Developing a corpus of plagiarized short answers [Clough and Stevenson, 2011]

Björn Rudzewitz
University of Tübingen

Hauptseminar Language Variation and Stylometrics
WS 15/16

December 16, 2015

1brzdwtz@sfs.uni-tuebingen.de
Developing a corpus of plagiarized short answers [Clough and Stevenson, 2011]

Björn Rudzewitz
University of Tübingen

Introduction
Plagiarism Typology
Corpus Creation
Data Analysis
   Individual Differences
   Data Observations
Automatic Plagiarism Detection
   N-Gram Overlap
   LCS
   Baselines
   L1 vs L2
   Classification
Conclusion
Discussion

To avoid the objection of plagiarism:

ideas and examples in this presentation are taken from Clough and Stevenson [2011]
Motivation

- correlation between availability of electronic resources and plagiarism
- plagiarism detection as a field suffering from lack of standardized evaluation resources
- previous corpus creation efforts suboptimal:
  - lack of data ('deception', how to find plagiarized text)
  - lack of gold labels (authors deny judgments)
  - lack of legal and ethical basis for data publication
  - lack of transparency in data preparation
    (→ Leech’s maximes for corpus creation)
Impact and application

Desired effects of the corpus:

- new resource for comparative evaluation and pedagogical methods
- enable new work on plagiarism detection and task strategies
Related work

- Microsoft Research Paraphrase Corpus [Dolan et al., 2004]
- Multiple-Translation Chinese Corpus [Pang et al., 2003]
- METER corpus [Gaizauskas et al., 2001]
- Corpus for plagiarism detection [Zu Eissen et al., 2007]
- PAN Plagiarism detection corpus [Eiselt and Rosso, 2009]

More related resources in Machine Translation evaluation and Short Answer Assessment.
Developing a corpus of plagiarized short answers [Clough and Stevenson, 2011]

Björn Rudzewitz
University of Tübingen

Introduction

Plagiarism Typology

Corpus Creation

Data Analysis
Individual Differences
Data Observations

Automatic Plagiarism Detection

N-Gram Overlap
LCS
Baselines
L1 vs L2 Classification

Conclusion
Discussion
References

High-level perspective on approaches

- extrinsic
  - comparison of source and (potentially) plagiarized text
  - authorship attribution approaches

- intrinsic
  - comparison of text passages in one document with each other
  - stylometric approaches

Problem: documents can plagiarize $n \in \mathbb{N}_0$ other documents in different ways

→ interaction between extrinsic and intrinsic analysis desirable
Plagiarism Techniques: How to plagiarize

Goal: produce an answer of 200-300 words to a question

- **Near copy**
  - copy-and-paste (parts of) Wikipedia article

Björn Rudzewitz
University of Tübingen

Introduction

Plagiarism Typology

Corpus Creation

Data Analysis

Individual Differences

Data Observations

Automatic Plagiarism Detection

N-Gram Overlap

LCS

Baselines

L1 vs L2 Classification

Conclusion

Discussion

References
Plagiarism Techniques: How to plagiarize

Goal: produce an answer of 200-300 words to a question

- Near copy
  - copy-and-paste (parts of) Wikipedia article

- Light revision
  - like light revision, but with possibility to replace words with synonyms, (lexical/morphosyntactic) paraphrasing
  - information structure preserved
Plagiarism Techniques: How to plagiarize

Goal: produce an answer of 200-300 words to a question

- **Near copy**
  - copy-and-paste (parts of) Wikipedia article

- **Light revision**
  - like light revision, but with possibility to replace words with synonyms, (lexical/morphosyntactic) paraphrasing
  - information structure preserved

- **Heavy revision**
  - rephrasing/paraphrasing of Wikipedia article, n-to-m sentence alignment
Plagiarism Techniques: How to plagiarize

Goal: produce an answer of 200-300 words to a question

- **Near copy**
  - copy-and-paste (parts of) Wikipedia article

- **Light revision**
  - like light revision, but with possibility to replace words with synonyms, (lexical/morphosyntactic) paraphrasing
  - information structure preserved

- **Heavy revision**
  - rephrasing/paraphrasing of Wikipedia article, n-to-m sentence alignment

- **Non-plagiarism**
  - no access to Wikipedia
  - participants read material, then answer question with their (partly freshly) acquired knowledge
Corpus Creation

- 19 participants, CS students
- each participant writing answer for each task (2 times non-plagiarism)
  → 95 answers + 5 articles = 100 documents (19,995 tokens)
- Graeco-Latin Square Design for systematic randomization and rotation of revision types per participant and question
- participant meta data: native language, familiarity with answer, perceived difficulty of task

\[ \mu_{\text{tok/aw}} = 208 \quad \sigma_{\text{tok/aw}} = 64.91 \]
\[ \mu_{\text{types/aw}} = 113 \quad \sigma_{\text{types/aw}} = 30.11 \]
Data Analysis: Individual Differences

- statistically significant difference ($p < 0.01$) between native and non-native speakers wrt. mean knowledge and perceived difficulty (two-sample t-test)
  → difference in population means of two independent samples
- Positive Pearson’s correlation of $r = 0.344$ between knowledge and perceived difficulty
Data Analysis: Observations

Question:
A. What is inheritance in object oriented programming?

