Project Debater

Noam Slonim
Distinguished Engineer
Project Debater Principal Investigator

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**)

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

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

Opening - 4 mins (x 2)

Rebuttal - 4 mins (x 2)

Summary - 2 mins (x 2)

Fully automatic debate

No human intervention

How does it work?

Subsidize pre-school

Topic Expansion

Offline Analysis

Online Analysis

Argument Mining

- Data Corpus: ~400M articles
- Cleansing, Wikification, NER …
- Sentence Level Indexing
- Claim Detection
- Evidence Detection
- Stance Classification

Corpus-based arguments

Debate Construction

- Redundancy Removal
- Clustering
- Theme Extraction
- Content Selection / Phrasing
- Expressive Text To Speech

Principled arguments

Argument Knowledge Base

- Detect Argument Class
- Authored Text Selection

Rebutted arguments

Principled leads & responses

Corpus-based leads & responses

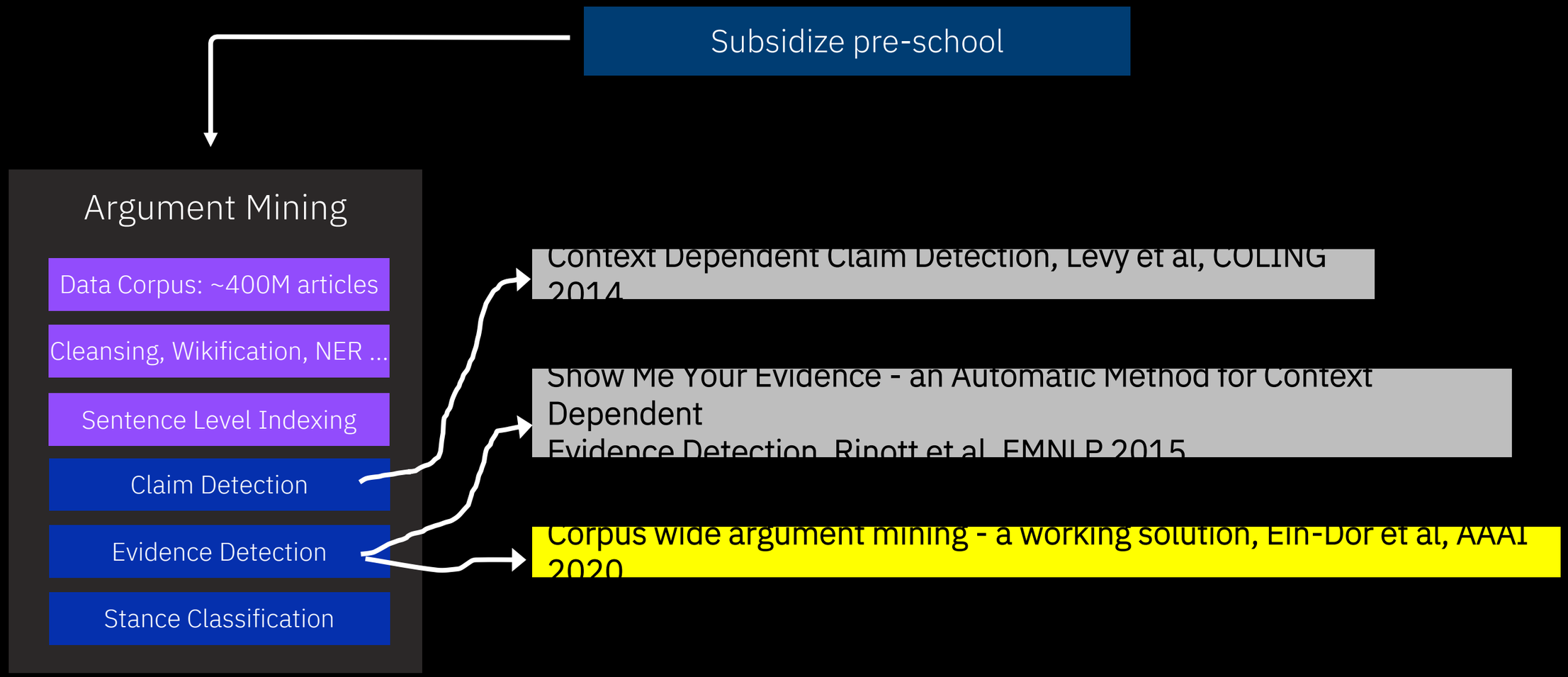
- Opening Speech
- Second Speech
- Summary Speech

Rebuttal

- Speech to Text
- Lead Detection
- Response Selection

Publications and Datasets are available at -

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


Subsidize pre-school
Argument Mining
Data Corpus: ~400M articles
Cleansing, Wikification, NER ...
Sentence Level Indexing
Claim Detection
Evidence Detection
Stance Classification
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
Corpus wide argument mining - a working solution, Ein-Dor et al, AAAI 2020

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

Why Evidence Detection is Hard?

Motion: **We should abandon Valentine's day**

Canadian firm surveyed Canadians and found that 62% agreed that Valentine's Day is a waste of time and money.

