Recent Developments in Processing Source Code and Natural Language

Dr. Steven Burrows

Web Technology and Information Systems Group
Bauhaus-Universität Weimar

21 February 2012
Seminar Overview

1. Personal introduction.

2. Topic 1: Authorship attribution (source code).

Personal Introduction

- Undergrad, Honours, PhD, teaching, RMIT University, 2000–2010.
- PostDoc, Bauhaus-Universität Weimar, 2011 to current.
Authorship Attribution

Disputed Document

Sandra’s Documents

Andrew’s Documents

?
Measuring Style

Natural language:
- Function words ("the", "and", "to", "of", "a", etc.).
- Parts of speech (nouns, verbs, conjunctions, etc).
- Metrics (average sentence length, etc).
- N-grams.

Source code:
- Programming language tokens (operators, keywords, etc).
- Software metrics (average function length, etc).
- N-grams.
Applying Information Retrieval

Seven steps:

2. Deidentification: Comply with ethics requirements.
3. Tokenisation: Extract meaningful features.
5. Indexing: Efficient authorship matching.
6. Querying: Identify likely authors.
7. Evaluation: How often is the correct author chosen?
int main(void) {
    IntList il;
    int anInt;
    ListMake(&il);
    while (scanf("", &anInt) == 1) {
        if (!ListInsert(&il, anInt)) {
            fprintf(stderr, "");
            break;
        }
    }
    ListDisplay(&il);
    ListFree(&il);
    exit(EXIT_SUCCESS);
}
N-Grams

```c
int ( void int ( & while ( ( , & ==
  if ( ! ( & , ) , break ( & ) & )
  ↓

in pa vo in pa am wh pa pa co am eq if
pa no pa am co pa co br pa am pa am pa
  ↓

1 in pa vo in pa am
2 pa vo in pa am wh
3 vo in pa am wh pa
4 in pa am wh pa pa
    ..... 
21 br pa am pa am pa
```
Evaluation

Evaluation measures:
- Chose P@1 over reciprocal rank and average precision.

Classification examples with P@1:
- Query ‘A3’ (fail):
  
<table>
<thead>
<tr>
<th>Rank</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A3</td>
</tr>
<tr>
<td>2</td>
<td>B1</td>
</tr>
<tr>
<td>3</td>
<td>J2</td>
</tr>
<tr>
<td>4</td>
<td>A4</td>
</tr>
<tr>
<td>5</td>
<td>B5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>60</td>
<td>C4</td>
</tr>
</tbody>
</table>

- Query ‘A4’ (success):
  
<table>
<thead>
<tr>
<th>Rank</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A4</td>
</tr>
<tr>
<td>2</td>
<td>A6</td>
</tr>
<tr>
<td>3</td>
<td>E1</td>
</tr>
<tr>
<td>4</td>
<td>A2</td>
</tr>
<tr>
<td>5</td>
<td>D4</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>60</td>
<td>J3</td>
</tr>
</tbody>
</table>
## Corpora

<table>
<thead>
<tr>
<th>Author Type</th>
<th>Language</th>
<th>ST-100</th>
<th>TIME</th>
<th>PLAN-C</th>
<th>PLAN-J</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>C</td>
<td>100</td>
<td>272</td>
<td>100</td>
<td>76</td>
</tr>
<tr>
<td>Student</td>
<td>C</td>
<td>1,597</td>
<td>1,632</td>
<td>1,095</td>
<td>453</td>
</tr>
<tr>
<td>Freelance</td>
<td>C/C++</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Freelance</td>
<td>Java</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Min Samples</th>
<th>Mean Samples</th>
<th>Max Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Authors</td>
<td>14</td>
<td>16</td>
<td>26</td>
</tr>
<tr>
<td>Total Samples</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>
Relative Timestamps in the Time Collection

1 2 3 4 5 6
Time
Task 1 Task 2 Task 1 Task 2 Task 1 Task 2

Semester 1
Programming Techniques
Task 1 Task 2

Semester 2
Algorithms and Analysis
Task 1 Task 2

Semester 3
Database Systems
Task 1 Task 2
Temporal Issue: Writing Style Changes Over Time

Dr. Steven Burrows (WebIS @ BUW)  Processing Code and Natural Language  21 February 2012 12 / 27
Other Semester Classification (TIME collection)
Results for Other Semester Matches (TIME collection)