Example 1:
Inheritance allows classes to be categorized, similar to the way humans categorize. It also provides a way to generalize due to the “is a” relationship between classes.

Example 2:
Generalisation also some time known as inheritance. The main reason behind this is a hierarchy structure of objects and classes. We can understand this mechanism by some examples: like fruit is a main class and mangoes apple, orange is child classes of the main class. So obviously inherit all the properties of fruit class.
Data Analysis: Observations

Question:
A. What is inheritance in object oriented programming?

Example 1:
Inheritance allows classes to be categorized, similar to the way humans categorize. It also provides a way to generalize du to the “is a” relationship between classes.

Example 2:
Generisation also some time known as inheritance. The main reason behind this is a hierarchy structure of objects and classes. We can understand this mechanism by some examples: like fruits as main class and mangoes apple orange is child classs of the main class. So obviously inherit all the properties of fruit class.
Data Analysis: Observations

Question:
A. What is inheritance in object oriented programming?

Example 1:
Inheritance allows classes to be categorized, similar to the way humans categorize. It also provides a way to generalize du to the “is a” relationship between classes.

Example 2:
Generalisation also some time known as inheritance. The main reason behind this is a hierarchy of objects and classes. We can understand this mechanism by some examples: like fruit is a main class and mangoes apple orange is child classes of the main class. So obviously, inherit all the properties of fruit class.

spelling mistake
missing predicate
missing subject
segmentation mistake
Data Analysis: Observations

Question:
A. What is inheritance in object oriented programming?

Example 1:
Inheritance allows classes to be categorized, similar to the way humans categorize. It also provides a way to generalize du to the “is a” relationship between classes.

Example 2:
Generalisation also some time known as inheritance. The main reason behind this is a hierarchy structure of objects and classes. We can understand this mechanism by some examples: like fruit is a main class and mangoes apple orange is child class of the main class. So obviously inherit all the properties of fruit class.

spelling mistake
missing predicate
missing subject
segmentation mistake

need for robust processing resources
Experimental Automatic Plagiarism Detection

2 classification tasks:

1. Prediction of plagiarism and plagiarism type: Predict a class $c$ with

   $c \in \{"near copy", "light revision", "heavy revision", "non-plagiarism"\}$

2. Binary classification of plagiarism: Predict a class $c$ with

   $c \in \{"plagiarism", "non-plagiarism"\}$

2 feature types: n-gram overlap, LCS
N-Gram Overlap

n-gram containment on document level

\[ c_n(A, B) = \frac{|S(A, n) \cap S(B, n)|}{|S(A, n)|} \]

\( n \in \mathbb{N}, 0 < n < 6 \) (window size)

A, B documents
Developing a corpus of plagiarized short answers [Clough and Stevenson, 2011]

Björn Rudzewitz
University of Tübingen

Introduction
Plagiarism Typology
Corpus Creation
Data Analysis
Individual Differences
Data Observations
Automatic Plagiarism Detection
N-Gram Overlap
LCS
Baselines
L1 vs L2 Classification
Conclusion
Discussion
References

Longest Common Subsequence (LCS)

- longest shared (possibly) non-continuous sequence
- compute minimum number of edit operations for transforming text A into B
- normalized lcs: normalize by length of answer text
Developing a corpus of plagiarized short answers [Clough and Stevenson, 2011]

Björn Rudzewitz
University of Tübingen

Introduction
Plagiarism Typology
Corpus Creation
Data Analysis
Individual Differences
Data Observations
Automatic Plagiarism Detection
N-Gram Overlap
LCS
Baselines
L1 vs L2 Classification
Conclusion
Discussion
References

**Figure:** Relation between Longest Common Subsequence and Edit Operations (from [Myers, 1986, page 253])
Figure: Relation between Longest Common Subsequence and Edit Operations (from [Myers, 1986, page 253])
## Baselines