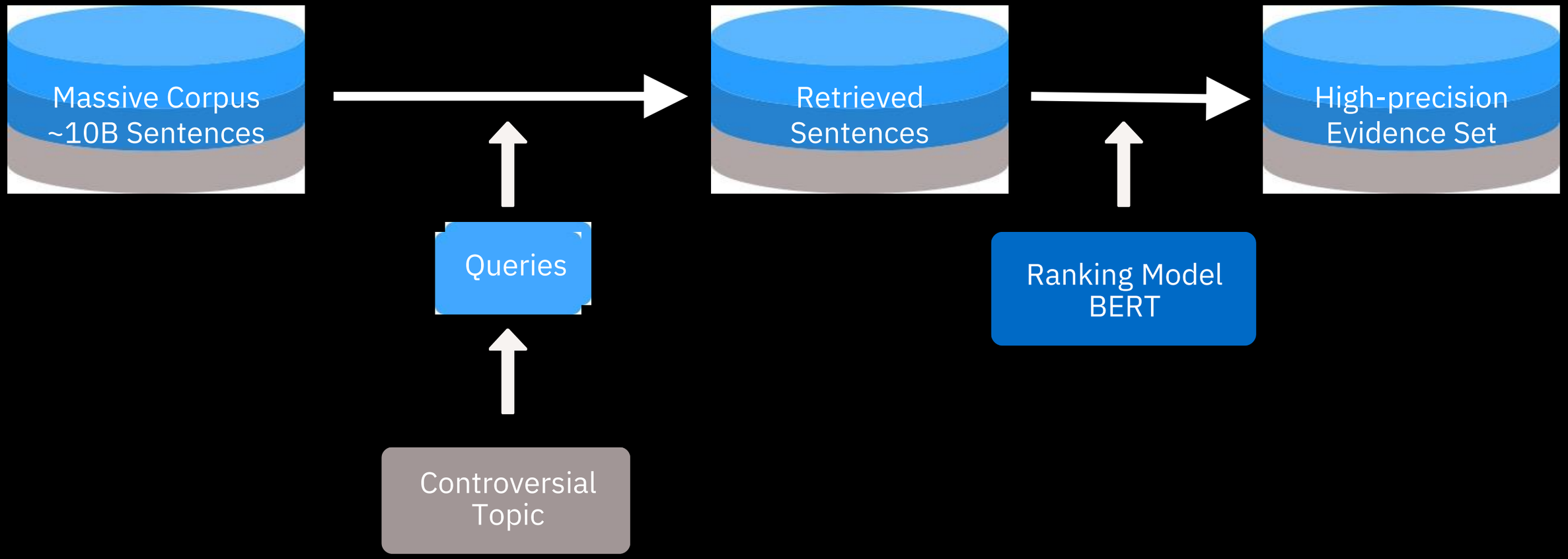
CONFIRMED

US Survey found that 59% of people said that if they were going to break up with someone, they would do so just before Valentine's Day to save money.

REJECTED

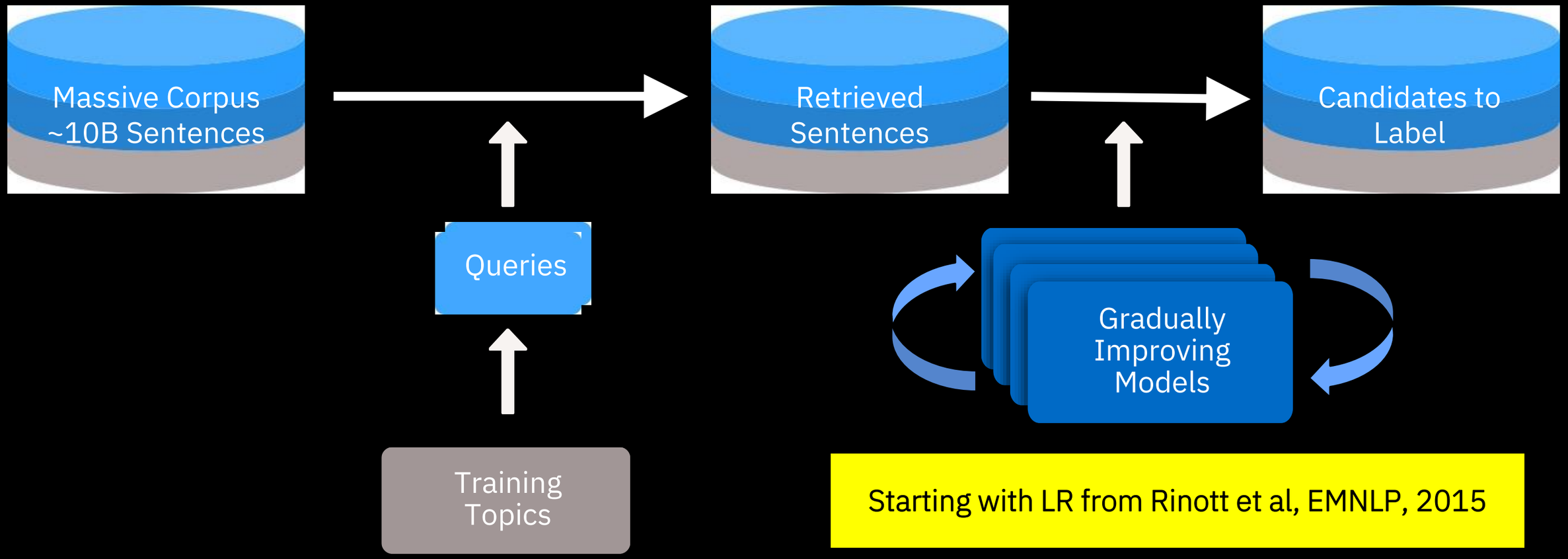
Corpus Wide Argument Mining - a Working Solution / System Architecture

