Human Language Technology Research Institute

Modeling Prompt Adherence in Student Essays

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

Automated Essay Scoring

- Important educational application of NLP
- Related research on essay scoring
  - Grammatical errors
  - Coherence
  - Thesis clarity
  - Organization
  - Prompt adherence

What is Prompt Adherence?

- refers to how related an essay's content is to the prompt for which it was written

What is Prompt Adherence?

- refers to how related an essay's content is to the prompt for which it was written
  - An essay with a high prompt adherence score consistently remains on topic introduced by the prompt and is free of irrelevant digressions

Goal

- Develop a model for scoring the prompt adherence of student essays

Related Work on Prompt Adherence Scoring

- Off-topic **sentence** detection (Higgins et al., 2004)
- Off-topic **essay** detection (Higgins et al., 2006)
- Off-topic **essay** detection with **short prompts** (Louis and Higgins, 2010)

Related Work on Prompt Adherence Scoring

- Off-topic **sentence** detection (Higgins et al., 2004)
- Off-topic **essay** detection (Higgins et al., 2006)
- Off-topic **essay** detection with **short prompts** (Louis and Higgins, 2010)

Binary decision (off-topic/on-topic)

Related Work on Prompt Adherence Scoring

- Off-topic **sentence** detection (Higgins et al., 2004)
- Off-topic **essay** detection (Higgins et al., 2006)
- Off-topic **essay** detection with **short prompts** (Louis and Higgins, 2010)

Binary decision (off-topic/on-topic)

Knowledge-lean

– Features derived from semantic similarity measures
  - Random indexing (RI) and Content Vector Analysis (CVA)

What are the differences between our work and previous work?

What are the differences between our work and previous work?

Task:

Score can range from 1-4 points

Supervised prompt adherence scoring

Approach:

Feature-rich

Plan for the Talk

- Corpus and Annotations
- Approach for scoring prompt adherence
- Evaluation

Plan for the Talk

- ➢ Corpus and Annotations
- Approach for scoring prompt adherence
- 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
  - annotate each essay with its prompt adherence score

Prompt Adherence Scoring Rubric

4 – essay fully addresses the prompt and **consistently stays on topic**

3 – essay mostly addresses the prompt or **occasionally wanders off topic**

2 – essay does not fully address the prompt or **consistently wanders off topic**

1 – essay does not address the prompt at all or is **completely off topic**

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

Inter-Annotator Agreement

- 707 of 830 essays scored by both annotators

Inter-Annotator Agreement

- 707 of 830 essays scored by both annotators

- Perfect agreement on 38% of essays
- Scores within 0.5 points on 66% of essays
- Scores within 1.0 point on 89% of essays

Inter-Annotator Agreement

- 707 of 830 essays scored by both annotators

- Perfect agreement on 38% of essays
- Scores within 0.5 points on 66% of essays
- Scores within 1.0 point on 89% of essays

- Whenever annotators disagree, use the average score rounded to the nearest half point

Plan for the Talk

- ✓ Corpus and Annotations
- ➢ Approach for scoring prompt adherence
- Evaluation

Approach for Scoring Prompt Adherence

- recast problem as a linear regression task
- train one regressor per prompt

Approach for Scoring Prompt Adherence

- recast problem as a linear regression task
- train one regressor per prompt
  - common problems students have writing essays for one prompt may not apply to essays written for another

Approach for Scoring Prompt Adherence

- recast problem as a linear regression task
- train one regressor per prompt
  - common problems students have writing essays for one prompt may not apply to essays written for another
- one training instance per training essay
  - "class" value: prompt adherence score
  - learner: LIBSVM
  - features: 7 feature types

Features

- 7 types of features
  - baseline features
  - 6 types of new features

Features

- 7 types of features
  - baseline features
  - 6 types of new features

Baseline Features

- Features based on Random Indexing (RI)
  - adapted from Higgins et al. (2004)

Baseline Features

- Features based on Random Indexing (RI)
  - adapted from Higgins et al. (2004)

- Random indexing
  - "an efficient and scalable alternative to LSI" (Sahlgren, 2005)
  - generates a semantic similarity measure between any two words
  - generalized to computing similarity between two groups of words (Higgins & Burstein, 2007)

Why Random Indexing (RI)?

- 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 RI model on the English Gigaword

5 Random Indexing Features

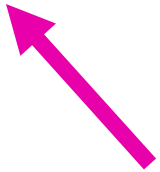
- The entire essay's similarity to the prompt
- The essay's highest individual sentence's similarity to the prompt
- The highest entire essay similarity to one of the prompt sentences
- The highest individual sentence similarity to an individual prompt sentence
- The essay's similarity to a manually rewritten version of the prompt that excludes extraneous material

Features

- 7 types of features
  - baseline features
  - 6 types of features

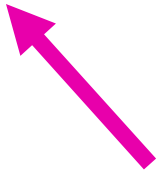
1. Thesis Clarity Keyword Features

1. Thesis Clarity Keyword Features



refers to how clearly an author explains the thesis of her essay

1. Thesis Clarity Keyword Features



refers to how clearly an author explains the thesis of her essay

- introduced in Persing & Ng (2013) for scoring the thesis clarity of an essay
- generated based on thesis clarity keywords

What are Thesis Clarity Keywords?

What are Thesis Clarity Keywords?

- "important" words in a prompt
  - important word: good word for a student to use when stating her thesis about the prompt

How to Identify Keywords?

How to Identify Keywords?

- By hand

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.

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.

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

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

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 student ignored the prompt component from which the value came

Thesis Clarity Keyword Features

- Though these features were designed for scoring thesis clarity, some of them are useful for prompt adherence scoring

2. Prompt Adherence Keyword Features

- **Motivation**: rather than relying on keywords for thesis clarity, why not hand-pick keywords for prompt adherence and create features from them?

2. Prompt Adherence Keyword Features

- Rather than relying on keywords for thesis clarity, why not hand-pick keywords for prompt adherence and create features from them?

An Illustrative Example

Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television.

An Illustrative Example

Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television.

- This question suggests that students discuss whether television is analogous to religion in this way

An Illustrative Example

Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television.

- This question suggests that students discuss whether television is analogous to religion in this way
  - prompt adherence keywords contain “religion”
  - thesis clarity keywords do not contain “religion”

An Illustrative Example

Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television.

- This question suggests that students discuss whether television is analogous to religion in this way
  - prompt adherence keywords contain “religion”
  - thesis clarity keywords do not contain “religion”
  - A thesis like “Television is bad” can be stated clearly without reference to “religion”

An Illustrative Example

Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television.

- This question suggests that students discuss whether television is analogous to religion in this way
  - prompt adherence keywords contain "religion"
  - thesis clarity keywords do not contain "religion"
  - A thesis like "Television is bad" can be stated clearly without reference to "religion"
    - essay with this thesis could have high thesis clarity score
      - But low adherence score: Religion should be discussed

Creating Features from Keywords

- Two types of features

Creating Features from Keywords

- Two types of features

- For each prompt component,
  1. take the RI similarity between the whole essay and the component's keywords
  2. compute the fraction of the component's keywords that appear in the essay

3. LDA Topics

- **Motivation**: the features introduced so far have trouble identifying topics that are related to but not explicitly mentioned in the prompt

3. LDA Topics

- **Motivation**: the features introduced so far have trouble identifying topics that are related to but not explicitly mentioned in the prompt

All armies should consist entirely of professional soldiers: there is no value in a system of military service

3. LDA Topics

- **Motivation**: the features introduced so far have trouble identifying topics that are related to but not explicitly mentioned in the prompt

> All armies should consist entirely of professional soldiers: there is no value in a system of military service

- An essay containing words like “peace”, “patriotism”, or “training” are probably not digressions and should not be penalized for discussing these topics

3. LDA Topics

- **Motivation**: the features introduced so far have trouble identifying topics that are related to but not explicitly mentioned in the prompt

> All armies should consist entirely of professional soldiers: there is no value in a system of military service

- An essay containing words like “peace”, “patriotism”, or “training” are probably not digressions and should not be penalized for discussing these topics
  - But the various measures of keyword similarities might not notice that anything related to the prompt is discussed

3. LDA Topics

- **Motivation**: the features introduced so far have trouble identifying topics that are related to but not explicitly mentioned in the prompt

> All armies should consist entirely of professional soldiers: there is no value in a system of military service

- An essay containing words like “peace”, “patriotism”, or “training” are probably not digressions and should not be penalized for discussing these topics
  - But the various measures of keyword similarities might not notice that anything related to the prompt is discussed
  - this might have effects like lowering the RI similarity scores

How to create LDA features?