Accuracy (%)

- All Matches (78.52%)
- Current Semester Matches (79.59%)
- Other Semester Matches (72.07%)

Assignment Task Number

Dr. Steven Burrows (WebIS @ BUW)
Definition of a paraphrase:

- Given two texts, $t_1$ and $t_2$, then $t_1$ is a paraphrase of $t_2$ if a fact $\phi$ that can be inferred from $t_1$ can also be inferred from $t_2$, and vice versa:

$$ (t_1 \land \alpha \models \phi) \iff (t_2 \land \alpha \models \phi) $$

The symbol $\alpha$ refers to a world (a domain theory or background knowledge) in the form of a set of relations that readers of $t_1$ and $t_2$ agree upon.
PAN Series

- Competition tasks in extrinsic plagiarism detection, intrinsic plagiarism detection, authorship identification, wikipedia vandalism detection, and wikipedia quality flaw prediction.
- Refer to pan.webis.de for history and current events.
- Of the paraphrases in the plagiarism corpora:
  - 2009: Computer-generated approaches only.
  - 2010: 3671 acquired paraphrases from Amazon’s Mechanical Turk.
  - 2011: 938 new acquired paraphrases from Amazon’s Mechanical Turk.
## Strategies for Machine-Generated Paraphrases

### Example paraphrases

<table>
<thead>
<tr>
<th>Original sentence</th>
<th>The quick brown fox jumps over the lazy dog.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random text operations</td>
<td>over The. the quick lazy dog (&lt;\text{context word})&gt; jumps brown fox over jumps quick brown fox The lazy. the brown jumps the. quick dog The lazy fox over</td>
</tr>
<tr>
<td>Semantic word variation</td>
<td>The quick brown dodger leaps over the lazy canine. The quick brown canine jumps over the lazy canine. The quick brown vixen leaps over the lazy puppy.</td>
</tr>
<tr>
<td>POS-preserving word shuffling</td>
<td>The brown lazy fox jumps over the quick dog. The lazy quick dog jumps over the brown fox. The brown lazy dog jumps over the quick fox.</td>
</tr>
<tr>
<td>Original</td>
<td>Paraphrase</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>The emigrants who sailed with Gilbert were better fitted for a crusade</td>
<td>The people who left their countries and sailed with Gilbert were more</td>
</tr>
<tr>
<td>than a colony, and, disappointed at not at once finding mines of</td>
<td>suited for fighting the crusades than for leading a settled life in the</td>
</tr>
<tr>
<td>gold and silver, many deserted; and soon there were not enough sailors</td>
<td>colonies. They were bitterly disappointed as it was not the America that</td>
</tr>
<tr>
<td>to man all the four ships. Accordingly, the Swallow was sent back to</td>
<td>they had expected. Since they did not immediately find gold and silver</td>
</tr>
<tr>
<td>England with the sick; and with the remainder of the fleet, well</td>
<td>mines, many deserted. At one stage, there were not even enough men to</td>
</tr>
<tr>
<td>supplied at St. John’s with fish and other necessaries, Gilbert (August</td>
<td>help sail the four ships. So the Swallow was sent back to England carrying</td>
</tr>
<tr>
<td>20) sailed south as far as forty-four degrees north latitude. Off Sable</td>
<td>the sick. The other fleet was supplied with fish and the other necessities</td>
</tr>
<tr>
<td>Island a storm assailed them, and the largest of the vessels, called</td>
<td>from St. John. On August 20, Gilbert had sailed as far as forty-four</td>
</tr>
<tr>
<td>the Delight, carrying most of the provisions, was driven on a rock and</td>
<td>degrees to the north latitude. His ship known as the Delight, which bore</td>
</tr>
<tr>
<td>went to pieces.</td>
<td>all the required supplies, was attacked by a violent storm near Sable</td>
</tr>
<tr>
<td></td>
<td>Island. The storm had driven it into a rock shattering it into pieces.</td>
</tr>
</tbody>
</table>

[Excerpt from “Abraham Lincoln: A History” by John Nicolay and John Hay.]
Paraphrase Corpora

Existing:
- MSRPC corpus: 5801 sentences.
- KMC corpus: 1067 sentences.
- Cohn et al. corpus: 900 sentences.
- METER corpus: 1717 journalism articles.

