Outline

1. FWC corpus
   • A new corpus of student writing

2. Automatic scoring
   • A topic-independent model for this type of writing
   • Present a model that can handle grading differences between teachers

3. Conclusion and future directions
Motivation

1. Provide feedback to help teachers evaluate students
   • Can automatic writing evaluation be used on classroom writing assignments?

2. Provide feedback to help teachers grade better
   • Can we overcome different grading tendencies between teachers?
Freshman Writing Corpus
Freshman Writing Corpus

1. Take-home essays
Freshman Writing Corpus

1. Take-home essays

2. Long(er)-form

<table>
<thead>
<tr>
<th></th>
<th>Kaggle</th>
<th>FWC</th>
</tr>
</thead>
<tbody>
<tr>
<td># essays</td>
<td>22k</td>
<td>25k</td>
</tr>
<tr>
<td>avg. # tokens</td>
<td>250</td>
<td>900</td>
</tr>
<tr>
<td>avg. # grafs</td>
<td>1.5</td>
<td>5.5</td>
</tr>
</tbody>
</table>
Freshman Writing Corpus

1. Take-home essays
2. Long(er)-form
3. Open-ended topic
Freshman Writing Corpus

1. Take-home essays

2. Long(er)-form

3. Open-ended topic

4. Aligned drafts
Freshman Writing Corpus

1. Take-home essays
2. Long(er)-form
3. Open-ended topic
4. Aligned drafts
5. Detailed rubric scores
Freshman Writing Corpus

1. Take-home essays
2. Long(er)-form
3. Open-ended topic
4. Aligned drafts
5. Detailed rubric scores
6. Teacher comments
Freshman Writing Corpus
Freshman Writing Corpus

Freshman Writing Corpus (FWC)

- English Composition I
- 4 writing assignments (“projects”)
- Students submitted Intermediate and Final drafts for each assignment
- Each draft graded
<table>
<thead>
<tr>
<th>Project</th>
<th>Target # words</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>600-770</td>
<td>A personal narrative that describes an experience and uses that experience to tell readers something important about the writer.</td>
</tr>
<tr>
<td>2</td>
<td>600</td>
<td>A bibliographic essay that asks you to understand the conversation surrounding your chosen topic by examining four relevant sources. …</td>
</tr>
<tr>
<td>3</td>
<td>600-800</td>
<td>A reflection that asks you to think carefully about how audience and purpose, as well as medium and genre, affect your choices as composers and reflect carefully on a new dimension of your topic.</td>
</tr>
<tr>
<td>4</td>
<td>1000-1200</td>
<td>A polished essay that asserts an arguable thesis that is supported by research and sound reasoning.</td>
</tr>
<tr>
<td>Category</td>
<td>Weight</td>
<td>Level</td>
</tr>
<tr>
<td>-------------</td>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>Focus</td>
<td>25%</td>
<td>Basics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Critical</td>
</tr>
<tr>
<td></td>
<td></td>
<td>thinking</td>
</tr>
<tr>
<td>Evidence</td>
<td>25%</td>
<td>Basics</td>
</tr>
<tr>
<td>Organization</td>
<td>25%</td>
<td>Basics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Critical</td>
</tr>
<tr>
<td></td>
<td></td>
<td>thinking</td>
</tr>
<tr>
<td>Style</td>
<td>25%</td>
<td>Basics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Critical</td>
</tr>
<tr>
<td>Format</td>
<td>5%</td>
<td>Basics</td>
</tr>
</tbody>
</table>
Weighted average of rubric scores corresponds to letter grade.
Scores

Intermediate drafts (M=2.4, SD=0.9)

[Graph showing density distribution of scores for different projects, with a peak around score 2.4 for intermediate drafts.]
Scores

Final drafts (M=3.0, SD=0.7)
Your introduction offers some general introduction to the topic. You introduce one of the sources completely. Your introduction requires a stronger thesis statement that draws the connections between the two.

Although I appreciate that you changed one of the sources, there still remains not much substance to summarize. Both sources are very brief, and the arguments are not complex.

