

Context-aware Argument Mining and Its Application in Education

University of Pittsburgh, April 14, 2017

Huy Nguyen

PhD Dissertation Defense

Committee

Dr. **Diane Litman** (*dissertation advisor*), Computer Science Department

Dr. **Rebecca Hwa**, Computer Science Department

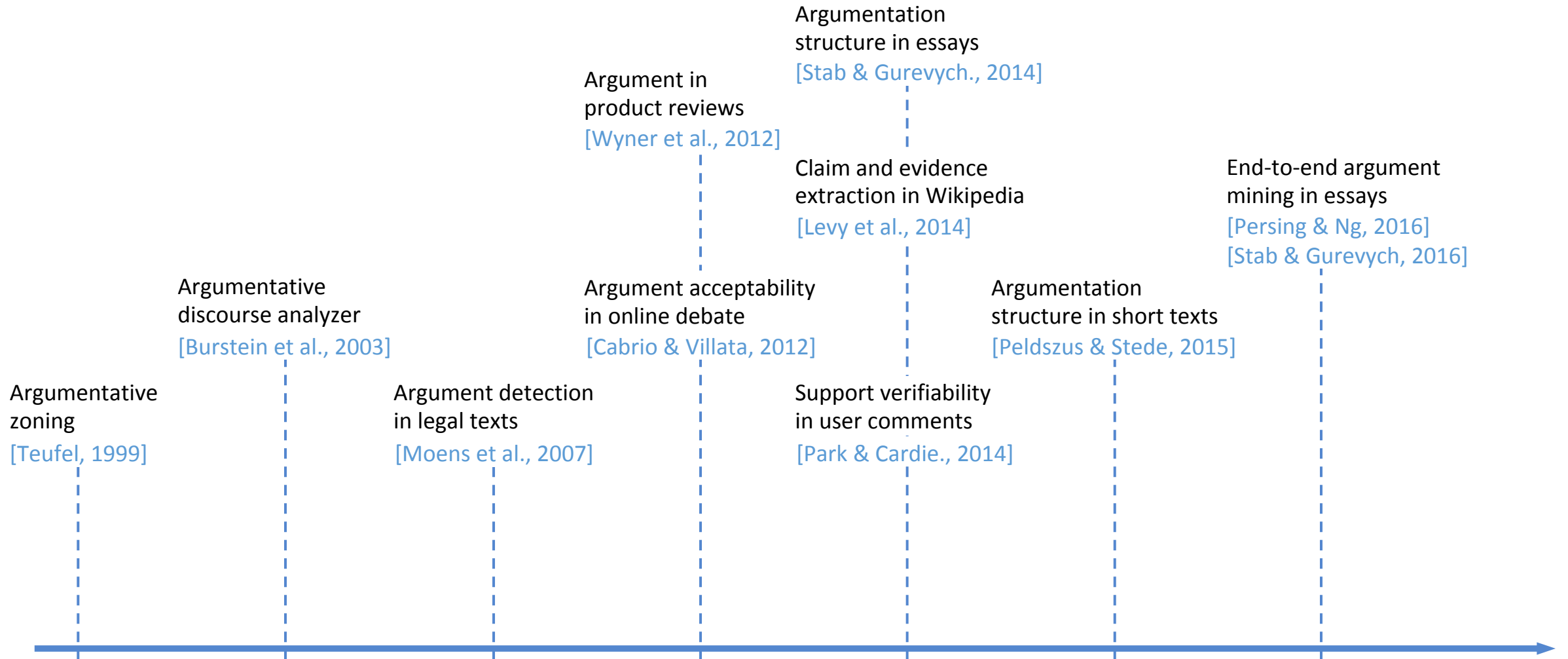
Dr. **Adriana Kovashka**, Computer Science Department

Dr. **Kevin Ashley**, Law School

A Not-very-short Introduction



A brief history of argument mining



Structured vs. abstract argumentation theories

- **Abstract argumentation**
 - **Argument as a primary element** without internal structure
 - Study relation between arguments for argument acceptability
- **Structured argumentation**
 - **Argument components** and their interactions
 - Typically employed in argument mining in texts
 - Models textual representation of arguments
 - Argument component types: premise/evidence, claim/conclusion

What is argument mining?

- “[...] the automatic discovery of an argumentative text portion, and the identification of the relevant components of the argument presented there.”

[Peldszus & Stede, 2013]

- An argument consists of a non-empty set of premises supporting a conclusion
- Argument component: argumentative discourse unit (ADU)
 - E.g., text segment, sentence, clause

Argument mining tasks

- **Argument component identification**
 - Separates argumentative from non-argumentative text units
 - Recognizes the boundaries of argument components

⁽¹⁾[Taking care of thousands of citizens who suffer from disease or illiteracy is more urgent and pragmatic than building theaters or sports stadiums]. ⁽²⁾As a matter of fact, [an uneducated person may barely appreciate musicals], whereas [a physical damaged person, resulting from the lack of medical treatment, may no longer participate in any sports games]. ⁽³⁾Therefore, [providing education and medical care is more essential and prioritized to the government].

Argument mining tasks (2)

- Argument component classification

- Labels argument components with their argumentative roles
- E.g., premise, claim

(1)[Taking care of thousands of citizens who suffer from disease or illiteracy is more urgent and pragmatic than building theaters or sports stadiums]**Claim**. (2)As a matter of fact, [an uneducated person may barely appreciate musicals]**Premise**, whereas [a physical damaged person, resulting from the lack of medical treatment, may no longer participate in any sports games]**Premise**. (3)Therefore, [providing education and medical care is more essential and prioritized to the government]**Claim**.

Argument mining tasks (3)

- Argumentative relation classification

- Recognizes if two argument components are argumentatively related or not
- Identifies argumentative functions of the relations, e.g., support, attack

(1)[Taking care of thousands of citizens who suffer from disease or illiteracy is more urgent and pragmatic than building theaters or sports stadiums]**Claim**. (2)As a matter of fact, [an uneducated person may barely appreciate musicals]**Premise**, whereas [a physical damaged person, resulting from the lack of medical treatment, may no longer participate in any sports games]**Premise**. (3)Therefore, [providing education and medical care is more essential and prioritized to the government]**Claim**.



