Argument Retrieval in Project Debater

Yufang Hou

IBM Research Europe, Dublin
IBM Research: History of Grand Challenges

1997
First computer to defeat a world champion in Chess (Deep Blue)

2011
First computer to defeat best human Jeopardy! players (Watson)

2019
First computer to successfully debate champion debaters (Project Debater)
Segments from a Live Debate (San Francisco, Feb 11th 2019)
Expert human debater: Mr. Harish Natarajan

Motion: We should subsidize preschool

Selected from test set based on assessment of chances to have a meaningful debate

Format: Oxford style debating

Fully automatic debate
No human intervention
Project Debater:
Media Exposure

- 2.1 Billion social media impressions
- 100 Million people reached
- Millions of video views
- Hundreds of press articles in all leading newspapers
• Full Live Debate, Feb-2019
  https://www.youtube.com/watch?v=m3u-1yttrVw&t=2469s

• “The Debater” Documentary
  https://www.youtube.com/watch?v=7pHaNMdWGsk&t=1383s
Outline

- System overview
- Argument retrieval in Project Debater
- Some retrospective thoughts
Current Publications Highlight Various Aspects of the System
Publications and Datasets are available at -

Outline

- System overview
- Argument retrieval in Project Debater
- Some retrospective thoughts
Related Work

• Lippi and Toroni, IJCAI, 2015
• Al-Khatib et al, NAACL 2016; Wachsmuth et al, Argument-Mining Workshop, 2017, ...
• Stab and Gurevych, EMNLP 2014; Stab et al, NAACL 2018, ...
• Recent reviews
  • Five years of argument mining: a data-driven analysis, Cabrio and Villata, IJCAI, 2018
  • Argumentation Mining, Stede and Schneider, Synthesis Lectures on HLT, 2018
  • Argument Mining: A Survey, Lawrence and Reed, CL, 2019
Wikipedia Stage


Wikipedia Stage

- Wikipedia Claim/Evidence Labeled Data – Labeling Process

- Controversial Topic
- Select Wikipedia Articles
- Find Claim Candidates per Article
- Confirm/Reject Each Claim Candidate
- Find Candidate Evidence per Claim
- Confirm/Reject Each Candidate Evidence

☑ 5 In-house Annotators Per Stage
☑ Exhaustive annotation
Wikipedia Stage

- Wikipedia Claim/Evidence Labeled Data - Results
  - 58 Controversial Topics selected from Debatabase
  - 547 relevant Wikipedia articles carefully labeled by in-house team
    - E.g., Ban the sale of Violent Video Games for Children
  - 2.6K Claims & 4.5K Evidence that support/contest the claims
    - Evidence length vary from one sentence to a whole paragraph
    - Three types of Evidence: Study, Expert, and Anecdotal
  - Pre-defined train/dev/test split
Wikipedia Stage

• System Design for Argument Mining

We should subsidize preschool

- Retrieve documents that directly address the topic and are likely to contain argumentative text segments

Claim Detection

- Simple logistic regression model with lots of carefully designed features
  - GrASP: Rich Patterns for Argumentation Mining, Shnarch et al., EMNLP 2017
- Static train/dev/test datasets
- Moderate success over a range of test topics
  - Only positive instances are annotated
  - Limited coverage
VLC (Very Large Corpus) Stage

Corpus wide argument mining - a working solution, Ein-Dor et al, AAAI 2020.
VLC (Very Large Corpus) Stage

Main Distinction from Prev. Work

• Sentence Level (SL) strategy, vs. Document Level used before
• SCALE
  • ~240 train/dev topics & ~100 test topics
  • ~200,000 sentences carefully annotated for train/dev → Retrospective Labeling Paradigm
  • ~10,000,000,000 Sentences - Reporting results over a massive corpus

Closer than ever to a working solution
VLC (Very Large Corpus) Stage

System Architecture

Massive Corpus ~10B Sentences

- Support flexible patterns to retrieve argumentative sentences
  - Topic terms
  - Evidence connectors
  - sentiment lexicon
  - NER

Queries

Controversial Topic

Retrieved Sentences

- Retrieve 12,000 sentences per evidence type per topic

Ranking Model BERT

Iteratively Collected Labeled-Data

High-precision Evidence Set

- Starting with LR from Rinott et al, EMNLP 2015
- Retrospective Labeling Paradigm
- An infrastructure that supports quick dynamic experiments and monitors annotation quality
VLC (Very Large Corpus) Stage

How to Collect Labeled Data?

• Collecting labeled data poses a two-fold challenge -
  • Low prior of positive examples
  • Annotation through crowd requires expertise – simple guidelines, careful monitoring...
  • BTW - Kappa of ~0.4 is actually quite good

• Developing corpus-wide argument mining poses another challenge
  • Imagine ~2,000 new predictions every week... → Associated infrastructure is a must

• Retrospective labeling of top predictions is a natural and effective solution
Motion: Blood donation should be mandatory

According to studies, blood donors are 88 percent less likely to suffer a heart attack...

Statistics ... show that students are the main blood donors contributing about 80 percent...
VLC (Very Large Corpus) Stage

Results

- Results by various BERT Models over a massive corpus of ~10B sentences
- BA baselines: BlendNet, Attention based bidirectional LSTM model [Shnarch et al. (2018)]
- High precision
- Wide coverage with diverse evidences (highly similar sentences are removed)
Outline

- System overview
- Argument retrieval in Project Debater
- Some retrospective thoughts
## Challenges to Consider while developing a Live Debate System

<table>
<thead>
<tr>
<th>Data-driven speech writing and delivery</th>
<th>Listening comprehension</th>
<th>Modeling human dilemmas</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Digest massive corpora</td>
<td>• Identify key claims hidden in long continuous spoken language</td>
<td>• Modeling the world of human controversy and</td>
</tr>
<tr>
<td>• Write a well-structured speech</td>
<td>• Compare to personal assistants - simple short commands</td>
<td>• Enabling the system to suggest principled arguments</td>
</tr>
<tr>
<td>• Deliver with clarity and purpose</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Argument retrieval is the first step to build such a system
The Problem: Many things need to succeed simultaneously and many things can go wrong...
Many things can go wrong... / Examples

• Getting the stance wrong means you support your opponent...

• Drifting from the topic – from *Physical Education* to *Sex Education* and back...

• The system is only as good as its corpus

→ ... *global warming will lead* *malaria virus* *to creep into hilly areas*...
Progress over time / Improvement in Precision of Detecting Claims

- Document level IR
- Exhaustive labelling of positive instances
- LR + Rich features

- Sentence level IR
- Very Large Corpus: 400 million articles (50 times larger than Wikipedia)
- Retrospective labelling
- Bert fine-tuning

- Claim Detection Macro-Avg Prec@50 over all Motions

- 2015 EOY
- 2016 EOY
- 2017 EOY
- 2017 EOY with DL
- 2019 with BERT

- Very large corpus
- Retrospective labelling
- Flexible query
- Attention-based Bi-LSTM with weak supervision
Beyond Project Debater

- Computational argumentation is emerging as an interesting research area
- “Argument mining” is the new keyword in the list of topics in recent *ACL conferences
Thanks!

Q&A