IOWA STATE UNIVERSITY

Applied Linguistics and Technology Program, Department of English

Automated rhetorical analysis: A hybrid approach to classification error analysis

Elena Cotos

Panel: Automatic Analysis of Complexity/Accuracy/Fluency

CALICO 2018

University of Illinois, Urbana-Champaign, May 31, 2018

Overview

- Automated rhetorical analysis (ARA) for genre-based automated writing evaluation
- Hybrid approach to classification error analysis
- Further exploration and implications

ARA for genre-based AWE

Genre: culturally recognized text type (affidavit, research article) with conventional discourse structures, communicative purposes, and rhetorical functions

- Genre in Machine Learning
  - Widely recognized class of texts defined by a communicative purpose or other functional traits, provided the function is connected to some formal cues and that the class is extensible (Kessler et al., 1997)
  - Detected by identifying functional styles of texts, provided the style markers are a set of pre-defined quantifiable measures (Stamatatos et al., 2000)

'Move'

'Step'

ARA for genre-based AWE

- Automated categorization of genre

> A general inductive process builds an automatic text classifier by learning, from a set of pre-classified documents, the characteristics of the categories of interest (Sebastiani, 2002)

- ➢ classifier ‘learns’ the characteristics of moves & steps in a human-coded corpus
- ➢ classifier identifies the move/step characteristics that new texts should have in order to be classified similarly to human coding
  - e.g., Naïve Bayes, Decision Tree, Rule-based, Neural Network, Maximum Entropy, Regression, Support Vector Machine

ARA for genre-based AWE

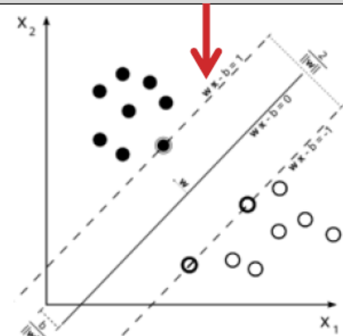
sentence → The operating system we had used for project development… — new

n-grams →
- the, operating, system, we, had, used, for, project, development
- operating system we, system we had, we had used, had used for, used for project, …

features →
- operating, system, used, project, development
- operating system we, system we had, we had used, used for project

representation → operating:1, system:1 , used:1, project:1, development:1, operating system we:1, system we had:1, we had used:1, used for project:1 (the:0, we:0, for:0, the operating system:0, for project development…)

