Argument Retrieval in Project Debater

Yufang Hou

IBM Research Europe, Dublin

IBM

Research AI

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*

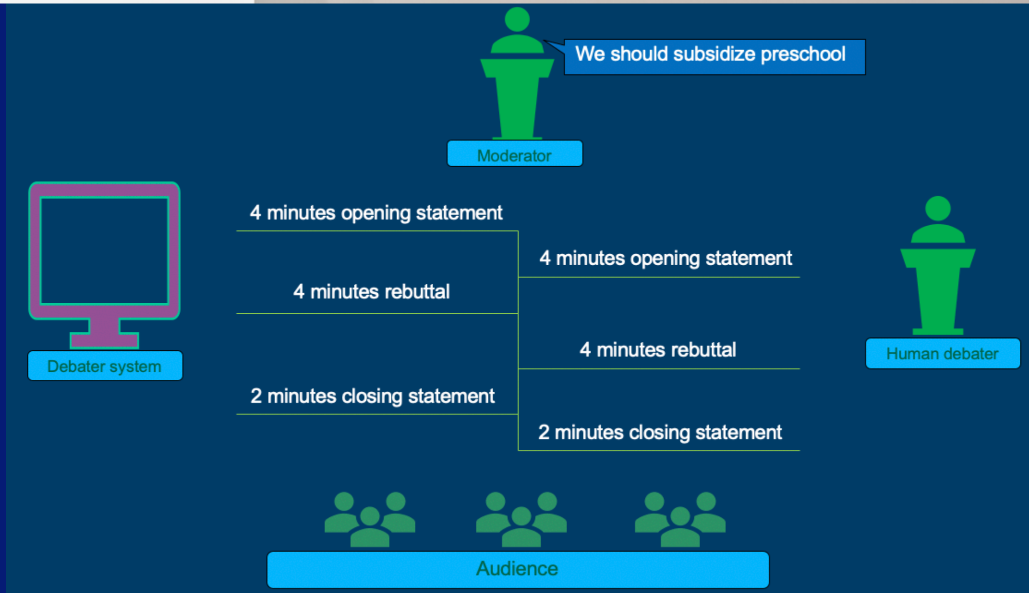
Motion: We should subsidize preschool

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

Format: Oxford style debating

Fully automatic debate

No human intervention



We should subsidize preschool

Moderator

Debater system

4 minutes opening statement

4 minutes opening statement

4 minutes rebuttal

4 minutes rebuttal

2 minutes closing statement

2 minutes closing statement

Human debater

Audience

Project Debater:

Media Exposure

2.1 Billion

social media impressions

100 Million

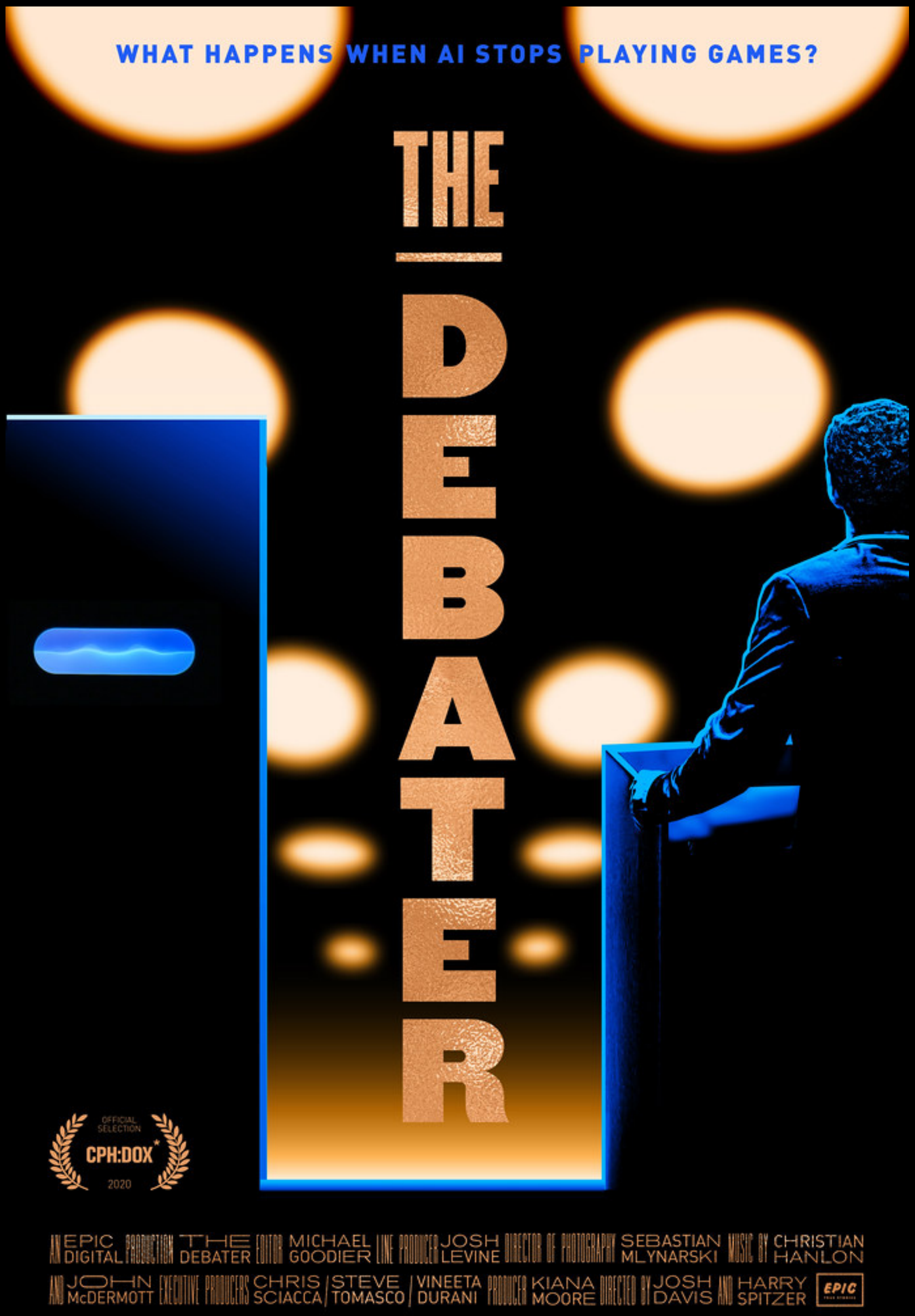
people reached

Millions

of video views

Hundreds

of press articles in all leading news papers


WHAT HAPPENS WHEN AI STOPS PLAYING GAMES?
THE DEBATER
OFFICIAL SELECTION CPH:DOX* 2020
AN EPIC DIGITAL PRODUCTION THE DEBATER EDITOR MICHAEL GOODIER LINE PRODUCER JOSH LEVINE DIRECTOR OF PHOTOGRAPHY SEBASTIAN MLYNARSKI MUSIC BY CHRISTIAN HANLON AND JOHN McDERMOTT EXECUTIVE PRODUCERS CHRIS SCIACCA / STEVE TOMASCO / VINEETA DURANI PRODUCER KIANA MOORE DIRECTED BY JOSH DAVIS AND HARRY SPITZER
EPIC