Ein-Dor et al, AAAI 2020



Corpus Wide Argument Mining - a Working Solution / Retrospective Labeling

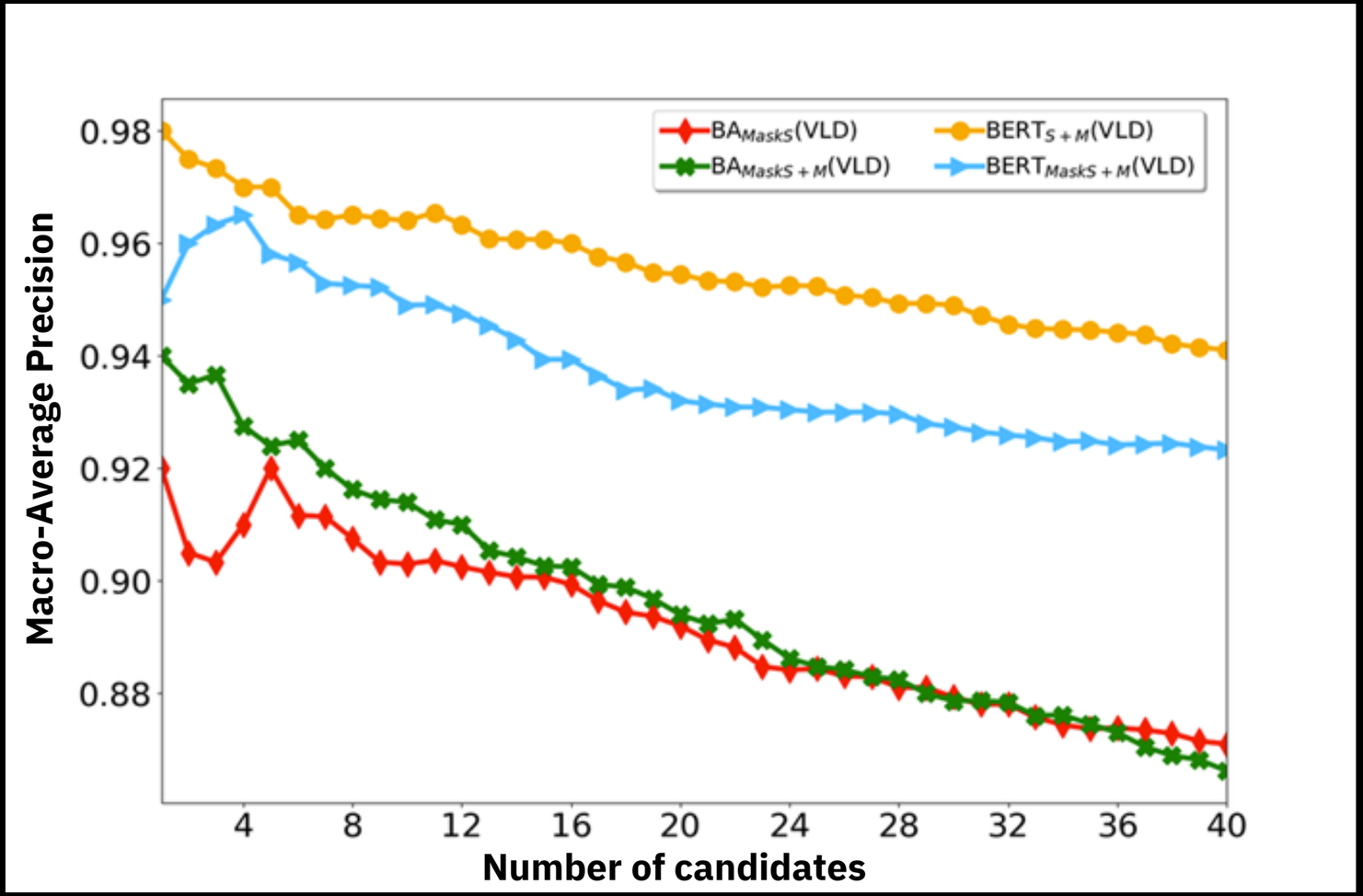
Ein-Dor et al, AAAI 2020



Corpus Wide Argument Mining - a Working Solution / Results

Ein-Dor et al, AAAI 2020

Results by various BERT Models over a massive corpus of ~10B sentences



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

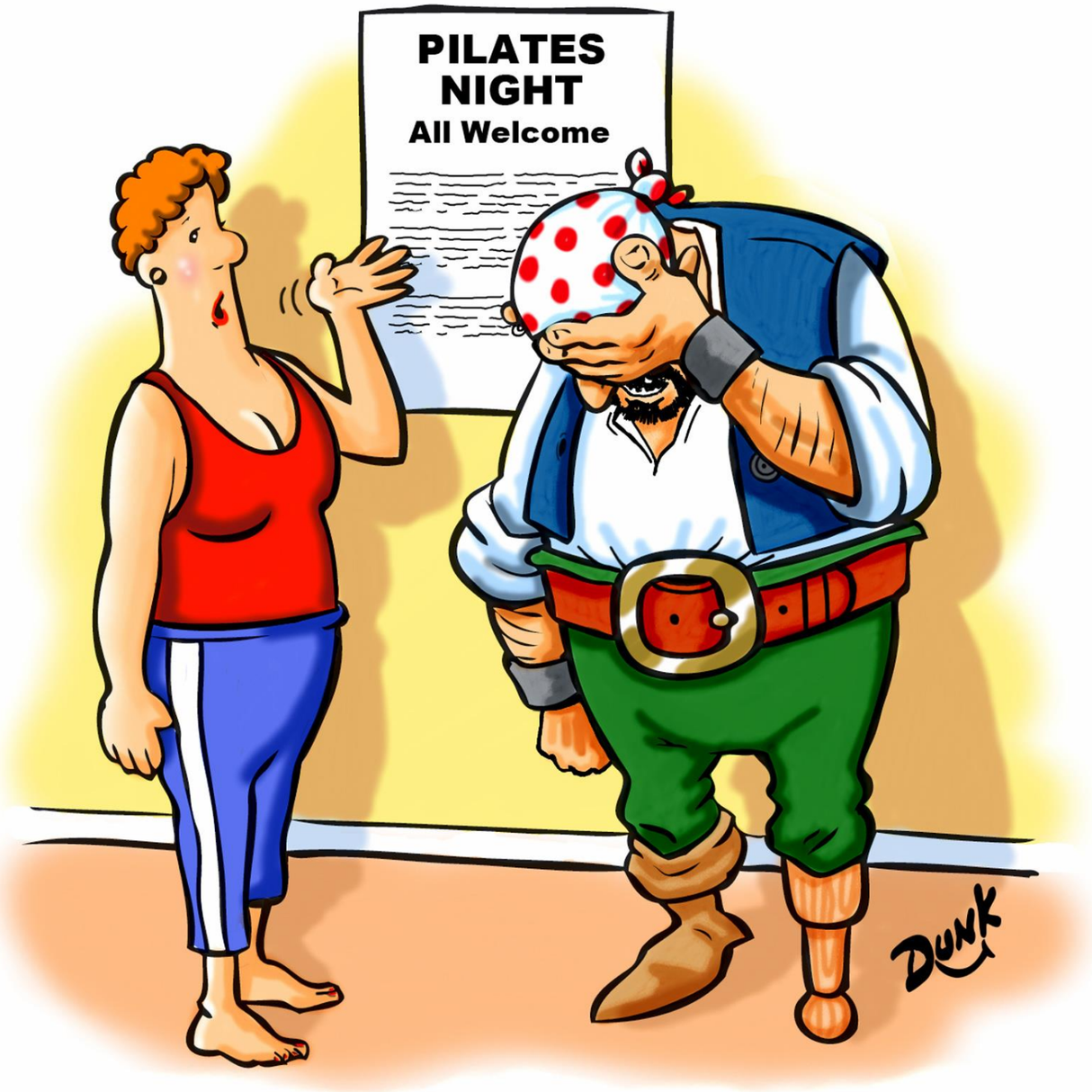
Listening comprehension

- 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

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



"Okay I'll admit you do look foolish but on the positive side you were only one letter out!"

CartoonStock.com

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…*

Many things can go wrong… / Examples - cont.

- Sarcasm is hard to detect
  … University "scientists" published a paper claiming that global warming is the root cause behind the increase in violence we're witnessing …

Many things can go wrong… / Examples - cont.

- Sarcasm is hard to detect
  … University "scientists" published a paper claiming that global warming is the root cause behind the increase in violence we're witnessing …

- Using “topic-tags” may be tricky

  → *People enjoy <gambling> therefore we should attempt to fix it rather than eliminate it…*

Many things can go wrong… / Examples - cont.

- Sarcasm is hard to detect
  … University "scientists" published a paper claiming that global warming is the root cause behind the increase in violence we're witnessing …

- Using “topic-tags” may be tricky

  → *People enjoy <gambling> therefore we should attempt to fix it rather than eliminate it…*

  → *People enjoy <assisted suicide> therefore we should attempt to fix it rather than eliminate it…*

Many things can go wrong… / Examples - cont.