How to create LDA features?

For each prompt,

1. collect all the essays in the ICLE corpus written in response to it, not just those we labeled

2. build an LDA of 1000 topics

How to create LDA features?

For each prompt,

1. collect all the essays in the ICLE corpus written in response to it, not just those we labeled

2. build an LDA of 1000 topics
   - Soft clustering of the words into 1000 sets

How to create LDA features?

For each prompt,

1. collect all the essays in the ICLE corpus written in response to it, not just those we labeled

2. build an LDA of 1000 topics
   - Soft clustering of the words into 1000 sets
     - E.g., for the most frequent topic for the military prompt, the five most important words are:

"man", "military", "service", "pay", and "war"

How to create LDA features?

For each prompt,

1. collect all the essays in the ICLE corpus written in response to it, not just those we labeled

2. build an LDA of 1000 topics
   - Model can tell us how much an essay spends on each topic
     - E.g.,

Topic1	25%
Topic2	45%
Topic3	15%
…	…
Topic1000	5%

How to create LDA features?

For each prompt,

1. collect all the essays in the ICLE corpus written in response to it, not just those we labeled

2. build an LDA of 1000 topics

3. construct 1000 features, one for each topic
   - Feature value encodes how much of the essay was spent discussing the topic

How to create LDA features?

For each prompt,

1. collect all the essays in the ICLE corpus written in response to it, not just those we labeled

2. build an LDA of 1000 topics

3. construct 1000 features, one for each topic
   - Feature value encodes how much of the essay was spent discussing the topic

     E.g., if an essay written for the military prompt spends

“man”, “military”, “service”, “pay”, “war”	45%
“fully”, “count”, “ordinary”, “czech”, “day”	55%

4. Manually Annotated LDA Topics

- **Motivation**: the regressor using LDA features may not be able to distinguish an infrequent topic that is adherent to the prompt and one that is an irrelevant digression

4. Manually Annotated LDA Topics

- **Motivation**: the regressor using LDA features may not be able to distinguish an infrequent topic that is adherent to the prompt and one that is an irrelevant digression
  - An infrequent topic may not appear enough in the training set for the regressor to make this judgment

4. Manually Annotated LDA Topics

- **Motivation**: the regressor using LDA features may not be able to distinguish an infrequent topic that is adherent to the prompt and one that is an irrelevant digression
  - An infrequent topic may not appear enough in the training set for the regressor to make this judgment

- Create manually annotated LDA features

How to create Manually Annotated LDA Features?

How to create Manually Annotated LDA Features?

1. For each set of essays written for a given prompt, build an LDA of 100 topics

How to create Manually Annotated LDA Features?

1. For each set of essays written for a given prompt, build an LDA of 100 topics

2. For each topic, inspect its top 10 words and hand-annotate it with a number from 0 to 5 representing how likely it is that the topic is adherent to the prompt
   - higher score → more adherent

How to create Manually Annotated LDA Features?

1. For each set of essays written for a given prompt, build an LDA of 100 topics

2. For each topic, inspect its top 10 words and hand-annotate it with a number from 0 to 5 representing how likely it is that the topic is adherent to the prompt
   - higher score → more adherent

3. For each essay, create 10 features from the labeled topics

10 Features from the labeled topics

- Five features encode the sum of contributions to an essay of topics annotated with the number 0, 1, …, 4, resp.

- Five features encode the sum of contributions to an essay of topics annotated with a number ≥ 1, ≥ 2, …, ≥ 5 resp.

10 Features from the labeled topics

- Five features encode the sum of contributions to an essay of topics annotated with the number 0, 1, …, 4, resp.

- Five features encode the sum of contributions to an essay of topics annotated with a number ≥ 1, ≥ 2, …, ≥ 5 resp.

- These features should give the regressor a better idea of how much of an essay is composed of prompt-related vs. prompt-unrelated discussions

5. Predicted Thesis Clarity Errors

- In previous work on thesis clarity essay scoring (Persing & Ng, 2013), we
  - score an essay w.r.t. the clarity of its thesis
  - determine which type(s) of errors an essay contains that detract from the clarity of its thesis

5 Types of Thesis Clarity Errors

- **Confusing Phrasing**
- **Missing Details**
- **Writer Position**
- **Incomplete Prompt Response**
- **Relevance to Prompt**

5 Types of Thesis Clarity Errors

- **Confusing Phrasing**
- **Missing Details**
- **Writer Position**
- **Incomplete Prompt Response**
- **Relevance to Prompt**

Features based on Error Types

- Introduced features for prompt adherence scoring that encode the error types an essay contains

Features based on Error Types

- Introduced features for prompt adherence scoring that encode the error types an essay contains
  - Though each essay was manually annotated with the errors it contains, in a realistic setting we won't have access to these manual annotations

Features based on Error Types

- Introduced features for prompt adherence scoring that encode the error types an essay contains
  - Though each essay was manually annotated with the errors it contains, in a realistic setting we won't have access to these manual annotations
    - Predict which of the 5 error types an essay contains
      - Recast as a multi-label classification task

Creating Predicted “Error Type” Features

- Add a binary feature indicating the presence or absence of each error type

6. N-gram features

- Can capture useful words and phrases related to a prompt

- 10K lemmatized unigrams, bigrams, and trigrams
  - selected according to information gain

Summary of Features

- Baseline features
  - Random indexing features
- Six types of features
  - Lemmatized n-grams
  - Thesis clarity keyword features
  - Prompt adherence keyword features
  - LDA topics
  - Manually annotated LDA topics
  - Predicted thesis clarity errors

Plan for the Talk

- ✓ Corpus and Annotations
- ✓ Model for scoring prompt adherence
- ➢ Evaluation

Evaluation

- **Goal**: evaluate our system for prompt adherence scoring

- 5-fold cross validation

Scoring Metrics

- Define 4 evaluation metrics

Scoring Metrics

- Define 4 evaluation metrics

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

probability that a system predicts the wrong score out of 7 possible scores (1, 1.5, 2, 2.5, 3, 3.5, 4)

Scoring Metrics

- Define 4 evaluation metrics

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

annotated scores

estimated scores

probability that a system predicts the wrong score out of 7 possible scores (1, 1.5, 2, 2.5, 3, 3.5, 4)

Scoring Metrics

- Define 4 evaluation metrics

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

(probability of error)

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

average absolute error

annotated scores

estimated scores

Scoring Metrics

- Define 4 evaluation metrics

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

(probability of error)

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

distinguishes near misses from far misses

annotated scores

estimated scores

Scoring Metrics

- Define 4 evaluation metrics

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

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

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

Scoring Metrics

- Define 4 evaluation metrics

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

(probability of error)

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

(average absolute error)

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

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

Scoring Metrics

- Define 4 evaluation metrics

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

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

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

PC: Pearson's correlation coefficient between A_i and E_i

S_1, S_2, S_3
- error metrics
- smaller value is better

Metrics

- Define 4 evaluation metrics

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

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

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

PC: Pearson's correlation coefficient between A_i and E_i

S_1, S_2, S_3

- error metrics
- smaller value is better

PC

- correlation coefficient
- larger value is better

- Define 4 evaluation metrics

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

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

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

PC: Pearson's correlation coefficient between A_i and E_i

Regressor Training

- SVM regressors are trained to maximize performance w.r.t. each scoring metric by tuning the regularization parameter on held-out development data

Results

System	S1	S2	S3	PC
Baseline	.517	.368	.234	.233

Results

System	S1	S2	S3	PC
Baseline	.517	.368	.234	.233

RI features

Results

probability of error

System	S1	S2	S3	PC
Baseline	.517	.368	.234	.233

Results

probability of error

average absolute error

System	S1	S2	S3	PC
Baseline	.517	.368	.234	.233

Results

System	S1 (probability of error)	S2 (average absolute error)	S3 (average squared error)	PC
Baseline	.517	.368	.234	.233