OFFICIAL
SELECTION
CPH:DOX*
2020

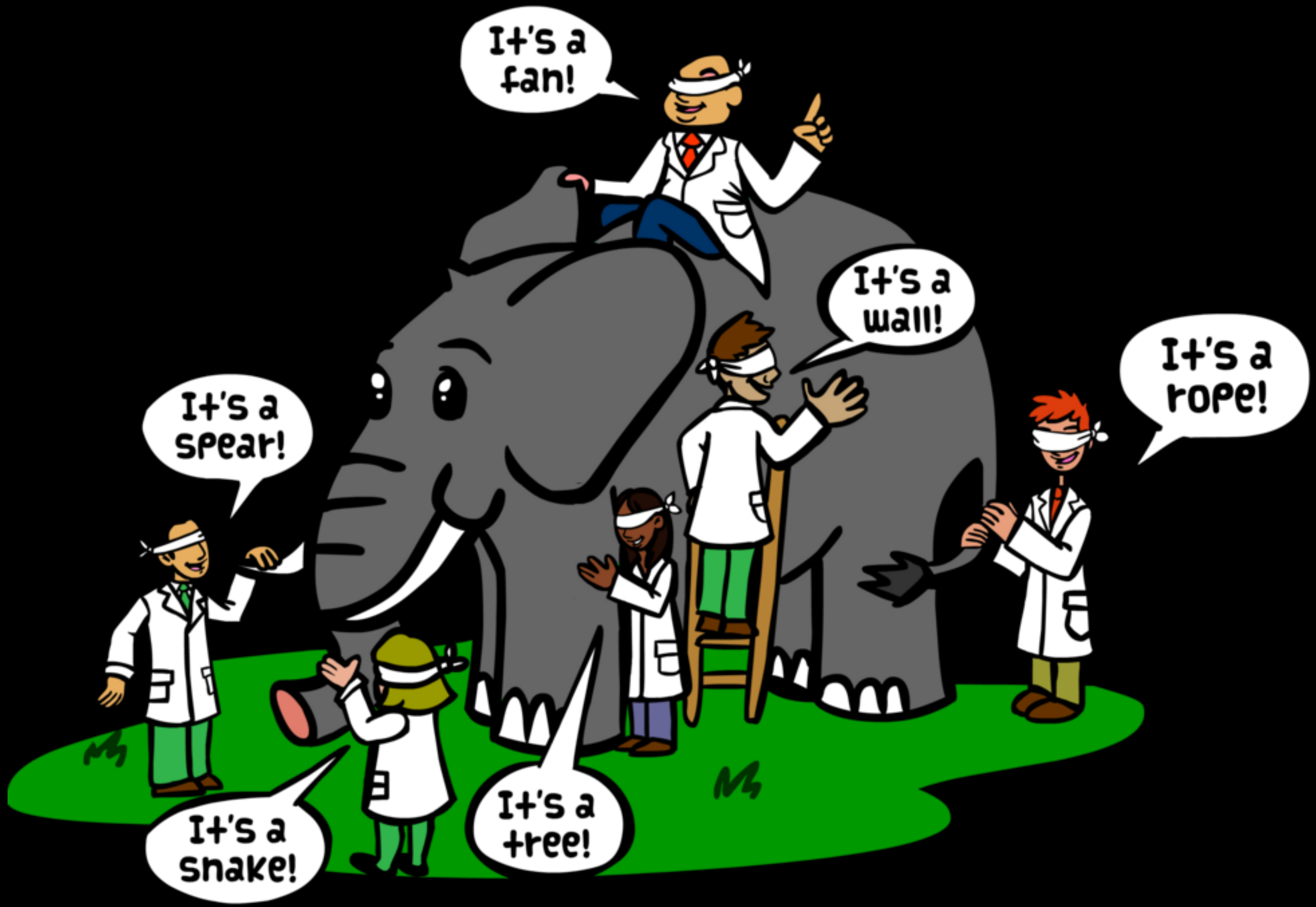
- Full Live Debate, Feb-2019
  https://www.youtube.com/watch?v=m3u-1yttrVw&t=2469s