- Sarcasm is hard to detect
  … University "scientists" published a paper claiming that global warming is the root cause behind the increase in violence we're witnessing …

- Using “topic-tags” may be tricky

  → *People enjoy <gambling> therefore we should attempt to fix it rather than eliminate it…*

  → *People enjoy <assisted suicide> therefore we should attempt to fix it rather than eliminate it…*

- Automatic debate-topic expansion is also challenging

  → *Let me discuss a welcome alternative to surrogacy. This is adoption.*

Many things can go wrong… / Examples - cont.

- Sarcasm is hard to detect
  … University "scientists" published a paper claiming that global warming is the root cause behind the increase in violence we're witnessing …

- Using “topic-tags” may be tricky

  → *People enjoy <gambling> therefore we should attempt to fix it rather than eliminate it…*

  → *People enjoy <assisted suicide> therefore we should attempt to fix it rather than eliminate it…*

- Automatic debate-topic expansion is also challenging

  → *Let me discuss a welcome alternative to surrogacy. This is adoption.*

  → *Let me discuss a welcome alternative to global warming. This is global cooling.*

Many things can go wrong… / Examples - cont.

- Sarcasm is hard to detect
  … University "scientists" published a paper claiming that global warming is the root cause behind the increase in violence we're witnessing …

- Using “topic-tags” may be tricky

  → *People enjoy <gambling> therefore we should attempt to fix it rather than eliminate it…*

  → *People enjoy <assisted suicide> therefore we should attempt to fix it rather than eliminate it…*

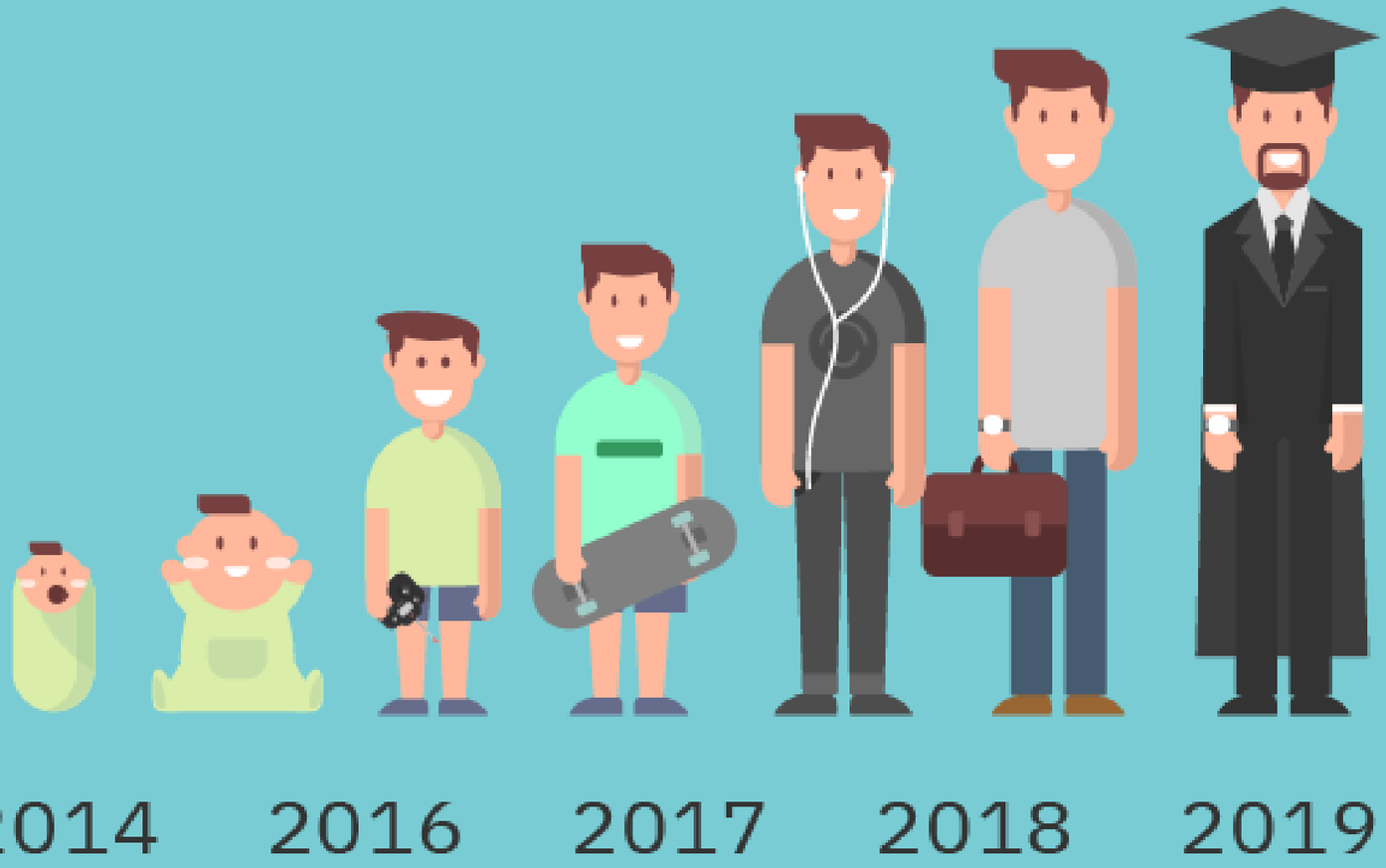
- Automatic debate-topic expansion is also challenging