Results

System	S1	S2	S3	PC
	probability of error	average absolute error	average squared error	Pearson's τ
Baseline	.517	.368	.234	.233

Results

System	S1	S2	S3	PC
Baseline	.517	.368	.234	.233
Our system	.488	.348	.197	.360

probability of error

average absolute error

average squared error

Pearson's τ

Results

System	S1	S2	S3	PC
	probability of error	average absolute error	average squared error	Pearson's τ
Baseline	.517	.368	.234	.233
Our system	.488	.348	.197	.360

- Improvements w.r.t. all four scoring metrics

Results

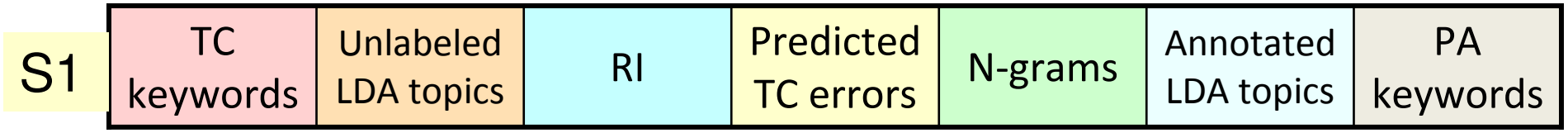
System	S1 (probability of error)	S2 (average absolute error)	S3 (average squared error)	PC (Pearson's τ)
Baseline	.517	.368	.234	.233
Our system	.488	.348	.197	.360

Significant differences

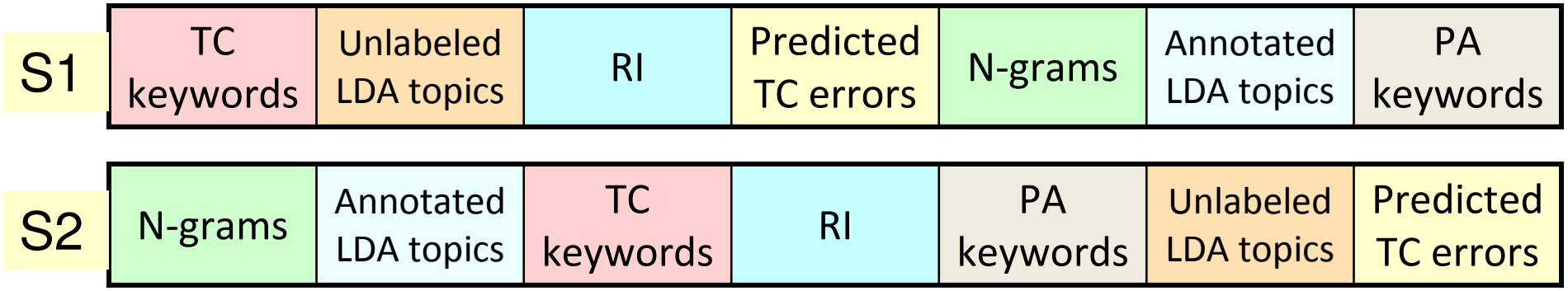
Feature Ablation

- **Goal**: examine how much impact each of the feature types has on our system's performance w.r.t. each scoring metric
  - Train a regressor on all but one type of features

