Modeling Thesis Clarity in Student Essays

Isaac Persing and **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas

Automated Essay Grading

- Important educational application of NLP
- Related research on essay scoring
  - Grammatical errors (Leacock et al.,2010)
  - Coherence (Miltsakaki and Kukich, 2004)
  - Relevance to prompt (Higgins et al., 2004)
  - Organization (Persing et al., 2010)
  - ➢ Little work done on modeling *thesis clarity*

What is Thesis Clarity?

- refers to how clearly an author explains the thesis of her essay
  - the position she argues for with respect to the topic on which the essay is written

What is Thesis Clarity?

- refers to how clearly an author explains the thesis of her essay
  - the position she argues for with respect to the topic on which the essay is written

  overall message of the **entire** essay
  - unbound from the concept of thesis sentences

Goals

- Develop a model for scoring the thesis clarity of student essays

Goals

- Develop a model for scoring the thesis clarity of student essays
- Develop a system for determining why an essay receives its thesis clarity score

Goals

- Develop a model for scoring the thesis clarity of student essays
- Develop a system for determining why an essay receives its thesis clarity score
  - Provides more informative feedback to a student

Goals

- Develop a model for scoring the thesis clarity of student essays
- Develop a system for determining why an essay receives its thesis clarity score
  - Provides more informative feedback to a student
  - Given a predefined set of common errors that impact thesis clarify, determine which of these errors occur in a given essay

Plan for the Talk

- Corpus and Annotations
- Model for identifying thesis clarity errors
- Model for scoring thesis clarity
- Evaluation

Plan for the Talk

- ➢ Corpus and Annotations
- Model for identifying thesis clarity errors
- Model for scoring thesis clarity
- Evaluation

Selecting a Corpus

- International Corpus of Learner English (ICLE)
  - 4.5 million words in more than 6000 essays
  - Written by university undergraduates who are learners of English as a foreign language
  - Mostly (91%) argumentative writing topics

- Essays selected for annotation
  - 830 argumentative essays from 13 prompts
  - 2 types of annotation: thesis clarity score and errors

Thesis Clarity Scoring Rubric

4 – essay presents a **very clear thesis** and requires little or no clarification

3 – essay presents a **moderately clear thesis** but could benefit from some clarification

2 – essay presents an **unclear thesis** and would greatly benefit from further clarification

1 – essay presents **no thesis of any kind** and it is difficult to see what the thesis could be

- Half-point increments (i.e., 1.5, 2.5, 3.5) allowed

Inter-Annotator Agreement

- 100 of 830 essays scored by both annotators

Inter-Annotator Agreement

- 100 of 830 essays scored by both annotators

- Perfect agreement on 36% of essays
- Scores within 0.5 point on 62% of essays
- Scores within 1.0 point on 85% of essays

5 Types of Thesis Clarity Errors

5 Types of Thesis Clarity Errors

- **Confusing Phrasing** (18%)
  - Thesis is phrased oddly, making it hard to understand writer’s point

5 Types of Thesis Clarity Errors

- **Confusing Phrasing** (18%)
  - Thesis is phrased oddly, making it hard to understand writer's point
- **Incomplete Prompt Response** (15%)
  - Thesis seems to leave part of a multi-part prompt unaddressed

5 Types of Thesis Clarity Errors

- **Confusing Phrasing** (18%)
  - Thesis is phrased oddly, making it hard to understand writer's point
- **Incomplete Prompt Response** (15%)
  - Thesis seems to leave part of a multi-part prompt unaddressed
- **Relevance to Prompt** (17%)
  - The apparent thesis's weak relation to the prompt causes confusion

5 Types of Thesis Clarity Errors

- **Confusing Phrasing** (18%)
  - Thesis is phrased oddly, making it hard to understand writer's point
- **Incomplete Prompt Response** (15%)
  - Thesis seems to leave part of a multi-part prompt unaddressed
- **Relevance to Prompt** (17%)
  - The apparent thesis's weak relation to the prompt causes confusion
- **Missing Details** (6%)
  - Thesis omits important detail needed to understand writer's point

5 Types of Thesis Clarity Errors

- **Confusing Phrasing** (18%)
  - Thesis is phrased oddly, making it hard to understand writer's point
- **Incomplete Prompt Response** (15%)
  - Thesis seems to leave part of a multi-part prompt unaddressed
- **Relevance to Prompt** (17%)
  - The apparent thesis's weak relation to the prompt causes confusion
- **Missing Details** (6%)
  - Thesis omits important detail needed to understand writer's point
- **Writer Position** (5%)
  - Thesis describes a position on the topic without making it clear that this is the position the writer supports

