Human Language Technology Research Institute

End-to-End Argumentation Mining in Student Essays

Isaac Persing and **Vincent Ng**

Human Language Technology Research Institute

University of Texas at Dallas

Argument Mining

- 2 subtasks

  1. Argument component identification (ACI)
  - Identify the locations and types of argument components
    - Major claims, claims, and premises

  2. Relation identification (RI)
  - Determine the relation that holds between components
    - Support, Attack

Example

I believe that we should attach more importance to cooperation during primary education. Through cooperation, children can learn about interpersonal skills which are significant in their future life. What we acquired from team work is how to achieve the same goal with others and get along with others.

Example

I believe that we should attach more importance to cooperation during primary education. Through cooperation, children can learn about interpersonal skills which are significant in their future life. What we acquired from team work is how to achieve the same goal with others and get along with others.

Major Claim

Example

I believe that we should attach more importance to cooperation during primary education. Through cooperation, children can learn about interpersonal skills which are significant in their future life. What we acquired from team work is how to achieve the same goal with others and get along with others.

Major Claim

Example

I believe that we should attach more importance to cooperation during primary education. Through cooperation, children can learn about interpersonal skills which are significant in their future life. What we acquired from team work is how to achieve the same goal with others and get along with others.

Major Claim

Claim

Example

I believe that we should attach more importance to cooperation during primary education. Through cooperation, children can learn about interpersonal skills which are significant in their future life. What we acquired from team work is how to achieve the same goal with others and get along with others.

Major Claim

supported by

Claim

Example

I believe that we should attach more importance to cooperation during primary education. Through cooperation, children can learn about interpersonal skills which are significant in their future life. What we acquired from team work is how to achieve the same goal with others and get along with others.

Major Claim

supported by

Claim

Example

I believe that we should attach more importance to cooperation during primary education. Through cooperation, children can learn about interpersonal skills which are significant in their future life. What we acquired from team work is how to achieve the same goal with others and get along with others.

Major Claim

supported by

Claim

Premise

Example

I believe that we should attach more importance to cooperation during primary education. Through cooperation, children can learn about interpersonal skills which are significant in their future life. What we acquired from team work is how to achieve the same goal with others and get along with others.

Major Claim

supported by

Claim

supported by

Premise

Why is argument mining challenging?

- Argument components (ACs) having the same type may not (lexically and semantically) resemble each other
- Accurate extraction of ACs is complicated by the fact that they are mostly clauses
- An AC cannot always be extracted independently of other ACs
  - Can we really decide whether a text segment is a premise without knowing what claims are being made?

Goal: End-to-End Argument Mining

- **Input**: raw text
- **Output**: text annotated with ACs and relations

Previous Argument Mining Systems

- rarely end-to-end
- Stab & Gurevych (2014)
  - **Argument component identification**
    - Assume as input gold AC boundaries and sentences that do not contain ACs
    - Classify each of them as Major Claim, Claim, Premise, or non-argumentative
  - **Relation identification**
    - Assume as input gold argument components

Previous Argument Mining Systems

- rarely end-to-end
- Stab & Gurevych (2014)
  - **Argument component identification**
    - Assume as input gold AC boundaries and sentences that do not contain ACs
    - Classify each of them as Major Claim, Claim, Premise, or non-argumentative
  - **Relation identification**
    - Assume as input gold argument components

Substantial simplification of the two tasks

Plan for the Talk

- Essay corpus
- End-to-end argument mining systems
  - Baseline system
  - Our approach
- Evaluation

Plan for the Talk

- Essay corpus
- End-to-end argument mining systems
  - Baseline system
  - Our approach
- Evaluation