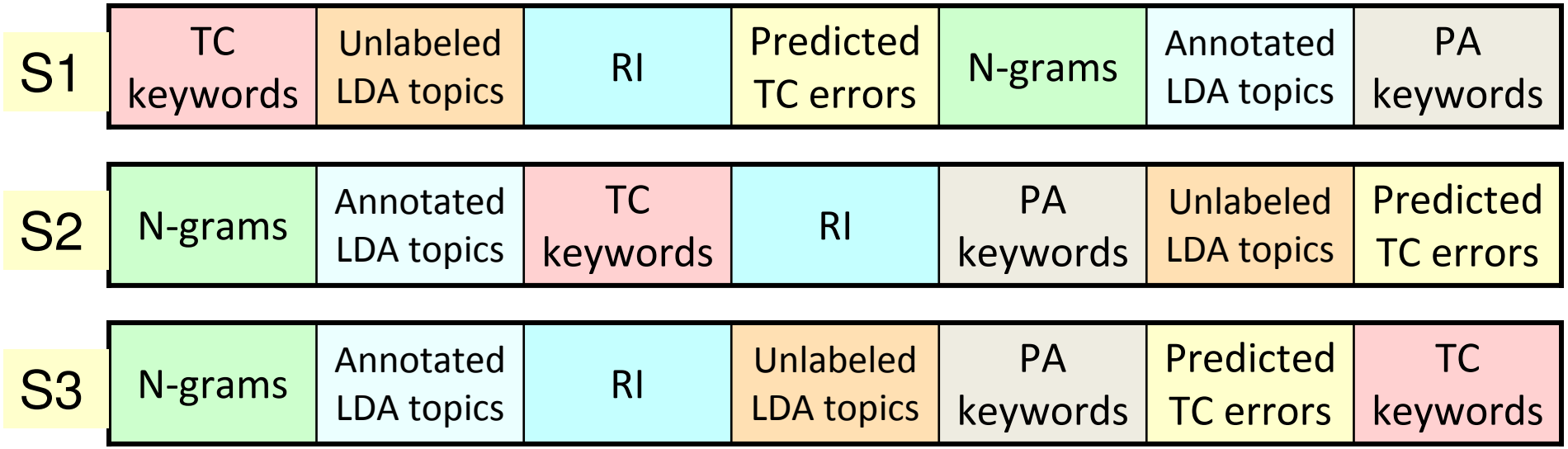
Feature Ablation Results



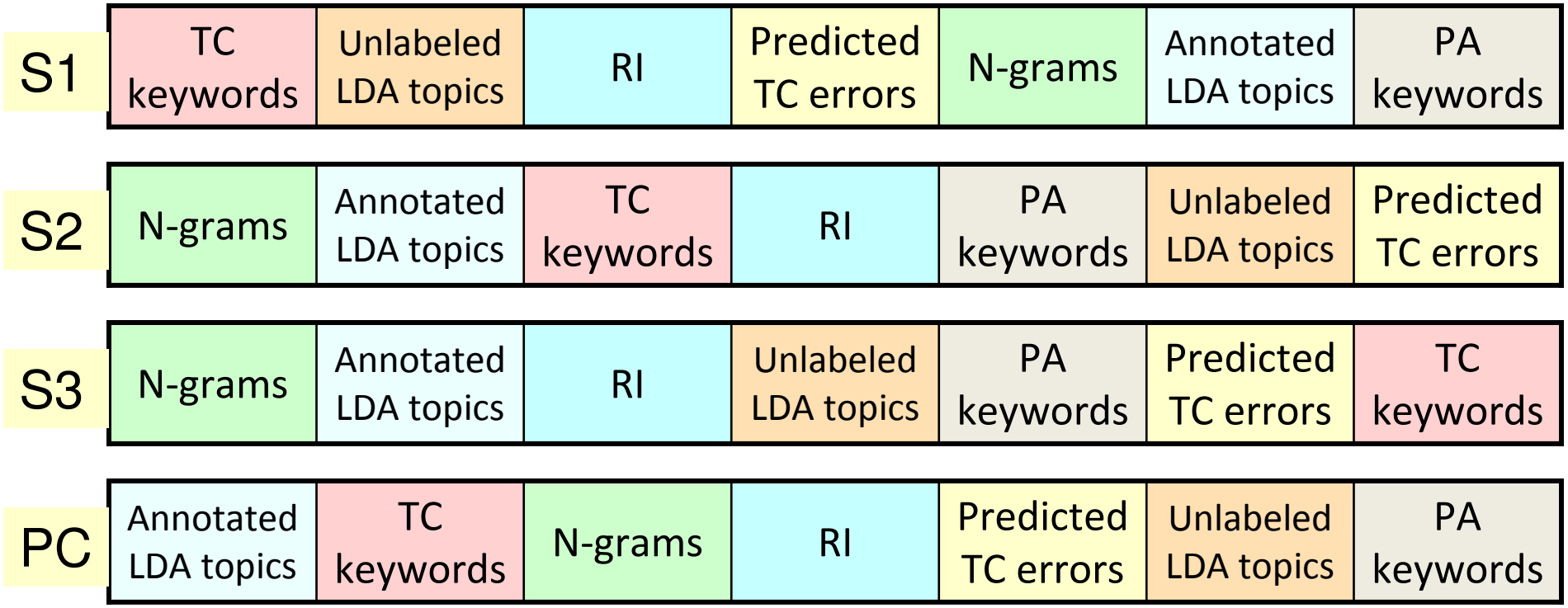
Feature Ablation Results

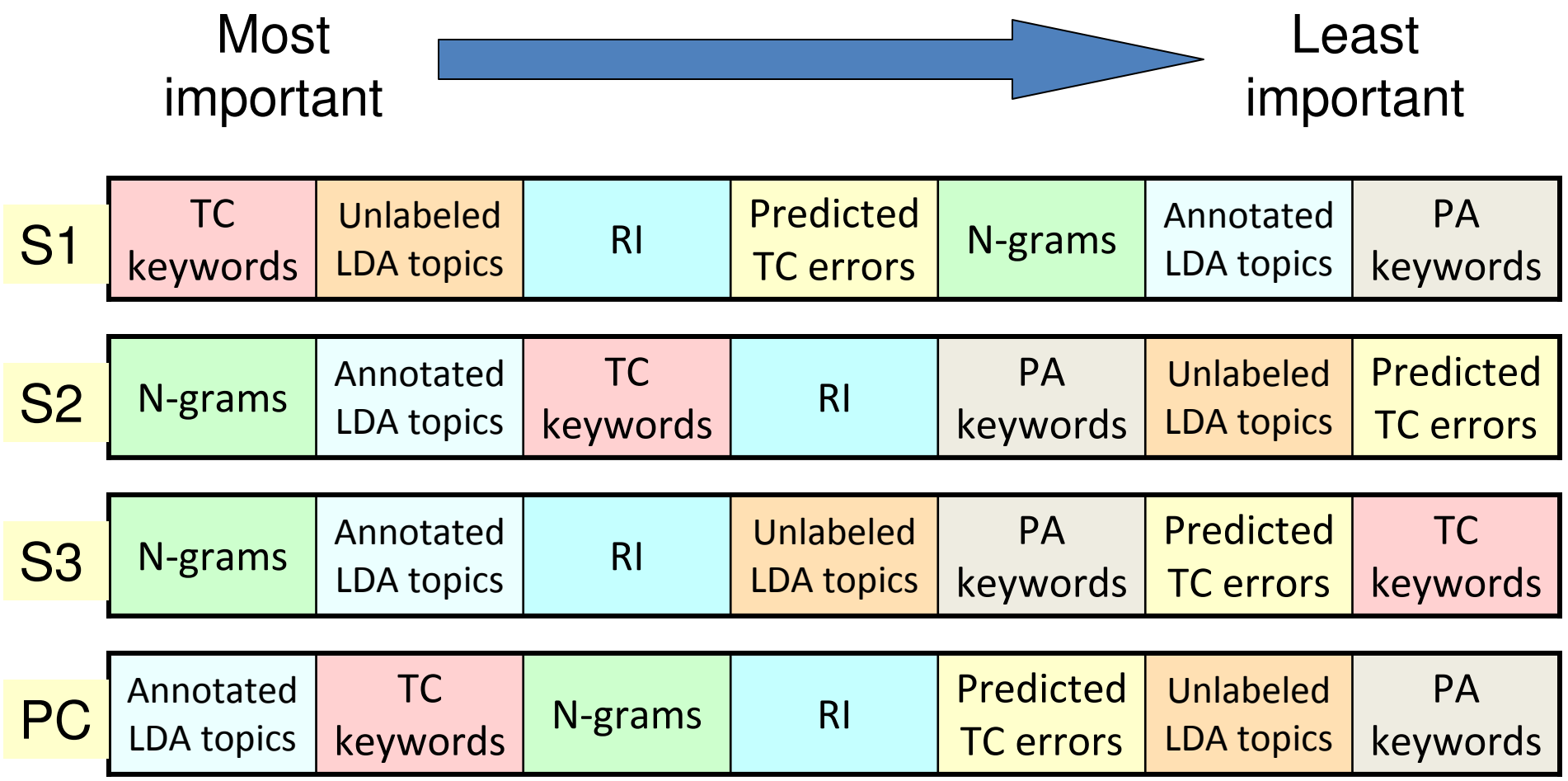


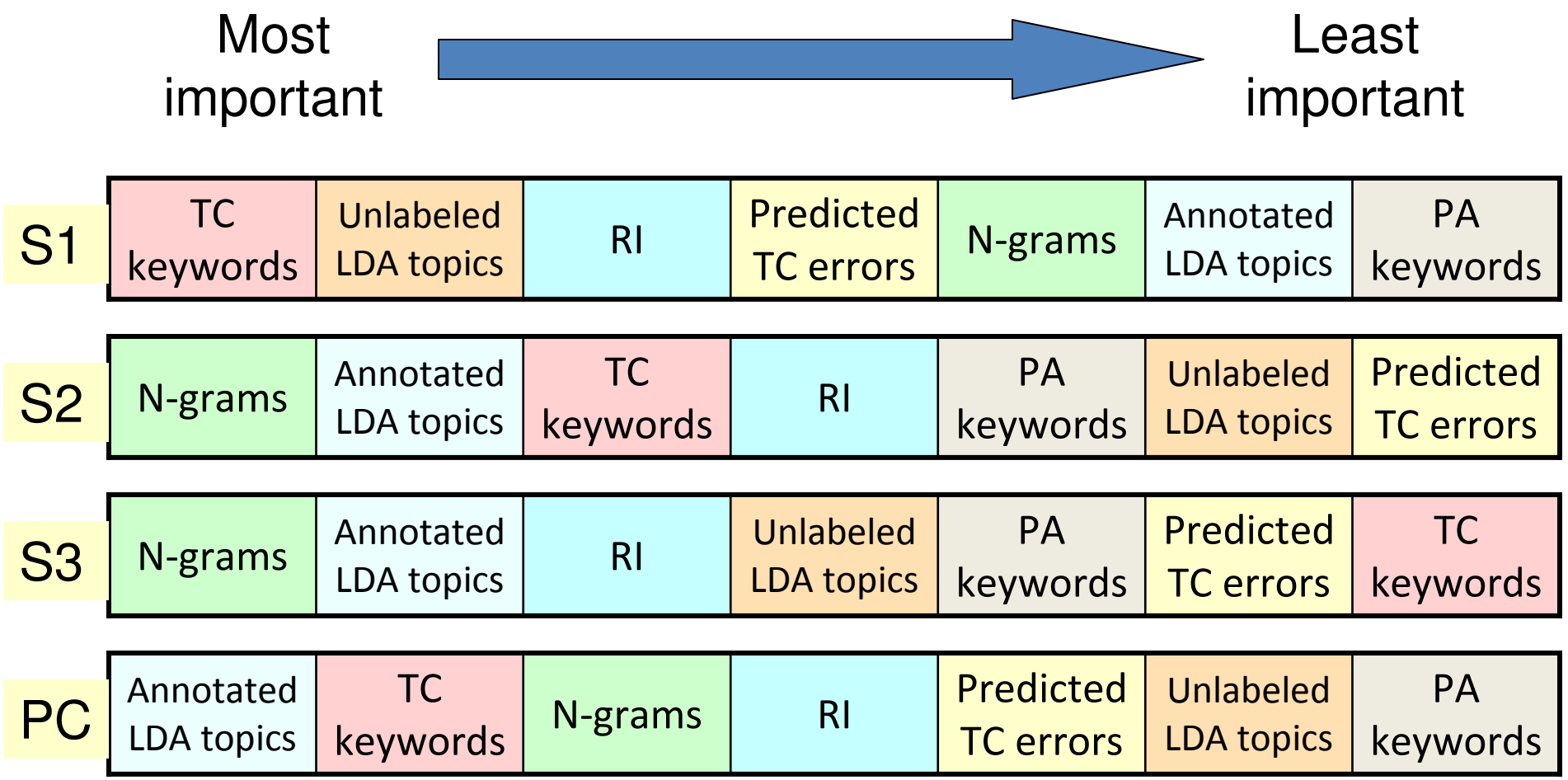
Feature Ablation Results



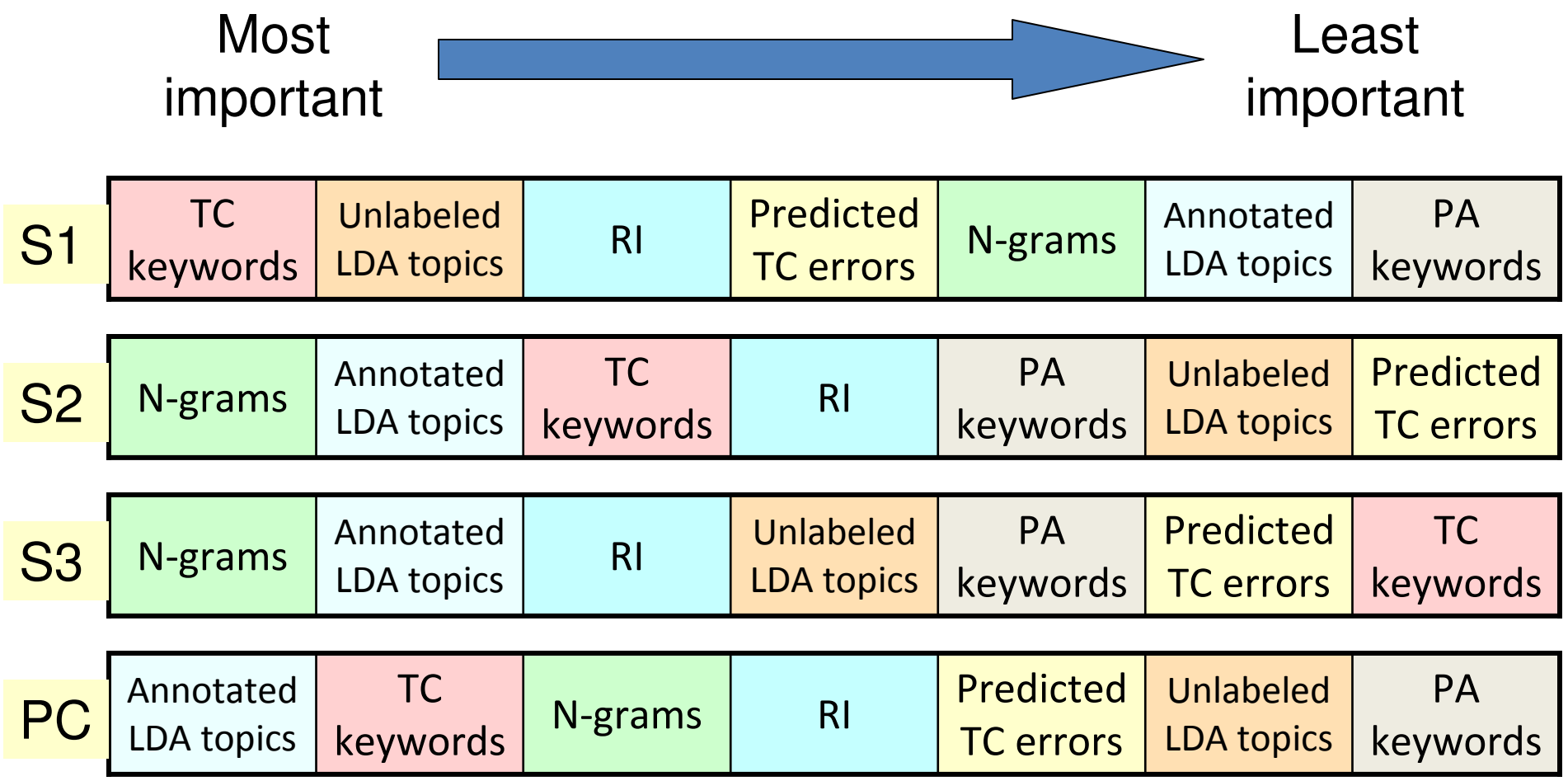
Feature Ablation Results



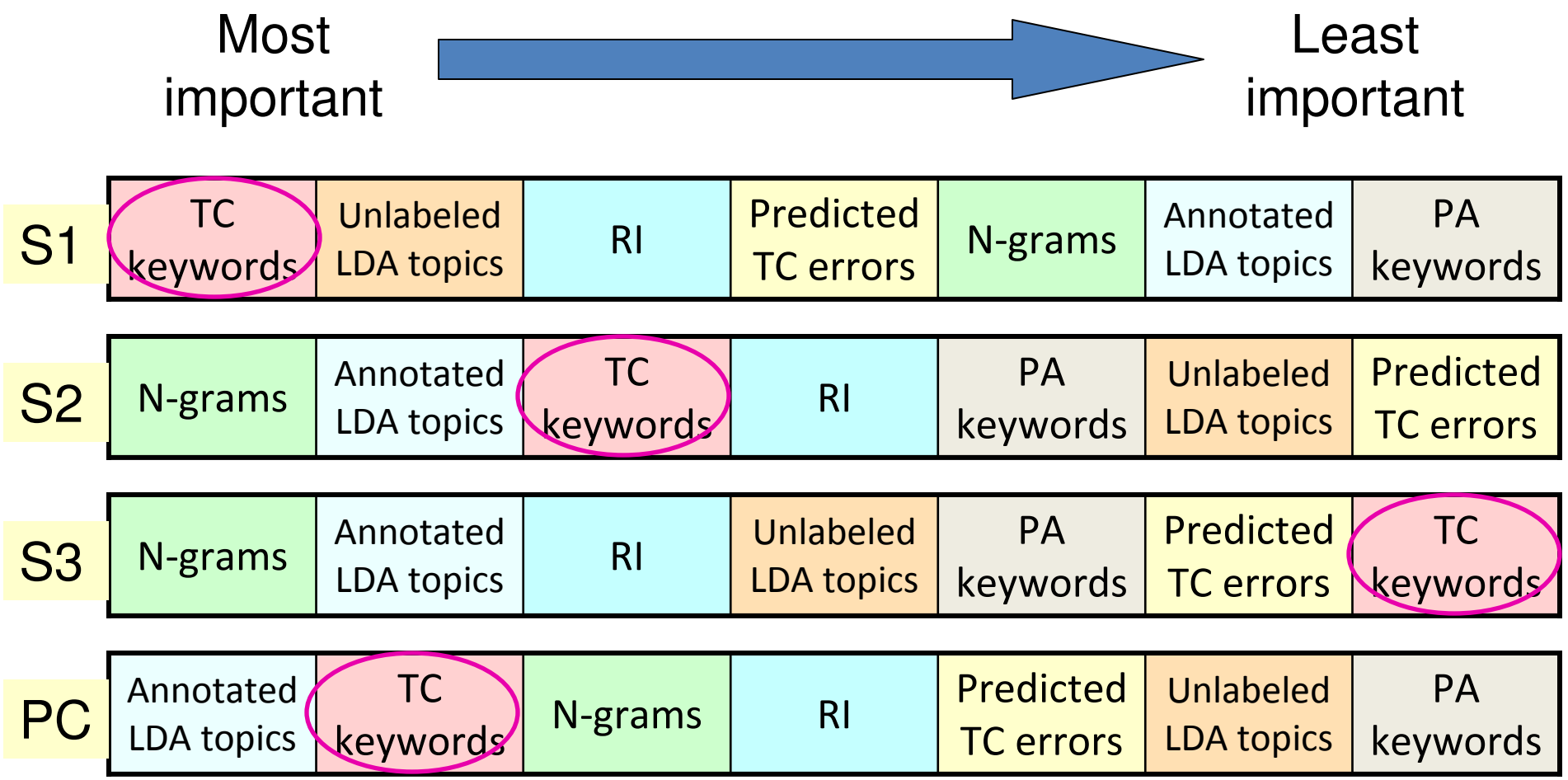

Most important
Least important
S1
TC keywords
Unlabeled LDA topics
RI
Predicted TC errors
N-grams
Annotated LDA topics
PA keywords
S2
N-grams
Annotated LDA topics
TC keywords
RI
PA keywords
Unlabeled LDA topics
Predicted TC errors
S3
N-grams
Annotated LDA topics
RI
Unlabeled LDA topics
PA keywords
Predicted TC errors
TC keywords
PC
Annotated LDA topics
TC keywords
N-grams
RI
Predicted TC errors
Unlabeled LDA topics
PA keywords