Premise(2.1) supports Claim(1)
Premise(2.1) supports Claim(3)
Premise(2.2) supports Claim(1)
Premise(2.2) supports Claim(3)
Claim(3) supports Claim(1)

Argument mining tasks (4)

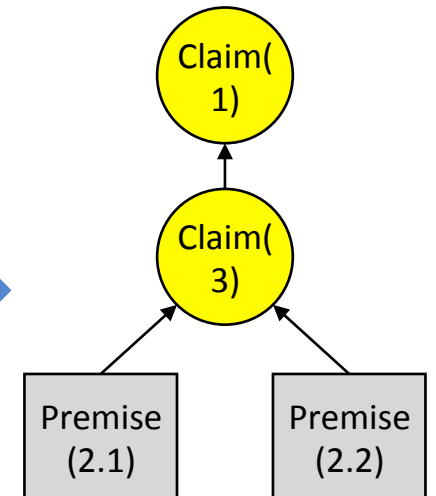
- Argumentation structure identification

- Constructs the graphical representation (i.e., tree) of arguments in which edges are **direct attachments** between argument components

(1)[Taking care of thousands of citizens who suffer from disease or illiteracy is more urgent and pragmatic than building theaters or sports stadiums]**Claim**. (2)As a matter of fact, [an uneducated person may barely appreciate musicals]**Premise**, whereas [a physical damaged person, resulting from the lack of medical treatment, may no longer participate in any sports games]**Premise**. (3)Therefore, [providing education and medical care is more essential and prioritized to the government]**Claim**.



Premise(2.1) supports Claim(1)
Premise(2.1) supports Claim(3)
Premise(2.2) supports Claim(1)
Premise(2.2) supports Claim(3)
Claim(3) supports Claim(1)



Context is crucial for resolving ambiguity

Example 1: **It** helps relieve tension and stress... **Exercising** improves self esteem and confidence.

Coreference resolution

Example 2: People who are addicted to games, especially **online games**, can eventually bear dangerous consequences...
Although it is undeniable that **computer** is a crucial part of human life, **it** still has its bad side

Topic context

Context-aware

Example 3: Firstly, **pictures can influence** the way people think. For example, nowadays horrendous **images** are displayed on the cigarette boxes to illustrate the consequences of smoking. As a result, statistics show a slight **reduction** in the number of smokers, indicating that they **realize** the effects of the negative habit.

Local context

Contextual features in prior studies

- Features extracted from surrounding sentences
 - Words, POS
 - Prediction labels of preceding/following components
 - Cosine similarity with the topic sentence
- Processed textual input isolatedly
 - Component classification: sentences, clauses
 - Relation classification: pairs of sentences and/or clauses
 - Did not investigate semantic relations between context sentences

⁽¹⁾Firstly, pictures can influence the way people think.

⁽²⁾For example, nowadays horrendous images are displayed on the cigarette boxes to illustrate the consequences of smoking.

⁽³⁾As a result, statistics show a slight reduction in the number of smokers, indicating that they realize the effects of the negative habit.

Context-aware argument mining

- Writing topics/prompts as an supervision to sublanguage identification
 - Argument and domain word extraction
 - Topic context (global)
- Surrounding text as a context-rich representation of the argument component
 - Window context (local)
- **Hypothesis**
 - *Argument mining can be improved w.r.t prediction performance by considering contextual information at both local and global levels when developing prediction features*

⁽¹⁾Firstly, pictures can influence the way people think.

⁽²⁾For example, nowadays horrendous images are displayed on the cigarette boxes to illustrate the consequences of smoking.

⁽³⁾As a result, statistics show a slight reduction in the number of smokers, indicating that they realize the effects of the negative habit.

Application in education

- Argumentation and argumentative writing receive increasing attention
 - Key focuses of Common Core Standard
 - Standardized tests, academic content courses
- Existing systems only considers grammar, vocabulary, mechanics, discourse structure
 - A demand for “argumentation-aware” automated writing evaluation systems [\[Beigman Klebanov et al., 2016\]](#)
- Emerging attention to evaluating argument aspect of essays [\[Persing & Ng, 2013\]](#)
 - Thesis clarity, evidence use , critical question, argument strength [\[Rahimi & Litman, 2014\]](#)
[\[Song et al., 2014\]](#)
[\[Persing & Ng, 2015\]](#)

Persuasive essay writing rubrics

- TOEFL iBT Independent Writing
 - “is well organized and well developed, using clearly **appropriate explanations, exemplifications and/or details**”
- Kaggle ASAP: automated student assessment prize
 - “Has fully **elaborated reasons with specific details**”
- Research Methods classes at Pitt
 - “Brief high-level overview of study design and **clear statement of hypotheses?**
Appropriate integration of **conflicting research findings** into a **convincing argument** for at least one hypothesis?”

- ...

Argument mining offers new capabilities that consider argumentation aspect

- Derive features from output of argument mining models
- Augment automated essay scoring systems
- Enable automated writing feedback

Research hypotheses

- **H1-1**: Proposed topic-context features improves argument component identification
- **H1-2**: Proposed topic-context and segment-context features improves argumentative relation classification
- **H2**: Prediction output of our argument mining models help improve automated argumentative essay scoring

Corpora

- Argument mining

- Persuasive essays: practice writings by ESL learners
- Academic essays: APA-style writings by college students

[Stab & Gurevych, 2014; 2016]

[Barstow et al., 2015]

- Automated essay scoring

- TOEFL11: real-test essays by ESL learners
- Kaggle ASAP essays: by 7 – 10 grade students

[Blanchard et al., 2013]

[Automated Student Assessment Prize, <https://www.kaggle.com/c/asap-aes>]

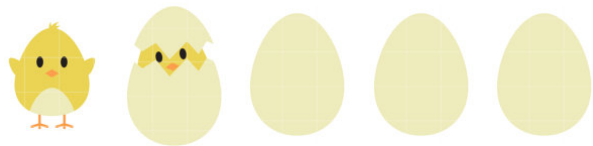
- Corpora are widely different

- Data sizes, annotation schemes
- Writer expertise, writing fluency & quality

Research outline

- Context-aware argument component classification
 - Argument and domain word extraction
- Context-aware argumentative relation mining
 - Context-window heuristics: window-size vs. text segmentation
- Argumentation features for improving automated essay scoring
 - Extrinsic evaluation of argument mining systems
 - Cross-domain AES
- Discussions and future work

Context-aware argument component classification



Introduction

- Problem statement
 - Given argument components (or argumentative sentences) as inputs, recognize their argumentative roles.
 - Assuming argument component boundaries are provided
- Data
 - Persuasive1 corpus: 90 persuasive essays by ESL learners [Stab & Gurevych, 2014]
 - Academic corpus: 150 academic writing from College Psychology classes in 2014 [Barstow et al., 2015]
- Models
 - Baseline: Stab & Gurevych, EMNLP 2014
 - Proposed: Nguyen & Litman, ARGMINING 2015; FLAIRS 2016

Data summary

- Persuasive1 corpus: 90 essays

Argumentative label	#instances	
<i>MajorClaim</i>	90	(4.8%)
<i>Claim</i>	429	(22.8%)
<i>Premise</i>	1033	(55.0%)
Non-argumentative	327	(17.4%)
Total	1879	(100%)

Krippendorff's $\alpha_U = 0.72$

My view is that the *government should give priorities to invest more money on the basic social welfares such as education and housing instead of subsidizing arts relative programs*. **MajorClaim**

Art is not the key determination of quality of life, but education is. **Claim**
 In order to make people better off, it is more urgent for governments to commit money to some fundamental help such as setting more scholarships in education section for all citizens. **Premise**