- “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 -

https://www.research.ibm.com/artificial-intelligence/project-debater/research/

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

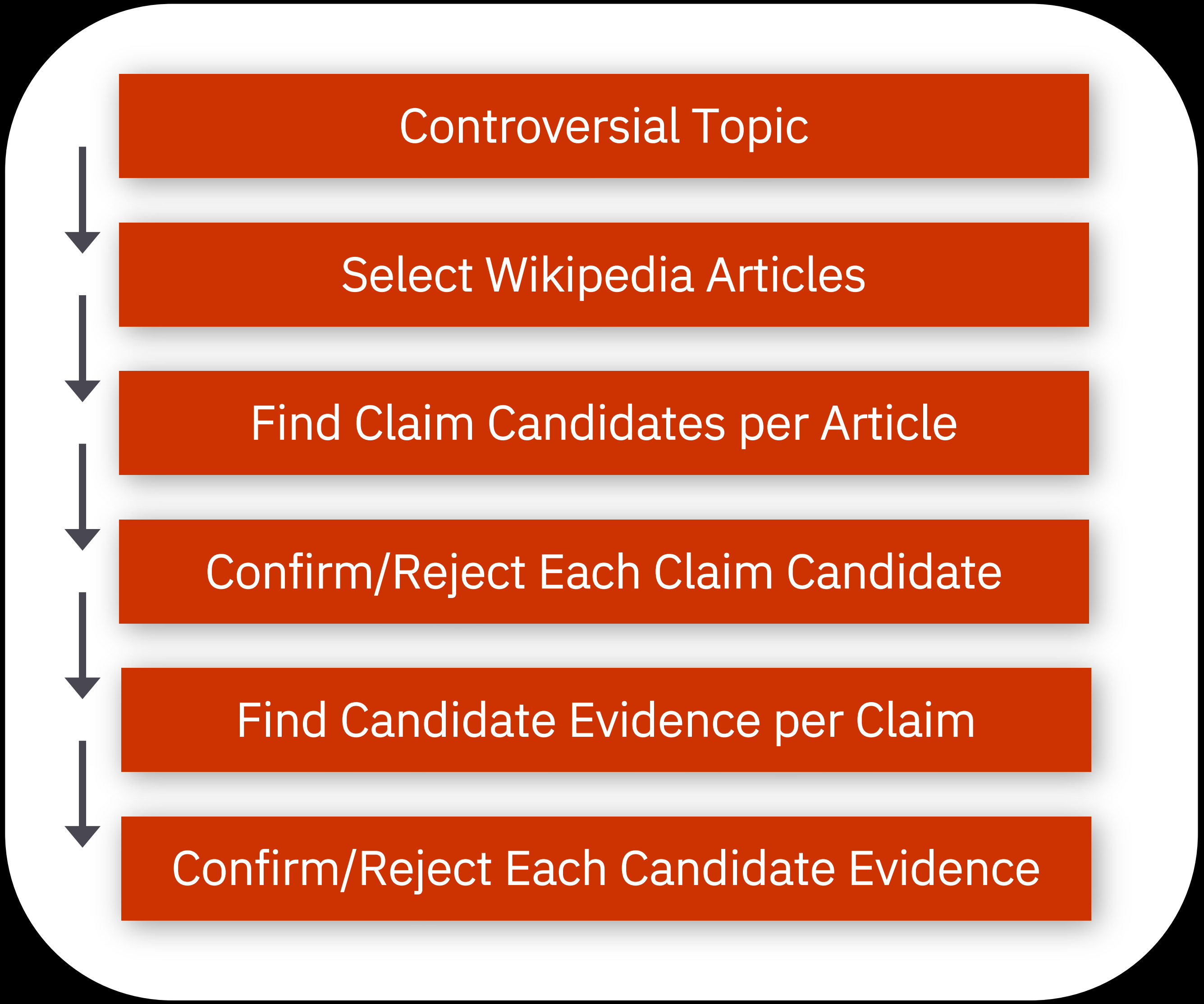
Wikipedia Stage

Context Dependent Claim Detection, Levy et al, COLING 2014.

Show Me Your Evidence - an Automatic Method for Context Dependent Evidence Detection, Rinott et al, EMNLP 2015.

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

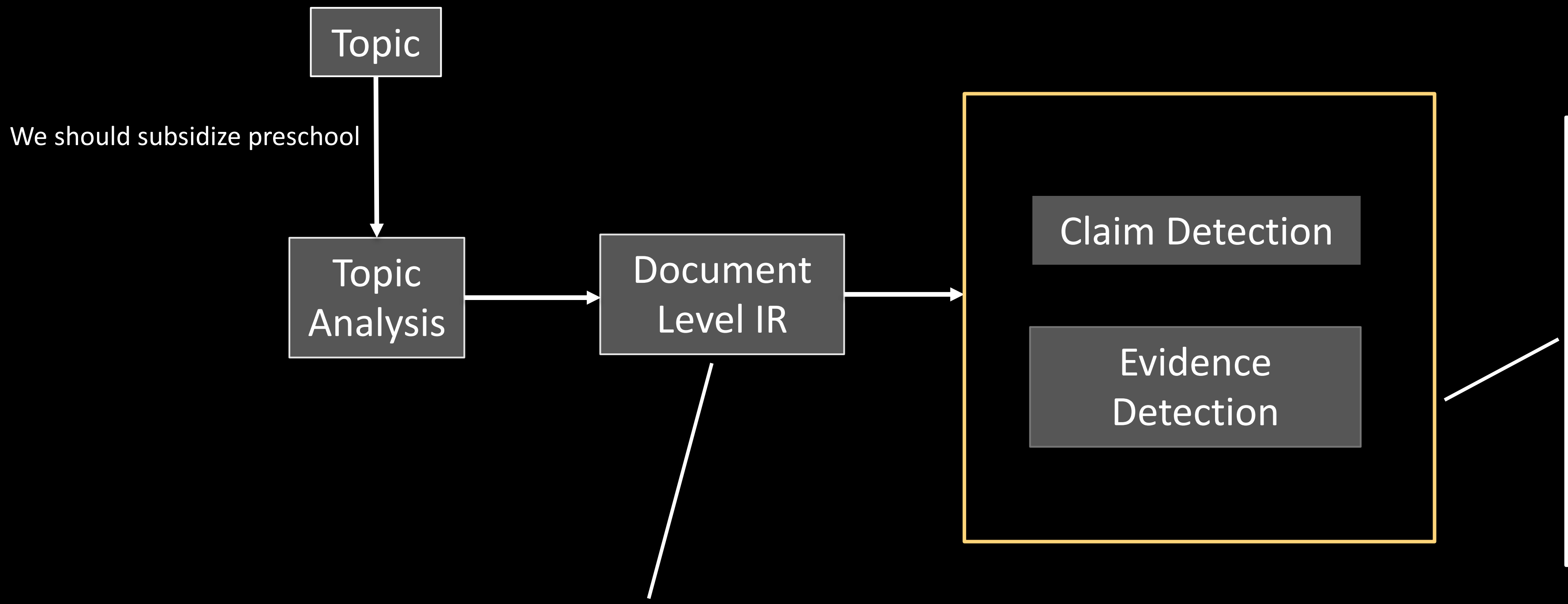
✓ 5 In-house Annotators Per Stage
✓ Exhaustive annotation