Comparison of answers with unrelated articles

<table>
<thead>
<tr>
<th>Task</th>
<th>$c_w(A, B)$ for $w$-gram</th>
<th>$\text{lcs}_{\text{norm}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>0.48</td>
<td>0.15</td>
</tr>
<tr>
<td>B</td>
<td>0.65</td>
<td>0.23</td>
</tr>
<tr>
<td>C</td>
<td>0.49</td>
<td>0.20</td>
</tr>
<tr>
<td>D</td>
<td>0.60</td>
<td>0.29</td>
</tr>
<tr>
<td>E</td>
<td>0.61</td>
<td>0.23</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.57</td>
<td>0.22</td>
</tr>
</tbody>
</table>

- high unigram overlap between topic-unrelated answers and Wikipedia articles
Baselines

Comparison of answers with related articles

<table>
<thead>
<tr>
<th>Category</th>
<th>$c_n(A, B)$ for n-gram</th>
<th>$lcs_{norm}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Near copy</td>
<td>0.95</td>
<td>0.89</td>
</tr>
<tr>
<td>Light revision</td>
<td>0.87</td>
<td>0.70</td>
</tr>
<tr>
<td>Heavy revision</td>
<td>0.81</td>
<td>0.52</td>
</tr>
<tr>
<td>Non-plagiarised</td>
<td>0.63</td>
<td>0.23</td>
</tr>
</tbody>
</table>

- high n-gram overlap between topic-related answers and Wikipedia articles
- less strong drop for higher n
- statistically significant differences between similarity of rewrite levels with articles (ANOVA with Bonferroni pos-hoc test)
Comparison of answers by question (‘task’)

- averaging over all (non)plagiarism types
- ‘most’ differences not significant

<table>
<thead>
<tr>
<th>Task</th>
<th>$c_n(A, B)$ for $n$-gram</th>
<th>$lcs_{norm}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>0.77</td>
<td>0.45</td>
</tr>
<tr>
<td>B</td>
<td>0.81</td>
<td>0.53</td>
</tr>
<tr>
<td>C</td>
<td>0.71</td>
<td>0.44</td>
</tr>
<tr>
<td>D</td>
<td>0.82</td>
<td>0.58</td>
</tr>
<tr>
<td>E</td>
<td>0.81</td>
<td>0.56</td>
</tr>
<tr>
<td>Avg</td>
<td>0.79</td>
<td>0.51</td>
</tr>
</tbody>
</table>
L1 vs L2

- higher n-gram containment scores for non-natives for heavier revision
- insignificant, though noticeably higher amount of lifting of material for participants writing in L2
Classification

- Naive Bayes Classifier from WEKA
- best result for binary classification: 94.3% accuracy
- best result for classification of 4 classes: 80.0% accuracy

\[ F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \]
Conclusion

- publicly available corpus of manually created plagiarized text: *Wikipedia Reuse Corpus*
- different types of plagiarism represented, authentic language
- simple features allowed plagiarism classification with 95% accuracy
Discussion

▶ Text of 200-300 words usually not considered a short answer in SAA (e.g. Burrows et al. [2015], Ziai et al. [2012])
▶ Probability of academics copying verbatim from Wikipedia?
▶ Are students working on a plagiarism project representative of the population of participants?
▶ Are measures of central tendency for very heterogeneous data justified?
▶ ”questionable gold standard annotation” [Zesch and Gurevych, 2012, page 174]?
Developing a corpus of plagiarized short answers [Clough and Stevenson, 2011]

Björn Rudzewitz
University of Tübingen

Introduction

Plagiarism Typology
Corpus Creation
Data Analysis
Individual Differences
Data Observations
Automatic Plagiarism Detection
N-Gram Overlap
LCS Baselines
L1 vs L2 Classification

Conclusion

Discussion

References


Robert Gaizauskas, Jonathan Foster, Yorick Wilks, John Arundel, Paul Clough, and Scott Piao. The meter corpus:
Developing a corpus of plagiarized short answers [Clough and Stevenson, 2011]

Björn Rudzewitz
University of Tübingen

Introduction

Plagiarism Typology

Corpus Creation

Data Analysis

Individual Differences

Data Observations

Automatic Plagiarism Detection

N-Gram Overlap

LCS

Baselines

L1 vs L2 Classification

Conclusion

Discussion

References


Developing a corpus of plagiarized short answers [Clough and Stevenson, 2011]

Björn Rudzewitz
University of Tübingen

Introduction

Plagiarism Typology

Corpus Creation

Data Analysis

Individual Differences

Data Observations

Automatic Plagiarism Detection

N-Gram Overlap

LCS

Baselines

L1 vs L2 Classification

Conclusion

Discussion

References
