

Advances in Debating Technologies

Part 1: Introduction

Noam Slonim

IBM Research AI

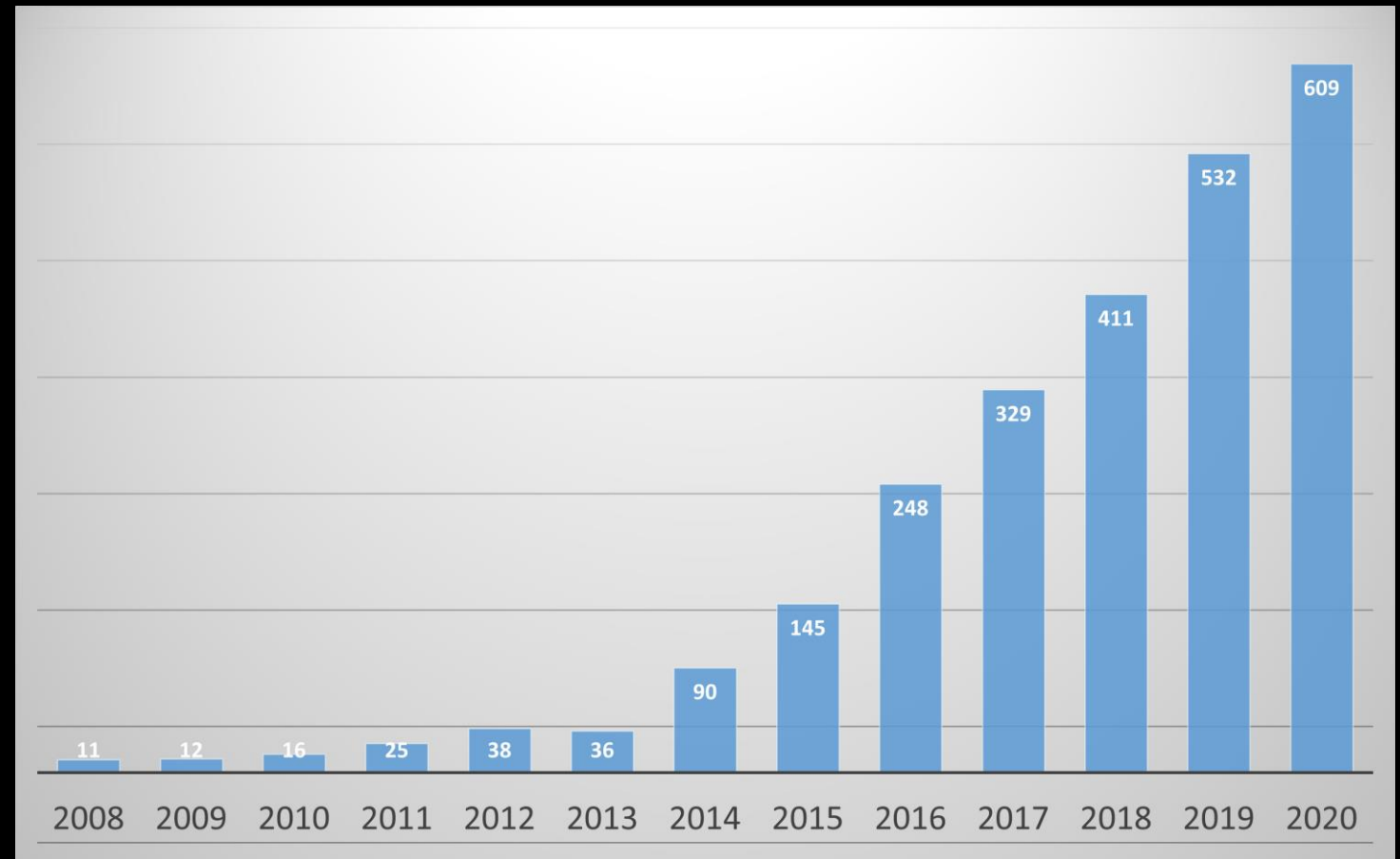


Research AI



Computational Argumentation

- We argue and debate a lot...
- **Computational Argumentation** is defined as “the application of computational methods for analyzing and synthesizing argumentation and human debate”
- And it is a rapidly emerging field...

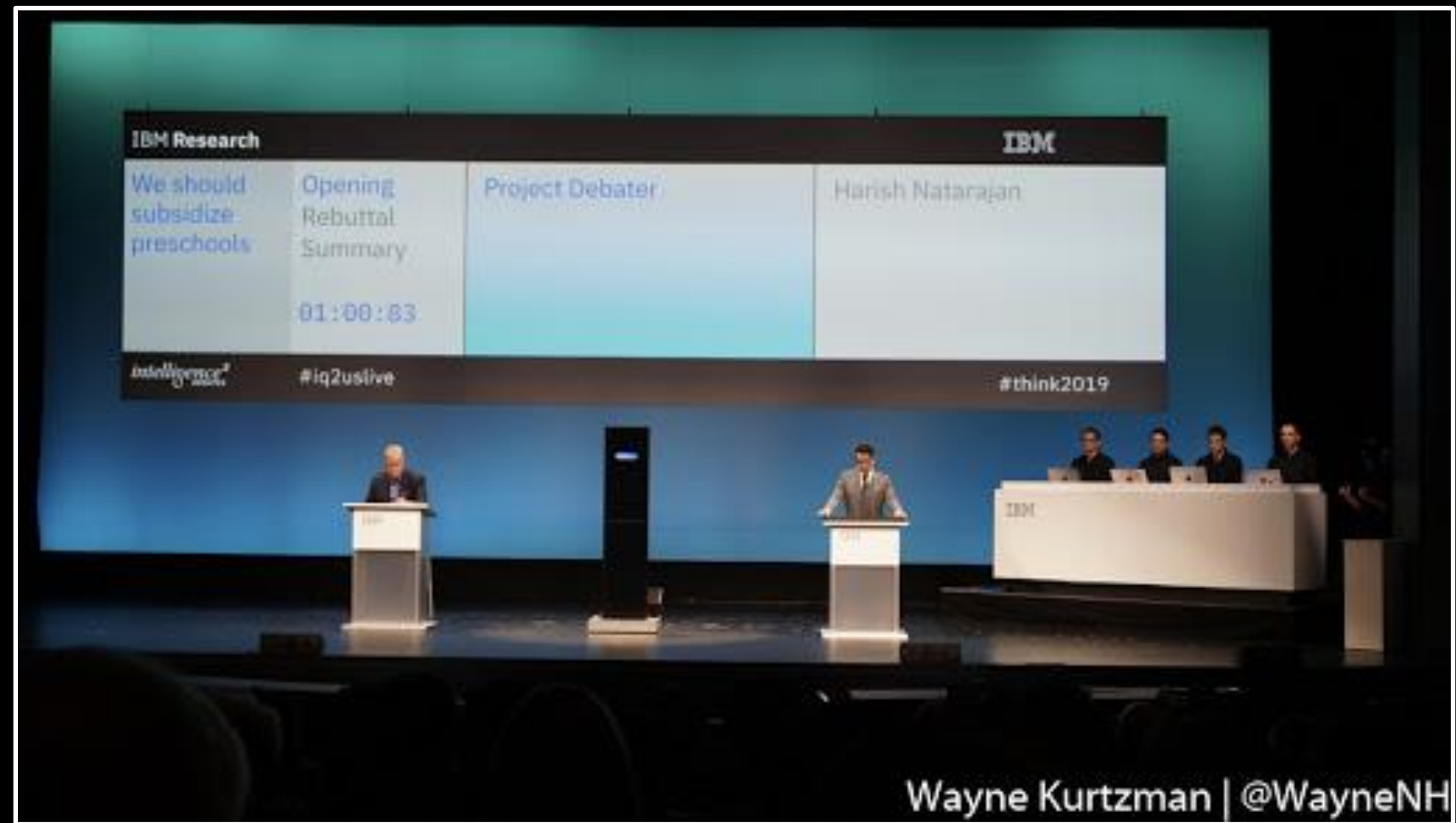


hits in Google Scholar
for “argument(ation) mining”
as of 27.6.21



Debating Technologies and Project Debater

- **Debating Technologies** are computational technologies developed directly to enhance, support, and engage with human debating
- We will focus on **Project Debater** – an AI system that was shown to be able to participate in a full live debate with an expert human debater in a meaningful manner



Outline

- **Part 1 - Introduction**
- **Part 2 - Argument Mining**
- **Part 3 - Argument Evaluation and Analysis**
- **Part 4 - Modeling Human Dilemma**
- **Part 5 - Listening Comprehension and Rebuttal**
- **Part 6 – Basic NLP Capabilities in Project Debater**
- **Part 7 - From Arguments to Narrative**
- **Part 8 - Summary and Moving Forward**
- **Part 9 - Demo Session - Using Debating Technologies in Your Application**



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



WHAT HAPPENS WHEN AI STOPS PLAYING GAMES?

THE DEBATER



AN EPIC PRODUCTION THE DEBATER EDITOR MICHAEL LYNE PRODUCER JOSH GOODIER DIRECTOR OF PHOTOGRAPHY SEBASTIAN MLYNARSKI MUSIC BY CHRISTIAN HANLON
EXECUTIVE PRODUCERS JOHN AND MCDERMOTT PRODUCERS CHRIS SCIACCA STEVE TOMASCO VINEETA DURANI PRODUCER KIANA MOORE DIRECTED BY JOSH DAVIS AND HARRY SPITZER

EPIC



- **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>



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





IBM Research AI - Project Debater

Grand Challenge Debate, San Francisco, Feb. 11, 2019

Motion:

We should subsidize preschools

Selected segments

Arguing for the motion: **Project Debater**

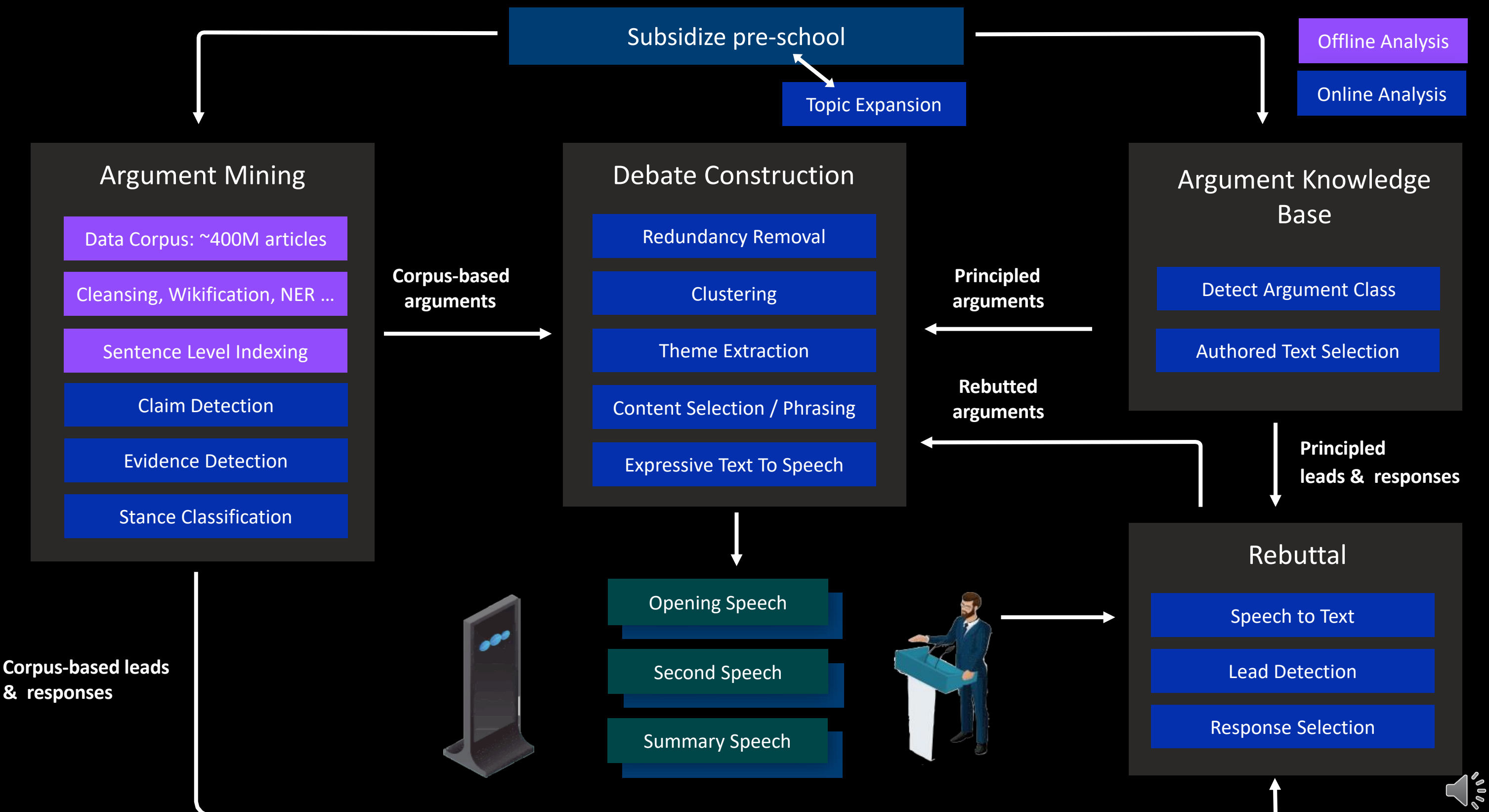
Arguing against the motion: **Mr. Harish Natarajan**