- Academic corpus: 115 essays

Argumentative label	#sentences	
<i>Hypothesis</i>	185	(5.6%)
<i>Finding</i>	131	(4.0%)
Non-argumentative	2998	(90.4%)
Total	3314	(100%)

Cohen's kappa = 0.79

(2)Although these studies demonstrate the bystander effect and diffusion of responsibility, other studies oppose these ideas. (3)One strong study that opposes the bystander effect was done in 1980 by Junji Harada that showed that increase in group size, even in a face to face proximity, did not decrease the likelihood of being helped (Harada, 1980). **Finding(Opposition)** ... (4)The hypothesis, based on the bystander effect demonstrated in Wegner's study (1978), is that with more people around, less people will take the time to help the girl pick up her papers (Wegner, 1978). **Hypothesis**

Baseline model (Stab14)

- Lexical features
 - 1-, 2-, 3-grams
 - Verbs, adverbs, presence of modal verb
 - Discourse connectives, singular first person pronouns
- Syntactic features:
 - Production rules, e.g., VP \rightarrow VBG NP
 - Tense of main verb
 - Number of subclauses, depth of parse tree
- Structural features
 - Numbers of tokens, punctuations
 - Sentence and paragraph positions
- Contextual features:
 - Preceding and following sentences
 - Numbers of tokens, punctuations, subclauses, and presence of modal verb

Limitation of prior studies

- **Large and sparse feature space** by n-grams and production rules
 - Model argumentative discourse, but have much noise
 - Feature selection helps, but still not efficient
- Indicator features are effective but have **limited coverage**
 - Stab & Gurevych used 55 discourse connectives
- **Sublanguage identification has not been applied** to argument mining
 - Separation of organizational content (shell) from topical content
 - Offer a better possibility to model argumentative discourse
 - Argument words vs. domain words

[Stab & Gurevych, 2014]

[Mochales & Moens, 2008]

[Madnani et al., 2012]

[Du et al., 2014]

[Seaghdha & Teufel, 2014]

Sublanguage identification in argumentative texts

- Shell language vs. content

[Madnani et al., 2012]

- Shell: sequence of words providing organization framework for an argument
- Supervised sequence model to determine shell boundaries

Example: **The argument states that** based on the result of the recent research, there probably were grizzly bears in Labrador... **There is a possibility that** they were a third kind of bear apart from black and grizzly bears.

- Rhetorical language model

[Seaghdha & Teufel, 2014]

- Word probabilities in a document-specific topic model or a rhetorical language model
- Unsupervised probabilistic topic model based on LDA

Example: Many algorithms that **compare** protein structures **can reveal** similarities that **suggest** related biological functions, even at great evolutionary distances. Proteins with related function **often** exhibit **differences in** binding specificity, but few algorithms identify structural variations that **effect** specificity.

Proposed argument and domain word extraction

[Nguyen & Litman, 2015]

- **Argument words**: words that signal the argumentative content, and are commonly used across different argument topics
 - E.g., *believe, view, should*
- **Domain words**: specific terminologies commonly used within the topic domain
 - E.g., *education, art*

Example: My **view is that the** government **should give** priorities to invest **more money on the** basic social welfares **such as** education and housing **instead of** subsidizing arts relative programs.

- Enable novel features for argument mining models

Extraction algorithm

- Requires a large essay set with writing prompts
- Starts with argument keyword set
 - Manually pre-selected
 - **Domain seed words** = prompt words – **argument keywords** – stop words
 - In-prompt frequency of domain seed words
- Post-processes LDA output
 - LDA topics approximate writing topics **but not completely**
 - **Identify** LDA of **argument words** and **maximize** its difference from **other LDA topics**
 - Weight of LDA topic = #argument keywords – Σ frequencies of domain seed words
 - Argument word list is the LDA topic with the largest weight

Development data for AD word extraction

- Persuasive development set
 - 6794 unannotated essays from essayforum.com
 - 10 argument seed words
 - Most frequent in prompts
 - agree, disagree, reason, support, advantage, disadvantage, think, conclusion, result, opinion
 - 3077 domain seed words
- Output (best $k = 36$)
 - 263 argument words
 - 1806 domain words
- Academic development set
 - 254 unannotated essays from classes in 2011 and 2013
 - 5 argument seed words
 - Specified in writing assignment
 - hypothesis, support, opposition, finding, study
 - 264 domain seeds
- Output (best $k = 11$)
 - 315 argument words
 - 1582 domain words

Example argument and domain words

Persuasive Set

Topic 1: reason example support agree think because disagree statement opinion believe therefore idea conclusion ...

Topic 2: city live big house place area small building apart town community factory urban ...

Topic 3: children parent school education teach kid adult grow childhood behavior taught ...

Academic Set

Topic 1: study research observe result hypothesis time finding however predict support expect opposition ...

Topic 2: response stranger group greet confederate individual verbal social size people sneeze ...

Topic 3: more gender women polite female male men behavior differ prosocial express gratitude ...

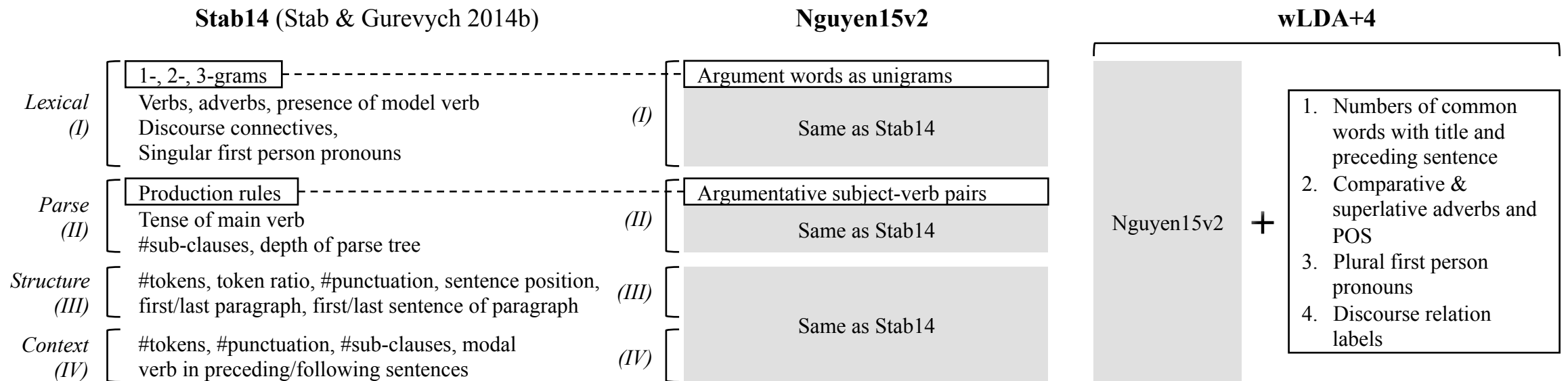
Only top words for each LDA topic are shown

