Improving Argument Mining in Student Essays
by
Learning and Exploiting Argument Indicators versus Essay Topics

Huy Nguyen¹ Diane Litman¹,²
¹Computer Science Department
²Learning Research & Development Center
University of Pittsburgh, Pennsylvania
Argumentative text is ubiquitous.

“Argumentation mining [...] involves automatically identifying argumentative structures within a document, [...] as well as argument-subargument and argument-counterargument relationships between pairs of arguments in the document.”

(The 1st Argument Mining Workshop at NAACL, 2014)
Overview of our research

Essay evaluation

Argumentative relation classification

Argument component identification

Students’ persuasive essays

Do arts and music improve the quality of life?

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. Art is not the key determination of quality of life, but education is. 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. 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.

Admittedly, art, to some extent, serve a valuable function about enriching one’s daily lives. For example, it could bring release one’s heavy burden of study pressure and refresh human bodies through a hard day from work. However, it is unrealistic to pursuit of this high standard of life in many developing countries, in which the basic housing supply has still been a huge problem with plenty of lower income family have squeezed in a small tight room. By comparison to these issues, the pursuit of art seems unimportant at all.

To conclude, art could play an active role in improving the quality of people's lives, but I think that governments should attach heavier weight to other social issues such as education and housing need because those are the most essential ways enable to make people a decent life.
Argument component identification

• Argument component: text portion with a specific role in forming the argument*

[...] To conclude, art could play an active role in improving the quality of people’s lives, but I think that governments should attach heavier weight to other social issues such as education and housing needs because those are the most essential ways enable to make people a decent life.

(Persuasive Essay Corpus, Stab & Gurevych 2014)

• The step before argumentative relation mining

• This study focuses on argument component identification in student essays

Prior argument component identification studies

• N-gram and production rule features (VP→VBG NP) [Stab & Gurevych 2014]
  ✗ Large and sparse feature space
  ✗ Have not considered abstraction of argument topic

• Lexicons of argument and domain words [Nguyen & Litman 2015]
  ✗ Lacked a quantitative evaluation

• Cross-fold validation
  ✗ Have not evaluated topic-independence of the models (e.g., train and test essays are of different topics)
Argument and domain word extraction [Nguyen & Litman 2015]

- 6794 un-annotated persuasive essays*

**Argument seeds:** agree, disagree, reason, support, advantage, disadvantage, think, conclusion, result, opinion

**Domain seeds:** title words that are not argument seeds or stop words

**Scoring algorithm:** looks for the most argumentative LDA topic, i.e., high argument weight and low domain weight

**Result:** 263 argument words and 1806 domain words (stemmed)

* www.essayforum.com
Example argument and domain words

**Argument seeds**: agree, disagree, reason, support, advantage, disadvantage, think, conclusion, result, opinion

**LDA topic 1**: reason example support agree think because disagree statement opinion believe therefor idea conclusion

**LDA topic 2**: city live big house place area small apart town build community factory urban

**LDA topic 3**: children parent school education teach kid adult grow childhood behavior taught

= **Argument seeds & variants, discourse connectives, stop words**
Baseline vs. Proposed models

<table>
<thead>
<tr>
<th>Stab14 (Stab &amp; Gurevych 2014b)</th>
<th>Nguyen15 (Nguyen &amp; Litman 2015)</th>
<th>This study (wLDA+4)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexical</strong> (I)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-, 2-, 3-grams</td>
<td>Argument words as unigrams</td>
<td></td>
</tr>
<tr>
<td>Verbs, adverbs, presence of model verb</td>
<td>Same as Stab14</td>
<td></td>
</tr>
<tr>
<td>Singular first person pronouns</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discourse connectives</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Parse</strong> (II)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Production rules</td>
<td>Argumentative subject-verb pairs</td>
<td></td>
</tr>
<tr>
<td>Tense of main verb</td>
<td>Same as Stab14</td>
<td></td>
</tr>
<tr>
<td>#sub-clauses, depth of parse tree</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Structure</strong> (III)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#tokens, token ratio, #punctuation, sentence position, first/last paragraph, first/last sentence of paragraph</td>
<td>Same as Stab14</td>
<td></td>
</tr>
<tr>
<td><strong>Context</strong> (IV)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#tokens, #punctuation, #sub-clauses, modal verb in preceding/following sentences</td>
<td>Same as Stab14</td>
<td></td>
</tr>
</tbody>
</table>

10-fold cross validation (data was randomly split into training and test sets) Cross writing-prompt validation (training and test essays are of different prompts)

1. Numbers of common words with title and preceding sentence
2. Comparative & superlative adverbs and POS
3. Plural first person pronouns
4. Discourse relation labels
Ablated models

• Replace argument and domain lexicons in wLDA+4 model

  • SEED model: uses only argument and domain seeds

  • woLDA model: does not use seed words or the two lexicons
    • Removes argument word features
    • Uses all subject-verb pairs
Persuasive Essay Corpus [Stab & Gurevych 2014]

- 90 persuasive essays*
  - MajorClaim
  - Claim
  - Premise

- 3 expert annotators
  - Accuracy 0.88
  - Krippendorff’s $\alpha_U$ 0.72

* www.essayforum.com
Evaluation method

• Cross-fold validation
  • Randomly: 10-fold cross validation
  • By-prompt: cross writing prompt validation

• In each folding
  • Select top 100 features in training folds (InfoGain + Ranking)
  • Train prediction model with top 100 features
  • Record prediction output on the test fold
### Experimental results: cross validation

<table>
<thead>
<tr>
<th></th>
<th>10-fold cross validation</th>
<th>Cross-prompt validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Stab14</td>
<td>Nguyen15</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.787*</td>
<td>0.792*</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.639*</td>
<td>0.649*</td>
</tr>
<tr>
<td>Precision</td>
<td>0.741*</td>
<td>0.745*</td>
</tr>
<tr>
<td>Recall</td>
<td>0.694*</td>
<td>0.698*</td>
</tr>
</tbody>
</table>

Best values in bold. +: p < 0.1, *: p < 0.05 by T-test when comparing with wLDA+4

Obtains comparable performances between two experiment settings

Proposed model (wLDA+4) performs the best in 10-fold cross validation

12 groups:
- 11 single-prompt groups (73 essays)
- 1 mixed group of minor prompts (17 essays)

Prompts: school, technologies, prepared food ...
## Experimental results: holdout test sets

<table>
<thead>
<tr>
<th></th>
<th>Stab’s test set</th>
<th>Nguyen’s test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Stab’s reported</td>
<td>wLDA+4</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.77</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.83</td>
</tr>
<tr>
<td>Kappa</td>
<td>--</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.69</td>
</tr>
<tr>
<td>F1</td>
<td>0.73</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.76</td>
</tr>
<tr>
<td>Precision</td>
<td>0.77</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.79</td>
</tr>
<tr>
<td>Recall</td>
<td>0.68</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.76</td>
</tr>
</tbody>
</table>
Feature evaluation

• Among all top features used to train the models
  • 49% are argument words
  • 8% are argumentative subject-verb pairs

• In the top-50
  • Common word counts
  • Comparative adverbs, and RBR part-of-speech
  • Person pronouns WE, OUR
  • Discourse labels Expansion, Contingency

LDA-enabled features in Nguyen15

Proposed features in this study
Conclusions and future work

• A study on argument component identification

• New features that model argument indicators and abstract over essay topics
  • A necessary supplement to the learned and noisy argument and domain words

• Cross-topic and 10-fold cross validations
  • Proposed model obtained comparable performances

• Our next study focuses on argumentative relation classification
Thank you!