Moderator: **John Donovan, Intelligence Squared US**

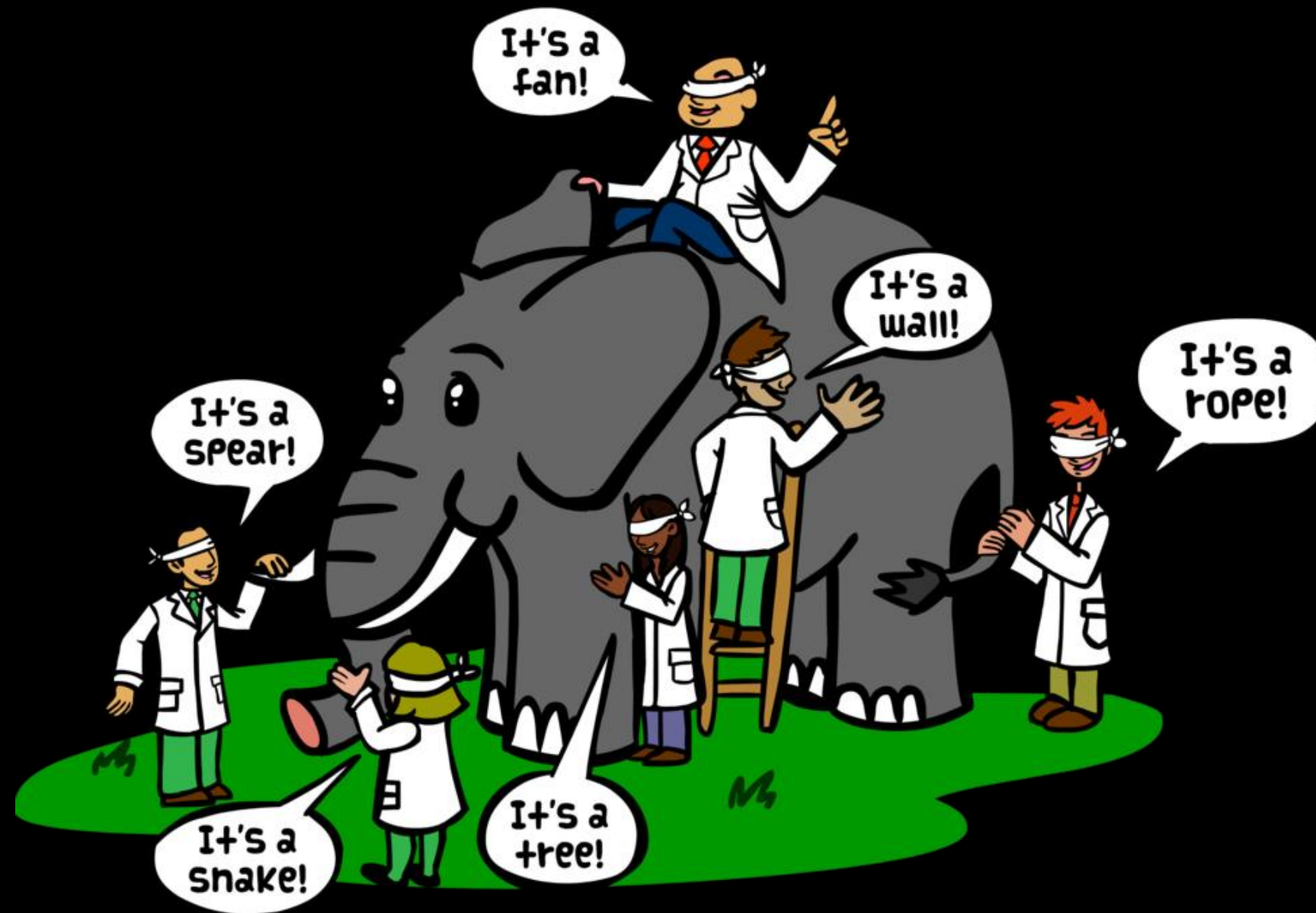


How does it work?



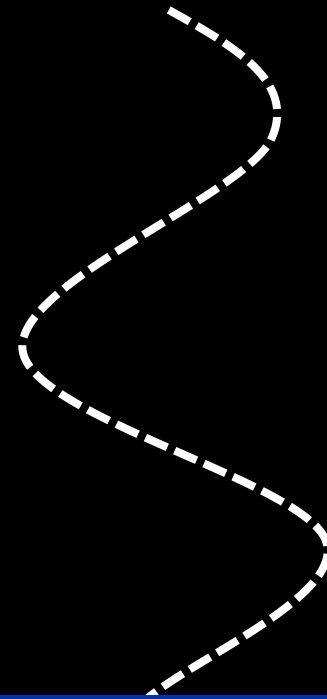


Current Publications Highlight Various Aspects of the System



Subsidize pre-school

INPUT



Expressive Text To Speech



OUTPUT

Opening Speech

Second Speech

Summary Speech



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.

Stance Classification of Context-Dependent Claims, Bar-Haim et al, EACL 2017.



Subsidize pre-school

Argument Mining

Data Corpus: ~400M articles

Cleansing, Wikification, NER ...

Sentence Level Indexing

Claim Detection

Evidence Detection

Stance Classification

Argument Knowledge Base

Detect Argument Class

Authored Text Selection

Inventing Arguments from First Principles, Bilu et al, ACL 2019.



Subsidize pre-school

Offline Analysis

Online Analysis

Argument Mining

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

Argument Knowledge Base

- Detect Argument Class
- Authored Text Selection

A Dataset of General-Purpose Rebuttal
Orbach et al, EMNLP 2019.

Principled
leads & responses

Listening Comprehension over Argumentative Content
Mirkin et al, EMNLP 2018.

Rebuttal

- Speech to Text
- Lead Detection
- Response Selection

Corpus-based leads
& responses

Listening for Claims: Listening Comprehension
using Corpus-Wide Claim Mining, Lavee et al,
Arg-Mining workshop, ACL 2019.

Leads from iDebate



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

Argument Knowledge Base

- Detect Argument Class
- Authored Text Selection

**From Surrogacy to Adoption; From Bitcoin to Cryptocurrency:
Debate Topic Expansion, Bar-Haim et al, ACL 2019.**

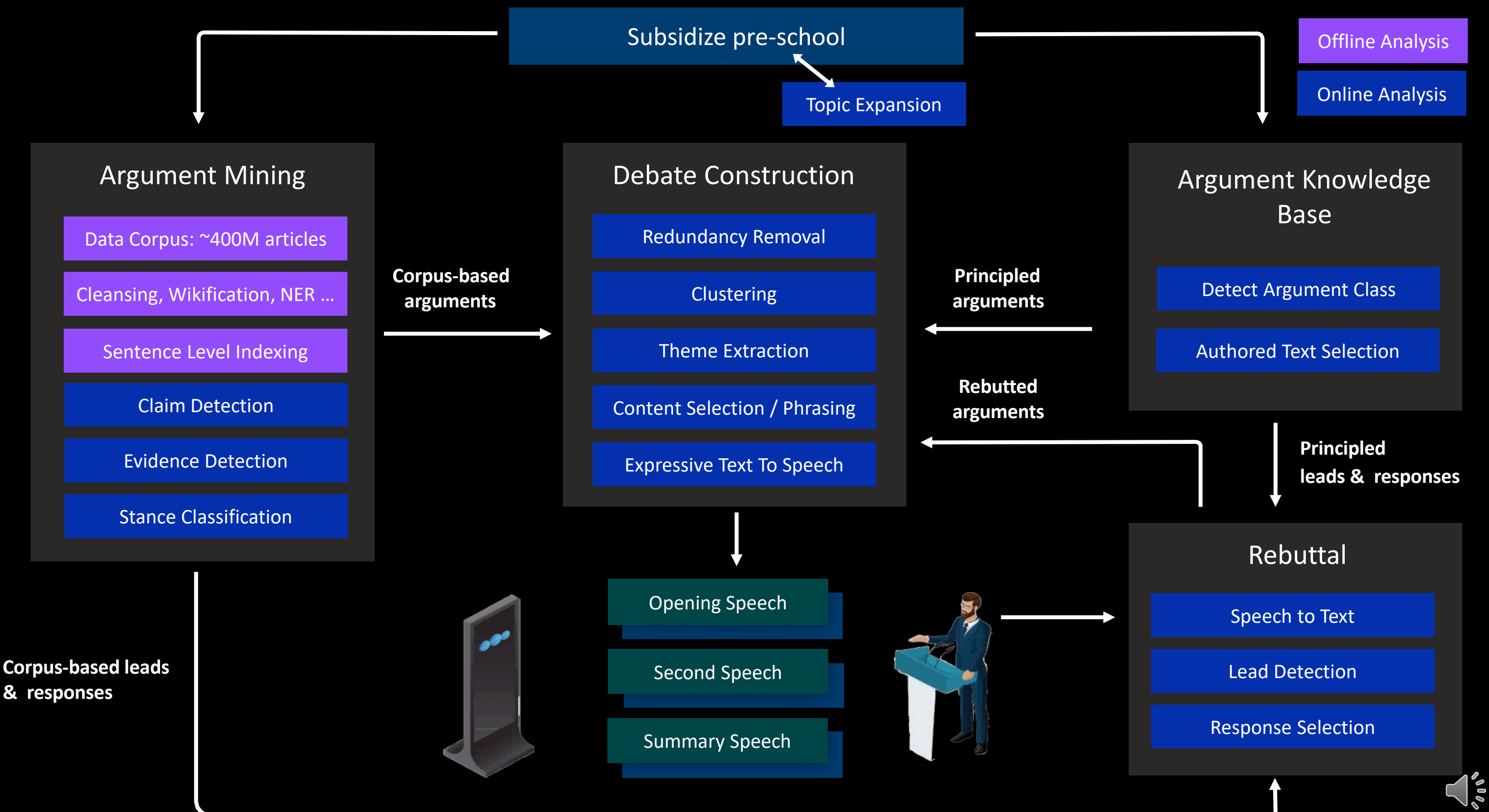
**Principled
leads & responses**

Rebuttal

- Speech to Text
- Lead Detection
- Response Selection

**Corpus-based leads
& responses**





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

Rebutted arguments

Argument Knowledge Base

Detect Argument Class

Authored Text Selection

Principled leads & responses

Rebuttal

Speech to Text

Lead Detection

Response Selection

Opening Speech

Second Speech

Summary Speech

Corpus-based leads & responses



Publications and Datasets are available at -



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



Advances in Debating Technologies 2: Argument Mining

Liat Ein-Dor

IBM Research AI

Argument Mining in Project Debater

Task definition

Given a topic, extract relevant arguments from a massive corpus
(billions of sentences)

Requirements:

- Extremely high precision
- Diverse set of arguments
- Wide range of debatable topics



Identifying Arguments is not Trivial

Task definition

Given a topic, extract relevant arguments from large text corpus



Motion: *Blood donation should be mandatory*

A **study** published in the American Journal of Epidemiology found that **blood donors** have **88-percent** less risk of suffering from a heart attack and stroke.

Statistics from the Nakasero Blood Bank show that students are the main **blood donors** contributing about **80 per cent** of the blood collected countrywide.



Identifying Arguments is not Trivial

Task definition

Given a topic, extract relevant arguments from large text corpus



Motion: *We should abandon Valentine's day*

The Canadian polling firm Insights West **surveyed** a representative sample of Canadians who are in a relationship and found that **62 percent** agreed that **Valentine's Day** is a waste of time and **money**.

A recent **survey** by Virgin Mobile USA found that **59 percent** 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**.



Argument Types

- **Claim** - a concise statement that has a clear stance towards the motion.
- **Evidence** - a single sentence that clearly supports or contests the motion, and provides an indication for whether a relevant belief or a claim is true.
 - **Study:** Evidence that includes a quantitative analysis of data.
 - **Expert:** Evidence that presents testimony by a relevant expert or authority.