...
Each start of the new school year, headlines bear the names of a handful of young, seemingly healthy athletes who die suddenly on the basketball court, the football field or the track. Most of the time the reason why, is unknown. This happens so often and yet no one ever sees it coming. Athletes train and work out for years with no problems, until one day they collapse and die. One minute Reggie Garrett was making a touchdown, and the next minute he collapsed and died, according to a NBC news report. About ten to twenty-five sports related sudden cardiac deaths in young athletes occur annually in the United States. Robin J Northcote who wrote the article, Sudden Cardiac Death in Sport, believes that these athletes had to have had a previous medical problem and that exercise along would not cause them to die. Dr. Milton Greenwich wrote an article on young athletes as well and also says that exercise alone would not cause death.
By the numbers

- Full corpus: 2 years of Composition I and II
  - Fall, Spring, Summer
  - > 25k essays
- This study: 1 semester of Composition I
  - 3,362 essays
  - 639 students, 55 sections, 21 instructors
By the numbers

<table>
<thead>
<tr>
<th>Draft</th>
<th>Count</th>
<th>Tokens</th>
<th>Sentences</th>
<th>Paragraphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter.</td>
<td>1200</td>
<td>840.3</td>
<td>35.6</td>
<td>5.2</td>
</tr>
<tr>
<td>Final</td>
<td>1762</td>
<td>938.5</td>
<td>39.6</td>
<td>5.7</td>
</tr>
</tbody>
</table>
Automatic Scoring
Previous work: Test writing

- Short answers or short essays
- In response to a prompt or passage
- Timed
- Limit on outside sources
Introduction

This work: Classroom writing

• Take-home assignments

• Open-ended topics

• Longer

• More polished (?)

• Different scoring criteria (?)
Experimental Setup

- Linear regression to predict score 0-4
  - Round to nearest 0.05
- Different models for Intermediate and Final drafts
- Data
  - Training: 1,200 Intermediate, 1762 Final essays
  - Testing: 100 Intermediate, 100 Final essays
Experimental Setup
Features

• Categories:
  • surface
  • structural
  • lexical
  • syntactic
  • grammatical
• 57 features + n-gram features
Results
Results

Gold standard (human performance)
Results

- Baseline
- LR

Graph showing % agreement for Intermediate Exact, Intermediate Adj., Final Exact, and Final Adj.
Challenges

639 students, 55 sections, 21 instructors
Instructors

- Standardized tests graded by multiple instructors
- Validated scores = trustworthy scores
- FWC scores are NOT validated
- Even when graders are trained and score on the same rubric, they *may* be inconsistent
Instructors

Do teachers grade differently?
Instructors

Intermediate draft scores by teacher

Final draft scores by teacher

Density

N = 1662   Bandwidth = 0.1391

N = 1131   Bandwidth = 0.1811
Instructors

-0.7  -0.6  -0.5  -0.4  -0.3  -0.2  -0.1  0  0.1  0.2  0.3  0.4  0.5

Avg. tokens/sent.  # tokens  Avg. coref chain length

% proper nouns

correlation with score

Avg. tokens/sent.

# tokens

Avg. coref chain length

% proper nouns

All

teacher
Instructors

Correlation with score

-0.7

-0.6

-0.5

-0.4

-0.3

-0.2

-0.1

0

0.1

0.2

0.3

0.4

0.5

Avg. tokens/sent.

# tokens

Avg. coref chain length

% proper nouns

Instructors
Instructors

Do teachers grade differently?

…maybe
Single-task vs. Multi-task Learning
Single-task Learning

• Original model was single-task
  • learns one task at a time
  • scoring all essays
Single-task Learning

• Other single-task variations:
  • Model each project separately
Single-task Learning

- Other single-task variations:
  - Model each project separately

![Bar chart showing Pearson's r for STL, STL-proj, and Other single-task variations.
  Intermediate: STL = 0.4, STL-proj = 0.6
  Final: STL = 0.4, STL-proj = 0.6]
Single-task Learning

• Other single-task variations:
  • Model each project separately
  • Model each teacher separately
Single-task Learning

- Other single-task variations:
  - Model each project separately
  - Model each teacher separately

Pearson's r

- STL
- STL-proj
- STL-teach

Intermediate

Final

Pearson's r

0.0
0.2
0.4
0.6
0.8
Multi-task Learning

- Multi-task learning
  - learns many problems at the same time
    - how each teacher scores
      - jointly models the scores given by each teacher
  - takes advantage of shared knowledge
Multi-task Learning

How?
Multi-task Learning

How?

• Enlarge the feature space
  • Extracted $m$ features for each essay
  • Add teacher-specific features
    • Each feature has a global copy and a teacher-specific copy
  • Now, $m \times (1 + \# \text{ teachers})$ features
Multi-task Learning

How?