Our work:
- Webis-CPC-11 corpus: 7859 passages (4067 positive and 3792 negative).
- See the www.webis.de/research/corpora/webis-cpc-11 corpus page.
Cost-Effective Paraphrase Acquisition

Observation:
- Manually checking worker submissions is expensive.

Question:
- Can machine learning effectively automate the submission filtering?
Paraphrase Similarity and Classification

Word-level metrics:
- Semantic similarity between individual words.
- WordNet::Similarity software [Pedersen et al., 2004].

Information retrieval metrics:
- Manhattan distance.
- Euclidean distance.
- Cosine similarity.
- Probabilistic models.

Paraphrase-specific metrics:
- N-grams.
Measuring Accepted and Rejected Paraphrases

(a) Accepted paraphrases

(b) Rejected paraphrases

- See Cordeiro et al. [2007a, 2007b] for descriptions of the metrics.
## Evaluating Five Classifiers

<table>
<thead>
<tr>
<th>Classifier Name</th>
<th>PPV</th>
<th>TPR</th>
<th>FPR</th>
<th>TNR</th>
<th>FNR</th>
<th>ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>c4.5 decision tree</td>
<td>0.80</td>
<td>0.91</td>
<td>0.25</td>
<td>0.75</td>
<td>0.09</td>
<td>0.84</td>
</tr>
<tr>
<td>k-nearest neighbor</td>
<td>0.81</td>
<td>0.90</td>
<td>0.22</td>
<td>0.78</td>
<td>0.10</td>
<td>0.84</td>
</tr>
<tr>
<td>maximum entropy</td>
<td>0.78</td>
<td>0.91</td>
<td>0.28</td>
<td>0.72</td>
<td>0.09</td>
<td>0.82</td>
</tr>
<tr>
<td>naive bayes</td>
<td>0.75</td>
<td>0.90</td>
<td>0.32</td>
<td>0.68</td>
<td>0.10</td>
<td>0.79</td>
</tr>
<tr>
<td>support vector machine</td>
<td>0.78</td>
<td>0.93</td>
<td>0.29</td>
<td>0.71</td>
<td>0.07</td>
<td>0.83</td>
</tr>
</tbody>
</table>
Applying Class Probability Estimates
Adjusting the Training Set Size

Above results for k-nearest neighbor classifier ($k = 50$).
## Cost Analysis

### Corpus Name and Properties

<table>
<thead>
<tr>
<th>Name</th>
<th>Webis-CPC-11</th>
<th>Repeat</th>
<th>Repeat</th>
<th>New</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training %</td>
<td>—</td>
<td>100</td>
<td>100</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>Precision</td>
<td>—</td>
<td>0.950</td>
<td>0.980</td>
<td>0.950</td>
<td>0.980</td>
</tr>
<tr>
<td>Recall</td>
<td>—</td>
<td>0.579</td>
<td>0.523</td>
<td>0.509</td>
<td>0.316</td>
</tr>
</tbody>
</table>

### Corpus Costs

| Task development | $577.20 | $577.20 | $577.20 | $577.20 | $577.20 |
| Accepted HITs    | $886.70 | $886.70 | $886.70 | $886.70 | $886.70 |
| Rejected HITs    | —       | $1,193.06 | $1,193.06 | $1,193.06 | $1,193.06 |
| Discarded HITs   | —       | $875.58   | $992.05   | $1,121.16 | $1,422.56 |
| Manual filtering | $2,782.92 | —       | —       | $695.73 | $695.73 |
| Machine learning | $409.44  | $136.48  | $136.48  | $136.48 | $136.48 |
| **TOTAL COST**   | **$4,636.26** | **$3,669.02** | **$3,785.49** | **$4,610.33** | **$4,911.73** |

### Savings from Webis-CPC-11

| Money Saved | — | $967.24 | $850.77 | $25.93 | —$275.47 |
| Time Saved  | — | 111h    | 111h    | 70.25h | 70.25h    |
The End

Further information:

- www.webis.de
- www.sites.google.com/site/stevenburrows81

Thankyou!
Questions?

Dr. Steven Burrows
steven.burrows@uni-wemar.de