Earlier Works

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

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

Solution approach:

- Dataset creation process
 - Relevant Wikipedia articles identified by labelers
 - Article's sentences are annotated
- Sentence classification based on hand-crafted features (logistic regression)



Deep-Learning Based Solutions

Solution Approach:

- Increase data size using sentence-level weak supervision
- DNN architecture: BiLSTM with inner attention

Levy et al, *Towards an argumentative content search engine using weak supervision*, ICCL 2018

Shnarch et al, *Will it Blend? Weak and Manual Labeled Data in a Neural Network for Argumentation Mining*, ACL 2018

Levy et al, *Unsupervised corpus-wide claim detection*, ACL 2017

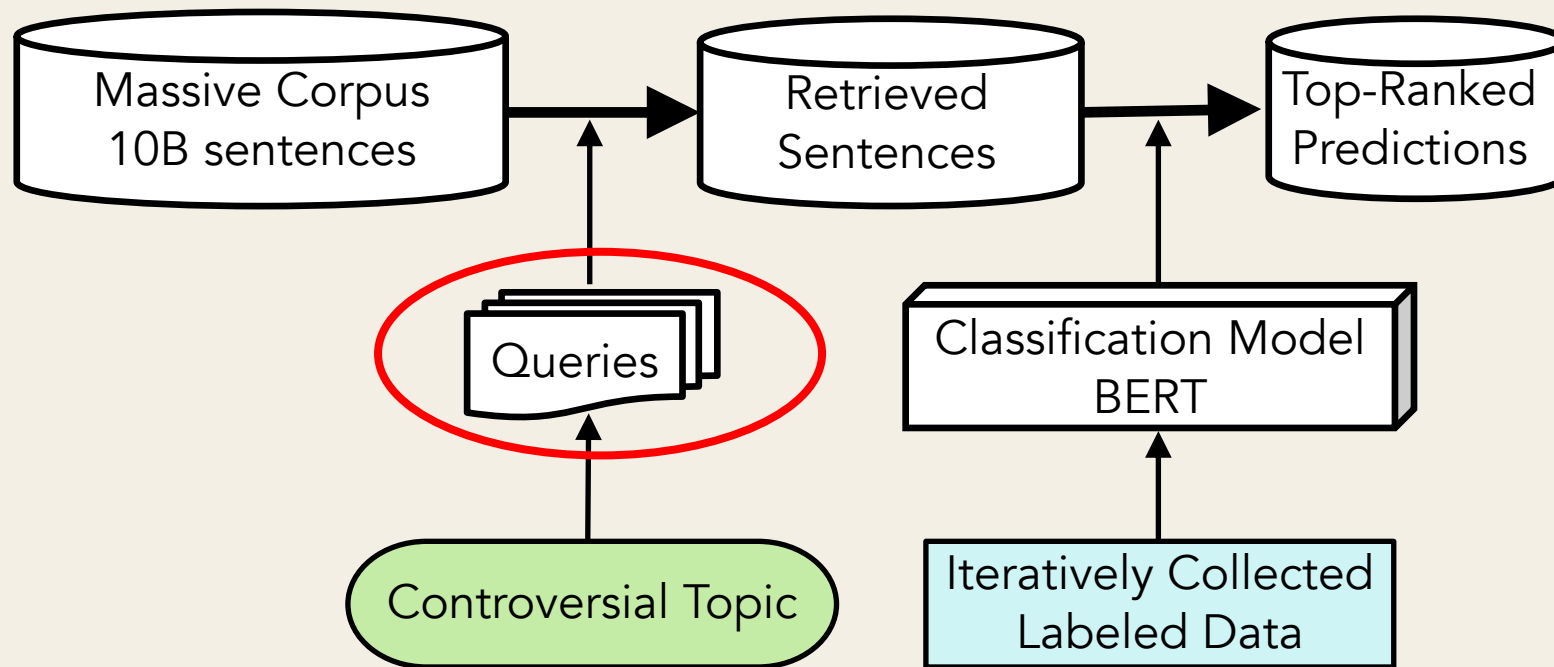


The Current Argument Mining Solution

Corpus Wide Argument Mining - a Working Solution,
Ein-Dor et al, AAAI 2020



The Current Argument Mining System



Massive Corpus – 400 million newspaper and journal articles provided by LexisNexis*

*<https://www.lexisnexis.com/en-us/home.page>



Document Level vs. Sentence Level Queries

	Document Level	Sentence Level
Topic coverage	Only highly controversial	Wide coverage
Content diversity	Limited to sentences appearing in documents that focus on the topic	Not limited to specific documents
Argument type	No control	Allows retrieval of specific types
Data skewness	Selected documents are enriched with argumentative texts.	Highly skewed



Sentence-Level Queries

Aim: Focus on relevant texts

- Dealing with textual data on a massive scale - reduce run-time
- Allows searching for arguments of a desired type

Examples: *Gambling should be banned*

Study lexicon → **that** → **Topic** → **Sentiment lexicon**

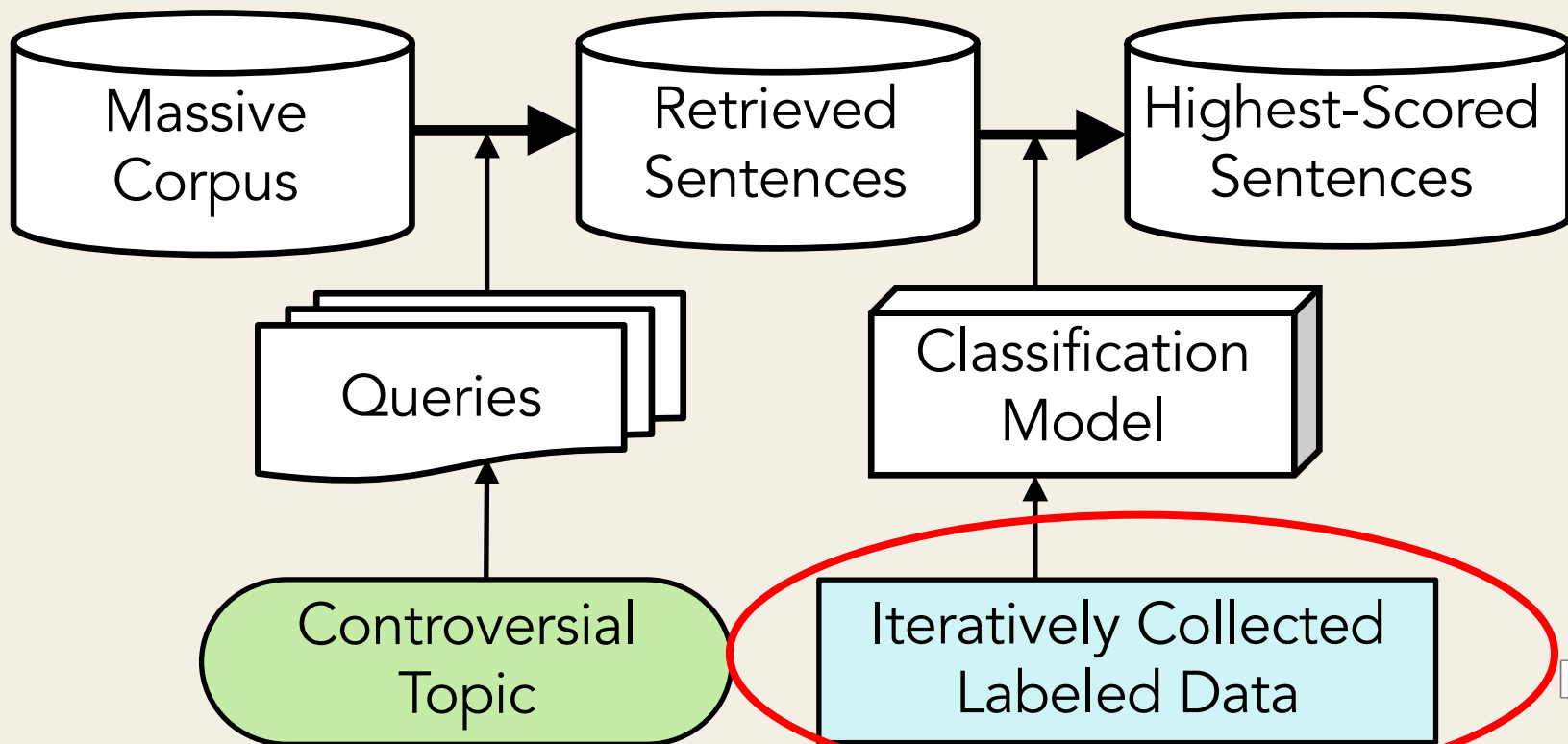
– The University of Glasgow and Healthy Stadia **research** warns **that gambling** is a public health issue with potential for **harm**

Numerical data → **Topic** → **Connector lexicon**

– More Americans now than ever, **69%**, say **gambling** is morally acceptable, **according to** a Gallup poll released this week

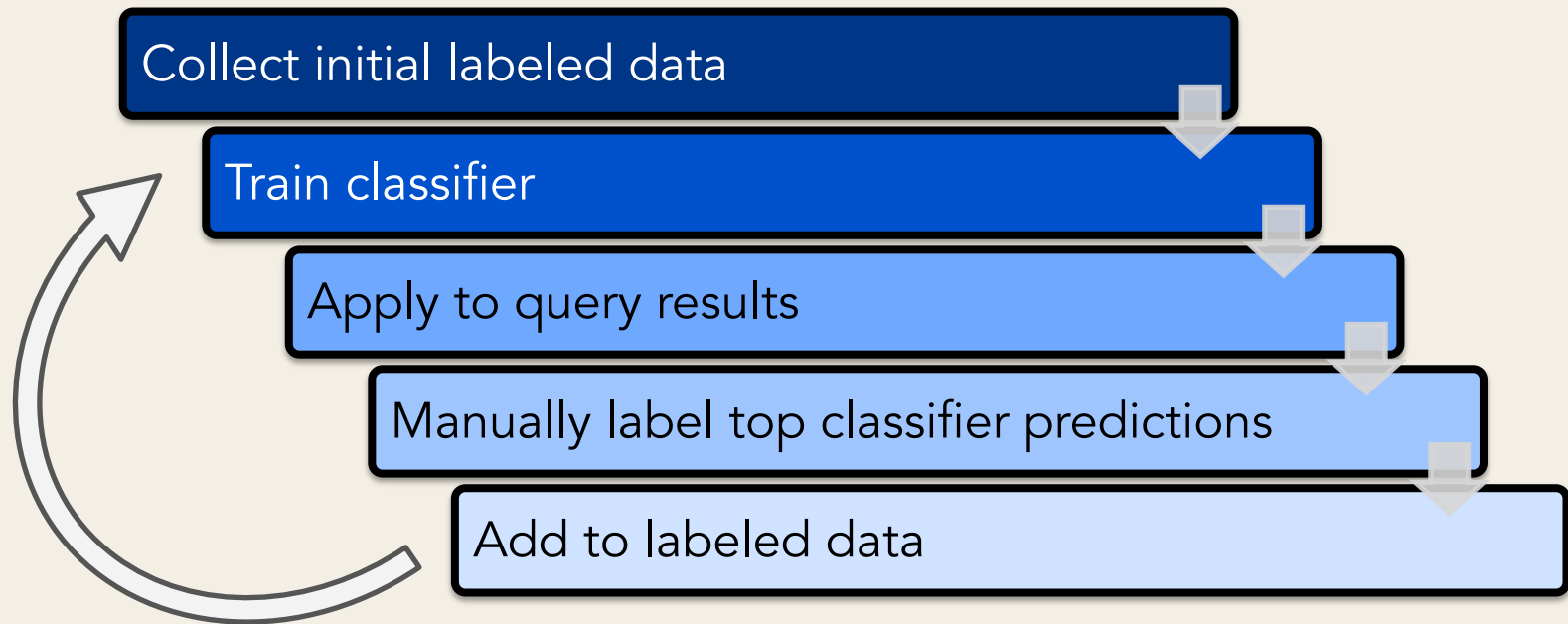


The New Labeling Paradigm



The New Labeling Paradigm

Aim: Overcome the rarity of positive examples to create large balanced data



Resultant Dataset

~200K manually labeled sentences, 33.5% positives

➤ Training set:

- 192 motions
- 154K sentences

➤ Dev set:

- 47 motions
- 45K sentences



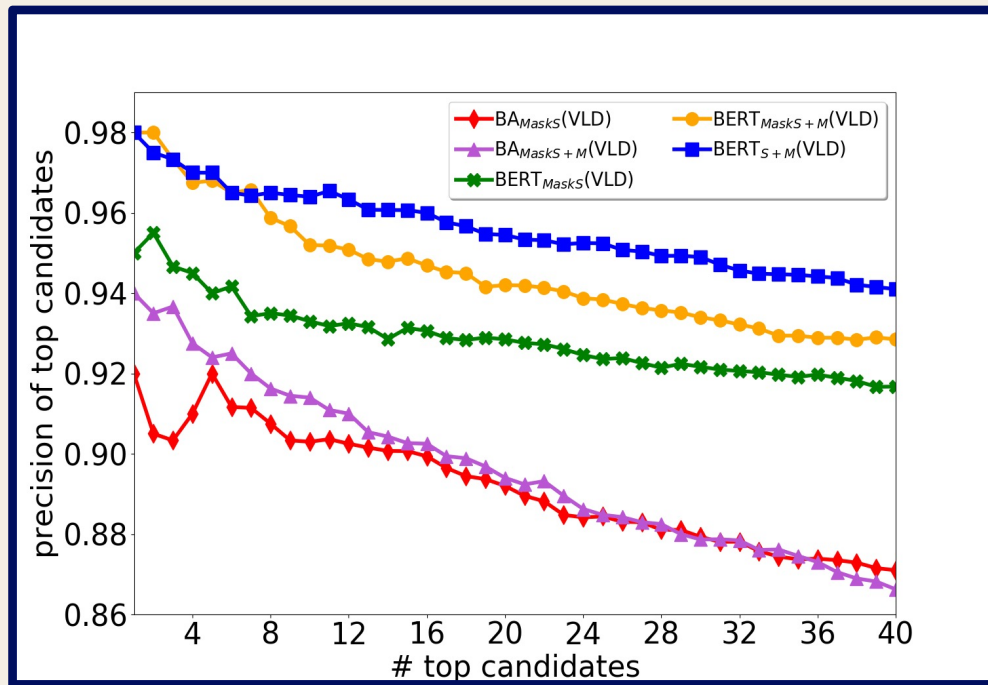
Results (for evidence)

Evaluation:

- Annotate top 40 candidates
- Perform macro average precision over the 100 test motions

Results:

- BERT with sentence & motion outperforms other models.
- High precision on a wide range of topics.
- Similar results for claims.