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



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

- ○ 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

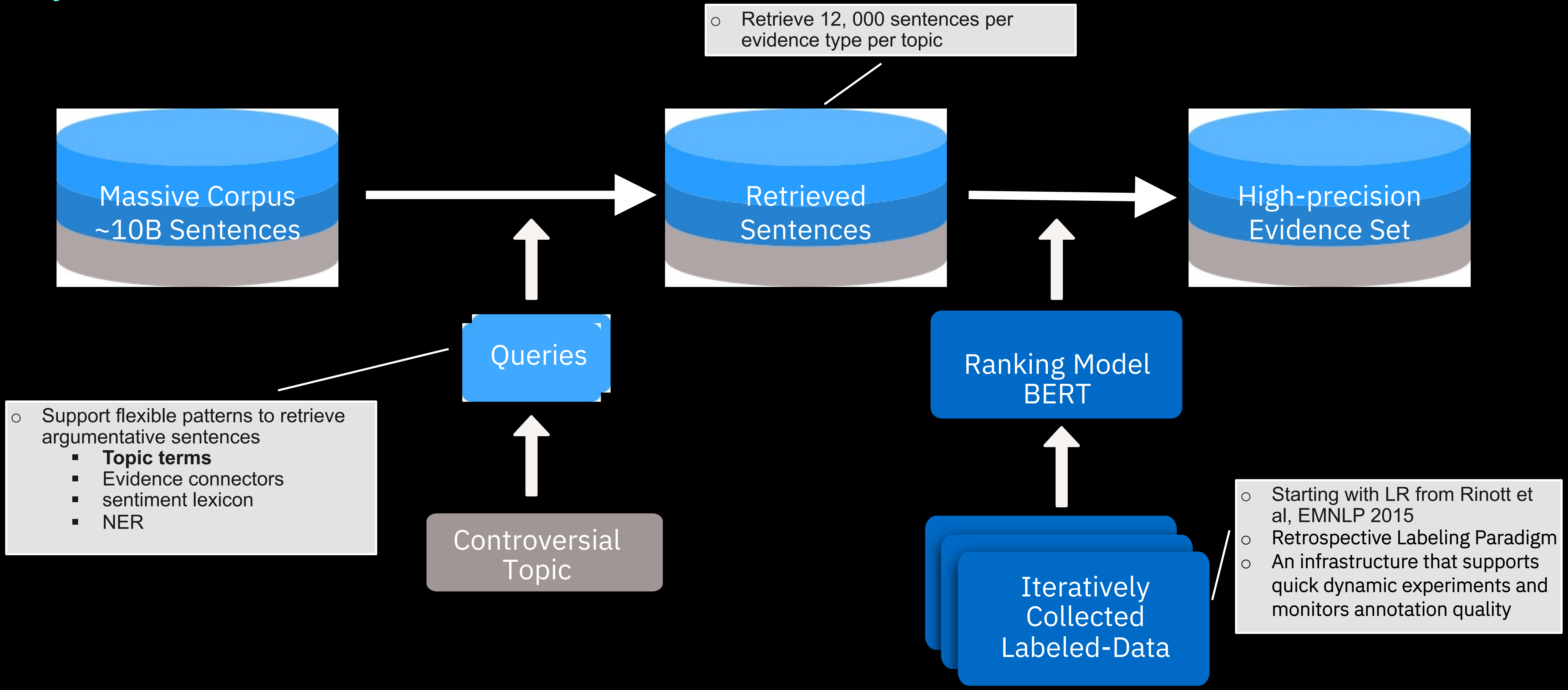
Main Distinction from Prev. Work

- 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



- Retrieve 12, 000 sentences per evidence type per topic

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

- 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

Why Evidence Detection is Hard?