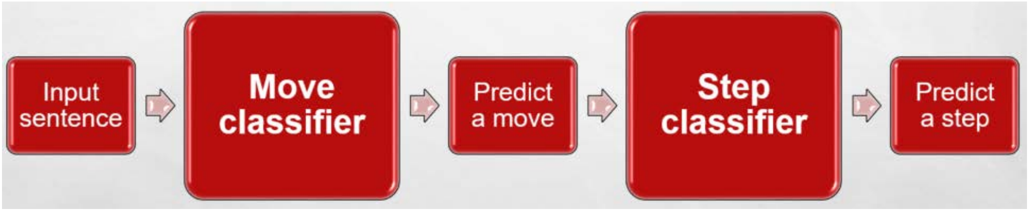
Move classifier →

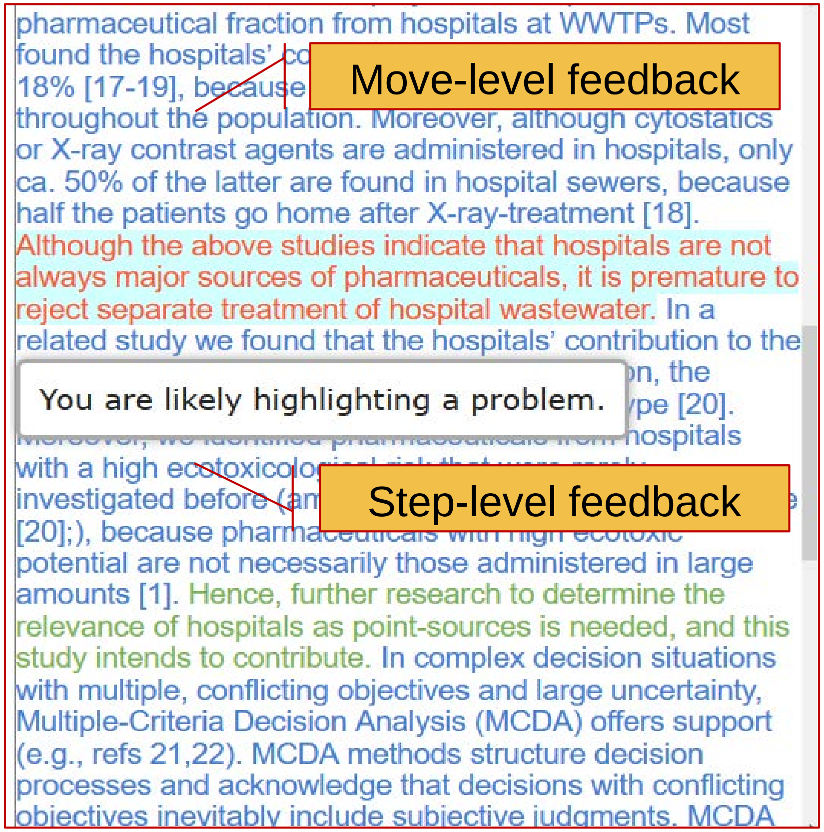


→ Move 3

Cotos & Pendar (2016)

ARA for genre-based AWE





Cotos (2016)

ARA performance

- Evaluation metrics vary across moves/steps
  (Cotos, Gilbert, & Sinapov, 2014; Cotos & Pendar, 2016; Cotos, Vajjala, Chapelle, & Kim, 2016)

e.g., Introduction

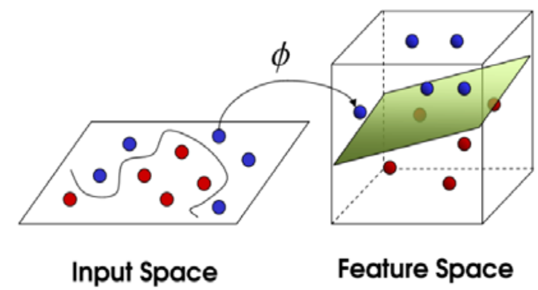
Move #	Move name	Precision (%)	Recall (%)	F1 Score (%)
1	Establishing a territory	73.3	**89.0**	**80.4**
2	Identifying a niche	59.2	37.3	45.8
3	Addressing the niche	**78.4**	57.2	66.1

Step #	Step name	Precision (%)	Recall (%)	F1 Score (%)
4 (Move2)	Indicating a gap	**75.2**	55.5	63.9
5 (Move2)	Highlighting a problem	64.7	**79.9**	**71.5**
6 (Move2)	Raising general questions	50.0	27.8	35.7
7 (Move2)	Proposing general hypotheses	66.3	50.0	57.0
8 (Move2)	Presenting a justification	68.9	66.2	67.5

- Data sparseness for some rhetorical categories
- Rhetorical meaning not clearly encoded in functional language
- Multiple rhetorical functions
- Semantic ambiguity