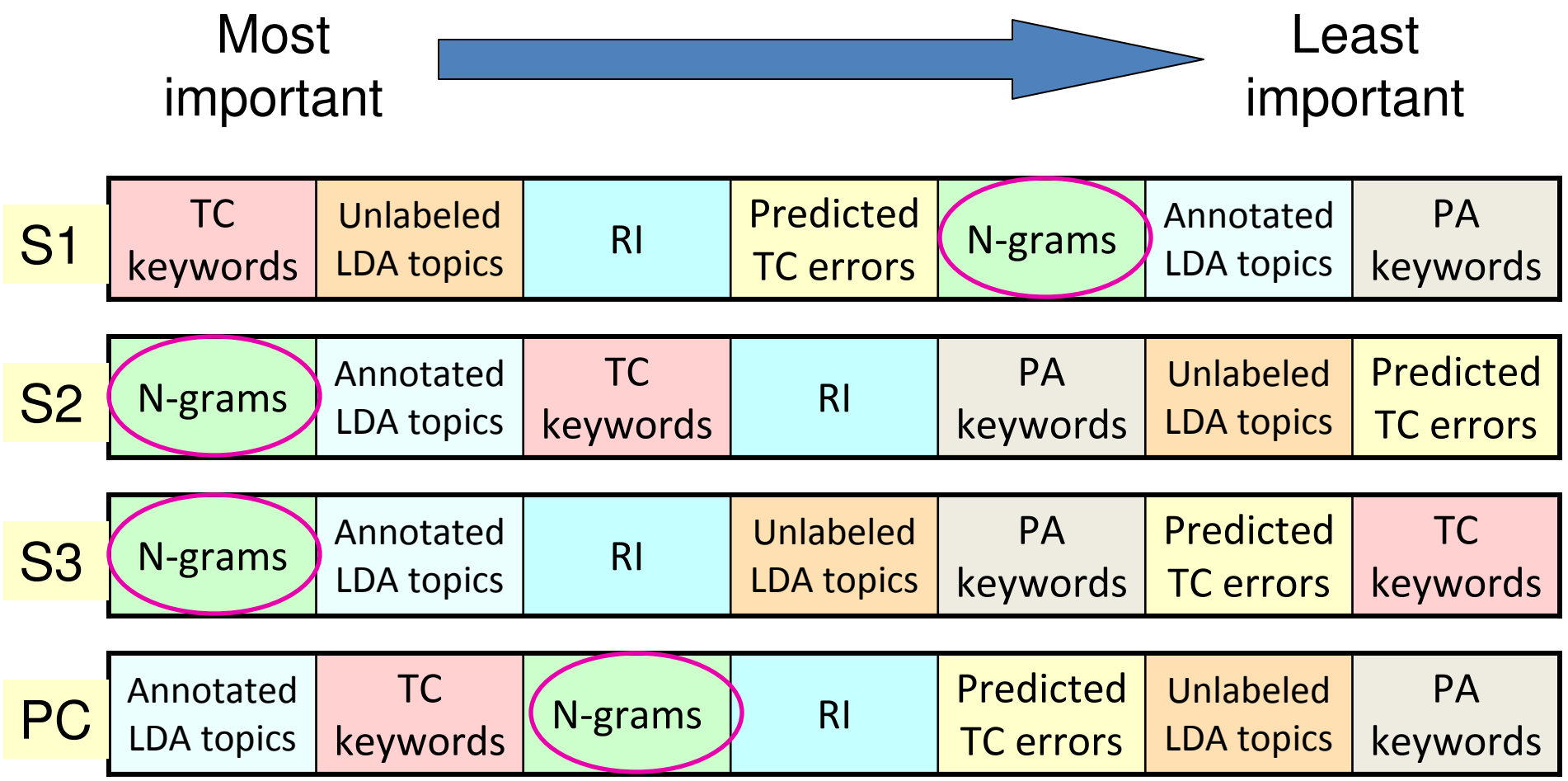
- Relative importance of features does not always remain consistent if we measure performance in different ways



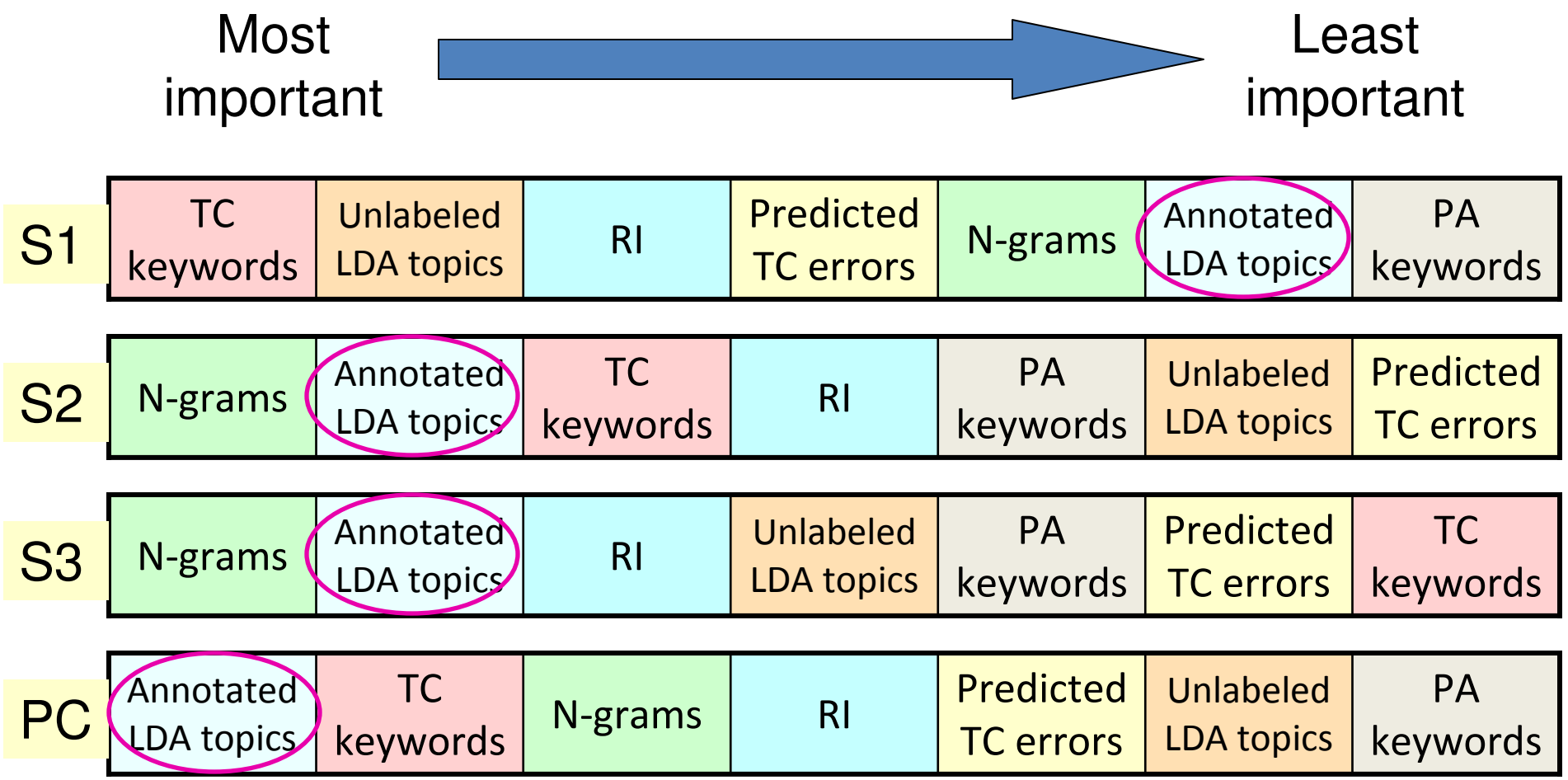
- But… there are feature that tend to be more important than the others in the presence of other features