Blue - BERT with sentence and motion
Red - BiLSTM with masked sentence (baseline)

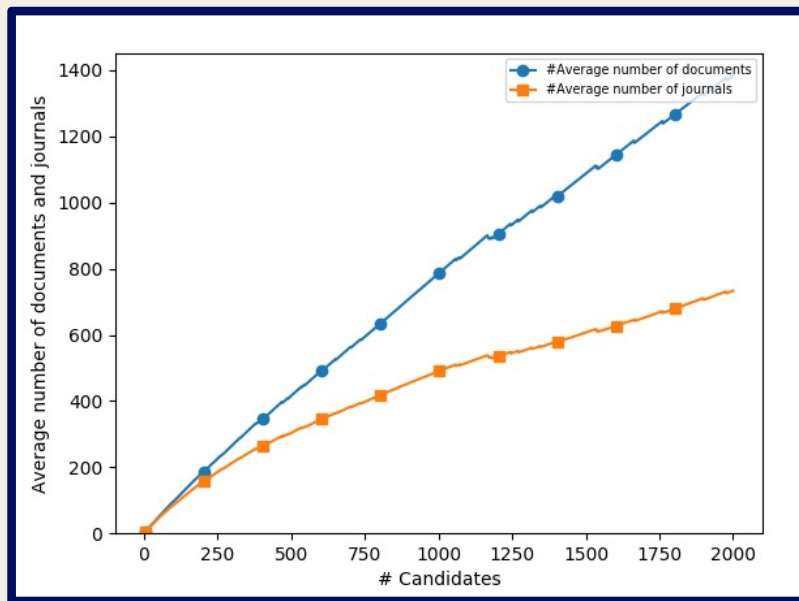


Analysis: Diversity

Do the system arguments originate from diverse documents/journals?

X- # of top-scored system candidates

Y - # of different **documents**, **journals** from which the candidates originate, averaged over 100 test motions



Top 500 sentences originate from ~450,300 different **documents**, **journal**

Debate Topic Expansion: Expanding the Boundaries of Discussion

Bar-Haim et al., ACL 2019

Given a debate topic, find:

Generalizations

Homeopathy \Rightarrow Alternative Medicine

Specializations

Gender inequality \Rightarrow Gender pay gap

Consistent expansions

stance preserving

Alternatives

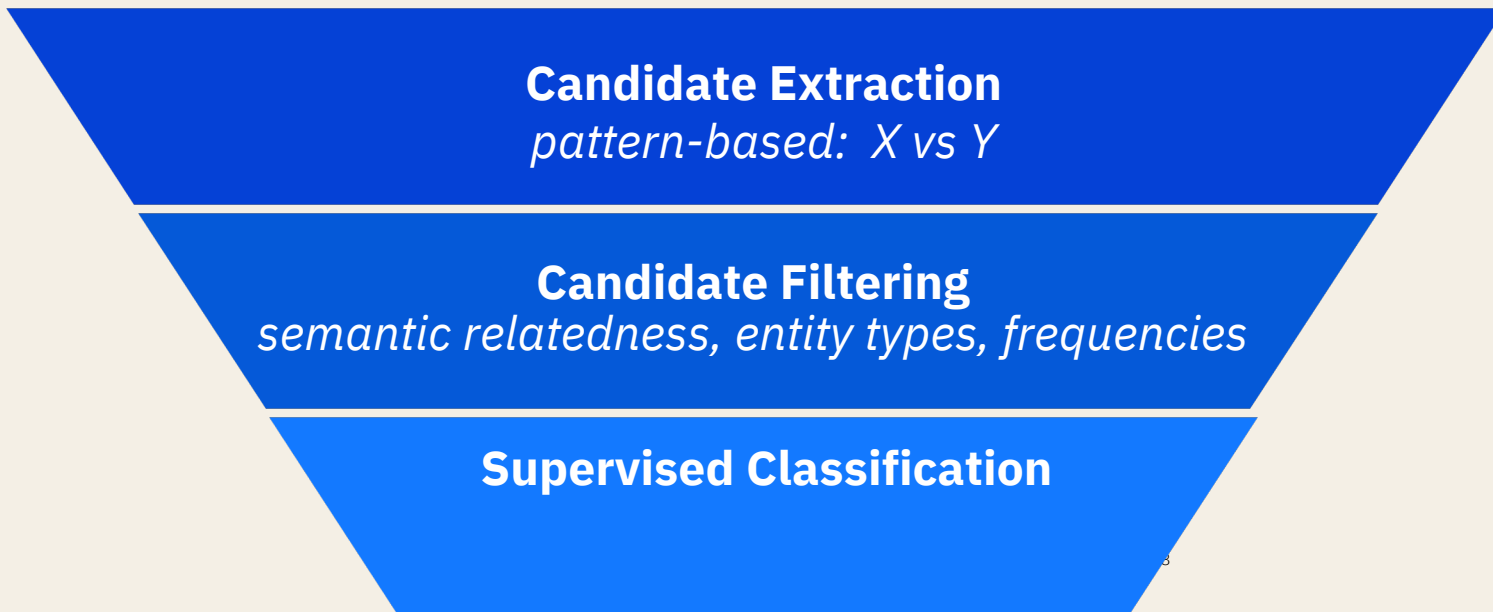
Two-party system \Rightarrow Multi-party system

Contrastive expansions

stance reversing



Debate Topic Expansion: Supervised Classification



Summary

- A first sentence-level end-to-end argument mining solution
- A new labeling paradigm, for coping with skewed label distribution
 - Used to generate a large and balanced argument classification data
- Significant improvements compared to previous models
- High precision on a wide range of topics
- Topic expansion is applied to increase arguments' scope



Related Work - Document-level based solution

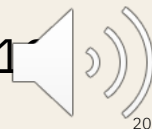
Document retrieval: Top Google-retrieved web pages using topic as query

Largest dataset: UKP corpus (Stab et al, EMNLP 2018)

- 8 controversial topics
- 50 documents per topic
- ~25K sentences
- labeled as argumentative or not

Models:

- Contextual BiLSTM(Stab et al, EMNLP 2018)
- BERT with topic information from external sources (Fromm et al, IEEE 2018)



Related Work – Token-level Argumentation Mining

Method: Classifying tokens into **B**eginning, **I**nside, or **O**utside an argument unit

Example: Nuclear energy may have horrific consequences if an accident occurs, but it has an enormous capacity for energy production with no carbon emissions

Dataset (Trautman et al 2020): 8 topics, 8000 sentences, 4973 arg. segments

Pros: Support more specific selections (aspects, stance)

Cons: Annotation more demanding and precision is lower

- **Ajjour et al**, Unit Segmentation of Argumentative Texts, 2017
- **Trautmann et al**, Fine-grained argument unit recognition and classification, 2020
- **Trautmann et al**, Relational and Fine-Grained Argument Mining, 2020



Recent Reviews

- Cabrio and Villata, IJCAI, 2018
- Stede and Schneider, Synthesis Lectures on HLT, 2018
- Argument Mining: A Survey, Lawrence and Reed, CL, 2019



Advances in Debating Technologies 3: Argument Evaluation and Analysis Stance Classification

Roy Bar-Haim

IBM Research AI

Mostly Based on...

- *Expert Stance Graphs for Computational Argumentation*
Toledo-Ronen et al., ArgMining 2016
- *Stance Classification of Context-Dependent Claims*
Bar-Haim et al., EACL 2017
- *Improving Claim Stance Classification with Lexical Knowledge Expansion and Context Utilization*
Bar-Haim et al., ArgMining 2017
- *Learning Sentiment Composition from Sentiment Lexicons*
Toledo-Ronen et al., COLING 2018
- *SLIDE - a Sentiment Lexicon of Common Idioms*
Jochim et al., LREC 2018

Stance Classification in Project Debater

- Select arguments that support our side
- Arguments with opposite polarity help rebuttal
- Both precision and coverage must be high
- Apply to unseen debate topics



A Brief History

Early Stages

- Complex arguments, manually found in Wikipedia
- Small amount of training data: $O(10^3)$
- Feature-based ML + extensive use of external knowledge

2019 Live Debate

- Simpler arguments, automatically extracted from Lexis Nexis corpus
- Larger amount of training data: $O(10^4)$
- Knowledge-Based + Neural Network

Current Implementation

- Automatically extracted + crowdsourced arguments; automatic data expansion
- Even larger amount of training data: $O(10^5)$
- BERT (Devlin et al., 2019)

Claim Stance Classification

Topic t: *We should embrace the free market*

Claim c: *Government intervention is important for safeguarding against monopoly formation*

Pro or Con?

Decomposing Stance Classification

Topic t: *We should embrace the free market*

Claim c: *Government intervention is important for safeguarding against monopoly formation*

Decomposing Stance Classification

Topic t: *We should embrace the **free market***

Claim c: ***Government intervention** is important for safeguarding against monopoly formation*

- 1. Identify the sentiment targets of topic and claim**

Decomposing Stance Classification

Topic t: *We should embrace the **free market*** 😊

Claim c: ***Government intervention** is important for safeguarding against monopoly formation* 😊

1. Identify the sentiment targets of topic and claim
2. **Identify the sentiment towards the targets**

Decomposing Stance Classification

Topic t: *We should embrace the **free market*** 😊



Claim c: ***Government intervention** is important for safeguarding against monopoly formation* 😊

1. Identify the sentiment targets of topic and claim
2. Identify the sentiment towards the targets
3. **Identify the relation between the targets (consistent/contrastive)**

Decomposing Stance Classification

Topic t: *We should embrace the **free market*** 😊



CON

Claim c: *Government intervention is important for safeguarding against monopoly formation* 😊

1. Identify the sentiment targets of topic and claim
2. Identify the sentiment towards the targets
3. Identify the relation between the targets (consistent/contrastive)
4. **Compute stance (negative sentiment and contrastive targets flip polarity)**

Decomposing Stance Classification

Topic t: *We should embrace the **free market*** 😊

CON

Claim c: *Government intervention is important for safeguarding against monopoly formation* 😊

1. Identify the sentiment targets of topic and claim
2. Identify the sentiment towards the targets
3. Identify the relation between the targets (consistent/contrastive)
4. Compute stance (negative sentiment and contrastive targets flip polarity)

⇒ **Dataset of 2,394 claims for 55 topics, annotated for the above subtasks**

Implementation: Target Extraction

classification of candidate phrases
Logistic Regression classifier with
syntactic, semantic & sentiment features

*Government intervention is important for safeguarding
against monopoly formation*

Implementation: Targeted Sentiment Analysis

using lexicons of *positive* and *negative* sentiment terms, and polarity *shifter* lists

Government intervention is important for safeguarding against monopoly formation

Implementation: Targeted Sentiment Analysis

using lexicons of *positive* and *negative* sentiment terms, and polarity *shifter* lists

Government intervention is important for safeguarding against monopoly formation

Knowledge Acquisition for Sentiment Analysis

Unigram/Bigram Sentiment

Automatic expansion of sentiment lexicon

approachable

overcrowded

forward thinking

ill mannered

Sentiment Composition

Learn lexical classes for sentiment composition

Reversers:

treat  

treat **cancer**

Propagators:

powerful  

powerful **adversary**

Idiom Sentiment

5,000 frequent idioms annotated via crowdsourcing

under fire

make a difference

over the moon

wide of the mark

And Now IBM Watson Understands Idioms...

two thumbs up

■ Neutral Entity ■ Positive Entity ■ Negative Entity

Edit Text

Sentiment

Analyze the sentiment toward specific target phrases and the sentiment of the document as a whole.

[Learn More](#) →

Extraction Classification Linguistics Custom

Sentiment Emotion Categories

Full Document POSITIVE 0.76

Keyword Sentiment Scores

thumbs POSITIVE 0.76

JSON</>

FEEDBACK

Implementation: Relation Classification

- Random forest classifier

Features:

- Lexical-semantic relations indicating similarity/contrast (WordNet, word2vec, PILSA...)
- Pattern-based relation mining from query logs:

“free market vs government intervention” ⇒ **contrastive**



Stance Classification of Expert Opinions with Background Knowledge

–Is this argument for or against ***Atheism***?

XXXXXX sums up his argument and states, "The temptation (to attribute the appearance of a design to actual design itself) is a false one, because the designer hypothesis immediately raises the larger problem of who designed the designer. The whole problem we started out with was the problem of explaining statistical improbability. It is obviously no solution to postulate something even more improbable."

Stance Classification of Expert Opinions with Background Knowledge

–Is this argument for or against ***Atheism***?

Dawkins sums up his argument and states, "The temptation (to attribute the appearance of a design to actual design itself) is a false one, because the designer hypothesis immediately raises the larger problem of who designed the designer. The whole problem we started out with was the problem of explaining statistical improbability. It is obviously no solution to postulate something even more improbable."