The S&G Essay Corpus

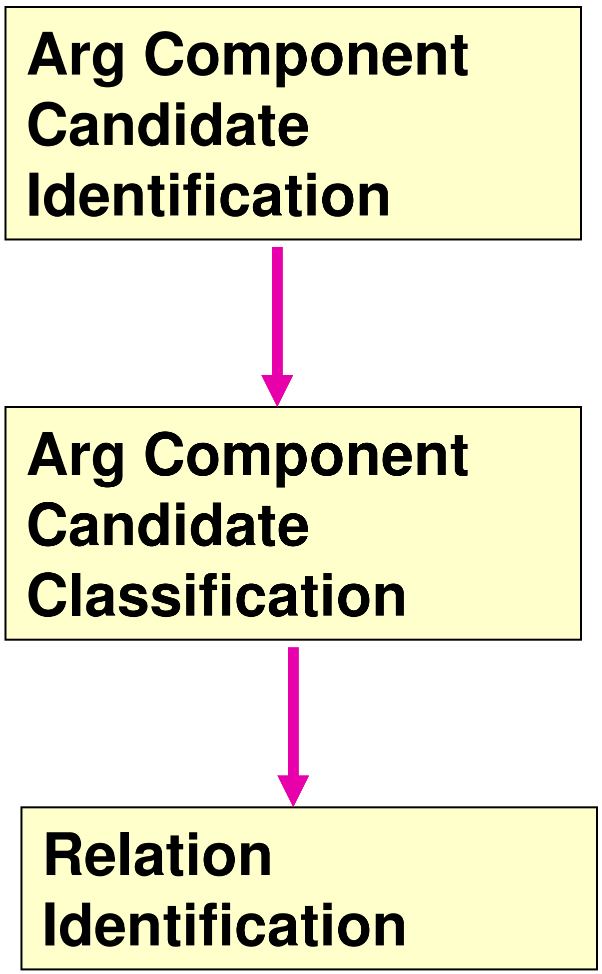
- 90 persuasive essays annotated by Stab & Gurevych (2014)

Argument Component Types	**Major Claim**	90
	Claim	429
	Premise	1033
Relations	**Support**	1312
	Attack	161

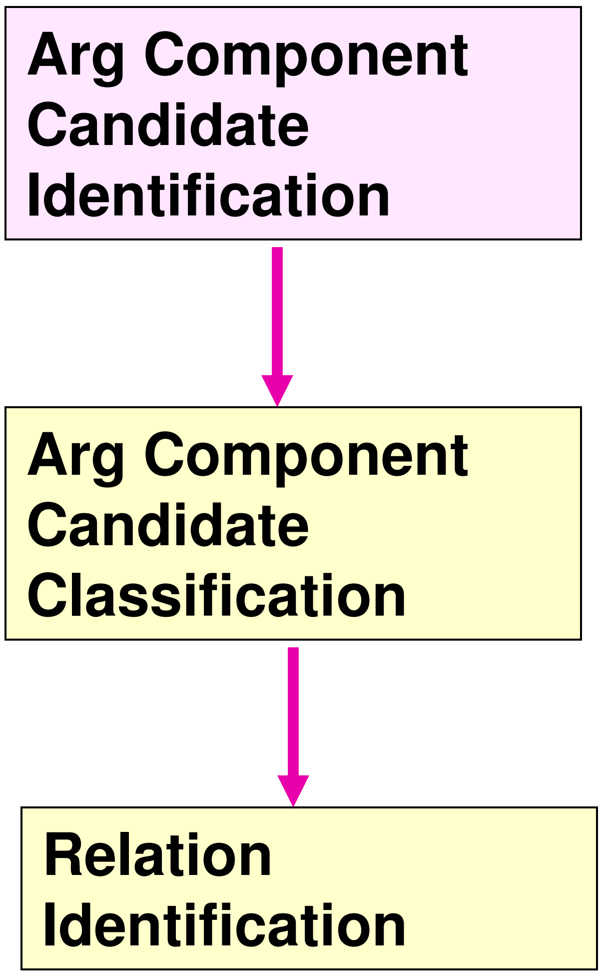
Plan for the Talk

- Essay corpus
- End-to-end argument mining systems
  - Baseline system: Pipeline Approach
  - Our approach
- Evaluation