Topic 1 is the argument word list

Proposed models

- [Nguyen15v2](#): Nguyen & Litman (2015) without argument and domain word counts
 - Replaces generic n-gram and production rules with AD word features
- [wLDA+4](#): Nguyen & Litman (2016a)
 - Introduces new features to model argument indicators and abstract over writing topics
- Ablated models
 - Evaluate the contribution of AD word extraction
 - [woLDA](#)
 - Disables argument/domain word-based features in wLDA+4
 - [Seed](#)
 - Replaces extracted AD words with argument keywords and domain seed words

Model summary



Experiment results

Stab14 performs worse than our two proposed models Nguyen15v2 and wLDA+4

Two ablated models are significantly worse than wLDA+4

- 10-fold cross validation
 - SVM learning algorithm + Top 100 features ranked by InfoGain algorithm

Persuasive1 Corpus

	Stab14	Nguyen15v2	woLDA	Seed	wLDA+4
Accuracy	0.787*	0.792*	0.780*	0.781*	0.805
Kappa	0.639*	0.649*	0.629*	0.632*	0.673
Precision	0.741*	0.745*	0.746*	0.740*	0.763
Recall	0.694*	0.698*	0.695*	0.695*	0.720

Academic Corpus

	Stab14	Nguyen15v2	woLDA	Seed	wLDA+4
Accuracy	0.934*	0.942+	0.933*	0.935*	0.941
Kappa	0.558*	0.635	0.528*	0.564*	0.629
Precision	0.804*	0.830+	0.829	0.826	0.825
Recall	0.628*	0.695	0.594*	0.637*	0.695

*: $p < 0.05$, +: $p < 0.1$ in comparison with wLDA+4

Experiment results (2)

- Cross-topic validation

Nguyen15v2 and wLDA+4 significantly outperform Stab14 and two ablated models



Persuasive1 Corpus
12 groups

	Stab14	Nguyen15v2	woLDA	Seed	wLDA+4
Accuracy	0.780*	0.796	0.774*	0.776*	0.807
Kappa	0.623*	0.654+	0.618*	0.623*	0.675
Precision	0.722*	0.757*	0.751	0.734	0.771
Recall	0.670*	0.695*	0.681*	0.686*	0.722

Academic Corpus
5 groups

	Stab14	Nguyen15v2	woLDA	Seed	wLDA+4
Accuracy	0.928*	0.939+	0.931*	0.935*	0.944
Kappa	0.491*	0.598+	0.474*	0.547*	0.630
Precision	0.768	0.832	0.866	0.839*	0.851
Recall	0.565*	0.664	0.551*	0.617*	0.686

Summary

- A novel semi-supervised algorithm to extract argument and domain words from argumentative essays
- Propose to use extracted argument and domain words as features and constraints
 - Efficiently replace generic n-grams and production rules for better performance
- New features to model argument indicators and abstract over writing topics
 - Improve cross-topic performance
- The results prove strongly our first hypothesis H1-1
 - Proposed topic-context features improve argument component classification

Context-aware argumentative relation mining



Introduction

- Problem statement
 - Given a **pair of source and target argument components**, determine whether a relation holds from the source to the target and **classify the argumentative function of the relation**
- Data
 - Persuasive1 corpus
 - Does not consider cross-paragraph pairs
 - Pairs are classified as Support vs. Not-support
 - Academic corpus
 - Relations are annotated regardless paragraph boundaries
 - Pairs are classified as Support vs. Opposition vs. None
- Models
 - Baseline: Stab14
 - Proposed: Nguyen & Litman, ACL 2016

Data summary

- Persuasive1 corpus

Label	#pairs	
<i>Support</i>	989	(16%)
<i>Not-support</i>	5341	(84%)
<i>Attack</i>	103	(2%)
<i>No-relation</i>	5238	(82%)
Total	6330	(100%)

Krippendorff's $\alpha = 0.81$

- 6330 ordered pairs

- Support vs. Not-support
- Source and target are in the same paragraph

- Academic corpus

Label	#pairs	
<i>Support</i>	50	(6%)
<i>Opposition</i>	82	(10%)
<i>No-relation</i>	702	(84%)
Total	834	(100%)

Cohen's kappa = 0.67

- 834 ordered pairs

- Support vs. Opposition vs. No-relation
- No paragraph constraint

Baseline model (Stab14)

- Lexical features
 - Pairs of words (from source and target), pair of first words
 - Number of words in common, presence of modal verb
- Syntactic features:
 - Production rules, e.g., VP → VBG NP
- Structural features
 - Numbers of words in source and target, word count difference
 - Sentence positions of source and target, position difference
 - Whether source and target are first or last sentences of paragraph
 - Does target occurs before source
- Indicator features
 - If source and target starts with a discourse connective
- Predicted type features
 - Predicted argumentative labels of source and target components

Target

Sentence ...

Sentence 1: Firstly, **picture can influence the way people think.**

Sentence 2: For example, **nowadays horrendous images are displayed on the cigarette boxes to illustrate the consequences of smoking.**

Source

Sentence 3: As a result, **statistics show a slight reduction in the number of smokers, indicating that they realize the effects of the negative habit.**

Sentence ...

Limitation of prior studies

- Depended on **heavy features**, e.g., word pairs, production rules
 - To be replaced by our proposed features
- **Topic-based features were not yet explored**
 - Widely used for argument component classification
- **Contextual features were used limitedly**
 - Words and POS in adjacent sentences
- **Single-sentence inputs** limit the use of semantic relation features
 - Semantic similarity and textual entailment between sentence sets
 - Discourse relations between the input and its surrounding text

[Levy et al., 2014]

[Nguyen & Litman, 2016]

[Peldzus & Stede, 2015]

[Brian & Rambow, 2011]

[Cabrio & Villata, 2012]

[Boltuzic & Snajder, 2014]

Proposed contextual features

- Topic-context features (global context)
 - Exploit argument and domain word lexicons
 - Model topically-related words
 - Domain words that share LDA topic(s)
- Window-context features (local context)
 - Group argument component with its adjacent sentences
 - Model content-relatedness by semantic relations
 - Discourse relations
 - Semantic textual similarity
 - Textual entailment

People who are addicted to games, especially **online games**, can eventually bear dangerous consequences...

Although it is undeniable that **computer** is a crucial part of human life, **it** still has its bad side

(1)Firstly, pictures can influence the way people think. (2)For example, nowadays horrendous images are displayed on the cigarette boxes to illustrate the consequences of smoking. (3)**As a result**, statistics show a slight reduction in the number of smokers, indicating that they realize the effects of the negative habit.