Motion: **Blood donation should be mandatory**

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

CONFIRMED

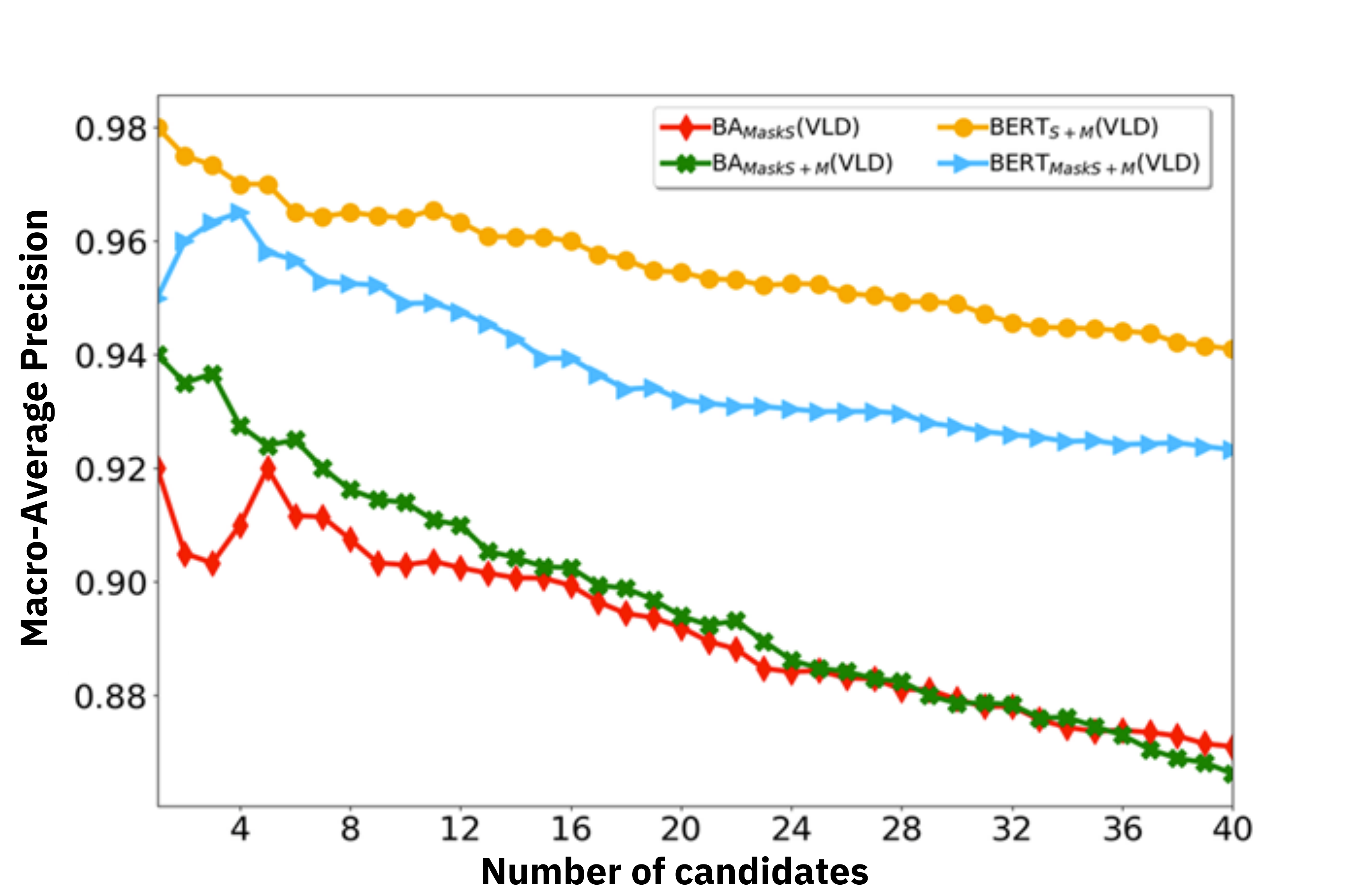
Statistics ... show that students are the main blood donors contributing about 80 percent...