Need to understand ARA classification output

HYBRID ERROR ANALYSIS

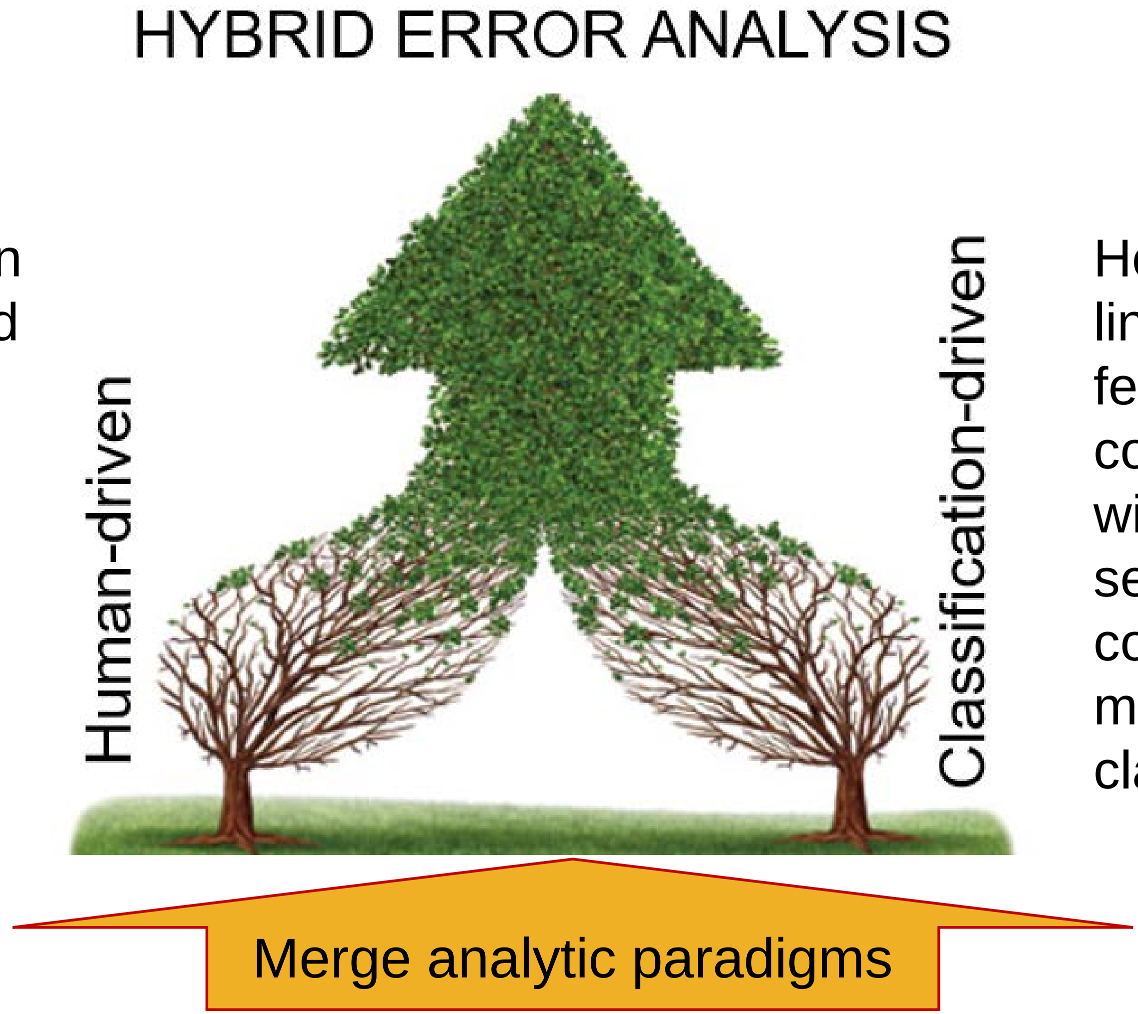


Merge analytic paradigms