Baseline System Architecture

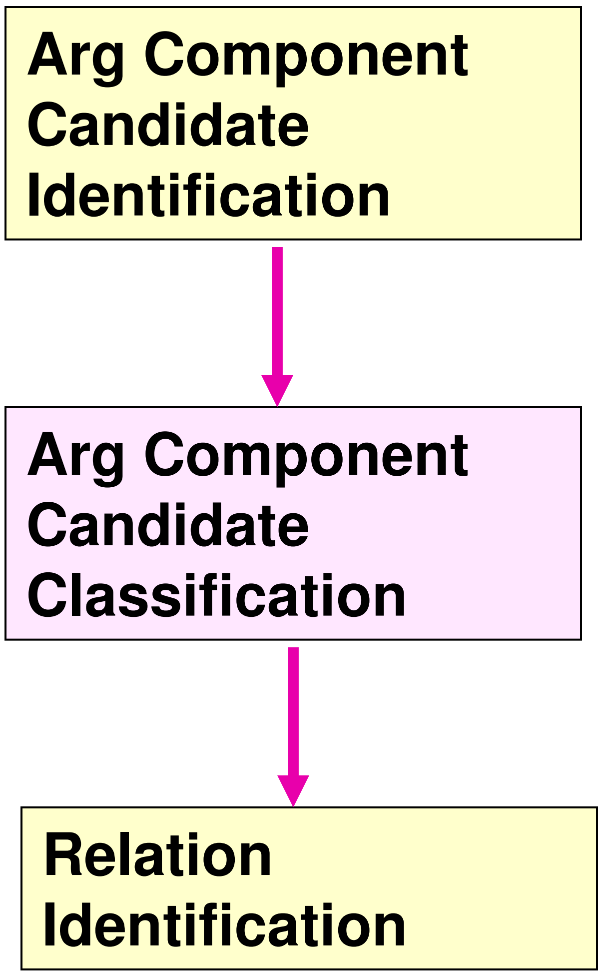


Baseline System Architecture



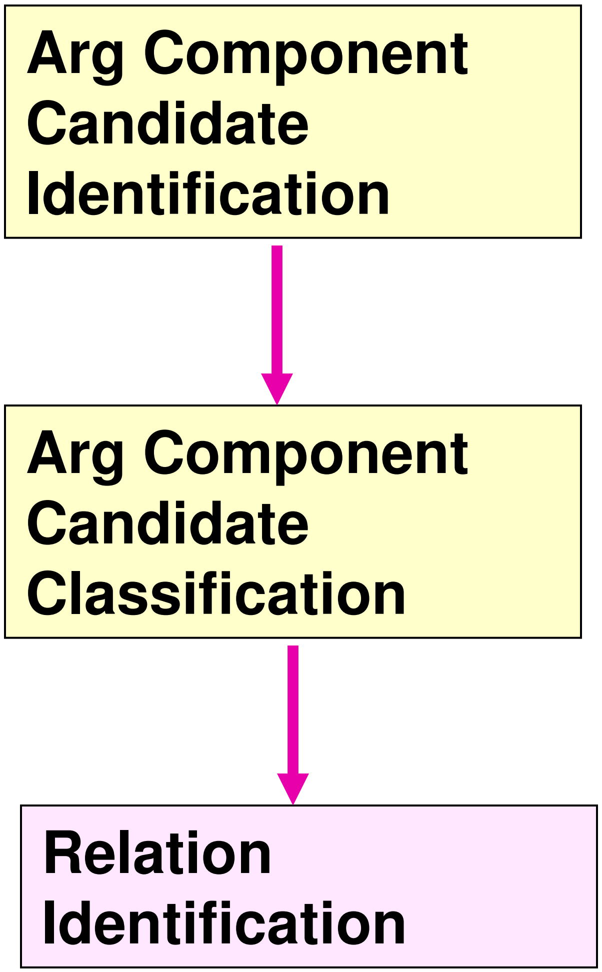
- Identifies AC candidates from raw text
  - heuristically (92% recall)

Baseline System Architecture



- Identifies AC candidates from raw text
  - heuristically (92% recall)
- Classifies each AC candidate as major claim, claim, premise, or non-argumentative
  - Train a MaxEnt classifier using Stab and Gurevych’s features

Baseline System Architecture



- Identifies AC candidates from raw text
  - heuristically (92% recall)
- Classifies each AC candidate as major claim, claim, premise, or non-argumentative
  - Train a MaxEnt classifier using Stab and Gurevych's features
- Classifies each pair of candidates as support, attack, or no relation
  - Train a MaxEnt classifier using Stab and Gurevych's features

Plan for the Talk

- Essay corpus
- End-to-end argumentation mining systems
  - Baseline system
  - Our approach
- Evaluation

Baseline: Pipeline Approach

The AC candidates are classified **independently** of each other

Problem 1

- Determining whether a text segment is an AC **cannot always** be done independently of other ACs