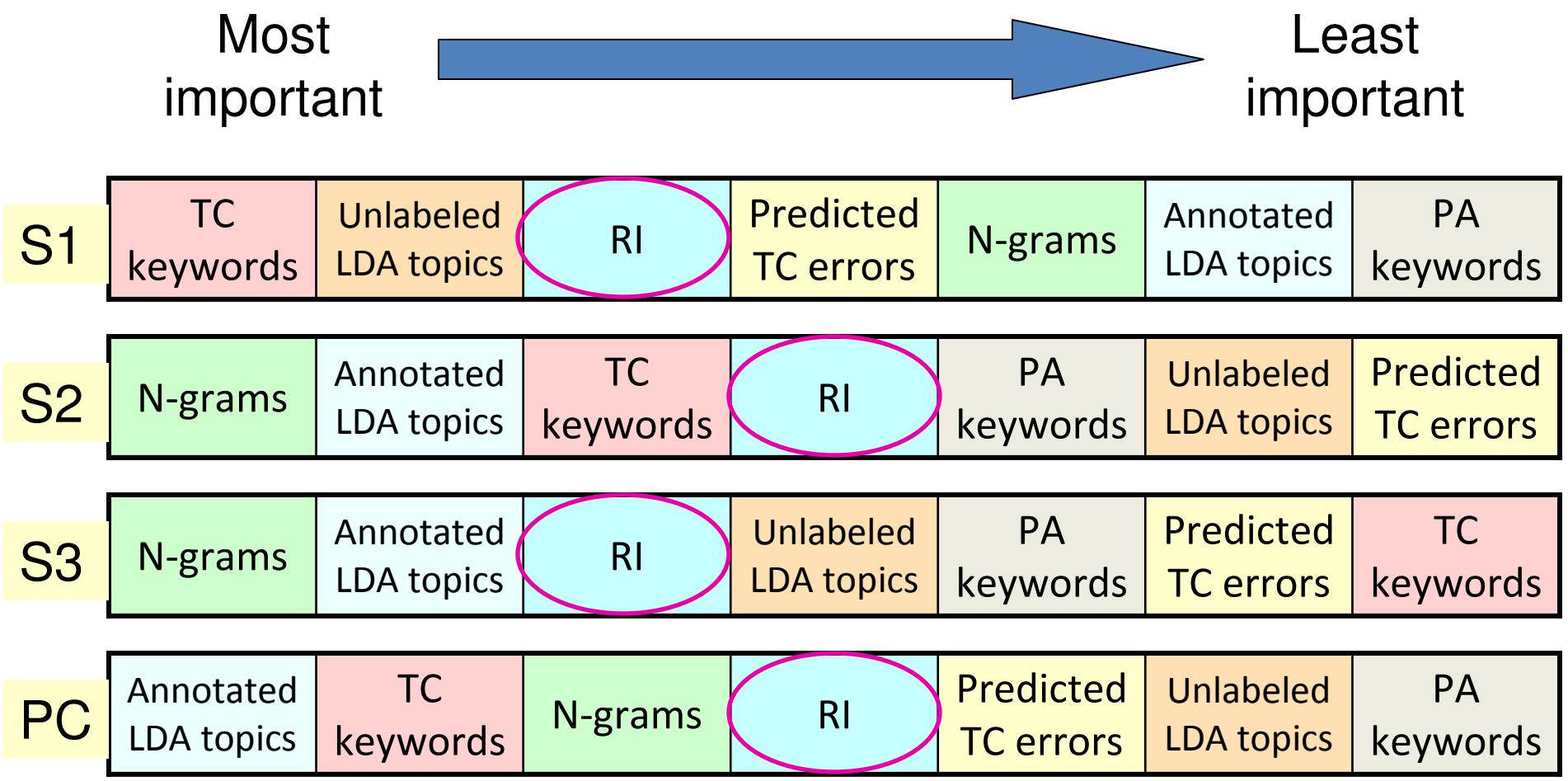
- most important: TC keywords



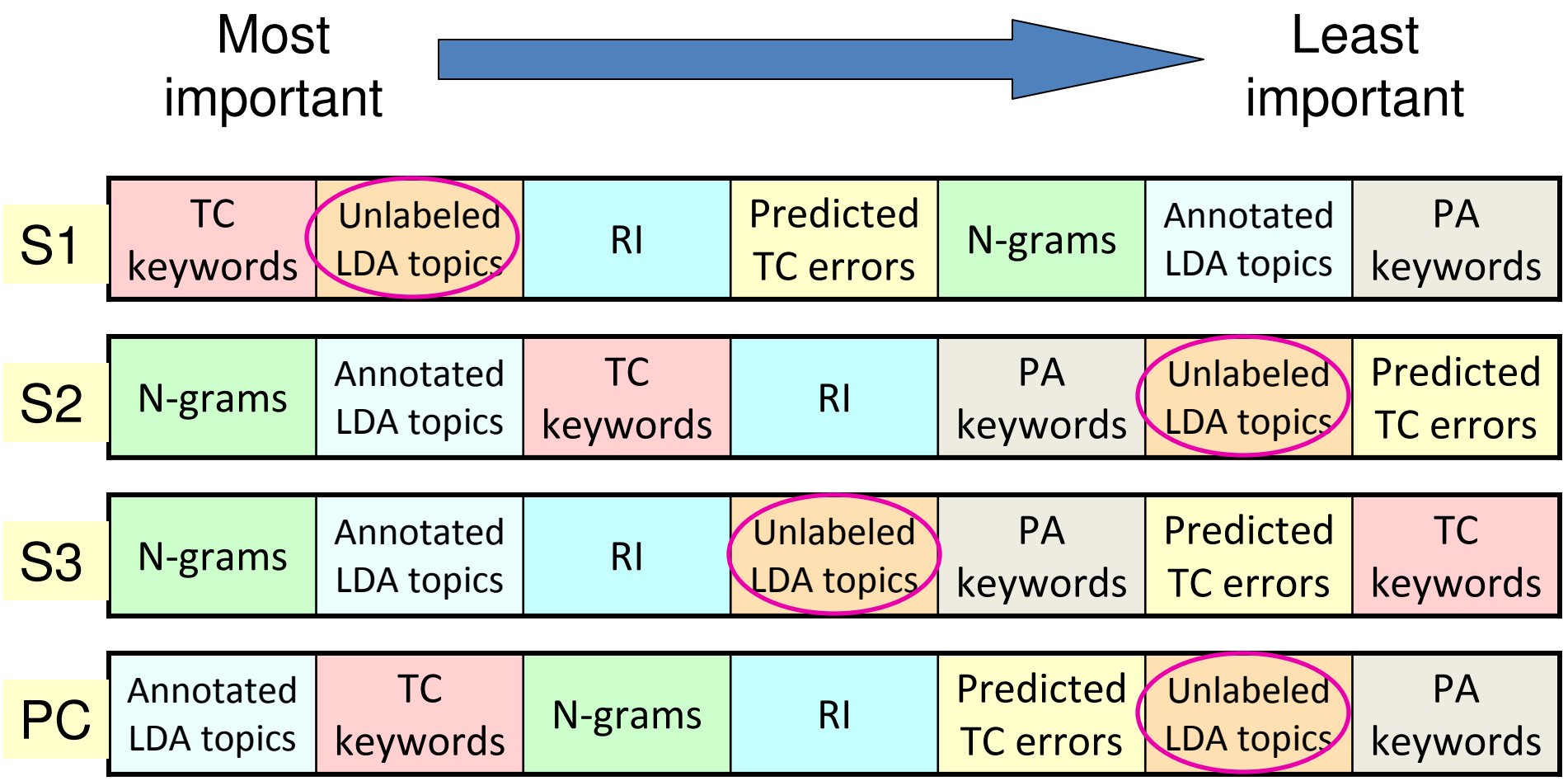
- most important: TC keywords, n-grams



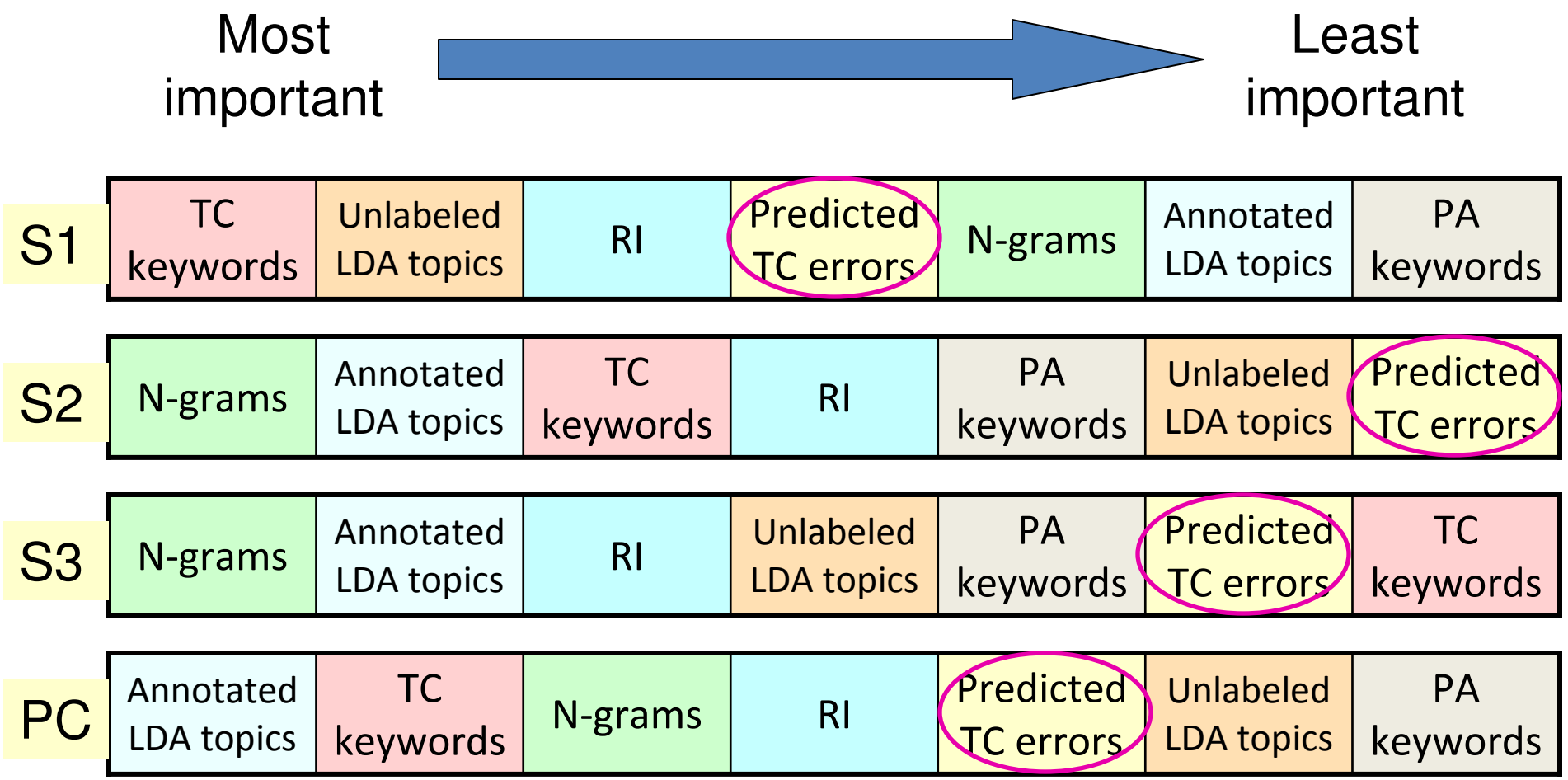
- most important: TC keywords, n-grams, annotated LDA topics



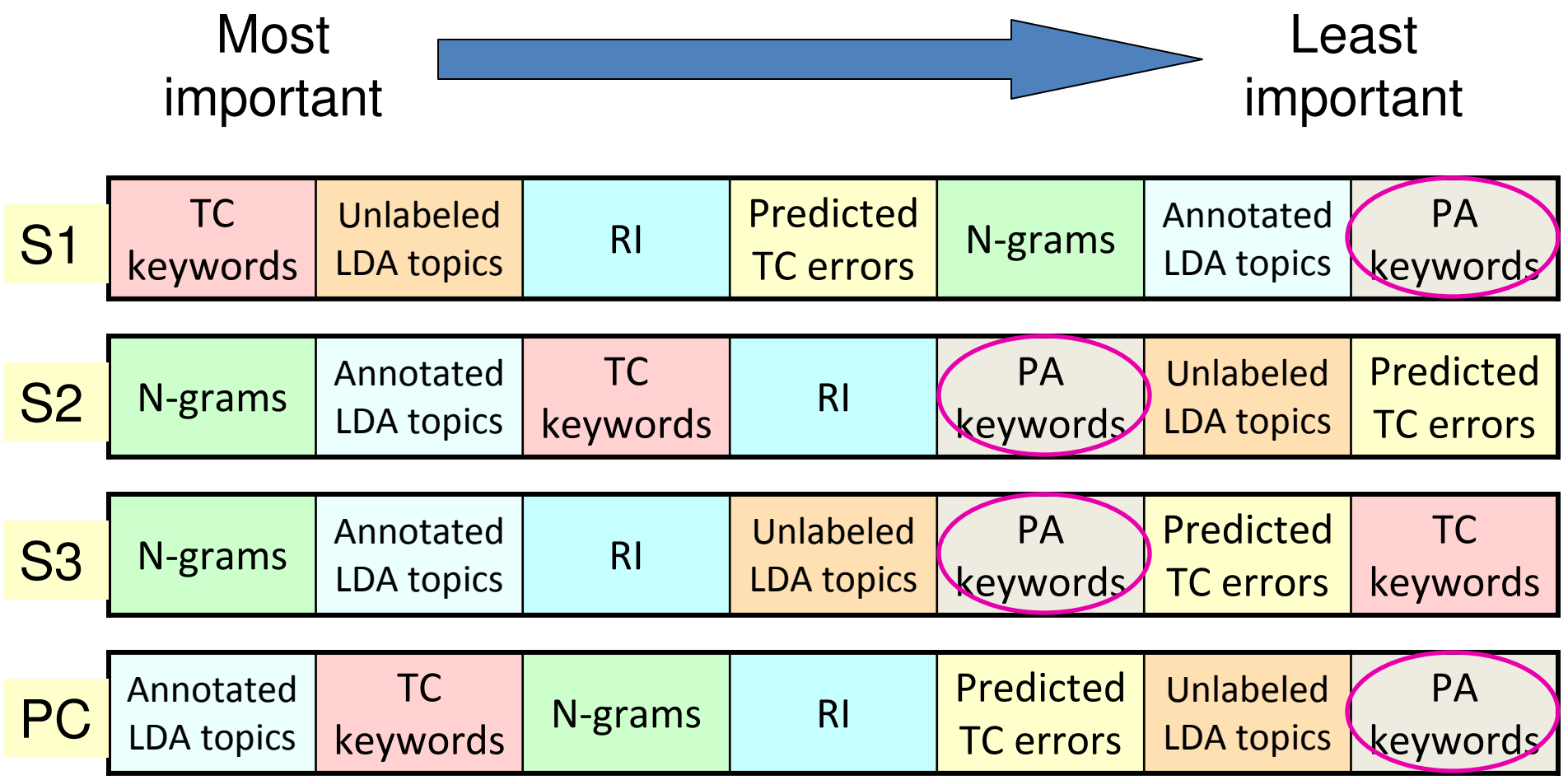
- most important: TC keywords, n-grams, annotated LDA topics
- middling important: RI



- most important: TC keywords, n-grams, annotated LDA topics
- middling important: RI, unlabeled LDA topics



- most important: TC keywords, n-grams, annotated LDA topics
- middling important: RI, unlabeled LDA topics
- least important: predicted TC errors



- most important: TC keywords, n-grams, annotated LDA topics
- middling important: RI, unlabeled LDA topics
- least important: predicted TC errors, PA keywords

Summary

- Examined the problem of prompt adherence scoring in student essays
  - feature-rich approach
- Released the annotations
  - prompt adherence scores
  - prompt adherence keywords
  - manually annotated LDA topics