Automated System for Cognitive Presence Coding

Presentation of the outcomes

Vitomir Kovanović, Srečko Joksimović, and Dragan Gašević
THE UNIVERSITY of EDINBURGH
s.joksimovic@ed.ac.uk
@s_joksimovic

UK Learning Analytics Network
London,
3rd February 2016
The Main Idea

Learning Analytics for Supporting Traditional DE/OL & MOOCs
Overall Goal and Study Approach

Through analytics:
- Drive pedagogical interventions
- Build models/frameworks of MOOC learning
- Empowering instructors and learners

Approach:
- Build Learning Analytics research on existing knowledge and models of Distance/Online education
- Use Learning Analytics to validate and extend existing educational theories
- Automate as much as possible
Community of Inquiry

1. Triggering event:
   Problem identification, sense of puzzlement

2. Exploration:
   Brainstorming, Idea exploration, divergence

1. Integration:
   Synthesis of relevant information

2. Resolution:
   Problem resolution, testing application

Garrison, Anderson, and Archer (1999)
Proposal: Automated CoI Content Analysis

Build a system for automated coding of discussion messages for the levels of cognitive presence

Advantages:
Enable for broader adoption of CoI model
Faster and cheaper adoption in research
Provide detailed operationalization of CoI coding scheme
Real-time feedback of learning in discussions
Enable for development of various analytics dashboards

Future work:
Eventually support other presences/models
### Proposed tasks and current implementation

<table>
<thead>
<tr>
<th>Task</th>
<th>Impl.</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implementation of lag features</td>
<td>![Completed]</td>
<td></td>
</tr>
<tr>
<td>Implementation of Coh-Metrix &amp; LIWC features</td>
<td>![Completed]</td>
<td></td>
</tr>
<tr>
<td>Implementation of Conditional Random Field classifier</td>
<td>![Completed]</td>
<td>Random Forest in the final version</td>
</tr>
<tr>
<td>Coding MOOC data</td>
<td>![In progress]</td>
<td>Extended the initial task</td>
</tr>
<tr>
<td>Evaluation on MOOC data</td>
<td>![Completed]</td>
<td>Extended the initial task</td>
</tr>
<tr>
<td>Dissemination of results &amp; reporting</td>
<td>![In progress]</td>
<td><strong>Papers:</strong> 2 published, 1 in preparation</td>
</tr>
</tbody>
</table>

**Completed**

**In progress**
## Data

### Course offerings statistics

<table>
<thead>
<tr>
<th>Course offer</th>
<th># Students</th>
<th># Msg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter 2008</td>
<td>15</td>
<td>212</td>
</tr>
<tr>
<td>Fall 2008</td>
<td>22</td>
<td>633</td>
</tr>
<tr>
<td>Summer 2009</td>
<td>10</td>
<td>243</td>
</tr>
<tr>
<td>Fall 2009</td>
<td>7</td>
<td>63</td>
</tr>
<tr>
<td>Winter 2010</td>
<td>14</td>
<td>359</td>
</tr>
<tr>
<td>Winter 2011</td>
<td>13</td>
<td>237</td>
</tr>
<tr>
<td><strong>Average (SD)</strong></td>
<td><strong>13.5 (5.1)</strong></td>
<td><strong>291.2 (192.4)</strong></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>81</strong></td>
<td><strong>1,747</strong></td>
</tr>
</tbody>
</table>

### Distribution of cognitive presence phases

<table>
<thead>
<tr>
<th>Phase</th>
<th># Msg</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other</td>
<td>140</td>
<td>8.0%</td>
</tr>
<tr>
<td>Triggering</td>
<td>308</td>
<td>17.6%</td>
</tr>
<tr>
<td>Exploration</td>
<td>684</td>
<td>39.2%</td>
</tr>
<tr>
<td>Integration</td>
<td>508</td>
<td>29.1%</td>
</tr>
<tr>
<td>Resolution</td>
<td>107</td>
<td>6.1%</td>
</tr>
<tr>
<td><strong>Average (SD)</strong></td>
<td><strong>349.4 (245.7)</strong></td>
<td><strong>20% (10%)</strong></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1,747</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>
Feature extraction

- **Context features**
  - Discussion context
    - Number of replies
    - Message depth
    - Cosine similarity to previous/next message
    - Start/end indicators

- **Linguistic features**

- **NLP features**

- **LSA**
  - Text similarity

- **Named entities**
  - Text annotation (DBPedia Spotlight)

- **Coh-Metrix**
  - 103 measures of text coherence

- **LIWC**
  - 93 measures representing psychological processes

Feature space
SMOTE - Synthetic Minority Over-sampling Technique

- Other: 140
- Trig.: 308
- Exp.: 684
- Integ.: 508
- Resol.: 107
Model selection and results