Proposed topic-context features

[Nguyen & Litman, 2016b]

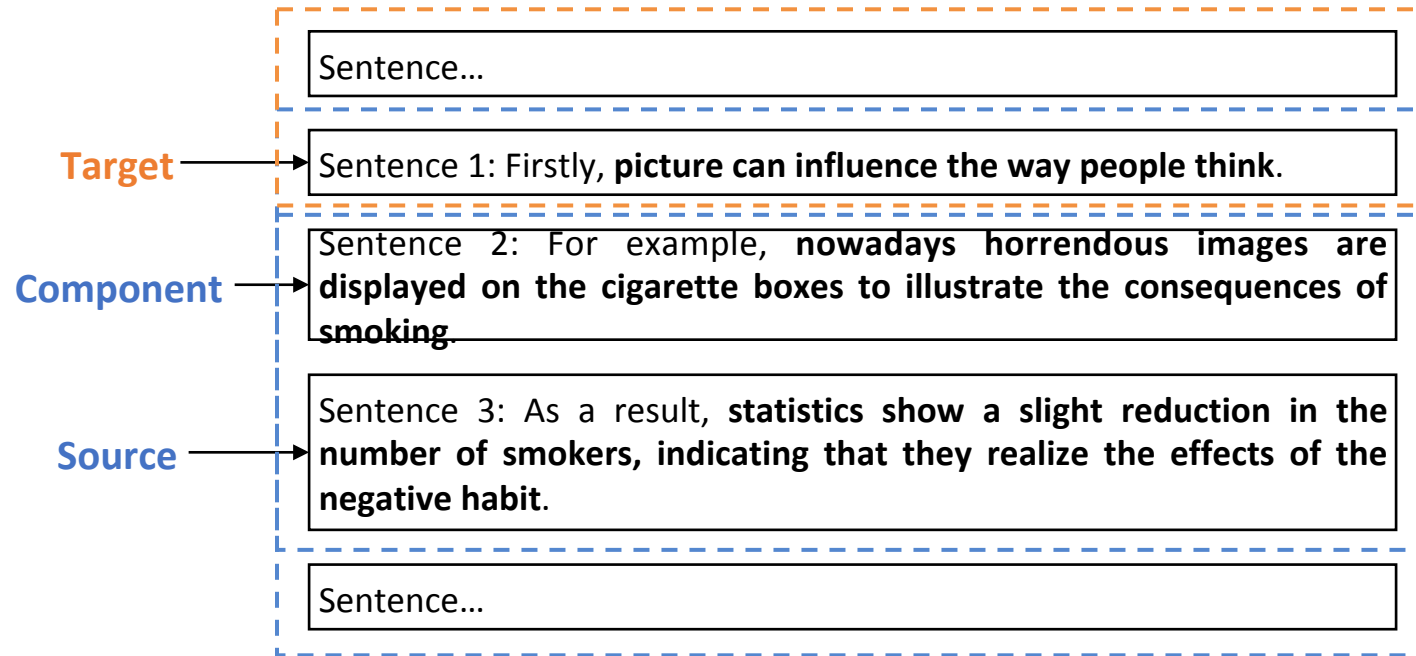
- Argument word features:
 - Argument words in source and target, pairs of argument words
 - Number of argument words in common, argument word count difference
- Domain word features:
 - Number of domain words in common, domain word count difference
 - Number of domain word pairs that share LDA topic(s), pairs that share no LDA topic
- Dependency parse features:
 - MainVerb-Subject dependency triples, e.g., *nsubj(believe, I)*

Proposed context-window framework

[Nguyen & Litman, 2016b]

- Definition: *Context-window of argument component*
 - Text segment formed by neighboring sentences and the component itself
 - The neighboring sentences (context sentences) must be in the same paragraph
- Context-window formation
 - Window-size heuristic
 - With half-size = n , form a window of size $2n$ at most
 - Text segmentation heuristic
 - Context window consists of sentences of the same paragraph and segment output of a text segment program
- Overlap resolution
 - Overlapping context sentences are assigned to source window

Context-window example



Context-window features

[Nguyen & Litman, 2016b]

- Word count
 - Number of words in common of covering sentences of source and target with preceding and following context sentences
- Discourse relation
 - Discourse relations between context sentences, and within covering sentences of source and target
 - Discourse relation between each pair of source and target sentences
- Discourse marker
 - Whether a discourse marker is present before the covering sentence or not

- Discourse features are enabled by PDTB and RST parsers

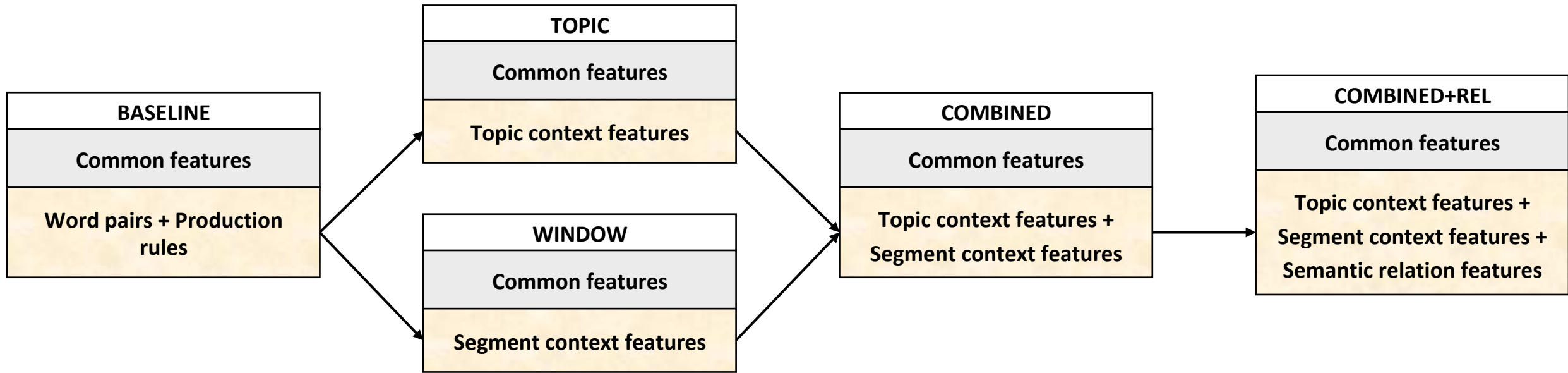
[Ji & Eisenstein, 2014]

[Wang & Lan 2015]

Proposed semantic relation features

- Semantic textual similarity (STS) features
 - STS score between source and target sentences
 - Max STS score among pairs of source and target context sentences
 - STS score between each source context sentence and target sentence
 - STS score between each target content sentence and source sentence
- Textual entailment (TE) features
 - TE score between source and target sentences
 - TE score between source context window and target sentence
 - Max TE score among pairs of source context sentence and target sentence
- Context-window enables aggregating score features for better performance

Model summary



- **Common features** (features in common among models)
 - Baseline features except word pairs and production rules
- **TOPIC, WINDOW** and **COMBINED**
 - To evaluate Topic-context and Window-context features in isolation and combination
- **FULL model** takes all features together

Experiment results in Persuasive1 corpus

[Stab & Gurevych, 2014]

- Data split: 80% training and 20% test
 - Compare with reported results in the prior study
- Window-size heuristic with best half-size = 3
 - Determined through cross-validation in training set

Combining Topic-context and Window-context features yields the best results

CHECKED

	Reported	BASELINE	TOPIC	WINDOW	COMBINED (half-size = 3)
Accuracy	<u>0.863</u>	0.869	<u>0.857</u>	<u>0.857</u>	0.870
F1	0.722	0.722	<u>0.703</u>	0.724	0.753*
Precision	<u>0.739</u>	0.758	<u>0.728</u>	<u>0.729</u>	0.754
Recall	0.705	0.699	<u>0.685</u>	0.720	0.752*
F1:Support	0.519	0.519	<u>0.488</u>	0.533	0.583*
F1:Not-support	<u>0.920</u>	0.925	<u>0.917</u>	<u>0.916</u>	<u>0.923</u>

*: $p < 0.05$ in comparison with Baseline. Values smaller than baseline are underlined. Best values are in boldface.