  → *Let me discuss a welcome alternative to surrogacy. This is adoption.*

  → *Let me discuss a welcome alternative to global warming. This is global cooling.*

  → *Let me discuss an alternative to suicide which has some advantages. This is homicide.*

Progress over time / From Toddler Level to University Level in Three Years



credit: “Vecteezy.com”

How to evaluate an autonomous debating system?

- Public debate approach – the audience votes before and after, and the side who pulled more votes to their side is declared the ‘winner’
- Limitations of this approach –
  - Unbalanced pre-debate vote will increase the burden on the leading side
  - Voting is subjective and affected by various factors that are difficult to quantify and control
  - Producing a live debate with an impartial large audience is complicated
  - Producing many such debates is even more so
- Still – reliable estimation is essential to evaluate system performance, compare to baselines, and track progress over time

An autonomous debating system

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian & Ranit Aharonov

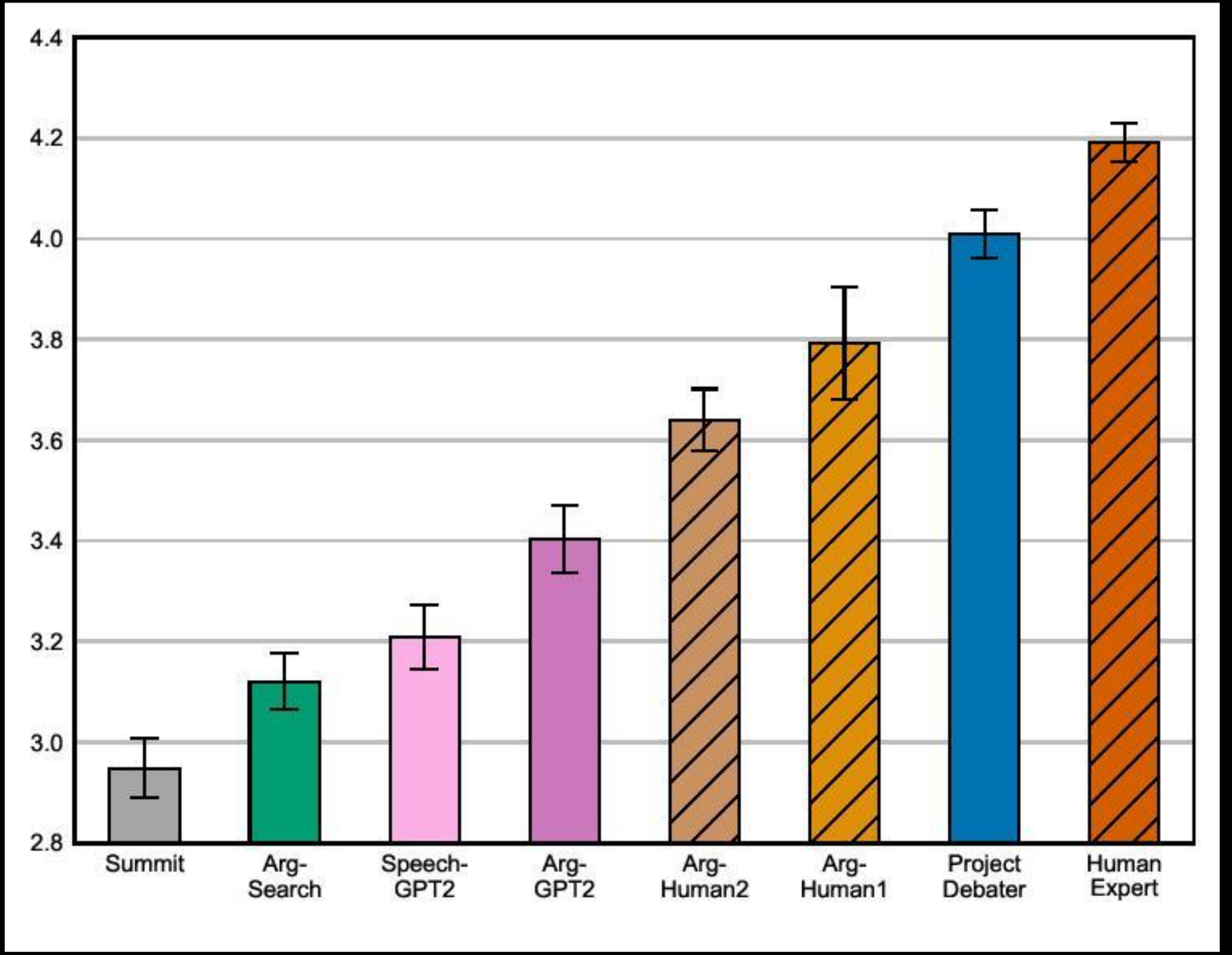
Nature, 2021

Comparison to baseline systems in producing an opening speech

- We are unaware of any other autonomous debating system
- Hence – focused on evaluating the *opening speech*, over an evaluation set of ~80 motions, comparing –
  1. Project Debater
  2. SUMMIIT - Multi-doc summarization (Feigenblat et al, SIGIR 2017)
  3. Speech-GPT2 – GPT2 fine-tuned on ~2k human debate speeches (Data from Orbach et al, ACL 2020)
  4. Arg-GPT2 – based on arguments generated by GPT2, fine-tuned on ~5k high quality human arguments (Gretz et al, EMNLP Findings, 2020)
  5. Arg-Search – based on arguments extracted via Argument-Text (Daxenberger et al, 2020)
  6. Arg-Human1 – based on arguments authored by humans (Data from Gretz et al, AAAI 2020)
  7. Arg-Human2 – based on arguments retrieved by Project Debater and manually curated (Ein-Dor et al, AAAI 2020)
  8. Human – two speeches delivered by expert human debaters

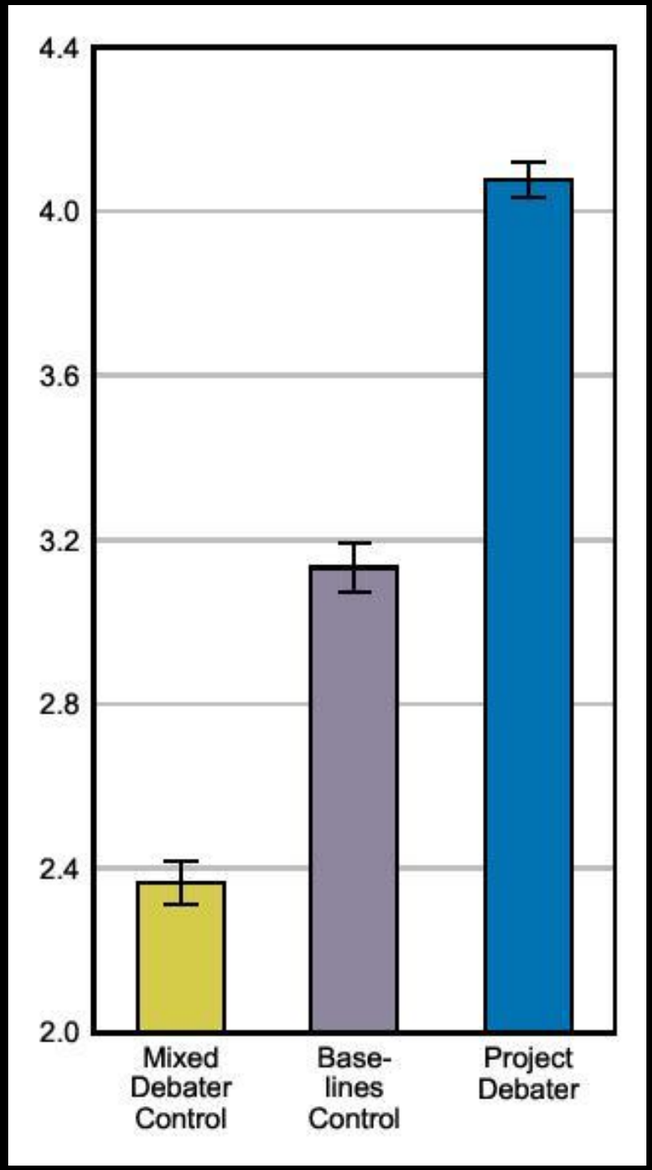
Comparison to baseline systems in producing an opening speech