Need to understand ARA classification output

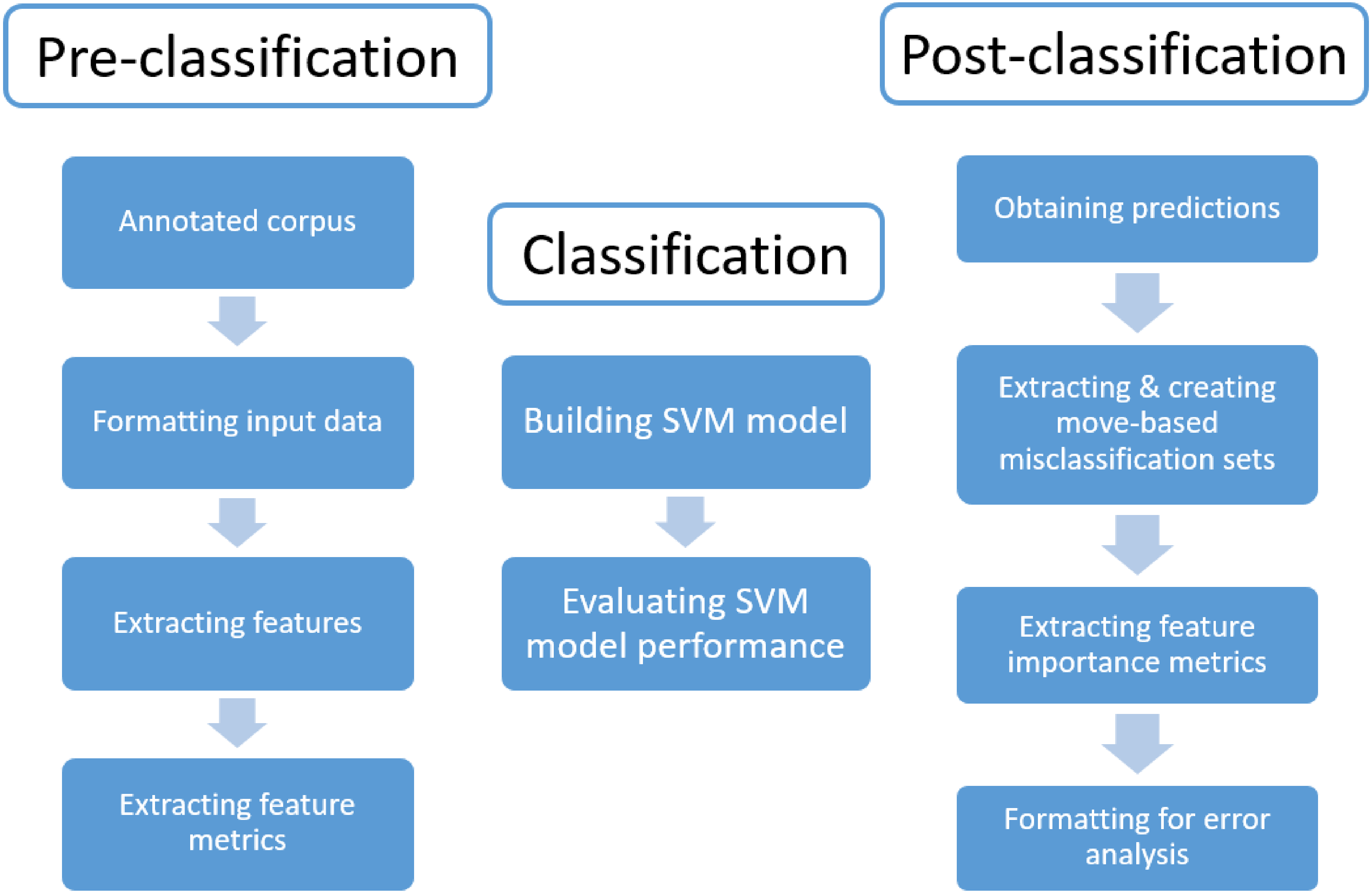
How do human and automated analyses compare?