Problem 2

- Within-task constraints cannot be enforced
  - E.g., for AC candidate classification, one constraint says that each essay has exactly one major claim

Problem 3

- Errors **propagate** from AC classifier to relation classifier
  - E.g., if the AC classifier misclassifies one or both ACs involved in a relation as non-argumentative, the relation classifier won't be able to identify their relationship
  - Problem arises because we are using 1-best outputs

Problem 3

- Errors **propagate** from AC classifier to relation classifier
  - E.g., if the AC classifier misclassifies one or both ACs involved in a relation as non-argumentative, the relation classifier won't be able to identify their relationship
  - Problem arises because we are using 1-best outputs
- **Solution**:
  - Use the n-best outputs from the AC classifier to create test instances for the relation classifier
    - More robust to errors made by the AC classifier

But… another problem could arise

- The output of the relation classifier may no longer be consistent with the output of the AC classifier
  - Relation classifier may posit a relation between A and B even if one of them is classified as non-argumentative

But… another problem could arise

- The output of the relation classifier may no longer be consistent with the output of the AC classifier
  - Relation classifier may posit a relation between A and B even if one of them is classified as non-argumentative

Need to enforce the cross-task consistency constraint:
A and B can be related only if both of them are ACs

How to enforce within-task and cross-task consistency constraints?

- Joint inference via Integer Linear Programming
  - Constrained optimization framework
    - Maximize an objective function subject to a set of linear constraints

- One ILP program **per essay**
  - Objective function involves decisions made for the AC classification task and the relation ident. task
  - four types of consistency constraints

Constraints on Major Claims

- Exactly one major claim per essay
- Major claim always occur in the first or last paragraph
- Major claims have no parents

Constraints derived from Stab & Gurevych's annotation guidelines

Constraints on Claims

- A claim can have no more than one parent
- If a claim has a parent, it must be a major claim

Constraints on Premises

- A premise has at least one parent

- A premise is only related to components in the same paragraph

Other Constraints

- The boundaries of the ACs don’t overlap
- Each paragraph must have at least one claim or major claim
- Each sentence may have at most two argument components

ILP Objective Function

- Sum of $X + Y$

$$X = \frac{1}{a}\sum_{i=1}^{a}\log(Cn_i Xn_i + Cp_i Xp_i + Cc_i Xc_i + Cm_i Xm_i)$$

$$Y = \frac{1}{|B|}\sum_{(i,j)\in B}\log(Dn_{i,j}Yn_{i,j} + Ds_{i,j}Ys_{i,j} + Da_{i,j}Ya_{i,j} + Drs_{i,j}Yrs_{i,j} + Dra_{i,j}Yra_{i,j})$$

ILP Objective Function

- Sum of $X + Y$

$$\boxed{X = \frac{1}{a}\sum_{i=1}^{a}\log(Cn_iXn_i + Cp_iXp_i + Cc_iXc_i + Cm_iXm_i)}$$

AC candidate classification

$$Y = \frac{1}{|B|}\sum_{(i,j)\in B}\log(Dn_{i,j}Yn_{i,j} + Ds_{i,j}Ys_{i,j} + Da_{i,j}Ya_{i,j} + Drs_{i,j}Yrs_{i,j} + Dra_{i,j}Yra_{i,j})$$

ILP Objective Function

- Sum of $X + Y$

$$\boxed{X = \frac{1}{a}\sum_{i=1}^{a}\log(Cn_i Xn_i + Cp_i Xp_i + Cc_i Xc_i + Cm_i Xm_i)}$$

Prob. classifications returned by MaxEnt AC cand. classifier

$$Y = \frac{1}{|B|}\sum_{(i,j)\in B}\log(Dn_{i,j}Yn_{i,j} + Ds_{i,j}Ys_{i,j} + Da_{i,j}Ya_{i,j} + Drs_{i,j}Yrs_{i,j} + Dra_{i,j}Yra_{i,j})$$

ILP Objective Function

- Sum of $X + Y$

$$\boxed{X = \frac{1}{a}\sum_{i=1}^{a} \log(Cn_i Xn_i + Cp_i Xp_i + Cc_i Xc_i + Cm_i Xm_i)}$$