REJECTED

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

Data-driven speech writing and delivery

- Digest massive corpora
- Write a well-structured speech
- Deliver with clarity and purpose

Listening comprehension

- Identify key claims hidden in long continuous spoken language
- Compare to personal assistants - simple short commands

Modeling human dilemmas

- Modeling the world of human controversy and
- Enabling the system to suggest principled arguments

Argument retrieval is the first step to build such a system

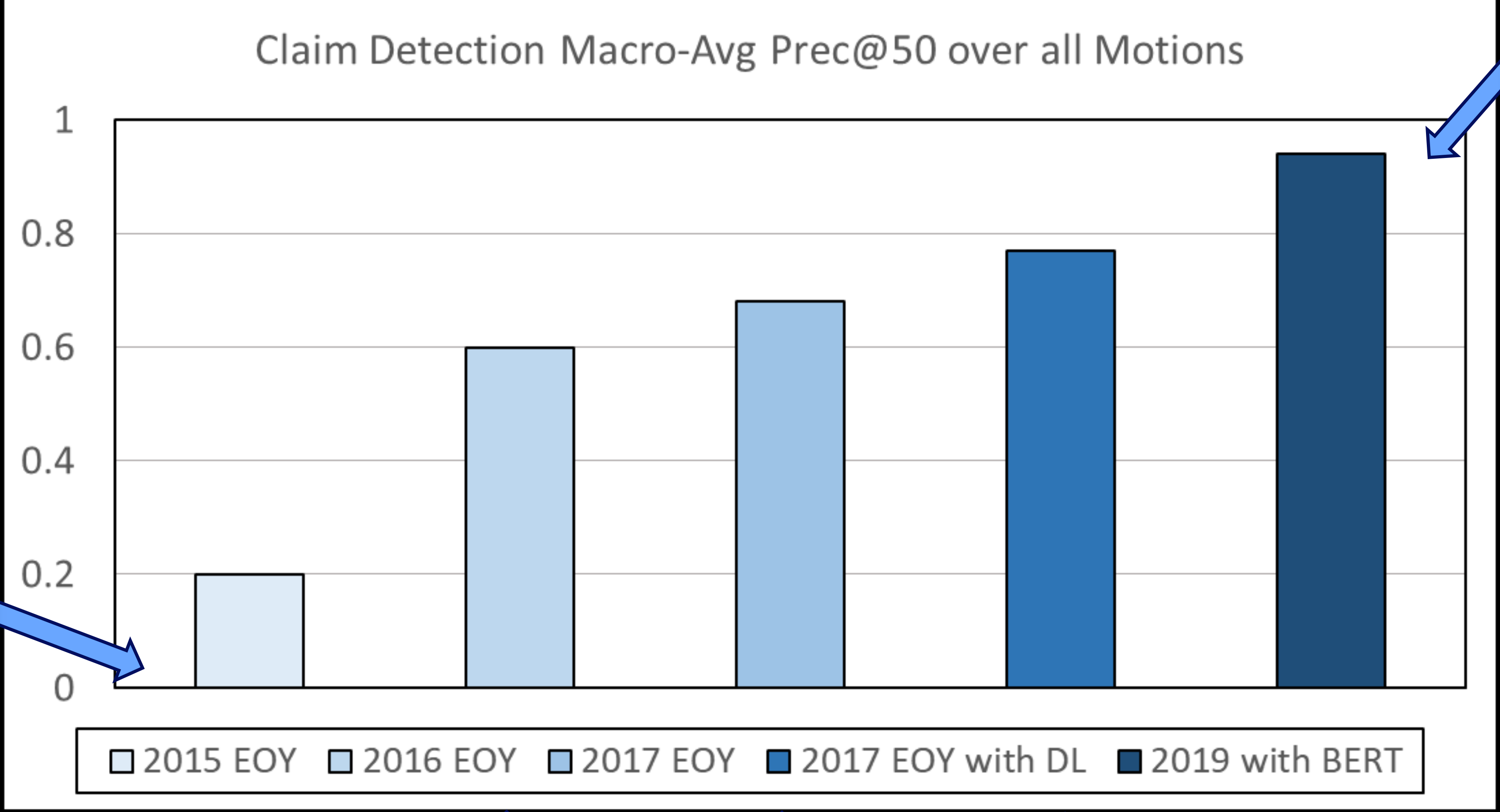
The Problem: Many things need to succeed simultaneously and many things can go wrong…

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
- Corpus: Wikipedia
- Exhaustive labelling of positive instances
- LR + Rich features