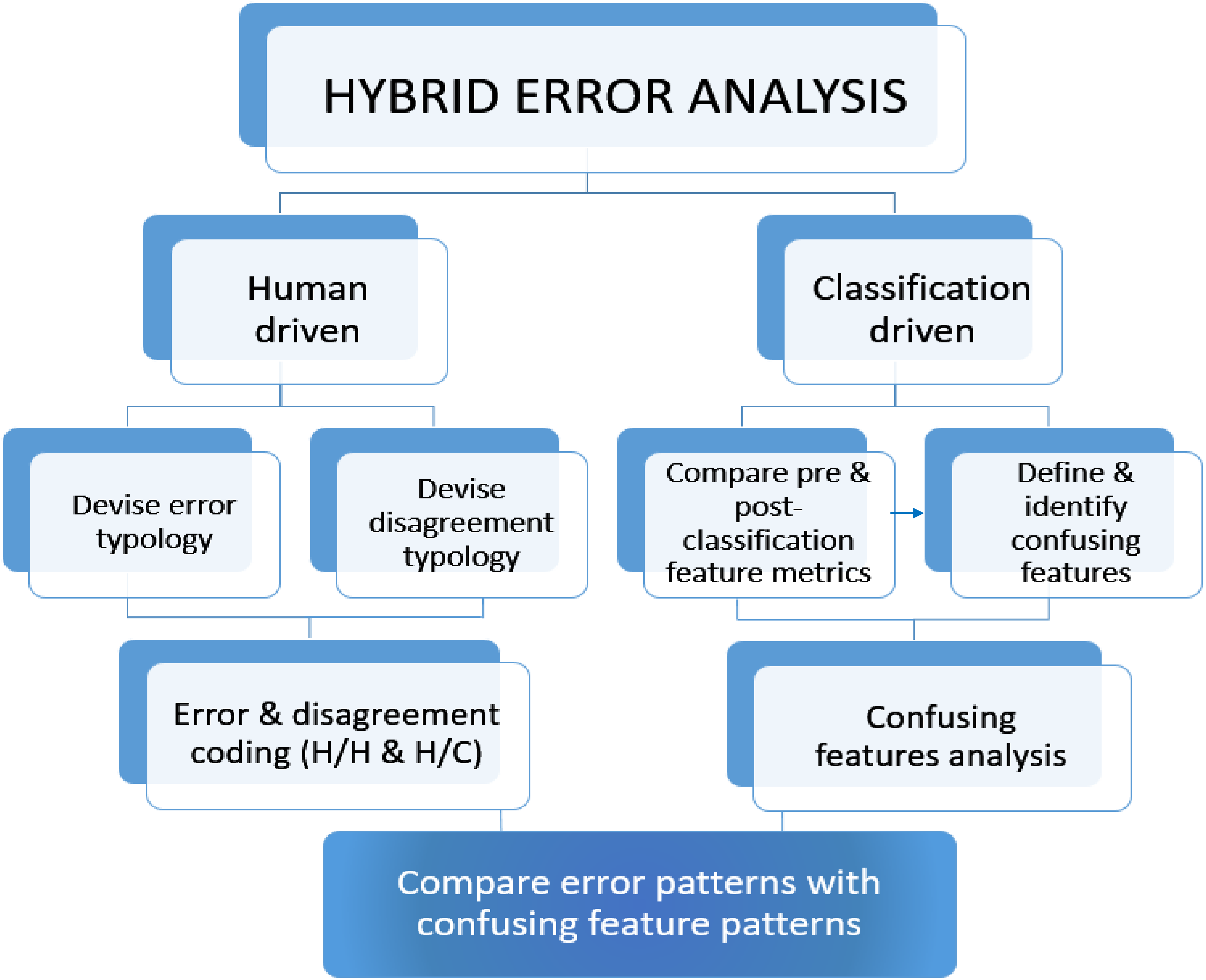
What are the causes of classification errors?



How do the linguistic features contained within a sentence contribute to its move classification?

Prerequisite to error analysis




HYBRID ERROR ANALYSIS
Human driven
Classification driven
Devise error typology
Devise disagreement typology
Compare pre & post-classification feature metrics
Define & identify confusing features
Error & disagreement coding (H/H & H/C)
Confusing features analysis
Compare error patterns with confusing feature patterns

Human-driven

Devising error categories

The error categories based on functional linguistic ‘signals’

→ error categories not necessarily mutually exclusive

- **Missing**: no explicit linguistic signal of the function
- **Unidentified**: a feature indicative of a function is present, but isn’t picked up on
- **Misleading**: a linguistic signal may have a function-related connotation, but doesn’t carry this function in the sentence
- **Ambiguous**: a linguistic signal is indicative of several functions, but the actual function can only be determined from the context
- **Underrepresented**: fewer linguistic signals that are indicative of the actual function than signals that are not
- **Competing**: several linguistic signals indicative of primary and secondary functions in a multi-functional sentence