The God Delusion, page 158

–Wikipedia says: “Dawkins is a noted **atheist**”

Extracting Expert Stance from Wikipedia Categories



Richard Dawkins
From Wikipedia, the free encyclopedia

For the archaeologist, see Richard MacGillivray Dawkins.

Clinton Richard Dawkins FRS FRS (born 26 March 1941) is an English ethologist, evolutionary biologist and writer. He is an emeritus fellow of New College, Oxford, and was the University of Oxford's Professor for Public Understanding of Science from 1995 until 2008.

Categories: Richard Dawkins | 20th-century biologists | 20th-century English writers | 21st-century biologists | 21st-century English writers | 1941 births | Alumni of Balliol College, Oxford | **Antitheists** | **Atheism activists** | Atheism in the United Kingdom | **Atheist feminists** | British secularists | **Critics of alternative medicine** | **Critics of creationism** | Critics of postmodernism | **Critics of religions** | Education activists | English activists | **English atheists** | English biologists | English feminists | English humanists | English memoirists | English republicans | English sceptics | English science writers | Ethologists | Evolutionary biologists | Evolutionary psychology | Fellows of New College, Oxford | Fellows of the Committee for Skeptical Inquiry | Fellows of the Royal Society | Fellows of the Royal Society of Literature | Former Anglicans | Living people | Male feminists | Memes | New College of the Humanities | People educated at Oundle School | People from Nairobi | Psychology writers | Recipients of the Medal of the Presidency of the Italian Republic | Science activists | Shakespeare Prize recipients | Simonyi Professors for the Public Understanding of Science | Stroke survivors | University of California, Berkeley faculty

Richard Dawkins
At Cooper Union, New York City (2010)

Born Clinton Richard Dawkins
26 March 1941 (age 75)
Nairobi, British Kenya

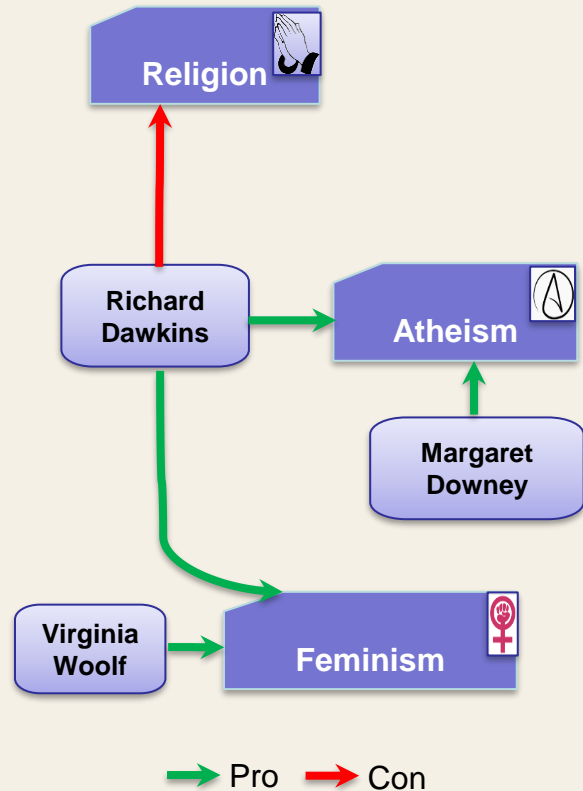
Education Oundle School

Alma mater Balliol College, Oxford
(BA, DPhil)

Thesis *Selective pecking in the domestic chick* (1967)

Constructing an Expert Stance Graph

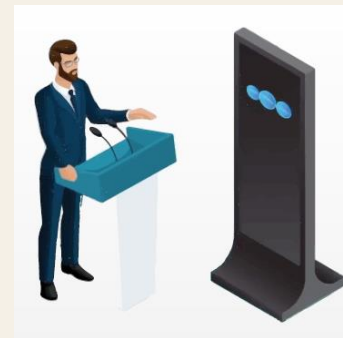
- Extract **Pro** and **Con** categories for controversial Wikipedia concepts
 - Manual annotation of automatically-extracted categories
- Persons (“experts”) in a category get its stance
- 114 concepts \Rightarrow 3.5K categories \Rightarrow 104K expert stances
- Rule-based classification of category names



Live Debate Implementation (2019)

Detect Target and Relation

- All arguments mention the debate concept (DC) \Rightarrow use Wikification to find it
- Rule-based identification of contrastive expressions around the DC:
 - The **alternatives** to **fossil fuel** are more expensive



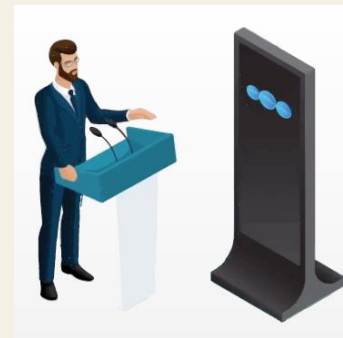
Live Debate Implementation (2019)

Detect Target and Relation

- All arguments mention the debate concept (DC) \Rightarrow use Wikification to find it
- Rule-based identification of contrastive expressions around the DC:
 - The *alternatives* to *fossil fuel* are more expensive

Targeted Sentiment Analysis

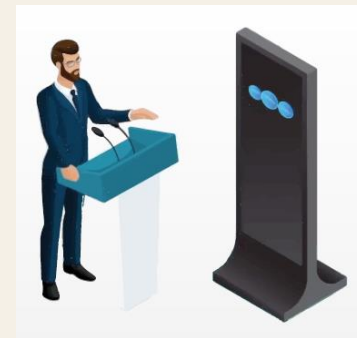
- Two sentiment analyzers:
 1. Knowledge-based (KB)
 2. DNN (GRU + Attention), enhanced with KB features (token and sentence-level)
- For claims use DNN only; for evidence KB and DNN must agree on stance



Live Debate Implementation (2019)

Analyze the internal structure of evidence arguments (rule-based)

A study conducted in Germany by von Salisch found no evidence that violent games caused aggression in minors

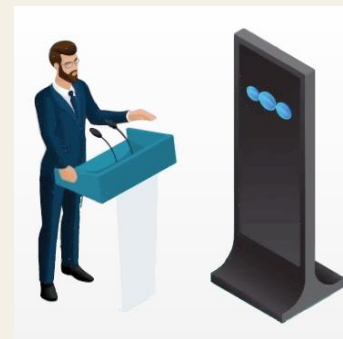


Live Debate Implementation (2019)

Analyze the internal structure of evidence arguments (rule-based)

*A study conducted in Germany by von Salisch found no evidence that **violent games caused aggression in minors***

Only analyze the sentiment of this part



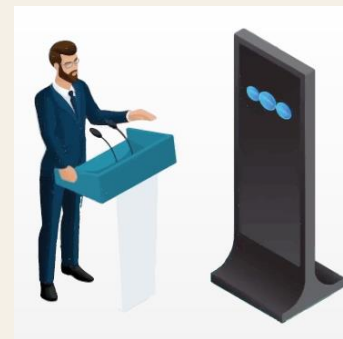
Live Debate Implementation (2019)

Analyze the internal structure of evidence arguments (rule-based)

Flip polarity

*A study conducted in Germany by von Salisch found **no evidence** that **violent games caused aggression in minors***

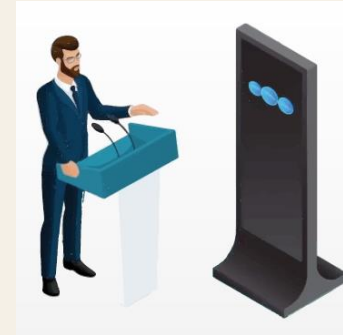
Only analyze the sentiment of this part



Live Debate Implementation (2019)

Performance - precision at coverage of 90%:

Claims	Expert Evidence	Study Evidence
0.95	0.90	0.91



Current Implementation

- We fine-tune BERT on argument-topic pairs, in two phases:
 - **Phase I** (~360K pairs): automatically-extracted arguments, augmented with topic expansion data [Bar-Haim et al., ACL 2019]
 - **Phase II** (~26K pairs): human-contributed arguments

- Accuracy of **~92%** on human-contributed arguments

Related Work : Stance Classification of...

Congressional floor debates

- Thomas et al., 2006; Yessenalina et al., 2010; Burfoot et al., 2011

Debating forums

- Somasundaran and Wiebe, 2009 & 2010; Walker et al., 2012a; Hasan and Ng, 2013

Public comments on proposed regulations

- Kwon et al., 2007

Student essays

- Faulkner, 2014

Argument components

- Stab et al., 2018

Related Work : Methods

A plethora of classification methods

- Rule based; traditional feature-based ML (SVM, Logistic Regression....), deep learning (LSTM, CNN, BERT...)

Features

- **Content-based:** sentiment, ngrams, dependency-based ...
- **Contextual:** agreement/disagreement between posts or speeches; author identity; conversation structure; discourse structure

Collective classification helps

- Thomas et al., 2006; Yessenalina et al., 2010; Burfoot et al., 2011; Hasan and Ng, 2013; Walker et al., 2012b; Sridhar et al., 2014

Related Work : Shared Tasks

Detecting Stance in Tweets [Mohammad et al., SemEval 2016]

- Detect the sentiment of tweets towards target entity

Related Work : Shared Tasks

Detecting Stance in Tweets [Mohammad et al., SemEval 2016]

- Detect the sentiment of tweets towards target entity

RumourEval [Derczynski et al., SemEval 2017; Gorrell et al., 2019]

- Stance classification as a subtask of claim validation
- Classify tweets in a conversation discussing a rumour as supporting/denying/querying/commenting

Advances in Debating Technologies 3: Argument Evaluation and Analysis

Argument Quality

Roy Bar-Haim

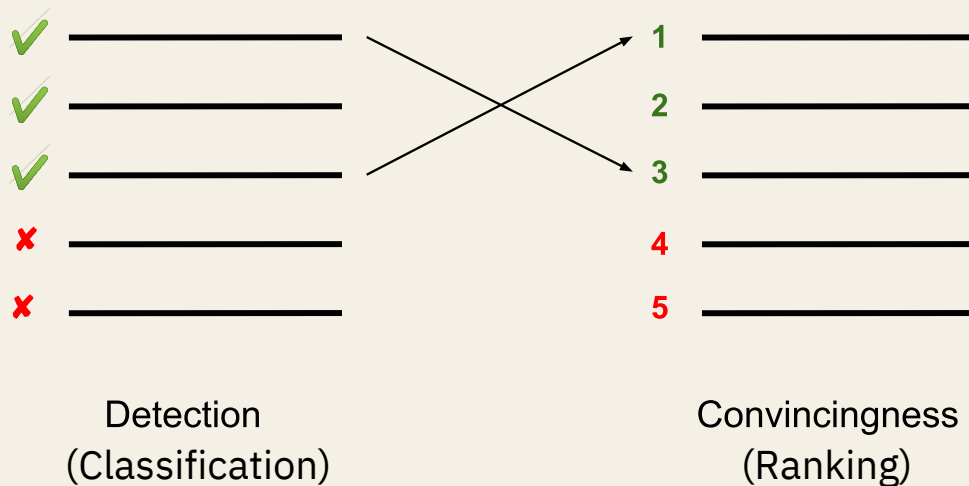
IBM Research AI

Mostly Based on...

- *Are You Convinced? Choosing the More Convincing Evidence with a Siamese Network*
Gleize et. al, ACL 2019
- *Automatic Argument Quality Assessment - New Datasets and Methods*
Toledo et al., EMNLP/IJCNLP 2019
- *A Large-Scale Dataset for Argument Quality Ranking: Construction and Analysis*
Gretz et al., AAAI 2020

Argument Quality in Project Debater: Evaluate Evidence Convincingness

Re-rank the output of the evidence detection model according to evidence convincingness



For the topic:

The use of AI should be abandoned

Our detection model gave a higher score to:

In China, which already has one of the highest concentrations of industrial robots, 53% of people agree that AI will create new jobs rather than lead to unemployment.

-Weak majority -Poll

Our convincingness model preferred:

The Wall Street Journal recently reported that evidence shows that the increased productivity AI provides will lead to more wealth, cheaper goods, greater spending power and ultimately, more jobs.

+Objective phrasing +Reputable source

Argument Quality Annotation: Pointwise vs. Pairwise

Claims	Pointwise	Pairwise
Task	<u>Given an argument A:</u> Would you recommend a friend to use it <i>as is</i> in a speech?	<u>Given a pair of arguments (A,B):</u> Which of the two arguments would have been preferred by most people to support/contest the topic?
# Annotations	$O(N)$	$O(N^2)$
Difficulty of Judgment	Hard – determining quality without reference	Easy – choosing between two options

Argument Quality Datasets in Project Debater

Gleize et al., ACL 2019

- 70 topics
- Evidence sentences from Wikipedia
- **Pairwise:** 5.7K pairs. Which evidence is more convincing?