Binary variables to be set by the ILP solver

$$Y = \frac{1}{|B|}\sum_{(i,j)\in B} \log(Dn_{i,j}Yn_{i,j} + Ds_{i,j}Ys_{i,j} + Da_{i,j}Ya_{i,j} + Drs_{i,j}Yrs_{i,j} + Dra_{i,j}Yra_{i,j})$$

ILP Objective Function

- Sum of $X + Y$

$$\boxed{X = \frac{1}{a}\sum_{i=1}^{a}\log(Cn_iXn_i + Cp_iXp_i + Cc_iXc_i + Cm_iXm_i)}$$

Unweighted average over all AC candidates

$$Y = \frac{1}{|B|}\sum_{(i,j)\in B}\log(Dn_{i,j}Yn_{i,j} + Ds_{i,j}Ys_{i,j} + Da_{i,j}Ya_{i,j} + Drs_{i,j}Yrs_{i,j} + Dra_{i,j}Yra_{i,j})$$

ILP Objective Function

- Sum of $X + Y$

$$X = \frac{1}{a} \sum_{i=1}^{a} \log(Cn_i Xn_i + Cp_i Xp_i$$
$$+ Cc_i Xc_i + Cm_i Xm_i)$$

$$\boxed{\begin{aligned} Y &= \frac{1}{|B|} \sum_{(i,j) \in B} \log(Dn_{i,j} Yn_{i,j} + Ds_{i,j} Ys_{i,j} \\ &+ Da_{i,j} Ya_{i,j} + Drs_{i,j} Yrs_{i,j} + Dra_{i,j} Yra_{i,j}) \end{aligned}}$$

Relation Identificatior

- We can now maximize this objective function using an ILP solver subject to our constraints

- But… we are still not happy

- We can now maximize this objective function using an ILP solver subject to our constraints

- But… we are still not happy

with the objective function

- We can now maximize this objective function using an ILP solver subject to our constraints

- But… we are still not happy

with the objective function

- ILP tries to maximize agreement with the two MaxEnt classifiers' probabilistic classifications
- But… we want an objective function that maximizes the average F-scores of the two tasks

F-score Maximizing Objective Function

$$F = \frac{2TP}{2TP + FP + FN}$$

- Problem
  - ILP can only handle linear combination of variables

F-score Maximizing Objective Function

$$F = \frac{2TP}{2TP + FP + FN}$$

- Problem
  - ILP can only handle linear combination of variables
- Solution
  - Maximize difference between numerator & denominator

F-score Maximizing Objective Function

$$\boxed{F = \frac{2TP}{2TP + FP + FN}}$$

- Maximize the following instead:

$$G = \alpha 2TP_e - (1-\alpha)(FP_e + FN_e)$$

F-score Maximizing Objective Function

$$F = \frac{2TP}{2TP + FP + FN}$$

- Maximize the following instead:

$$\boxed{G = \alpha 2TP_e - (1-\alpha)(FP_e + FN_e)}$$

F-score Maximizing Objective Function

$$F = \frac{2TP}{2TP + FP + FN}$$

- Maximize the following instead:

$$G = \alpha 2TP_e - (1 - \alpha)(FP_e + FN_e)$$

F-score Maximizing Objective Function

$$F = \frac{2TP}{2TP + FP + FN}$$

- Maximize the following instead:

$$G = \alpha 2TP_e - (1-\alpha)(FP_e + FN_e)$$

F-score Maximizing Objective Function

$$F = \frac{2TP}{2TP + FP + FN}$$

- Maximize the following instead:

$$G = \alpha 2TP_e - (1 - \alpha)(FP_e + FN_e)$$

F-score Maximizing Objective Function

$$F = \frac{2TP}{2TP + FP + FN}$$

- Maximize the following instead:

$$G = \alpha 2TP_e - (1 - \alpha)(FP_e + FN_e)$$

F-score Maximizing Objective Function

$$F = \frac{2TP}{2TP + FP + FN}$$

- Maximize the following instead:

$$G = \alpha 2TP_e - (1-\alpha)(FP_e + FN_e)$$

F-score Maximizing Objective Function

$$F = \frac{2TP}{2TP + FP + FN}$$

- Maximize the following instead:

$$G = \alpha 2TP_e - (1-\alpha)(FP_e + FN_e)$$

F-score Maximizing Objective Function

$$F = \frac{2TP}{2TP + FP + FN}$$

- Maximize the following instead:

$$G = \alpha 2 TP_e - (1 - \alpha)(FP_e + FN_e)$$

estimated TPs, FPs and FNs

F-score Maximizing Objective Function

$$F = \frac{2TP}{2TP + FP + FN}$$

- Maximize the following instead:

$$G = \alpha 2 TP_e - (1 - \alpha)(FP_e + FN_e)$$

estimated TPs, FPs and FNs

How to estimate these values?