Experiment results in Academic corpus

- 10x5-fold cross validation
- Window size heuristic with half-size = 1
 - No half-size optimization due to small data

All 3 proposed models significantly outperform BASELINE **CHECKED**

	BASELINE	TOPIC	WINDOW	COMBINED (half-size = 1)
Accuracy	0.828	<u>0.823*</u>	<u>0.819*</u>	0.829
F1	0.493	0.540*	0.521*	0.553*
Precision	0.529	0.560*	0.536*	0.575*
Recall	0.472	0.525*	0.510*	0.536*
F1:Support	0.260	0.399*	0.300*	0.405*
F1:Opposition	0.307	0.317	0.360*	0.344*
F1:No-relation	0.912	<u>0.904*</u>	<u>0.904*</u>	<u>0.909*</u>

Model performance with text segmentation heuristic

- Compare COMBINED's performance with two heuristics for context-window
 - 10x5-fold cross validation
 - Text segmentation: Bayesian Topic Segmentation algorithm
 - Window-size heuristic with best half-size n
 - $n = 3$ for Persuasive1 corpus
 - $n = 8$ for Academic corpus

[Eisenstein & Barzilay, 2008]

• Text seg. outperforms the best half-size in Persuasive1

• but performs worse than half-size in Academic

CHECKED

	Persuasive1 corpus		Academic corpus	
	COMBINED Half-size = 3	COMBINED Text seg.	COMBINED Half-size = 8	COMBINED Text seg.
Accuracy	0.871	0.873*	0.836*	0.829
F1	0.743	0.746*	0.571*	0.545
Precision	0.758	0.762*	0.590*	0.567
Recall	0.731	0.734*	0.556*	0.530

*: $p < 0.05$

Model performance with semantic relation features

- 10x5-fold cross validation

Semantic relation features help improve performance with different base models

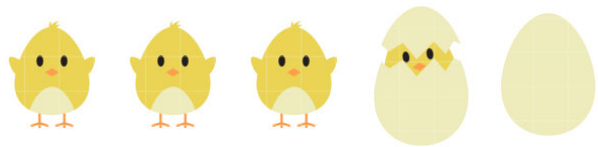
CHECKED

	Persuasive1 corpus				Academic corpus			
	Half-size = 3		Text segmentation		Half-size = 8		Text segmentation	
	COMBINED	COMBINED + REL	COMBINED	COMBINED + REL	COMBINED	COMBINED + REL	COMBINED	COMBINED + REL
Accuracy	0.871	0.872	0.873	0.873	0.836	0.837	0.829	0.833*
F1	0.743	0.745	0.746	0.747	0.571	0.573	0.545	0.556*
Precision	0.758	0.759	0.762	0.762	0.590	0.593	0.567	0.578*
Recall	0.731	0.733	0.734	0.735	0.556	0.558	0.530	0.540*

Summary

- Proposed context-aware argumentative relation mining
 - Makes use of contextual features from topic and context window
 - Context windows enables aggregated semantic relation features for better performance
- Proposed two heuristics for forming context windows
 - Window-size is simple but needs size optimization
 - Text segmentation shows promising result when it outperform window-size in persuasive essays
- The results prove strongly our second hypothesis H1-2
 - Proposed topic- and window-context features improve argumentative relation classification

Argumentation features for improving automated essay scoring



Introduction

- Application of Argument Mining (AM) in Automated Essay Scoring (AES)
 - End-to-end AM
 - Impact of AM accuracy to AES performance
 - Argumentation features for improving AES performance
- AES data
 - TOEFL11: real-test essays by ESL learners [Blanchard et al., 2013]
 - Kaggle ASAP essays: by 7 – 10 grade students [Automated Student Assessment Prize, <https://www.kaggle.com/c/asap-aes>]
- Argument mining models
 - Base models: Stab & Gurevych, 2014
 - Our proposed models

Prior studies

- Argumentation features improve AES
 - Input of AM is gold-standard argument components
 - Simplified argumentative relation classification
 - Word-count baseline
- Argumentation features from output of end-to-end AM
 - Insignificant improvement to a word-count baseline
- Sequences of argument components as features
 - Significantly improved an ngram baseline
 - Trait score: organization, argument strength

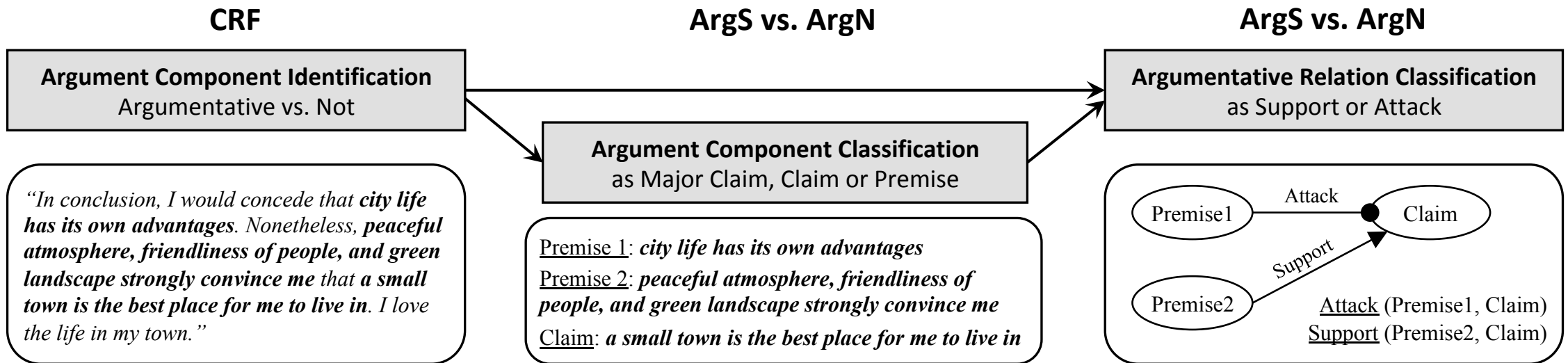
[Ghosh et al., 2016]

[Beigman Klebanov et al., 2016]

[Wachsmuth et al., 2016]

Argument mining pipeline

- ArgS: Stab & Gurevych 2014
- ArgN: our proposed context aware AM
- Stacked with a CRF model for argument component identification



Argument component identification

- Sentences are segmented into components
 - Ex: “It’s true that technology and computers do make their jobs easier but it cannot definitely replace them.”