Toledo et al., EMNLP 2019

- 11 topics
- Arguments collected from large audiences
- **Pointwise:** 6.3K/5.3K args. Quality score - fraction of positive annotations
- **Pairwise:** 14K/9.1K pairs. Good agreement with pointwise annotations, esp. for large score differences

Gretz et al., AAAI 2020

- 71 topics
- Arguments collected from large audiences via crowdsourcing
- **Pointwise:** 30K args
- Use Weighted-Average (WA) and MACE probability (MACE-p) to derive quality scores from binary labels

Experimental Results

Gretz et al., AAAI 2020

- 30K arguments
- Split by topic: train (49), dev (7), test (15)
- Evaluate Pearson (r) and Spearman (ρ) correlation
- And the winner is...

BERT-FT_{TOPIC}

- Fine tuned BERT, adapted for regression
- Input: (argument, topic)

	WA		MACE-P	
	r	ρ	r	ρ
Arg-Length	0.21	0.22	0.22	0.23
SVR BOW	0.32	0.31	0.33	0.33
Bi-LSTM GloVe	0.44	0.41	0.43	0.42
BERT-V	0.48	0.43	0.49	0.48
BERT-FT	0.51	0.47	0.52	0.50
BERT-FT_{TOPIC}	0.52	0.48	0.53	0.52

Related Work: Argument Convincingness

Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. Habernal and Gurevych, ACL 2016

Tasks

- Given a pair of arguments, which one is more convincing?
- Rank a set of arguments based on their convincingness

Dataset

- Arguments from debate portals
- 16K annotated argument pairs
- 1K ranked arguments. Ranking is derived from pairwise annotations

Experiments

- Using SVM and bi-LSTM

Related Work: Argument Convincingness

Finding Convincing Arguments Using Scalable Bayesian Preference Learning.
Edwin Simpson and Iryna Gurevych, TACL 2018

- A follow-up work to [Habernal and Gurevych, 2016]
- Improved results using Gaussian Process Preference Learning (GPPL)
- GPPL is shown to be particularly advantageous with small, noisy datasets, and in an active learning set-up

Related Work: Argument Quality Assessment

Computational Argumentation Quality Assessment in Natural Language

Henning Wachsmuth

Bauhaus-Universität Weimar
Weimar, Germany

henning.wachsmuth@uni-weimar.de

Nona Naderi

University of Toronto
Toronto, Canada

nona@cs.toronto.edu

Yufang Hou

IBM Research
Dublin, Ireland

yhou@ie.ibm.com

Yonatan Bilu

IBM Research
Haifa, Israel

yonatanb@il.ibm.com

Vinodkumar Prabhakaran

Stanford University
Stanford, CA, USA

vinod@cs.stanford.edu

Tim Alberdingk Thijm, Graeme Hirst

University of Toronto
Toronto, Canada

{thijm, gh}@cs.toronto.edu

Benno Stein

Bauhaus-Universität Weimar
Weimar, Germany

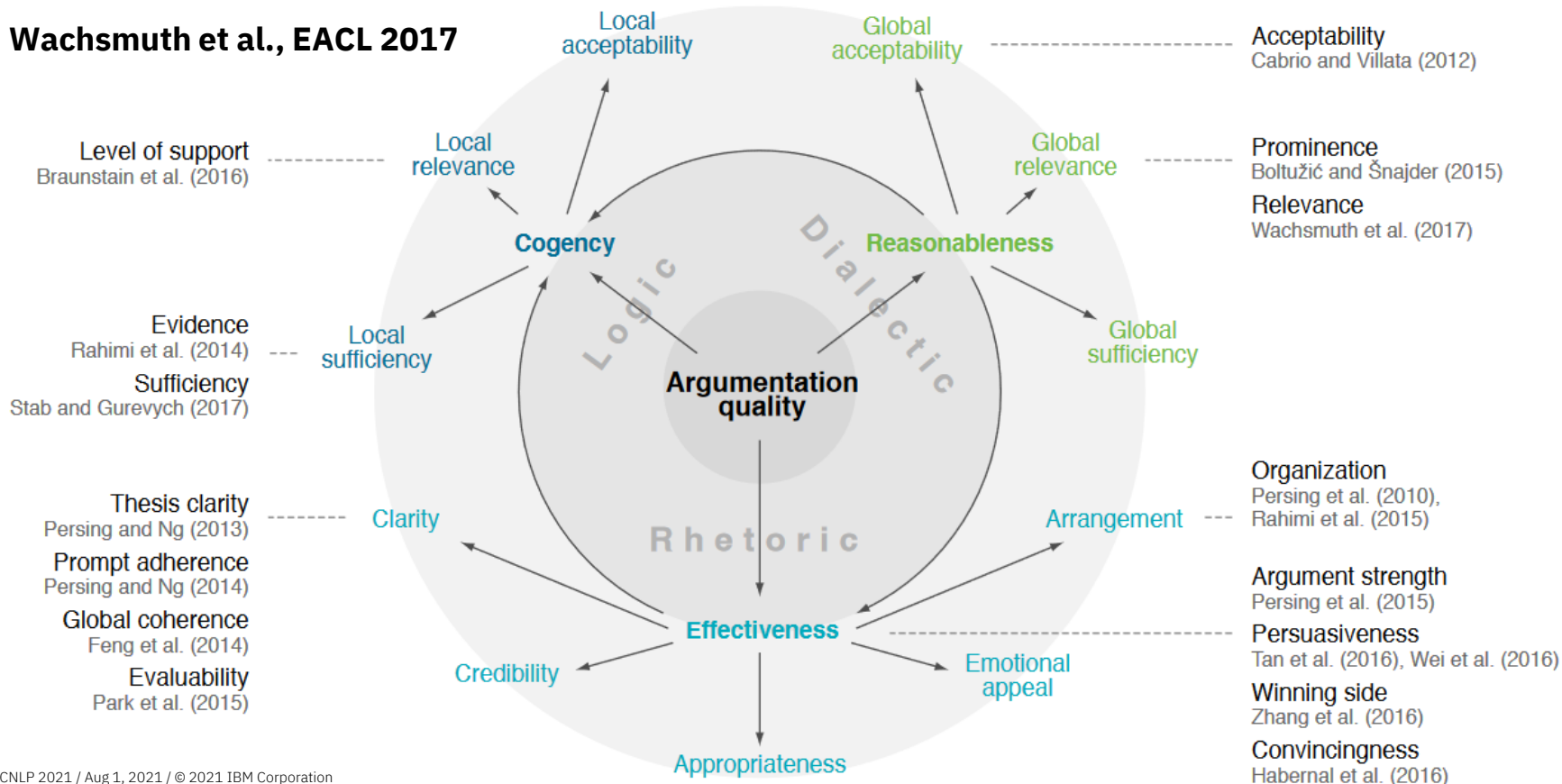
benno.stein@uni-weimar.de

EACL 2017

- A comprehensive survey of research on argumentation quality assessment
- A taxonomy of all major quality dimensions of natural language argumentation
- An annotated corpus for computational argumentation quality assessment

Related Work: Argument Quality Assessment

Wachsmuth et al., EACL 2017



Related Work: Efficient Argument Quality Annotation

Efficient Pairwise Annotation of Argument Quality.

Lukas Gienapp, Benno Stein, Matthias Hagen, Martin Potthast, ACL 2020

- Sampling strategy for saving up to 93% of the comparisons in pairwise annotation
- A model for transforming pairwise annotations to scalar quality scores
- A new argument quality corpus
 - Rhetorical, logical, dialectical, and overall quality scores for 1.3K arguments
 - Inferred from 40K pairwise annotations

Advances in Debating Technologies

4: Principled Arguments

Matan Orbach

IBM Research AI

Principled Arguments

Mostly based on...

Argument Invention from First Principles, ACL 2019

Yonatan Bilu, Ariel Gera, Daniel Hershcovich, Benjamin Sznajder, Dan Lahav, Guy Moshkovich, Anael Malet, Assaf Gavron, Noam Slonim



Ariel
Gera



Daniel
Hershcovich



Assaf
Gavron



Yonatan
Bilu

What does Argument Mining* miss?

Motion: We should increase the use of **telemedicine**

Claim: People always worry about new technologies, but they are the basis for mankind's advancement and prosperity.

- Does not mention telemedicine explicitly
- Touches upon a fundamental aspect of the debate
- Attempts to frame the debate
- Common to many debates about technology (but not all):
 - **We should limit the use of social media**
 - **Autonomous cars will bring more good than harm**

(*when it's based on sentence-level queries)

Working hypothesis

Principled arguments, which are recurring in debates, can be written independently of any specific motion.

We can create a knowledge base of texts, such that when presented with a new motion, there will usually be relevant texts within this knowledge base.

Advantages of hand-crafted texts

- Author arguments discussing the principles underlying the motion
- Texts can frame the debate in a favorable way
- Texts are well written
- Supplement argumentative texts with non-argumentative ones
- Relations among texts are clear

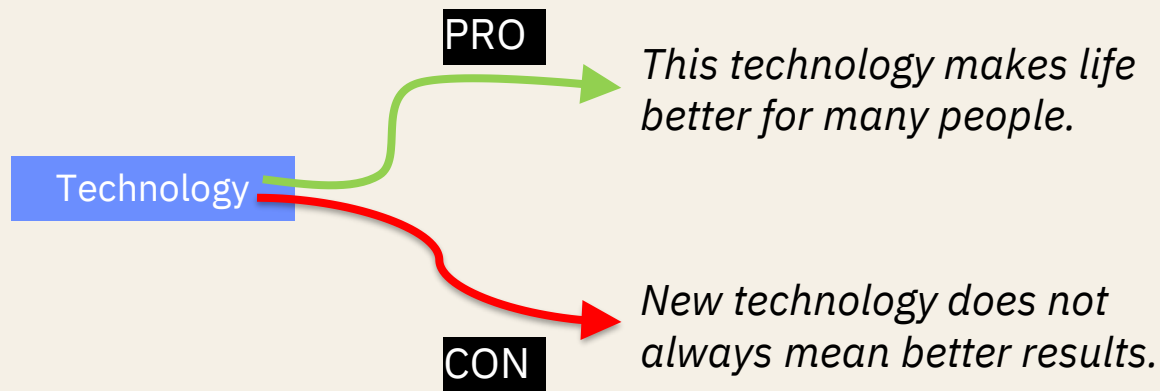
Challenge: given a novel motion, determine which texts are relevant

How did we do it?

- Define classes of recurring debate themes (commonplace, principles).
- Define a typology of texts, e.g. speech opening, proactive claim, quote.
- For each class, author texts of each type (and annotate them).

- Develop a mechanism that, given a novel motion, determines which classes are relevant.
- Suggest texts from matched classes to the debate construction component.

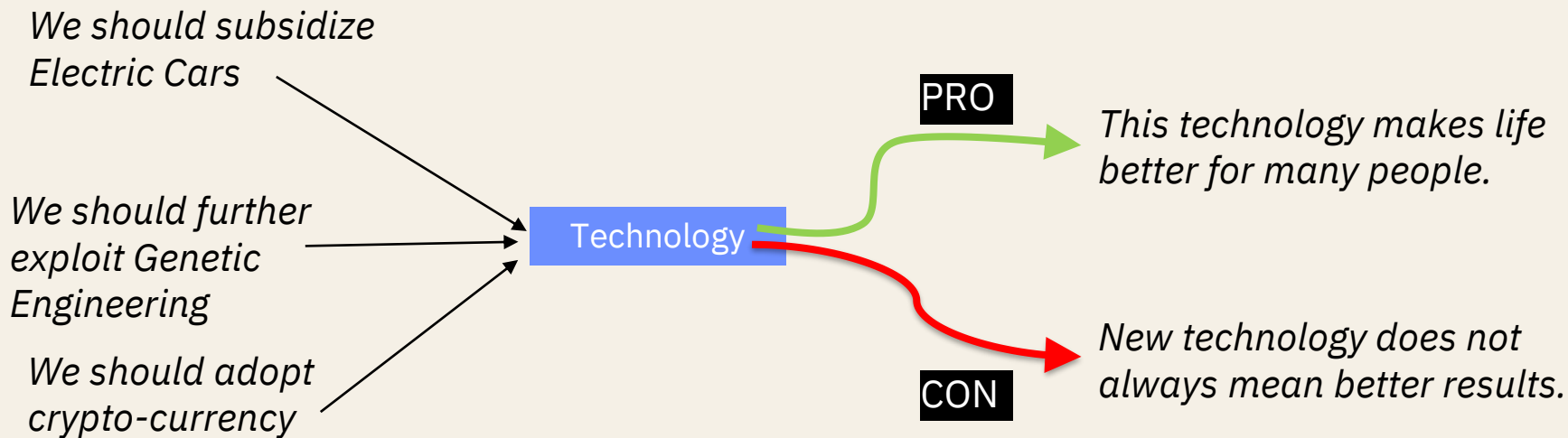
Classes of Principled Arguments (CoPAs)



CoPA

Principled Arguments

Classes of Principled Arguments (CoPAs)