Inter-Annotator Agreement

- 100 of 830 essays scored by 2 annotators
- Compute Cohen's Kappa on each error type from the two sets of annotations

Inter-Annotator Agreement

- 100 of 830 essays scored by 2 annotators
- Compute Cohen's Kappa on each error type from the two sets of annotations
- Average Kappa: 0.75

Plan for the Talk

- ✓ Corpus and Annotations
- ➢ Model for identifying thesis clarity errors
- Model for scoring thesis clarity
- Evaluation

Error Identification

- **Goal**: assign zero or more of the five error types to each essay

Error Identification

- **Goal**: assign zero or more of the five error types to each essay

- **Approach**:
  - recast problem as a set of 5 binary classification tasks
    - train five binary classifiers, each of which predicts whether a particular type of error exists in an essay

Learning the Binary Classification Tasks

- Goal: train a classifier c_i for identifying error type e_i

- Training data creation
  - create one training instance from each training essay
  - label the instance as
    - positive if essay has e_i as one of its labels
    - negative otherwise

- Learning algorithm
  - SVM^{light}

Features

- 7 types of features
  - 2 types of baseline features
  - 5 types of new features

Features

- 7 types of features
  - 2 types of baseline features
  - 5 types of new features

N-gram features

- Lemmatized unigrams, bigrams, and trigrams
  - only the top k n-gram features selected according to information gain is used for each classifier
    - k is determined using validation data

Features based on Random Indexing

- Random indexing
  - "an efficient and scalable alternative to LSI" (Sahlgren, 2005)
  - generates a semantic similarity measure between any two words

Why Random Indexing?

- May help identify Incomplete Prompt Response and Relevance to Prompt errors
  - May help find text in essay related to the prompt even if some of its words have been rephrased
    - E.g., essay talks about “jail” while prompt has “prison”
- Train a random indexing model on English Gigaword

4 Random Indexing Features

- The entire essay’s similarity to the prompt
- The essay’s highest individual sentence’s similarity to the prompt
- The highest entire essay similarity to one of the prompt sentences
- The highest individual sentence similarity to an individual prompt sentence

Features

- 7 types of features
  - 2 types of baseline features
  - 5 types of new features

Misspelling Feature

- Motivation
  - When examining the information gain top-ranked features for the Confusing Phrasing error, we see some misspelled words at the top of the list

Misspelling Feature

- Motivation
  - When examining the information gain top-ranked features for the Confusing Phrasing error, we see some misspelled words at the top of the list
- This makes sense!
  - A thesis sentence containing excessive misspellings may be less clear to the reader

Misspelling Feature

- Motivation
  - When examining the information gain top-ranked features for the Confusing Phrasing error, we see some misspelled words at the top of the list
- This makes sense!
  - A thesis sentence containing excessive misspellings may be less clear to the reader
- Introduce a misspelling feature
  - Value is the number of spelling errors in an essay's most-misspelled sentence

Keyword Features

- **Observations**
  - If an essay doesn't contain words that are semantically similar to the important words in the prompt (i.e., **keywords**), it could have a Relevance to Prompt error
  - If an essay doesn't contain words semantically similar to the keywords from every part of a multi-part prompt, it could have an Incomplete Prompt Response error

Keyword Features

- **Observations**
  - If an essay doesn't contain words that are semantically similar to the important words in the prompt (i.e., **keywords**), it could have a Relevance to Prompt error
  - If an essay doesn't contain words semantically similar to the keywords from every part of a multi-part prompt, it could have an Incomplete Prompt Response error
- **Hypothesis**: could identify these two types of errors by
  1. Hand-picking keywords for each part of each prompt
  2. Designing features that encode how similar an essay's words are to the keywords

Step 1: Hand-Selecting Keywords

- Hand-segment each multi-part prompt into parts
- For each part, hand-pick the most important (primary) and second most important (secondary) words that it would be good for a writer to use to address the part

Step 1: Hand-Selecting Keywords

- Hand-segment each multi-part prompt into parts
- For each part, hand-pick the most important (primary) and second most important (secondary) words that it would be good for a writer to use to address the part

The prison system is outdated. No civilized society should punish its criminals: it should rehabilitate them.

Step 1: Hand-Selecting Keywords

- Hand-segment each multi-part prompt into parts
- For each part, hand-pick the most important (primary) and second most important (secondary) words that it would be good for a writer to use to address the part

The prison system is outdated. No civilized society should punish its criminals: it should rehabilitate them.