It	's	true	that	technology	and	computers	do	make	their	jobs	easier	but	it	cannot	definitely	replace	them	.
B	I	I	I	I	I	I	I	I	I	I	I	O	B	I	I	I	I	O

- CRF model
 - Structural: token position, punctuation, sentence position
 - Syntactic: POS, lowest common ancestor, lexico-syntactic
 - Conditional probability
 - Argument/Domain words

[Stab & Gurevych, 2016]

[Proposed in our study]

Training data for AM pipelines

- ArgS was trained follow the original implementation
 - Persuasive1 corpus with 90 essays
- ArgN was trained with the new persuasive corpus
 - Using the training set of 322 essays
- CRF model was trained with the 322-essay set

[Stab & Gurevych, 2014]

[Stab & Gurevych, 2016]

Persuasive1 corpus

Class label	#instances	
<i>MajorClaim</i>	90	(6%)
<i>Claim</i>	429	(28%)
<i>Premise</i>	1033	(66%)
<i>Support</i>	989	(16%)
<i>Not-support</i>	5341	(84%)
Essays	90	

Persuasive2 corpus, training set

Class label	#instances	
<i>MajorClaim</i>	598	(12%)
<i>Claim</i>	1202	(25%)
<i>Premise</i>	3023	(53%)
<i>Support</i>	2846	(16%)
<i>Not-support</i>	14404	(84%)
Essays	322	

AES data

- TOEFL11
 - TOEFL test essays, written by ESL learners
 - Essay grades were categorized: A, B, C
 - TE107 set: 107 essays of two prompts
 - Annotated for AM
- Persuasive essays of Kaggle ASAP data
 - Prompt 1: 1783 essays about **good vs. bad effects of computers**
 - Prompt 2: 1800 essays about **ensorship in libraries**


[Blanchard et al., 2013]

[Ghosh et al., 2016]

Study 1: AM accuracy

- AM pipelines are tested on TE107 data

- ArgN significantly outperforms ArgS
- AM performance is greatly affected by the segmentation accuracy



TE107 data

Class label	#instances	
<i>MajorClaim</i>	105	(9%)
<i>Claim</i>	468	(40%)
<i>Premise</i>	603	(51%)
<i>Support</i>	507	(11%)
<i>Not-support</i>	4179	(89%)

Test performance, input are true argument components. *: $p < 0.05$

AM Pipeline	F1:MajorClaim	F1:Claim	F1:Premise	F1:Support	F1:Not-support
<i>ArgS</i>	0.453	0.295	0.710	0.148	0.917
<i>ArgN</i>	0.622*	0.508*	0.751*	0.211*	0.915

Argument component identification

F1	Precision	Recall	F1:B	F1:I	F1:O
0.578	0.575	0.591	0.436	0.757	0.540

Test performance, input are predicted argument components. *: $p < 0.05$

AM Pipeline	F1:MajorClaim	F1:Claim	F1:Premise	F1:Support	F1:Not-support
<i>ArgS</i>	0.078	0.226	0.343	0.088	0.962*
<i>ArgN</i>	0.156*	0.258*	0.404*	0.126*	0.947

Study 2: Impact of AM accuracy to AES performance

- Extrinsically evaluate ArgS and ArgN in a AES tasks
 - Te107 data
 - 10x10-fold cross-validation with Logistic Regression
 - Report quadratic-weighted-kappa (qwk)
- Features extracted from true argument components and argumentative relations
- Hypothesis
 - Feature from more accurate AM can predict essay score more accurately

Argumentation Feature Sets

- 33 features in 5 categories
 - from prior studies
 - proposed in this study

[Persing & Ng, 2015]

[Wachsmuth et al., 2016]

[Ghosh et al., 2016]

[Beigman Klebanov et al., 2016]

Argumentation Feature sets	
Argument Component Features (AC)	
1, 2	Number and fraction of argument components over total number of sentences in essay
3, 4	Number and fraction of argumentative sentences
5	Total number of words in argument components
6	Number of paragraphs containing argument components
7	Whether the essay has paragraph without any argument component
Argument Component Label Features (CL)	
8	Number of Major Claims
9, 10	Number and fraction of Claims over total number of sentences
11, 12	Number and fraction of Premises
13	Average number of Premises per Claim
Argument Flow Features (AF)	
14	Number of paragraphs that contain Major Claims and Claims
15	Number of paragraphs that contain Major Claims and Premises
16	Number of paragraphs that contain Claims and Premises
17 – 24	Frequency of 8 typed bigrams of argument components
Argumentative Relation Features (RL)	
25	Number of supported Claims
26	Number of dangling Claims
27	Number of supporting Premises
28	Number of paragraphs that have support relations
Argumentation Structure Typology Features (TS)	
29	Number of Chain-structures
30	Number of Tree-structures
31	Number of Tree-structures with height = 1
32	Number of paragraphs that contain Chain-structures
33	Number of paragraphs that contain Tree-structures

Argument flow features

[Wachsmuth et al., 2016]

- Sequence of typed components
 - Ex: Claim – Premise – Premise
- Bigrams of typed components
- 3 types of components
- We use 8 bigrams
 - Major Claim – Major Claim is ignored

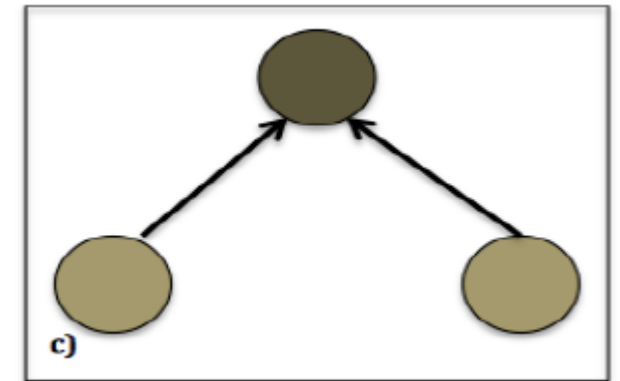
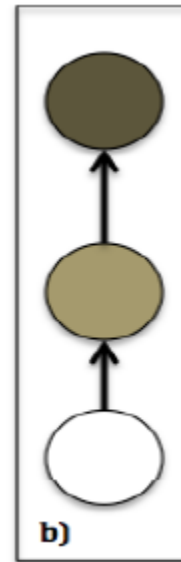
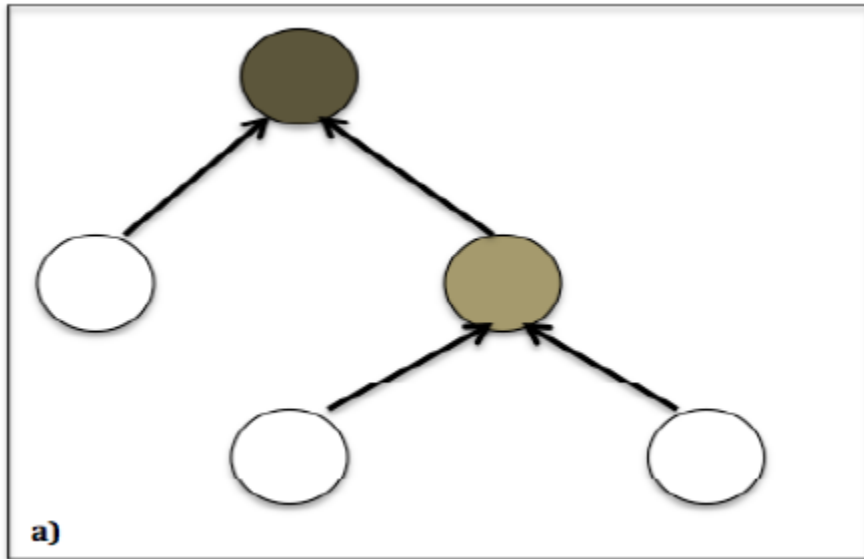
My view is that the *government should give priorities to invest more money on the basic social welfares such as education and housing instead of subsidizing arts relative programs* **MajorClaim**.