• each feature has a global feature and a teacher-specific feature for each teacher

• replicate feature values for the teacher-specific features if that teacher graded the essay (0 otherwise)

• STL: $m$ features

• MTL: $m \times (1 + \# \text{ teachers})$ features

• dimensionality reduction with PCA

• linear regression
Multi-task Learning
Multi-task Learning
Multi-task Learning

global \{ 
\begin{array}{c}
\frac{1}{0} \\
\frac{1}{5} \\
\frac{2}{1} \\
\frac{1}{2} \\
\frac{1}{4}
\end{array}
\}
Multi-task Learning

global

\{ \begin{align*} 
&1 \\
&0 \\
&5 
\end{align*} \}

teacher A

\begin{align*} 
&1 \\
&2 \\
&1 \\
&5 \\
&2 \\
&4 
\end{align*}
Multi-task Learning

global

\{ 

0

5

teacher A

\{ 

teacher B

\{ 

\{ 

1

0

5

2

1

5

1

2

4
Multi-task Learning

global

teacher A

teacher B

teacher A
Multi-task Learning

global

teacher A

teacher B

teacher A
Multi-task Learning

global

teacher A

teacher B

teacher A
Multi-task Learning

global

\begin{align*}
\text{teacher A} & \quad \begin{bmatrix}
1 \\
0 \\
5 \\
1 \\
0 \\
5 \\
0 \\
0 \\
0
\end{bmatrix} \\
\text{teacher B} & \quad \begin{bmatrix}
1 \\
2 \\
4
\end{bmatrix}
\end{align*}

\text{teacher A}
Multi-task Learning

global

teacher A

teacher B

teacher A
### Multi-task Learning

The diagram illustrates the concept of multi-task learning with three different teachers: global, teacher A, and teacher B. Each teacher has a set of tasks associated with it, represented by numbers in each column. The tasks are distributed across the columns, with teacher A having tasks in columns 1 and 2, teacher B in column 3, and the global tasks in column 4.

#### Teacher A
- Column 1: 1, 0, 5
- Column 2: 2, 1, 5

#### Teacher B
- Column 3: 0, 0, 0

#### Global
- Column 4: 1, 2, 4
Multi-task Learning

global

teacher A

teacher B

1 0 5
1 0 5
0 0 0

2 1 5
2 1 5
0 0 0

1 1 0 2 4

teacher B
Multi-task Learning

- global
- teacher A
- teacher B

![Diagram showing multi-task learning with teachers A and B and their corresponding tasks and outputs.](image)
Multi-task Learning

global

teacher A

teacher B

teacher B
Multi-task Learning
MTL Results

Pearson's r

- Baseline
- STL
- STL-teach
- STL-proj
- MTL

Baseline Results

Intermediate

Final
MTL Results

Pearson's $r$ values for Intermediate and Final phases with different methods: Baseline, STL, STL-teach, STL-proj, and MTL.
MTL Results

- Baseline
- STL
- MTL

% agreement

- Intermediate Exact
- Intermediate Adj.
- Final Exact
- Final Adj.

Bar chart showing comparison between Baseline, STL, and MTL for different stages (Intermediate Exact, Intermediate Adj., Final Exact, Final Adj.).
MTL Results

- Baseline
- STL
- MTL

% agreement

Intermediate Exact
Intermediate Adj.
Final Exact
Final Adj.

0 0.2 0.4 0.6 0.8
Other experiments

Can we predict...

• specific rubric scores?

• the improvement/decline between aligned drafts?

• scores given by unseen teachers?
Pearson's r

Predict rubric scores

Overall  Focus  Evidence  Organization  Style

Format

Intermediate

Final

Pearson's r

0.8

0.6

0.4

0.2

0
Can we predict the score change between aligned drafts?

- Train: 794 draft pairs
- Test: 50 pairs
Calculate the difference between the paired feature vectors:

<table>
<thead>
<tr>
<th>Intermediate</th>
<th>Final</th>
<th>Delta</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>-2</td>
</tr>
</tbody>
</table>
Predict improvement

Pearson's $r$

- Baseline
- STL, no content*
- STL
- MTL, no content*
- MTL

* no token unigram or trigram features
Unseen teachers

Pearson's r

Baseline   STL    MTL    STL-LOO    MTL-LOO

0.8

0.6

0.4

0.2

0.0

Intermediate    Final
Potential applications

- Examine feature weights across individual teacher models
- Potentially share this information to help teachers grade more consistently
Potential application
Summary

- A new corpus of student essays
  - more representative of college writing
- Multi-task learning to account for differences across teachers
Future Work

This task

• Tailor features for specific rubric categories
• Better model for unseen teachers
• Validate scores
• Test MTL on different writing corpora

This corpus

• Examine types of revisions made
• Categorize teacher comments
• Align teacher comments to spans of text
Thank you