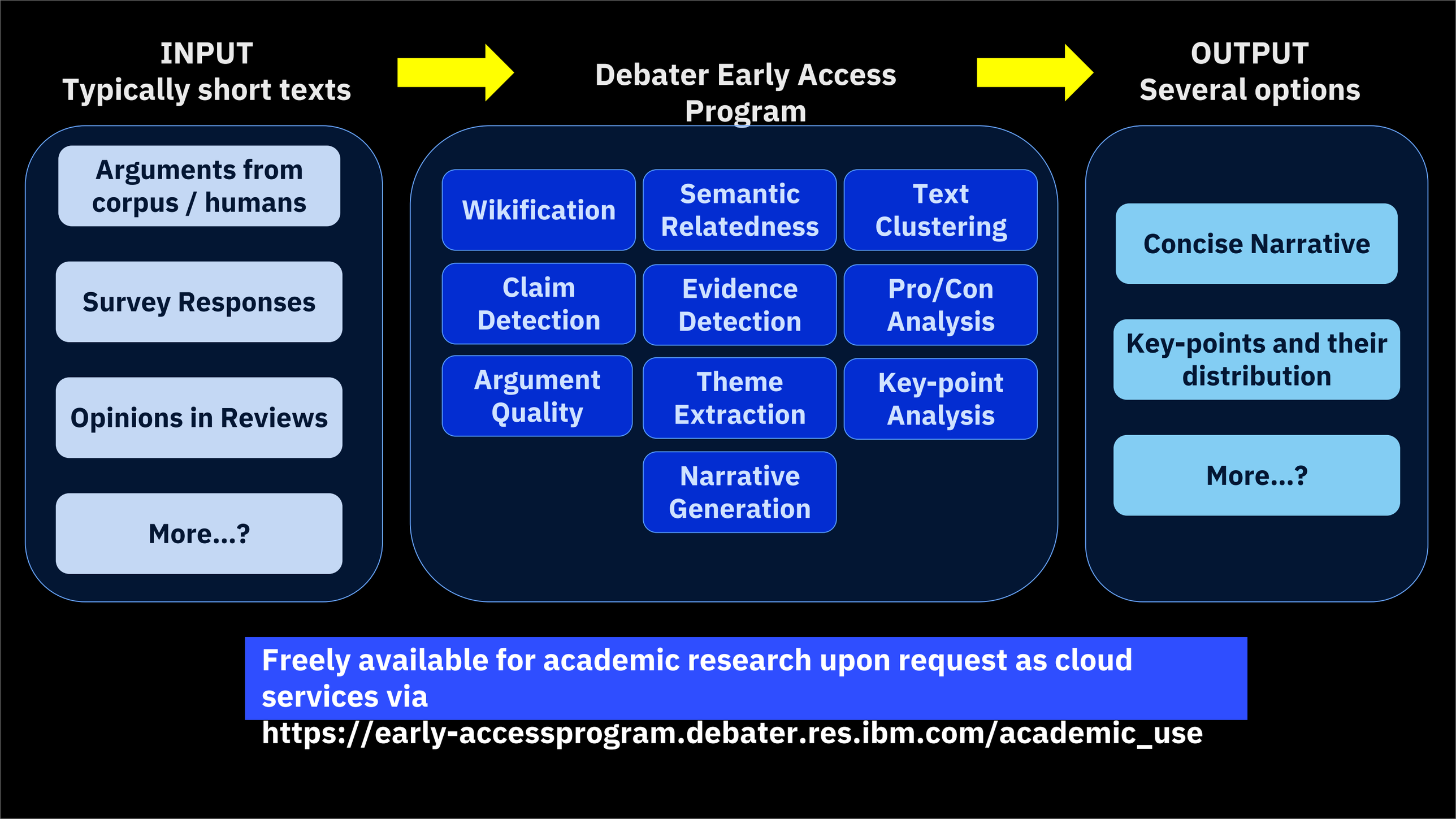
- Each speech was evaluated by 15 experienced crowd annotators on the Appen platform
- Do you Disagree/Agree (from 1 to 5) to the following statement – *This speech is a good opening speech for supporting the topic*



Evaluation of the final system

- Each triplet of speeches was evaluated by 20 experienced crowd annotators on the Appen platform
- Do you Disagree/Agree (from 1 to 5) to the following statement –
  The first speaker is exemplifying a decent performance in this debate
- In 96% of the motions the avg score was > 3
- In 64% of the motions the avg score was >= 4
- Limitations –
  - Considering only S1 and S3
  - Comparing to simple controls, not to human expert
  - Rely on reading as opposed to attending a live debate




INPUT
Typically short texts
Arguments from corpus / humans
Survey Responses
Opinions in Reviews
More...?
Debater Early Access Program
Wikification
Semantic Relatedness
Text Clustering
Claim Detection
Evidence Detection
Pro/Con Analysis
Argument Quality
Theme Extraction
Key-point Analysis
Narrative Generation
OUTPUT
Several options
Concise Narrative
Key-points and their distribution
More...?
Freely available for academic research upon request as cloud services via
https://early-accessprogram.debater.res.ibm.com/academic_use

Why pursue a Grand Challenge?

- **Advancing Science, pushing the boundaries of AI**
  - ~50 papers in EMNLP/ACL/NAACL/EACL & associated workshops
  - Freely available high-quality data sets & Debater Early Access Program
  - Workshops & Tutorials → **Debater tutorial in ACL-2021**
- **Pioneering Research on new problems**
  - Context-Dependent Claim/Evidence Detection:
    Levy et al, COLING, 2014; Rinott et al, EMNLP 2015.
  - Principled Arguments – Bilu et al, ACL 2019.
  - Rebuttal…
- Many use cases of interest to IBM Customers

A lighthouse in the dark

From Checkers to Debate and beyond…

From Checkers to Chess & Go in ~70 years -
All in the ‘comfort zone’ of AI -

- Easy to know who won → facilitating RL techniques
- Moves are well-defined and their values can be quantified → enabling game solving techniques
- Massive available data – e.g., many games played by humans
- AI can win via tactics humans do not comprehend

→ A new territory for AI Grand Challenges?

IBM Research


Thank you!