Art is not the key determination of quality of life, but education is **Claim**. In order to make people better off, it is more urgent for governments to commit money to some fundamental help such as setting more scholarships in education section for all citizens **Premise**. This is simply because knowledge and wisdom is the guarantee of the enhancement of the quality of people's lives for a well-rounded social system **Premise**.

Argumentation structure typology features

[Ghosh et al., 2016]

- Claims are linked with supporting premises
- Classify 3 tree-like structures



Essay score prediction results (qwk)

- 10x10-fold cross validation
- TrueLabel: argumentation features are extracted from true labels of components and relations

- Component-based features predict essay scores better than relation-based features
- ArgN's features perform better than ArgS' features, except TS



TE107 data statistics

#essays	107
#prompts	2
Low score (C)	31
Medium score (B)	36
High score (A)	40

Input of AM are true components. **: higher with $p < 0.01$, †: smaller with $p < 0.01$

Feature set	AC	CL	AF	RL	TS	All
TrueLabel	0.765	0.768**	0.686	0.747**	0.620**	0.636
ArgN	0.765	0.744	0.695	0.454	0.139 †	0.608
ArgS	0.765	0.729 †	0.577 †	0.423 †	0.165	0.559 †

Input of AM are predicted components. **: $p < 0.01$, *: $p < 0.05$

Feature set	AC	CL	AF	RL	TS	All
ArgN	0.716	0.633	0.512**	0.312*	0.057	0.536
ArgS	0.716	0.603	0.423	0.259	0.189**	0.514

Study 3: Improving AES with argumentation features

- Base AES model

- Enhanced AI Scoring Engine (EASE)

<https://github.com/edx/ease>

- Features:

- Length: counts of words, characters, punctuations, average word length
 - Prompt: count and fraction of words in commons with prompts
 - Bag of words: unigrams, bigrams
 - Part-of-speech: count and fraction of “good” POS sequences

- Proposed model

- EASE augmented with argumentation features

- Persuasive essays of Kaggle ASAP data

- Prompt 1: 1783 essays about good vs. bad effects of computers
 - Prompt 2: 1800 essays about censorship in libraries

In-domain performance

- 5-fold cross validation with a stochastic gradient boosting classification algorithm
- Results in [Phandi et al., 2015] are reported

Argumentation features help improve a competitive base AES model

CHECKED

Kaggle ASAP data statistics

	Essay set 1	Essay set 2
#essays	1783	1800
Average length	350	350
Score range	2–12	1–6
Median	8	3

*: significant difference from EASE ($p < 0.05$)

Feature set	Essay set 1		Essay set 2	
	Kappa	QWK	Kappa	QWK
[Phandi et al., 2015]	n/a	0.781	n/a	0.621
EASE	0.316	0.792	0.463	0.663
ARG	0.308	0.763	0.414	0.612
EASE + ARG	0.328*	0.797	0.475*	0.676

Cross-domain AES

- Domain adaptation in AES as a remedy for the lack of data
- Different machine learning approaches have been proposed
 - Correlated linear regression, automatic features using neural network
- This study is not about domain adaptation
 - But evaluates AM features in cross-domain settings
 - Demonstrates the domain-independence characteristic of argumentation features
- Cross-domain experiment
 - Essay scores of source (training) domain are scaled to range $[-1, 1]$
 - AES models are trained with new score range
 - Predicted scores of target (test) domain are scaled to original range

Phandi et al., 2015

Dong & Zhang, 2016

Phandi et al., 2015

Cross-domain results

- EASE in regression mode
- Predicted scores are rounded to compute Kappa measures
- Experiment with different feature combinations for an upper-bound performance
 - Set 1→2: EASE + AC + CL + TS
 - Set 2→1: EASE + AC + RL + TS

- Argumentation features help improve cross-domain performance
- Both component-based and relation-based features show up in the best combination

CHECKED

Feature set	Set: 1→2		Set: 2→1	
	Kappa	QWK	Kappa	QWK
[Phandi et al., 2015]	n/a	0.545	n/a	n/a
[Dong & Zhang, 2016]	n/a	0.569	n/a	n/a
EASE	0.234	0.585	0.048	0.491
EASE + ARG	0.298	0.622	0.049	0.493
Our best	0.336	0.649	0.053	0.529

Summary

- This study brings up a strong proof of application of argument mining in AES
 - Output of end-to-end argument mining for improving holistic score prediction in persuasive essay
 - The largest argumentation feature set have ever been studied
 - A competitive AES model as the baseline
 - In- and cross-domain evaluations
- Extrinsic evaluation of AM models using AES task
 - Argumentation features of more accurate AM model can predict essay scores more accurately

Conclusions and future work



Contribution summary

- Novel contextual features for improving argument mining
 - Algorithm for argument and domain word extraction
 - Context-window framework
- Extensive studies on improving automated essay scoring with argument mining output
 - Extrinsic evaluation of argument mining systems
 - Comprehensive analysis of a large set of argumentation features
 - In- and cross-domain validation of AES
- AM and AES models are evaluated comprehensively for strongest proofs
 - Cross-fold, cross-topic, end-to-end validation
 - Different corpora, annotation schemas, writing fluency and quality

Future work

- Improve and compare our AD extraction algorithm with related work
 - Lexicon quality, impact on argument mining
- Implement joint-prediction of argument components and argumentative relations
 - Advantage of mutual-information between the two problems
 - Successes proved in prior studies
- Deploy and expand our research
 - Automated assessment of argumentative writing in peer-review/tutoring systems
 - Application in different text genres, e.g., user comments, online debates

Thank you!

HAPPY
EASTER