Step 1: Hand-Selecting Keywords

- Hand-segment each multi-part prompt into parts
- For each part, hand-pick the most important (primary) and second most important (secondary) words that it would be good for a writer to use to address the part

> The prison system is outdated. No civilized society should punish its criminals: it should rehabilitate them.

Primary: rehabilitate
Secondary: society

Step 2: Designing Keyword Features

- **Example**: in one feature, we
  1. compute the random indexing similarity between the essay and each group of primary keywords taken from parts of the essay's prompt
  2. assign the feature the lowest of these values

Step 2: Designing Keyword Features

- **Example**: in one feature, we
  1. compute the random indexing similarity between the essay and each group of primary keywords taken from parts of the essay's prompt
  2. assign the feature the lowest of these values

- A low feature value suggests that the essay may have an Incomplete Prompt Response error

Aggregated Word N-gram Features

- **Motivation**: Regular N-gram features have a problem
  - It is infrequent for the exact same useful phrase to occur frequently
    - May render useful phrases less useful

Aggregated Word N-gram Features

- **Motivation**: Regular N-gram features have a problem
  - It is infrequent for the exact same useful phrase to occur frequently
    - May render useful phrases less useful

- **Solution**: Construct aggregate versions of the word N-gram features

Aggregated Word N-gram Features

- **Motivation**: Regular N-gram features have a problem
  - It is infrequent for the exact same useful phrase to occur frequently
    - May render useful phrases less useful

- **Solution**: Construct aggregate versions of the word N-gram features

How?

Aggregated Word N-gram Features

- For each error type e_i, we create two aggregated word n-gram features, $Aw+_i$ and $Aw-_I$

Aggregated Word N-gram Features

- For each error type e_i, we create two aggregated word n-gram features, $Aw+_i$ and $Aw-_I$

  $Aw+_i$ counts the number of word n-grams we believe indicate that the essay contains e_i

Aggregated Word N-gram Features

- For each error type e_i, we create two aggregated word n-gram features, $Aw+_i$ and $Aw-_l$

$Aw+_i$ counts the number of word n-grams we believe indicate that the essay contains e_i	$Aw-_i$ counts the number of word n-grams we believe indicate that the essay does **not** contain e_i

Aggregated Word N-gram Features

- For each error type e_i, we create two aggregated word n-gram features, $Aw+_i$ and $Aw-_I$

$Aw+_i$ counts the number of word n-grams we believe indicate that the essay contains e_i	$Aw-_i$ counts the number of word n-grams we believe indicate that the essay does **not** contain e_i

- To compute $Aw+_i$ and $Aw-_i$, we need to create two sets of word n-grams for each error type e_i

Aggregated Word N-gram Features

- For each error type e_i, we create two aggregated word n-gram features, $Aw+_i$ and $Aw-_l$

$Aw+_i$ counts the number of word n-grams we believe indicate that the essay contains e_i	$Aw-_i$ counts the number of word n-grams we believe indicate that the essay does **not** contain e_i

- To compute $Aw+_i$ and $Aw-_i$, we need to create two sets of word n-grams for each error type e_i
  - word n-grams whose presence suggest essay has e_i
  - word n-grams whose presence suggest essay doesn't have e_i

How to create these two sets?

How to create these two sets?

- For each error type e_i,
  - sort the list of all word n-gram features occurring at least 10 times in the training set by information gain
  - by inspecting the top 1000 features, manually create
    - **a positive set**
    - **a negative set**

How to create these two sets?

- For each error type e_i,
  - sort the list of all word n-gram features occurring at least 10 times in the training set by information gain
  - by inspecting the top 1000 features, manually create
    - **a positive set**
      - word n-grams whose presence suggest essay has e_i
    - **a negative set**
      - word n-grams whose presence suggest essay doesn’t have e_i

Aggregated Word N-gram Features

- May help identify the two minority error types, Missing Details and Writer Position

Aggregated Word N-gram Features

- May help identify the two minority error types, Missing Details and Writer Position
  - e.g., for Missing Details
    - positive set may contain phrases like “there is something” or “this statement”
    - negative set may contain words taken from an essay’s prompt

Aggregated POS N-gram Features

- Computed in the same way as the aggregated word n-gram features, except that POS n-grams (n = 1, 2, 3 and 4) are used
  - Two sets, the positive set and the negative set, are created manually for each error type i