Human-driven

Devising disagreement categories

Primary and secondary annotations from human coders

Probabilities from classifier

Function-level

- Agreement primary (AP): same primary step
- Agreement secondary (AS): same secondary step
- Disagreement primary (DP): different primary step
- Disagreement secondary (DS): different secondary step
- Flipped agreement (FA): same step, but primary and secondary switched

Overall

- Complete agreement: AP, AP+AS
- Partial agreement: AP+DS, DP+AS, FA
- Complete disagreement: DP, DP+DS

Human-driven

Devising error categories

Devising disagreement categories

Sentence-level coding

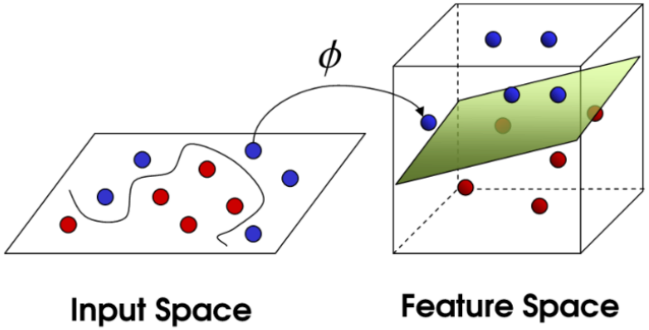
M2 predicted as M1:

However, desegregative busing quickly became broadly unpopular.

				Agreement		
Actual	**Predicted**	**Error 1**	**Error 2**	**Function 1**	**Function 2**	**Overall**
m2, highlighting a problem	m1, providing general background	Unidentified (however, unpopular)	Competing (quickly, broadly)	Disagreement primary	Disagreement secondary (additional step classified)	Complete disagreement

Human-driven

- So far:
  - Understanding of the nature of errors and disagreement
- Explore further:
  - Which error types are most pervasive/serious?
  - How do error and disagreement patterns relate?
  - How can human-identified error patterns help explain misclassifications?



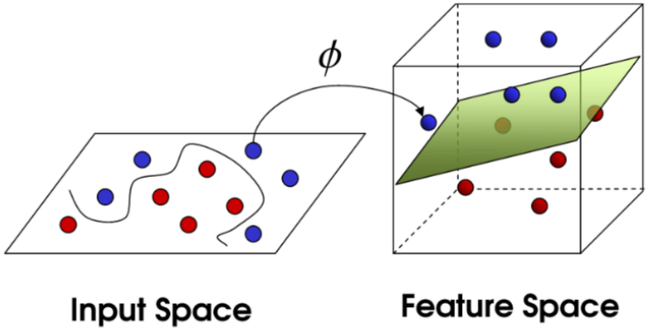
Classification-driven

Comparing pre-post classification metrics per feature & per sentence

M2 predicted as M1:

Airway functional abnormalities, ranging from persistent increases in airway resistance and hyperresponsiveness to asthma, may develop following acute viral infections, especially in young children.

N-gram	Pre-classification		Post-classification
feature	m2_OR	m1_OR	Feature weight
mai	0.251	0.178	-0.817
persist	-0.101	0.322	-0.258
infect	-0.184	0.577	-0.175
especi	0.099	0.339	-0.173
rang	-0.449	0.697	-0.165
resist	-0.210	0.492	-0.161
follow	-1.036	0.305	-0.113
develop	-0.421	0.559	-0.065
to	-0.622	1.285	-0.025
from	-0.544	0.631	0.054
increas	-0.459	0.726	0.179
young	-0.489	0.511	0.243
children	-0.263	0.615	0.325