<table>
<thead>
<tr>
<th>Study</th>
<th>Classifier</th>
<th>Accuracy</th>
<th>Cohen’s Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kovanović et al. (2014)</td>
<td>Support Vector Machines</td>
<td>53.38%</td>
<td>0.41</td>
</tr>
<tr>
<td>Waters et al. (2015)</td>
<td>Conditional Random Fields</td>
<td>64.20%</td>
<td>0.48</td>
</tr>
<tr>
<td>Kovanović et al. (2016)</td>
<td>Random Forest</td>
<td>70.30%</td>
<td>0.63</td>
</tr>
</tbody>
</table>
### Variable importance

<table>
<thead>
<tr>
<th>Variable</th>
<th>Other</th>
<th>Triggering</th>
<th>Exploration</th>
<th>Integration</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of words</td>
<td>55.41 (61.06)</td>
<td>80.91 (41.56)</td>
<td>117.71 (67.23)</td>
<td>183.30 (102.94)</td>
<td>280.68 (189.62)</td>
</tr>
<tr>
<td>Number of named entities</td>
<td>13.44 (15.36)</td>
<td>21.67 (10.55)</td>
<td>28.84 (16.93)</td>
<td>44.75 (24.85)</td>
<td>64.18 (32.54)</td>
</tr>
<tr>
<td>Lexical diversity, all words</td>
<td>0.85 (0.12)</td>
<td>0.77 (0.09)</td>
<td>0.71 (0.10)</td>
<td>0.65 (0.09)</td>
<td>0.58 (0.09)</td>
</tr>
<tr>
<td>Position within discussion</td>
<td>2.39 (1.13)</td>
<td>1.00 (0.90)</td>
<td>1.84 (0.97)</td>
<td>1.87 (0.94)</td>
<td>2.00 (0.68)</td>
</tr>
<tr>
<td>Lexical diversity, content words</td>
<td>0.95 (0.06)</td>
<td>0.90 (0.06)</td>
<td>0.86 (0.08)</td>
<td>0.82 (0.07)</td>
<td>0.78 (0.07)</td>
</tr>
<tr>
<td>Avg. givenness of each sentence</td>
<td>0.10 (0.07)</td>
<td>0.14 (0.06)</td>
<td>0.18 (0.07)</td>
<td>0.21 (0.06)</td>
<td>0.24 (0.06)</td>
</tr>
<tr>
<td>Number of question marks</td>
<td>0.27 (0.85)</td>
<td>1.84 (1.63)</td>
<td>0.92 (1.26)</td>
<td>0.58 (0.82)</td>
<td>0.38 (0.55)</td>
</tr>
<tr>
<td>Similarity with previous message</td>
<td>0.20 (0.17)</td>
<td>0.06 (0.13)</td>
<td>0.22 (0.21)</td>
<td>0.30 (0.24)</td>
<td>0.39 (0.19)</td>
</tr>
<tr>
<td>Lexical diversity, VOCD</td>
<td>12.92 (33.93)</td>
<td>28.99 (50.61)</td>
<td>53.57 (54.68)</td>
<td>83.47 (43.00)</td>
<td>97.16 (28.95)</td>
</tr>
<tr>
<td>Avg. number of paragraphs sent.</td>
<td>4.26 (2.98)</td>
<td>6.37 (2.76)</td>
<td>7.49 (4.11)</td>
<td>10.17 (5.64)</td>
<td>14.05 (8.88)</td>
</tr>
</tbody>
</table>
Project Insights

• Classification performance
  - Substantial level of agreement

• Feature Modeling
  - surface features (e.g., unigrams, bigrams, POS-bigrams)
  - contextual features
  - domain-independent features

• “Class Balancing”
  - Synthetic Minority Over-sampling Technique
Future Work

• Classification improvement
  - “Quoting” problem
  - Class balancing – different approaches
  - Extending feature set

• MOOC data – work in progress

• Learning Analytics platform
  - Integrate classification module
  - Provide associated probabilities
Team

Vitomir Kovanović
1. Third year PhD student (School of Informatics, UoE)
2. Msc Software Engineering
3. Four years industry experience as software developer
4. Best paper award at LAK 2015

Srećko Joksimović
1. Third year PhD student (School of Education, UoE)
2. Msc Information systems
3. Six years industry experience as software developer

Dragan Gašević (supervisor)
1. Professor at the University of Edinburgh (Informatics & Education)
2. President of SOLAR - Society for Learning Analytics research
3. PhD Information Systems
References


Automated System for Cognitive Presence Coding

Q&A

Vitomir Kovanović, Srećko Joksimović, and Dragan Gašević
THE UNIVERSITY of EDINBURGH
s.joksimovic@ed.ac.uk
@s_joksimovic