Aggregated Frame-based Features

- For each sentence in an essay,
  1. identify each semantic frame occurring in it as well as the associated frame elements using SEMAFOR
     - **frame**: describes an event mentioned in a sentence
     - **frame element**: person/object participating in the event

Aggregated Frame-based Features

- For each sentence in an essay,
  1. identify each semantic frame occurring in it as well as the associated frame elements using SEMAFOR
     - **frame**: describes an event mentioned in a sentence
     - **frame element**: person/object participating in the event

"They said they don't believe the prison system is outdated"

Aggregated Frame-based Features

- For each sentence in an essay,
  1. identify each semantic frame occurring in it as well as the associated frame elements using SEMAFOR
     - **frame**: describes an event mentioned in a sentence
     - **frame element**: person/object participating in the event

  "They said they don't believe the prison system is outdated"
     - **frame**: Statement
     - **frame element**: they with the semantic role Speaker

Aggregated Frame-based Features

- For each sentence in an essay,
  1. identify each semantic frame occurring in it as well as the associated frame elements using SEMAFOR
     - **frame**: describes an event mentioned in a sentence
     - **frame element**: person/object participating in the event

  "They said they don't believe the prison system is outdated"
     - **frame**: Statement
     - **frame element**: they with the semantic role Speaker

  2. create a frame-based feature by pairing the frame with the frame element and its role
     - Statement-Speaker-they

Aggregated Frame-based Features

- After collecting all frame-based features, create **aggregated** frame-based features
  - Computed in the same way as aggregated word/POS n-gram features, except that frame-based features are used
    - Two sets, the positive set and the negative set, are created manually for each error type i

Aggregated Frame-based Features

- After collecting all frame-based features, create **aggregated** frame-based features
  - Computed in the same way as aggregated word/POS n-gram features, except that frame-based features are used
    - Two sets, the positive set and the negative set, are created manually for each error type i
- May help identify Writer Position errors
  - e.g., positive set may contain Statement-Speaker-they
    - It tells us the writer is attributing the statement made to someone else

Features for Training the Error Identification Classifiers

- Two types of baseline features
  - Lemmatized n-grams
  - Random indexing features
- Five types of novel features
  - Misspelling feature
  - Keyword features
  - Aggregated word n-gram features
  - Aggregated POS n-gram features
  - Aggregated frame-based features

Plan for the Talk

- ✓ Corpus and Annotations
- ✓ Model for identifying thesis clarity errors
- ➢ Model for scoring thesis clarity
- Evaluation

Score Prediction

- **Goal**:
  - predict the thesis clarity score for an essay

Score Prediction

- **Goal**:
  - predict the thesis clarity score for an essay
- **Approach**:
  - recast problem as a linear regression task

Score Prediction

- **Goal**:
  - predict the thesis clarity score for an essay
- **Approach**:
  - recast problem as a linear regression task
  - One training instance created from each training essay
    - "class" value: thesis clarity score
    - features: same as those used for error identification
    - learner: SVM^{light}

Plan for the Talk

- ✓ Corpus and Annotations
- ✓ Model for identifying thesis clarity errors
- ✓ Model for scoring thesis clarity
- ➢ Evaluation

Evaluation

- **Goal**: evaluate our systems for
  - error identification
  - scoring

- 5-fold cross validation

Evaluation

- **Goal**: evaluate our systems for
  - error identification
  - scoring

Evaluation Metrics

- Recall, precision, micro F, and macro F aggregated over the 5 error types
  - Micro F: places more importance on frequent classes
  - Macro F: places equal importance on all classes

Results: Error Identification

System	Prec.	Recall	Micro F	Macro F
Baseline	24.8	44.7	31.1	24.0

Results: Adding Misspelling Feature

System	Prec.	Recall	Micro F	Macro F
Baseline	24.8	44.7	31.1	24.0
+ Misspelling feature	24.2	44.2	31.2	25.3

Results: Adding Misspelling Feature

System	Prec.	Recall	Micro F	Macro F
Baseline	24.8	44.7	31.1	24.0
+ Misspelling feature	24.2	44.2	31.2	25.3

- small, insignificant improvements in micro and macro F
  - Though designed to improve Confusing Phrasing, it has more of a positive impact on Missing Details and Writer Position

Results: Adding Keyword Features

System	Prec.	Recall	Micro F	Macro F
Baseline	24.8	44.7	31.1	24.0
+ Misspelling feature	24.2	44.2	31.2	25.3
+ Keyword features	29.2	44.2	34.9	26.7