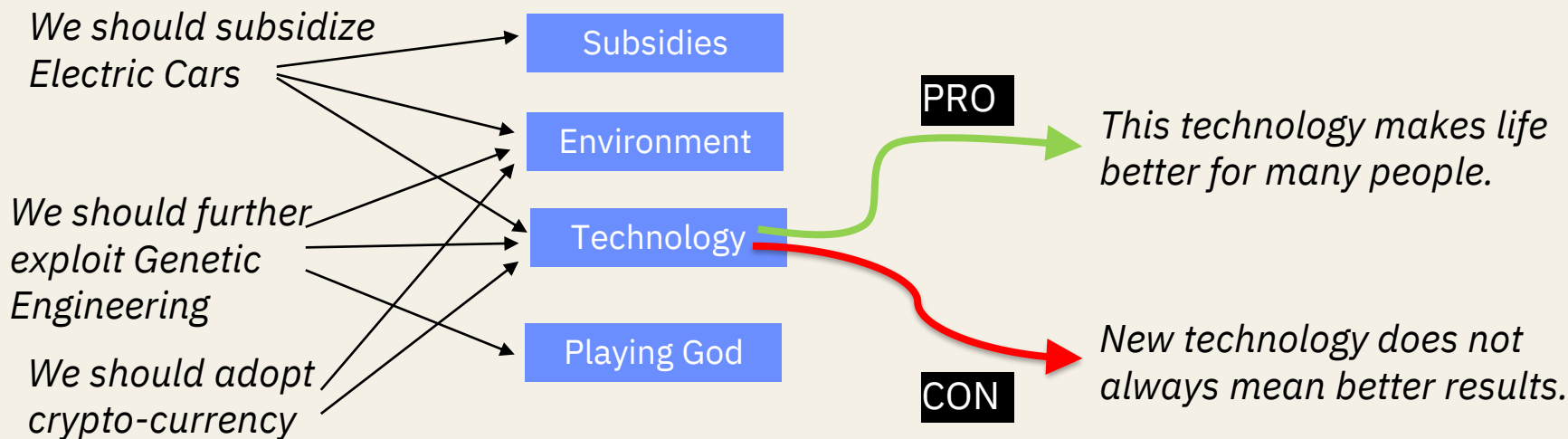


Motions

CoPAs

Principled Arguments

Classes of Principled Arguments (CoPAs)

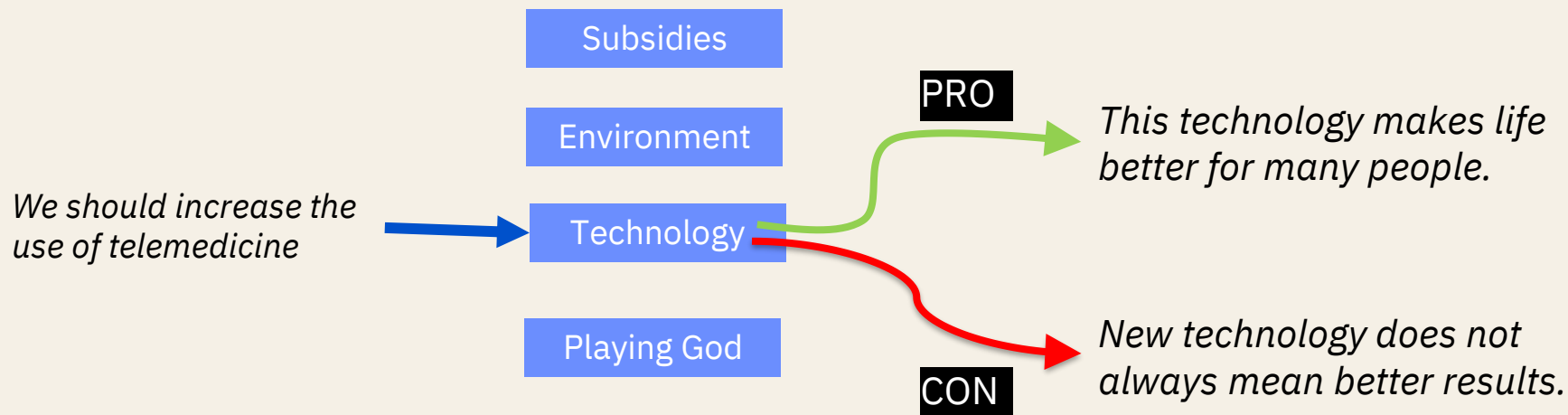


Motions

CoPAs

Principled Arguments

Classes of Principled Arguments (CoPAs)



Novel Motion

CoPAs

Principled Arguments

Modeling Goals

Define a set of principled arguments with the following properties:

1. **Coherent** – People can easily decide if an argument is relevant for a motion.
2. **Non trivial** – Not too general, not too specific, not a partition.
3. **Coverage** – Most motions belong to at least one class.
4. **Real world** – Texts tend to be used in real debates.
5. **Computer friendly** – Classes can be automatically matched to novel motions

CoPA construction

Defined ~50 classes

Defined in the context of 100 motions

Later expanded to 689 motions

Exhaustively annotated all motions vs. all CoPAs

Adolescent rights (9 motions)	
Many adolescents can not make responsible decisions	Adolescents are as capable as adults
Animal rights ^t (21 motions)	
Animals should not be treated as property	There is nothing wrong with using animals to further human interests
Big government (21 motions)	
Public utility is best served by actions coordinated by central government	Public interest is best served and propelled by voluntary interactions, and not ones dictated by government
Black market ^t (35 motions)	
Prohibiting products and activities makes them less visible and available, and thus less harmful	Prohibition is counterproductive and only leads to increased demand

Coherent – People can easily decide if an argument is relevant for a motion.

Crowd annotation of a sample of (motion, CoPA argument) pairs

Topic: We should ban homeopathy

1. Homeopathy is considered offensive by various religious groups.

Is the claim reasonable to use when discussing the topic? (required)

- Yes
- No

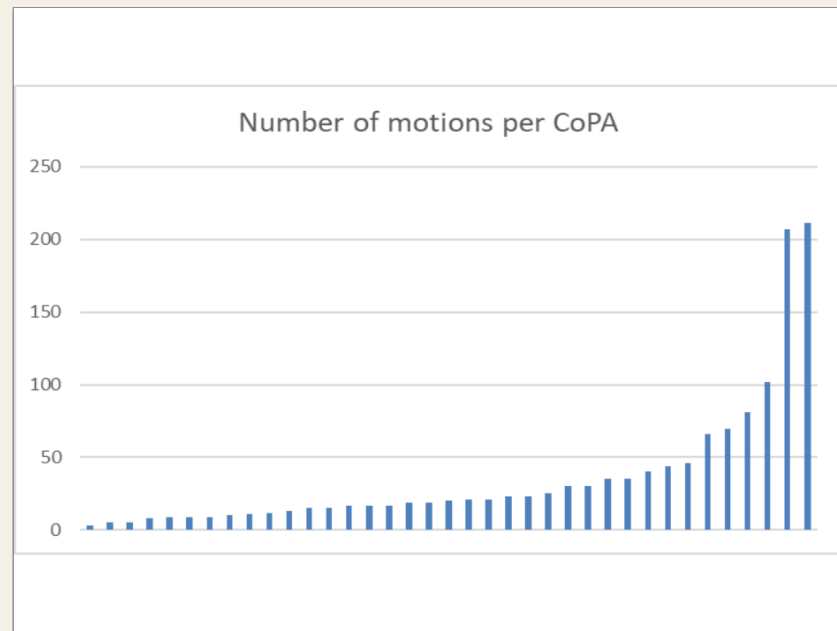
Cohen's Kappa between majority vote and ground truth: 0.76

Mean inter annotator kappa: 0.60

Not trivial – CoPAs are not huge, tiny or a partition of motions

Good coverage – Most motions belong to at least one CoPA

- 87% of motions belong to at least one CoPA.
- On average ~2 CoPAs per motion.
- Maximum of 6 CoPAs per motion.
- On average, a CoPA intersect with 12 others.
- On average, intersections cover 20% of the motions in a class.



Real World – used in actual speeches (but not always)

184 debate speeches from Mirkin et al. 2018

Speeches made by professional debaters, unaware of CoPAs

Showed crowd annotators the entire speech

Asked them if CoPA-Claims (from matching CoPAs) are made in speech

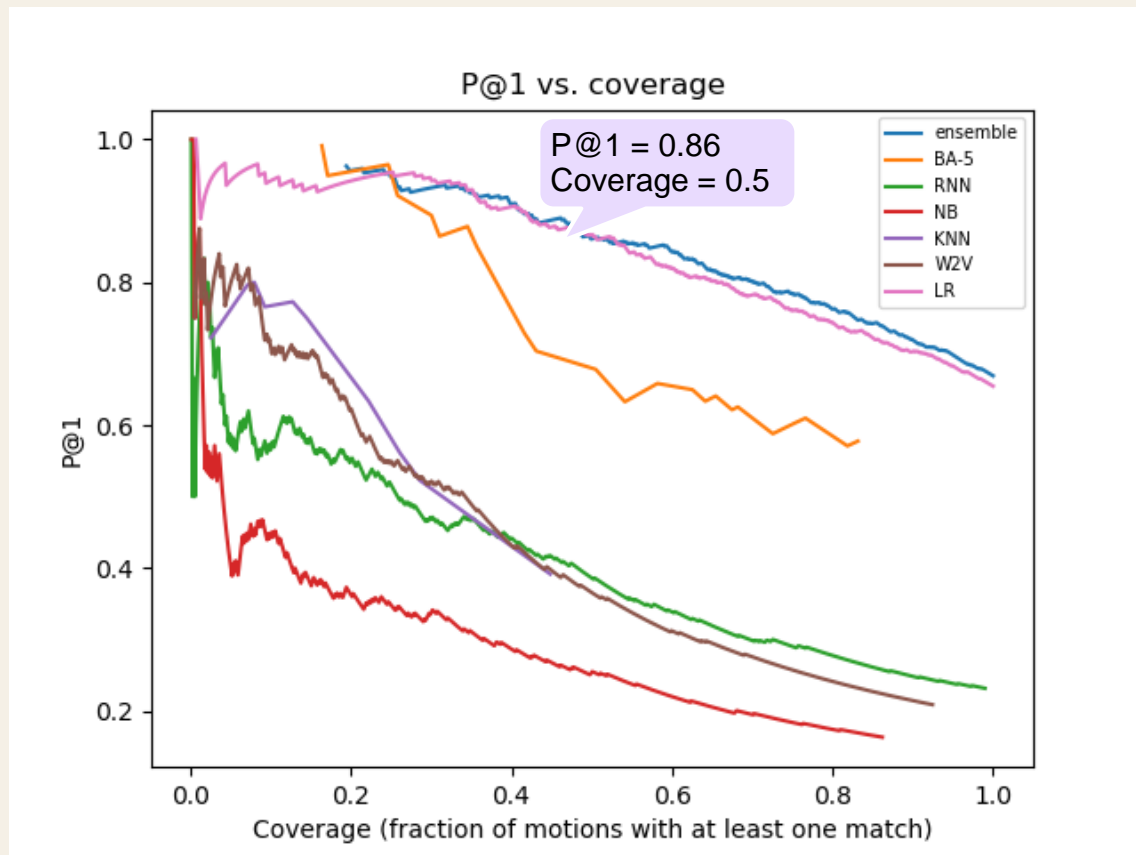
- 87% of speeches referenced at least one CoPA-Claim (mostly implicit)
- 66% of (speech, claim) pairs annotated as matching
- CoPA-claims are commonplace, but not trivial

Computer friendly – given a novel motion, which CoPAs are relevant?

Examined various classification methods – LR, NB, KNN, RNN...

Most successful is LR

Examined in a leave-one-motion-out analysis



Computer Friendly – but not so easy...

Black Market

We should legalize cannabis

We should ban sex work

We should ban internet cookies

We should legalize polygamy

Debate: We should increase the use of telemedicine

We are discussing today the merits of technology and its importance to mankind's ongoing advancement and prosperity. I am a true believer in the power of technology, as I should be, *being myself a prime example of its power.*

The use of **telemedicine** is reliable. It is a new technology that is more reliable than the so-called "conventional" old-school alternatives.

[... arguments from argument mining engine ...]

In today's debate you are likely to hear the other side express concerns about new technologies, their reliability, their track record and so on. They will explain how the old methods are good enough and don't need replacing. I will say in response: don't be afraid, and let us move on with the changing world!

What else is inside a CoPAs?

- **Claims**
 - **Extended Claims**
 - **Evidence**
 - **Opening / Framing**
 - **Conclusion referencing opening**
 - **Proactive arguments**
 - **Quotes**
 - **Humor**
 - **Closing words**
 - **Canonical examples**
 - **Rebuttal units**
- Nearly all types of texts for all CoPAs – lots of authoring work!**

Related Work

Argumentation Schemes

- Douglas Walton, Christopher Reed, and Fabrizio Macagno. *Argumentation schemes*. 2008.

Related Work

Argumentation Schemes

- Douglas Walton, Christopher Reed, and Fabrizio Macagno. *Argumentation schemes*. 2008.

Framing (social science)

- Holli A Semetko and Patti M Valkenburg. 2000. Framing European politics: A content analysis of press and television news.
- Claes H. de Vreese. 2005. News framing: Theory and typology.

Related Work

Argumentation Schemes

- Douglas Walton, Christopher Reed, and Fabrizio Macagno. *Argumentation schemes*. 2008.

Framing (social science)

- Holli A Semetko and Patti M Valkenburg. 2000. Framing European politics: A content analysis of press and television news.
- Claes H. de Vreese. 2005. News framing: Theory and typology.

