is] [a] [text]

Phoneme Layer: /ðɪs ɪz ə tɛkst/

Part-Of-Speech Layer: This_DT is_VBZ a_DT text_NN

Semantics Layer:
Features for IPD

- Features should be extracted from any meaningful layer

- Statistically spoken: one/very few layers are not sufficient for the task

- Some possible features:

<table>
<thead>
<tr>
<th>Symbol Layer</th>
<th>Token Layer</th>
<th>POS Layer</th>
<th>POS Phrase Layer</th>
<th>Semantics Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio of vowels per word: (e, a, u, o, i, y)</td>
<td>Ratio of content-words/all words</td>
<td>Ratio of nouns, verbs, adverbs adjectives and all words</td>
<td>Ratio of NP’s, VP’s, PP’s and all sentences</td>
<td>Ratio of synonym usage and the corresponding synset</td>
</tr>
</tbody>
</table>
Features for IPD

- Other (more reliable) features...
Feature Extraction

• Once we know which features are promising for the task, the next step would be to “get” them.

• Solution: split document into non-overlapping regions.

• Apply feature extraction $features(r_i)$ on each region.

• Example...
Feature Extraction

\[ r_1 \rightarrow \text{features}(r_1) = (0.43, 0.68, 0.09, \ldots) \]

- **content words**
- **all words**
- **Noun Phrases**
- **all sentences**
- **long words**
- **all words**
Feature Extraction

- Computing feature vectors is not enough - how should we detect outliers?

- Simple approach: Compute the document mean vector, which is the mean of all the feature vectors.

- Once we have the mean vector, we can start to measure similarities...
Similarity Functions

• Loose definition:
  - Numerical measure of how alike two data objects are
  - Is higher when objects are more alike
  - Often falls in the range \([0 ; 1]\)

• Many similarity functions can be used in IPD
  - Overlap Coefficient
  - Cosine Similarity
  - Bray–Curtis dissimilarity
  - Jaccard Index
  - Minkowski distance
Similarity Functions

- Example:

- Overlap Coefficient, defined by:
  \[
  \text{over}(X, Y) = \frac{\sum_{i=0}^{n} \min(x_i, y_i)}{\min\left(\sum_{i=0}^{n} x_i, \sum_{i=0}^{n} y_i\right)}
  \]

- Assume the following vectors are given:

  - \( X = (2, 2, 2) \)
  - \( Y = (7, 1, 0) \)

\[
\text{over}(X, Y) = \frac{\min(2,7) + \min(2,1) + \min(2,0)}{\min(2+2+2, 7+1+0)}
\]

\[
\text{over}(X, Y) = \frac{2 + 1 + 0}{\min(6, 8)} = \frac{3}{6} = \frac{1}{2}
\]
Putting It All Together...

\[ F_1 = \text{features}( r_1 ) \]
\[ F_2 = \text{features}( r_2 ) \]
\[ F_3 = \text{features}( r_3 ) \]
\[ F_4 = \text{features}( r_4 ) \]
\[ F_5 = \text{features}( r_5 ) \]

similarity( \( F_i \), mean )

Threshold

Outlier(s)

Note: Outlier(s) = plagiarized region(s)!
Intrinsic Plagiarism Detection:

...is an intensively discussed research discipline (especially since the last decade), but in comparison to EPD, comparatively a poorly investigated area.

...involves a lot of complex techniques (in this talk: focus on the “basic idea”)

...requires a lot of creativity (and patience) 😊
Thanks for your attention!
Questions...?
References


Sven Meyer zu Eissen, Benno Stein, and Marion Kulig,
Decker and Lenz (Eds.): Advances in Data Analysis
Selected Papers from the 30th Annual Conference of the German Classification Society (GfKl)
http://www.springerlink.com/content/w87475j2m5413220/

GENET 418/518 - Human genetics, Winter 2010 - Dr. Heather McDermid
www.biology.ualberta.ca/people/heather_mcdermid/genet418/UnderstandingPlagiarism.pdf
References

[4] “Prof. Dr. Benno Stein”,

Picture taken from the Web Technology & Information Systems (Uni Weimar) website:

www.uni-weimar.de/cms/medien/webis/people/benno-stein.html


A bronze sculpture made by Auguste Rodin, held in the Musée Rodin in Paris, France.


BRYAN DALTON / MISTAKE THE BEAUTIFUL. PRODUCES FREELANCE PHOTO-ILLUSTRATION, DESIGN AND ANIMATION IN PORTLAND, OR.

http://www.mistakethebeautiful.com/thankyou.html
References

[7] “Introduction to Data Mining”,
Lecture Notes, by: Tan, Steinbach, Kumar
www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap2_data.pdf

[8] “Current World Population and World Population Growth Since the Year One”,
www.geography.about.com/od/obtainpopulationdata/a/worldpopulation.htm

Philosophical Studies 83 (1996): 1–27
www.tedsider.org/papers/intrinsic_properties.pdf

All external links have been accessed on: 30.10.2011 (some links updated...)