Results: Adding Keyword Features

System	Prec.	Recall	Micro F	Macro F
Baseline	24.8	44.7	31.1	24.0
+ Misspelling feature	24.2	44.2	31.2	25.3
+ Keyword features	29.2	44.2	34.9	26.7

- Significant gains in micro F; insignificant gains in macro F
  - due to large improvements in Incomplete Prompt Response and Relevance to Prompt

Results: Adding Aggregated Word n-grams

System	Prec.	Recall	Micro F	Macro F
Baseline	24.8	44.7	31.1	24.0
+ Misspelling feature	24.2	44.2	31.2	25.3
+ Keyword features	29.2	44.2	34.9	26.7
+ Aggregated word n-grams	28.5	49.6	35.5	31.4

Results: Adding Aggregated Word n-grams

System	Prec.	Recall	Micro F	Macro F
Baseline	24.8	44.7	31.1	24.0
+ Misspelling feature	24.2	44.2	31.2	25.3
+ Keyword features	29.2	44.2	34.9	26.7
+ Aggregated word n-grams	28.5	49.6	35.5	31.4

- Significant gains in macro F; insignificant gains in micro F
  - due to large improvements in Missing Details and Writer Position

Results: Adding Aggregated POS n-grams

System	Prec.	Recall	Micro F	Macro F
Baseline	24.8	44.7	31.1	24.0
+ Misspelling feature	24.2	44.2	31.2	25.3
+ Keyword features	29.2	44.2	34.9	26.7
+ Aggregated word n-grams	28.5	49.6	35.5	31.4
+ Aggregated POS n-grams	34.2	49.6	40.4	34.6

Results: Adding Aggregated POS n-grams

System	Prec.	Recall	Micro F	Macro F
Baseline	24.8	44.7	31.1	24.0
+ Misspelling feature	24.2	44.2	31.2	25.3
+ Keyword features	29.2	44.2	34.9	26.7
+ Aggregated word n-grams	28.5	49.6	35.5	31.4
+ Aggregated POS n-grams	34.2	49.6	40.4	34.6

- Significant gains in both micro and macro F
  - due to large improvements in Confusing Phrasing, Incomplete Prompt Response, and Missing Details

Results: Adding Aggregated Frames

System	Prec.	Recall	Micro F	Macro F
Baseline	24.8	44.7	31.1	24.0
+ Misspelling feature	24.2	44.2	31.2	25.3
+ Keyword features	29.2	44.2	34.9	26.7
+ Aggregated word n-grams	28.5	49.6	35.5	31.4
+ Aggregated POS n-grams	34.2	49.6	40.4	34.6
+ Aggregated frames	33.6	54.4	41.4	37.6

Results: Adding Aggregated Frames

System	Prec.	Recall	Micro F	Macro F
Baseline	24.8	44.7	31.1	24.0
+ Misspelling feature	24.2	44.2	31.2	25.3
+ Keyword features	29.2	44.2	34.9	26.7
+ Aggregated word n-grams	28.5	49.6	35.5	31.4
+ Aggregated POS n-grams	34.2	49.6	40.4	34.6
+ Aggregated frames	33.6	54.4	41.4	37.6

- Significant gains in macro F; insignificant gains in micro F
  - due to very large improvements in Missing Details and Writer Position

Results: Adding Aggregated Frames

System	Prec.	Recall	Micro F	Macro F
Baseline	24.8	44.7	31.1	24.0
+ Misspelling feature	24.2	44.2	31.2	25.3
+ Keyword features	29.2	44.2	34.9	26.7
+ Aggregated word n-grams	28.5	49.6	35.5	31.4
+ Aggregated POS n-grams	34.2	49.6	40.4	34.6
+ Aggregated frames	33.6	54.4	41.4	37.6

- Full system improves the baseline by 13.3% in macro F and 10.3% in micro F

Results: Adding Aggregated Frames

System	Prec.	Recall	Micro F	Macro F
Baseline	24.8	44.7	31.1	24.0
+ Misspelling feature	24.2	44.2	31.2	25.3
+ Keyword features	29.2	44.2	34.9	26.7
+ Aggregated word n-grams	28.5	49.6	35.5	31.4
+ Aggregated POS n-grams	34.2	49.6	40.4	34.6
+ Aggregated frames	33.6	54.4	41.4	37.6

- Full system improves the baseline by 13.3% in macro F and 10.3% in micro F
- No consistent pattern to how precision and recall changed as more features are added

Evaluation

- **Goal**: evaluate our systems for
  - error identification
  - scoring

Scoring Metrics

- Define 3 evaluation metrics:

Scoring Metrics

- Define 3 evaluation metrics:

$$S_1 = \frac{1}{N} \sum_{A_i \neq E_i} 1$$

measures frequency at which a system predicts the wrong score out of 7 possible scores

A_i and E_i are annotated and estimated scores

Scoring Metrics

- Define 3 evaluation metrics:

$$S_1 = \frac{1}{N} \sum_{A_i \neq E_i} 1$$

(frequency of error)

$$S_2 = \frac{1}{N} \sum_{i=1}^{N} |A_i - E_i|$$

measures the average distance between a predicted score and a correct score

A_i and E_i are annotated and estimated scores

Scoring Metrics

- Define 3 evaluation metrics:

$$S_1 = \frac{1}{N} \sum_{A_i \neq E_i} 1$$

(frequency of error)

$$S_2 = \frac{1}{N} \sum_{i=1}^{N} |A_i - E_i|$$

distinguishes near misses
from far misses

A_i and E_i are annotated and estimated scores

Scoring Metrics

- Define 3 evaluation metrics:

$$S_1 = \frac{1}{N} \sum_{A_i \neq E_i} 1$$ (frequency of error)

$$S_2 = \frac{1}{N} \sum_{i=1}^{N} |A_i - E_i|$$ (average error distance)

$$S_3 = \frac{1}{N} \sum_{i=1}^{N} (A_i - E_i)^2$$ measures average square of the distance between correct score and predicted score

A_i and E_i are annotated and estimated scores

Scoring Metrics

- Define 3 evaluation metrics:

$$S_1 = \frac{1}{N} \sum_{A_i \neq E_i} 1$$ (frequency of error)

$$S_2 = \frac{1}{N} \sum_{i=1}^{N} |A_i - E_i|$$ (average error distance)

$$S_3 = \frac{1}{N} \sum_{i=1}^{N} (A_i - E_i)^2$$

prefer systems whose estimations are not too often far away from correct scores

A_i and E_i are annotated and estimated scores

Results: Scoring

System	**S1**	**S2**	**S3**
Baseline	.658	.517	.403

Results: Adding Misspelling Feature

System	S1	S2	S3
Baseline	.658	.517	.403
+ Misspelling feature	.654	.515	.402

Results: Adding Misspelling Feature

System	S1	S2	S3
Baseline	.658	.517	.403
+ Misspelling feature	.654	.515	.402

- small, insignificant improvements in scoring according to all 3 metrics

Results: Adding Keyword Features

System	S1	S2	S3
Baseline	.658	.517	.403
+ Misspelling feature	.654	.515	.402
+ Keyword features	.663	.490	.369

Results: Adding Keyword Features

System	S1	S2	S3
Baseline	.658	.517	.403
+ Misspelling feature	.654	.515	.402
+ Keyword features	.663	.490	.369

- S2’s and S3’s scores are improved significantly
- insignificant impact on S1’s score

Results: Adding Aggregated Word n-grams

System	S1	S2	S3
Baseline	.658	.517	.403
+ Misspelling feature	.654	.515	.402
+ Keyword features	.663	.490	.369
+ Aggregated word n-grams	.651	.484	.374

Results: Adding Aggregated Word n-grams

System	**S1**	**S2**	**S3**
Baseline	.658	.517	.403
+ Misspelling feature	.654	.515	.402
+ Keyword features	.663	.490	.369
+ Aggregated word n-grams	.651	.484	.374

- S2’s score is improved significantly
- insignificant impact on the other two metrics

Results: Adding the Remaining Features

System	S1	S2	S3
Baseline	.658	.517	.403
+ Misspelling feature	.654	.515	.402
+ Keyword features	.663	.490	.369
+ Aggregated word n-grams	.651	.484	.374
+ Aggregated POS n-grams	.671	.483	.377
+ Aggregated frames	.672	.486	.382

Results: Adding the Remaining Features

System	**S1**	**S2**	**S3**
Baseline	.658	.517	.403
+ Misspelling feature	.654	.515	.402
+ Keyword features	.663	.490	.369
+ Aggregated word n-grams	.651	.484	.374
+ Aggregated POS n-grams	.671	.483	.377
+ Aggregated frames	.672	.486	.382

- Adding aggregated POS n-grams and aggregated frame-based features do not improve any scores

Summary

- Examined the problem of determining thesis clarity errors and scores in student essays
  - Proposed new features for use in these tasks
  - Lots of room for improvement
- Released the thesis clarity annotations