Fill missing data with expected values:
the probabilistic classifications provided by MaxEnt

Plan for the Talk

- Essay corpus
- End-to-end argumentation mining systems
  - Baseline system: Pipeline approach
  - Our approach: Joint inference
- Evaluation

Experimental Setup

- 5-fold cross-validation on S&G’s 90-essay corpus

Evaluation Metrics

Argument Component Identification

- recall, precision, and F-score based on
  - Exact match
    - consider an AC correctly extracted if its boundaries and type are exactly the same as those of a gold AC
  - Approximate match
    - Consider an AC correctly extracted if its type is the same as that of a gold AC and shares at least half of its tokens

Evaluation Metrics

Relation Identification

- recall, precision, and F-score based on
  - Exact and approximate match
    - a relation is correct if its ACs have an exact/approximate match with those of a gold relation and their types match

Results: AC Identification

Approximate Match				
	MajClaim	Claim	Premise	Overall
Baseline	11.1	26.9	51.9	44.0
Our Approach	22.2	42.6	66.0	57.2

- Overall improvement: 13.2% absolute F-score

Results: Relation Identification

Approximate Match			
	Support	Attack	Overall
Baseline	6.1	0.8	5.8
Our Approach	21.3	1.1	20.4

- Overall improvement: 14.6% absolute F-score

Results: Average over the two tasks

Baseline	24.9
Our Approach	38.8

- Overall improvement
  - 13.9% absolute F-score (18.5% relative error reduction)

Ablation Results: Avg of the two tasks

ALL	38.8
No features for the AC classifier	27.7
No features for the relation classifier	38.2
No ILP	27.0
No ILP constraints on Major Claims	32.4
No ILP constraints on Claims	34.5
No ILP constraints on Premises	37.0
No ILP other constraints	30.7
No ILP f-score optimizing function	33.4

Ablation Results: Avg of the two tasks

ALL	38.8
No features for the AC classifier	27.7
No features for the relation classifier	38.2
No ILP	27.0
No ILP constraints on Major Claims	32.4
No ILP constraints on Claims	34.5
No ILP constraints on Premises	37.0
No ILP other constraints	30.7
No ILP f-score optimizing function	33.4

Ablation Results: Avg of the two tasks

ALL	38.8
No features for the AC classifier	27.7
No features for the relation classifier	38.2
No ILP	27.0
No ILP constraints on Major Claims	32.4
No ILP constraints on Claims	34.5
No ILP constraints on Premises	37.0
No ILP other constraints	30.7
No ILP f-score optimizing function	33.4

Ablation Results: Avg of the two tasks

ALL	38.8
No features for the AC classifier	27.7
No features for the relation classifier	38.2
No ILP	27.0
No ILP constraints on Major Claims	32.4
No ILP constraints on Claims	34.5
No ILP constraints on Premises	37.0
No ILP other constraints	30.7
No ILP f-score optimizing function	33.4

Ablation Results: Avg of the two tasks

ALL	38.8
No features for the AC classifier	27.7
No features for the relation classifier	38.2
No ILP	27.0
No ILP constraints on Major Claims	32.4
No ILP constraints on Claims	34.5
No ILP constraints on Premises	37.0
No ILP other constraints	30.7
No ILP f-score optimizing function	33.4

Ablation Results: Avg of the two tasks

ALL	38.8
No features for the AC classifier	27.7
No features for the relation classifier	38.2
No ILP	27.0
No ILP constraints on Major Claims	32.4
No ILP constraints on Claims	34.5
No ILP constraints on Premises	37.0
No ILP other constraints	30.7
No ILP f-score optimizing function	33.4

Summary

- Presented the first results on end-to-end argument mining in persuasive essays
  - Using a pipeline approach
  - Using ILP-based joint inference in combination with a F-score optimizing objective function

- The joint inference approach yields a 18.5% relative error reduction over the pipeline system when evaluated on 90 essays