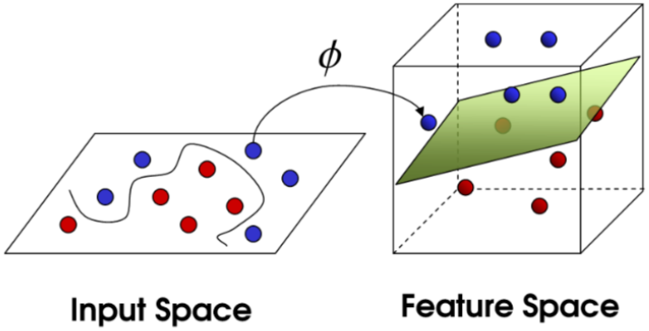
Classification-driven

Defining & identifying confusing features per sentence

Low OR for actual move & high feature weight

→ potentially confusing

N-gram feature	Pre-classification		Post-classification
	m2_OR	m1_OR	Feature weight
surviv	-0.690	0.824	0.648
note	-0.317	0.788	0.636
last	-1.184	0.994	0.627
found	-0.984	1.020	0.619
advantag	-0.743	1.020	0.617
length	-0.765	1.020	0.615
hold	-0.635	1.020	0.613
inclus	-0.372	1.020	0.608
origin	-0.724	1.020	0.586
studi_of_the	-0.177	1.020	0.572
demonstr_that_the	-1.022	1.020	0.552
aspect	-0.354	1.020	0.537
shown	-1.069	1.020	0.533
et_al._1997	-1.064	1.020	0.500
the_us_of	-0.476	1.020	0.493



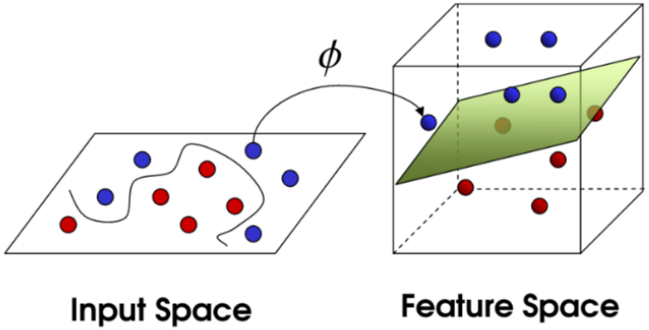
Classification-driven

Defining & identifying confusing features per sentence

High OR for actual move & low feature weight

→ potentially useful

N-gram feature	Pre-classification		Post-classification
	m2_OR	m1_OR	Feature weight
unclear	1.270	-0.666	-1.219
is_difficult_to	1.147	-0.544	-1.043
have_not_been	0.831	-0.302	-1.749
unknown	0.801	-0.239	-0.679
difficult	0.731	-0.294	-1.265
ha_not_been	0.671	-0.154	-0.958
howev_the	0.573	-0.100	-0.607
lack	0.454	0.011	-0.891
challeng	0.451	-0.028	-0.839
complic	0.286	0.186	-0.737
mai	0.251	0.178	-0.817
might	0.211	-0.055	-0.900
fail	0.131	0.330	-0.915
constraint	0.065	0.140	-0.670



Classification-driven

- So far:
  - Features with low/negative log OR (actual class) and high/positive feature weights → ‘confusing’
- Explore further:
  - Would removing ‘confusing’ features from the pre-classification feature set enhance performance?
  - What can be learned from features weighted sum per sentence?
  - How to compare/integrate with human-driven error analysis results ?

Implications

"[W]e have to ask whether genre can be reliably identified by means of computationally tractable cues" (Kessler et al.,1997, p. 1)

- Augmented ARA
  - Knowledge-based approach & human-generated hand-written rules (e.g., Madnani et al., 2012)
  - Feature engineering
  - Ranking of classification decisions based on higher probabilities to distinguish bw primary & secondary functions
- Voting algorithm that would pass final classification decisions considering the output of several independent analyzers (e.g., Burstein et al., 2003)

Research assistants Erin Todey and Ziwei Zhou

Elena Cotos, ecotos@iastate.edu
https://works.bepress.com/elena_cotos
https://cce.grad-college.iastate.edu/about-us/directory/elena-cotos