Very large corpus
Retrospective labelling

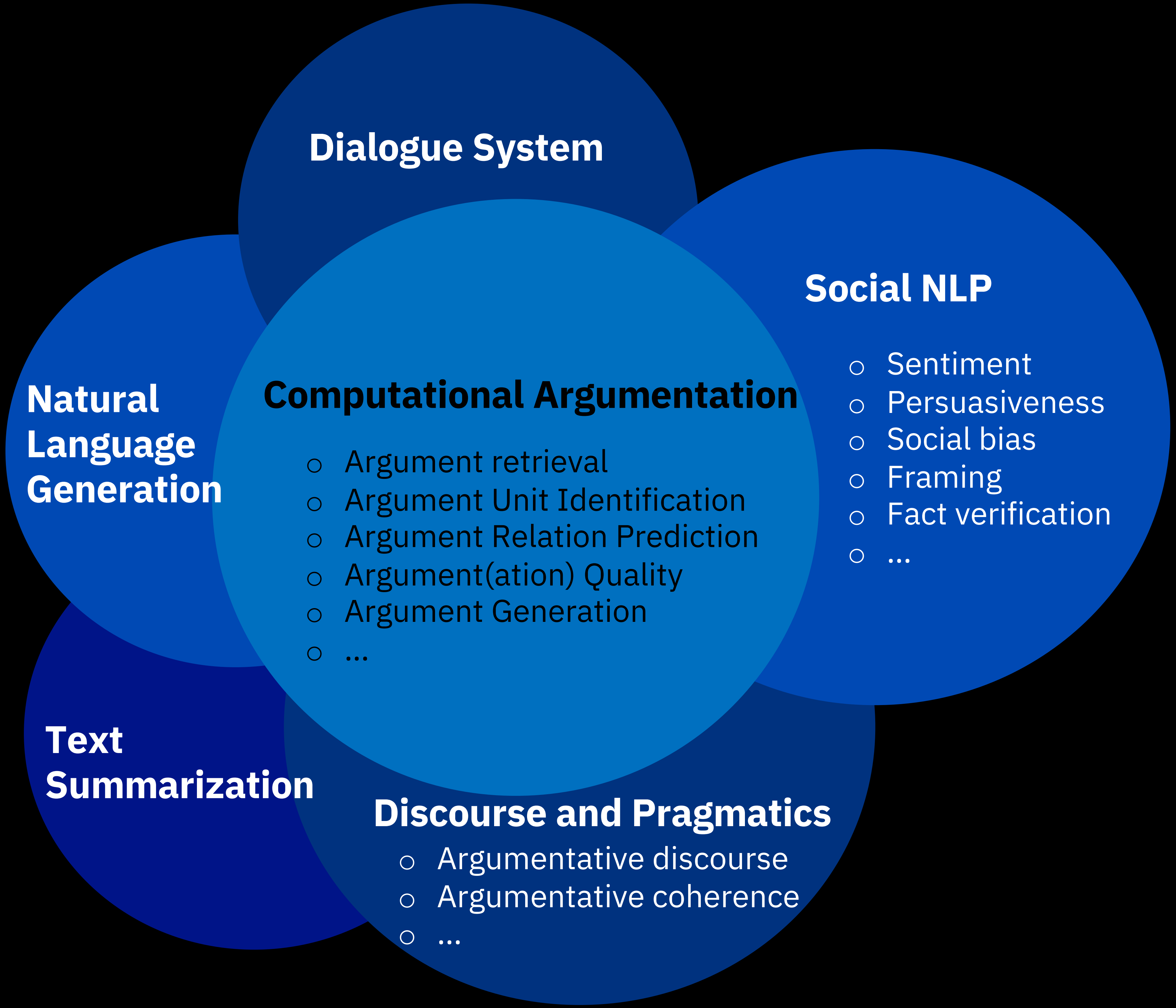
Sentence level IR
Flexible query

Attention-based Bi-LSTM
with weak supervision

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

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



Dialogue System

Natural Language Generation

Text Summarization

Computational Argumentation

- Argument retrieval
- Argument Unit Identification
- Argument Relation Prediction
- Argument(ation) Quality
- Argument Generation
- ...

Social NLP

- Sentiment
- Persuasiveness
- Social bias
- Framing
- Fact verification
- ...

Discourse and Pragmatics

- Argumentative discourse
- Argumentative coherence
- ...

Thanks!

Q&A