Framing (NLP)

- Dallas Card, Amber E. Boydston, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues.
- Nona Naderi and Graeme Hirst. 2017. Classifying frames at the sentence level in news articles.
- Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation.
- Benjamin Schiller, Johannes Daxenberger and Iryna Gurevych. 2020. Aspect-controlled neural argument generation.

Summary

- Defined CoPAs – collections of texts based around a recurring debate theme.
- Manually annotated 689 motions for associated CoPAs.
- Verified that this annotation has desired properties – *explicit modeling of recurring arguments!*
- Trained a classifier to determine whether a (motion, CoPA) pair is a match
- Given a novel motion, go over all CoPAs and determine which of them match
- Use texts from these CoPAs in the speech

Advances in Debating Technologies

5: Listening Comprehension and Rebuttal

Matan Orbach

IBM Research AI

Mostly based on...

Shahar Mirkin et al., *A Recorded Debating Dataset*, LREC 2018.

Shahar Mirkin et al., *Listening Comprehension over Argumentative Content*, EMNLP 2018.

Tamar Lavee et al. *Towards Effective Rebuttal: Listening Comprehension using Corpus-Wide Claim Mining*, ArgMining@ACL 2019

Matan Orbach et al., *A Dataset of General-Purpose Rebuttal*, EMNLP 2019

Tamar Lavee et al., *Crowd-sourcing annotation of complex NLU tasks: A case study of argumentative content annotation*, AnnoNLP@EMNLP 2019

Matan Orbach et al., *Out of the Echo Chamber: Detecting Countering Debate Speeches*, ACL 2020



Shachar
Mirkin



Tamar
Lavee



Lili
Kotlerman



Lena
Dankin

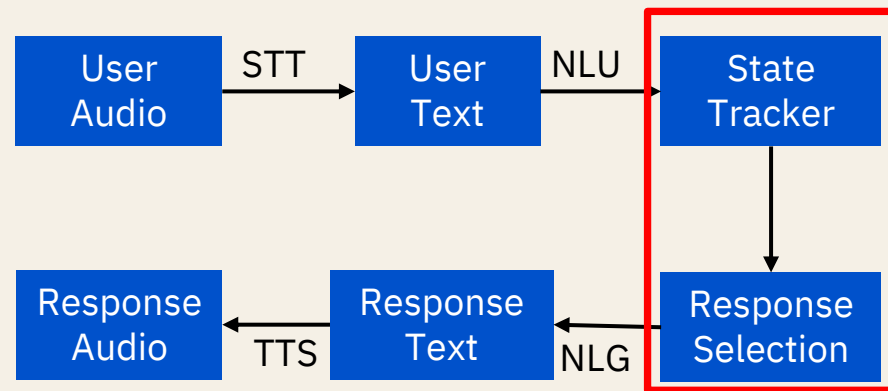
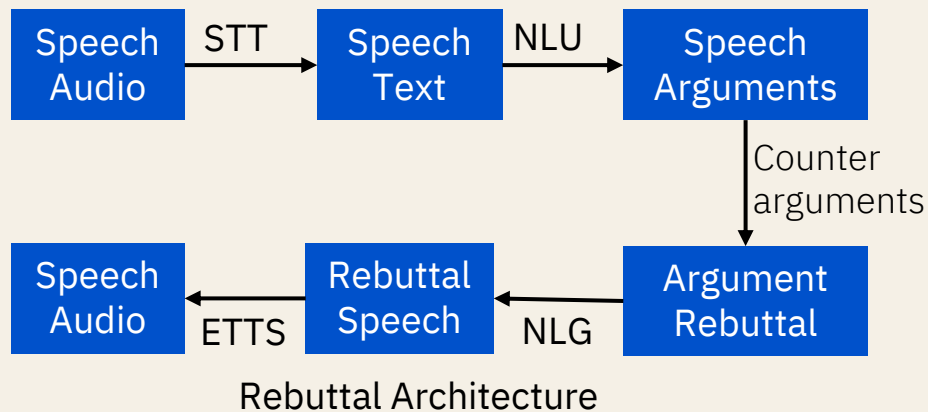


Yoav
Kantor



Yonatan
Bilu

Speech Rebuttal Task



Dialogue Systems

Task-oriented system:

- Order a table at a restaurant, get bus route information.
- Usually based on slot filling – e.g. date, time, restaurant's name, number of people
- Often rule based, recent works suggest neural components

Chit-chat systems (Blender from Facebook, Meena from Google)

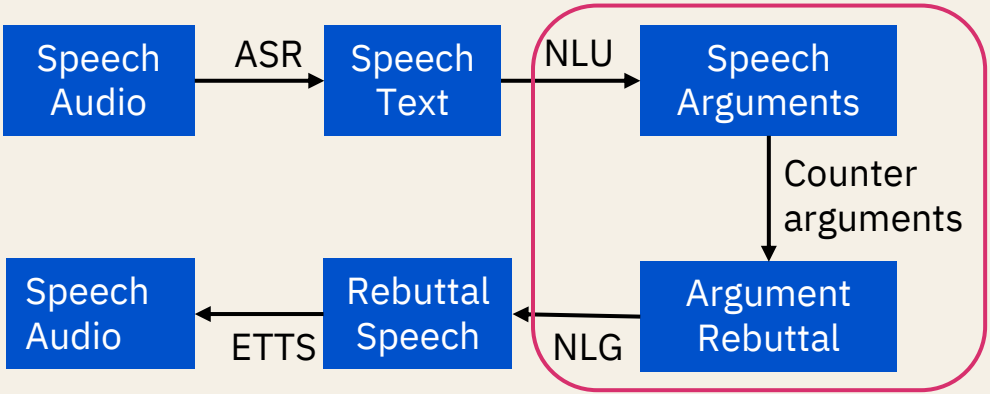
- Keep the user engaged
- Modern system are end-to-end transformer-based, LM-like.

Hybrid systems

Rebuttal vs. Dialogue

	Task Oriented	Chit Chat	Rebuttal
→ Topic	Defined in advance	Open	Controversial topic
→ Goal	Well defined	Engagement?	Persuade audience
→ Length	Short utterances, Few acts	Short utterances, open ended	Long speech, 3 rounds
→ User	Cooperative	Curious	Opposing
→ NLU	Recognize slot fillers	Remain coherent	Main points (some of them)
→ Data	Limited when new	Lots	Limited (esp. in context)

Speech Rebuttal Task



Rebuttal Architecture

The Rebuttal Unit

1. Query – What the opponent might claim
2. Quoted Claim – How to refer to the opponent's claim
3. Rebuttal Argument – How to respond to the opponent's claim

Example:

Debate Topic: We should increase the use of Telemedicine

Rebuttal Unit:

1. Degraded care
2. The care offered by telemedicine is not as good as traditional care
3. Comparative studies have found no correlation between telemedicine and patients' prognosis.

The Rebuttal Unit

1. Query – What the opponent might claim
2. Quoted Claim – How to refer to the opponent's claim
3. Rebuttal Argument – How to respond to the opponent's claim

Algorithm:

1. Check if the opponent mentioned any of the queries
2. Decide which matching Rebuttal Units to embed in rebuttal speech
3. Use Quoted Claim and Rebuttal Argument in rebuttal speech

Challenge: How to construct Rebuttal Units when the topic is not known in advance?

Advantage: All potential responses are known in advance.

Multiple Types of Rebuttal

CoPA-based – Text is based on recurring debate themes.

Example:

1. Query: Black Market
2. Quoted: A Black Market will arise
3. Rebuttal: Even if there is some truth to that, it should not stop us from doing what is right

If a CoPA is matched to the debate topic, all its rebuttal units are considered.

Multiple Types of Rebuttal

Argument Mining based

1. Query: Mined Claim with an opposing stance
2. Quoted: Same as Query
3. Rebuttal: Evidence of an opposing stance, rebutting the claim (mentioning the same concepts)

Motion: “We should increase the use of telemedicine”

Query / Quoted claim: Telemedicine is unpleasant for both patients and doctors

Rebuttal: Dr. Amin A. Muhammad, Niagara Health's Interim Chief of Mental Health and addictions, says telemedicine has been a positive experience for patients, their families and staff.

Challenges

1. How to determine whether a *Query* was mentioned by the opponent?
2. How to evaluate the Rebuttal System?
3. How to collect data, and how to label it?

Data Collection

1. Team of 10 academic debaters, from the U.S. and Israel
2. Defined a set of motions
3. For each motion, recorded several opening speeches
4. For each recorded speech, recorded several response speeches

3562 such speeches available for download, Orbach et al. 2020, ACL.

Annotation

Task: Given a speech and a set of claim-queries, mark which claims are mentioned in the speech.

Feasible with an expert work force and a small set of claims.

Challenging at scale – crowd workers want to be done quickly.

Label (Query, Speech Sentence) pairs?

- Too many pairs for exhaustive annotation
- Small fraction of positives
- Difficult to annotate correctly without speech context

Annotation: Rise to the challenge

Task: Given a speech and a set of claim-queries, mark which claims are mentioned in the speech.

Labels: mentioned explicitly, mentioned implicitly, no mentioned.

- Cultivate a suitable crowd workforce.
- Post tasks both in a general channel and a dedicated channel (Appen).
- Identify good annotators and add them to the dedicated channel.
 - Look at shared IPs, annotator correlations, fraction of positives...
 - Use test-questions in post processing.
- Give bonuses based on post-annotation evaluation.

Lavee et al. 2019, AnnoNLP@EMNLP

Annotation: Example (Mined Claims)

Motion: We should support Water Fluoridation

“It's essential that something is done to ensure that people don't have dental problems later in life. Water fluoridation is so cheap it's almost free. There are no proven side effects, the FDA and comparable groups in Europe have done lots and lots of tests and found that water fluoridation is actually a net health good, that there's no real risk to it”

Water fluoridation has no side-effects

Explicit

Water fluoridation is safe and effective

Implicit

Most people wouldn't get regular fluoride treatment at the dentist

**Not
Mentioned**

Matching Queries to Speeches

Task: Given a speech and a set of Queries, determine algorithmically which Queries are matched in the speech.

For CoPA-based Queries, we can compute a-priori probability that a Query will be mentioned:

Speeches Query labeled positive

Speeches Query was a candidate

Baseline: pick rebuttal unit with highest a-priori probability

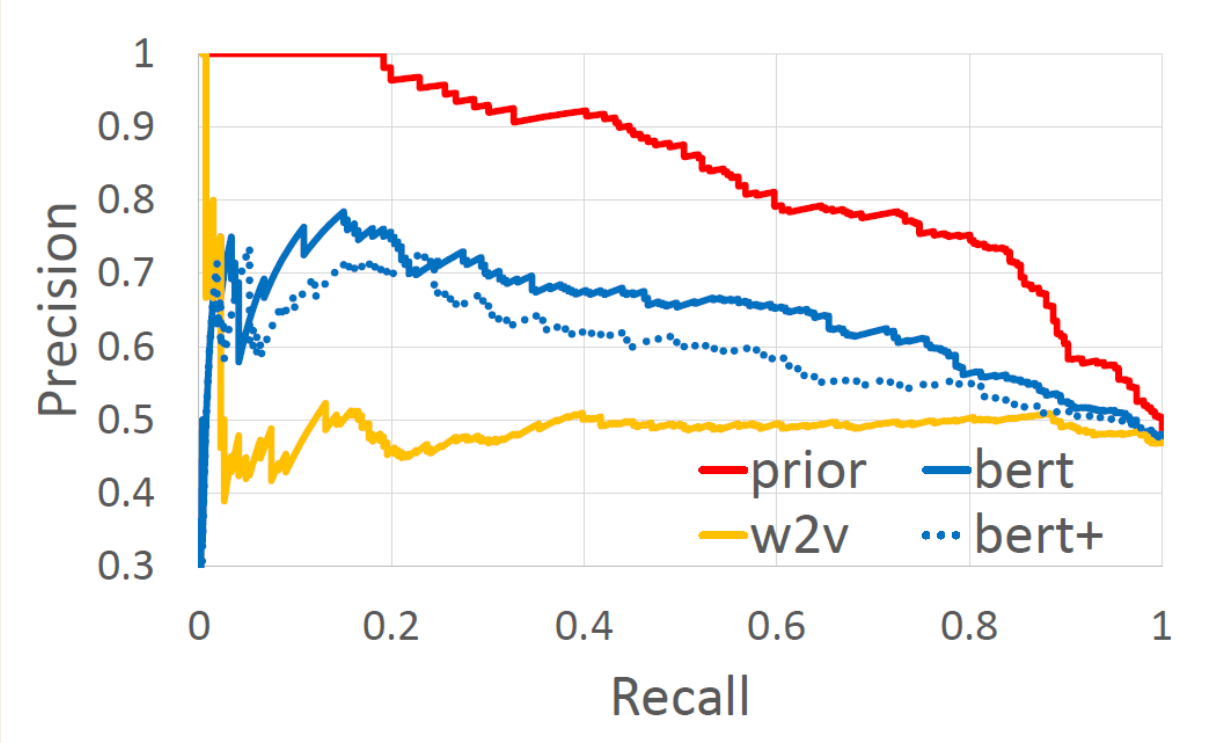
Matching Queries to Speeches

Task: Given a speech and a set of Queries, determine algorithmically which Queries are matched in the speech.

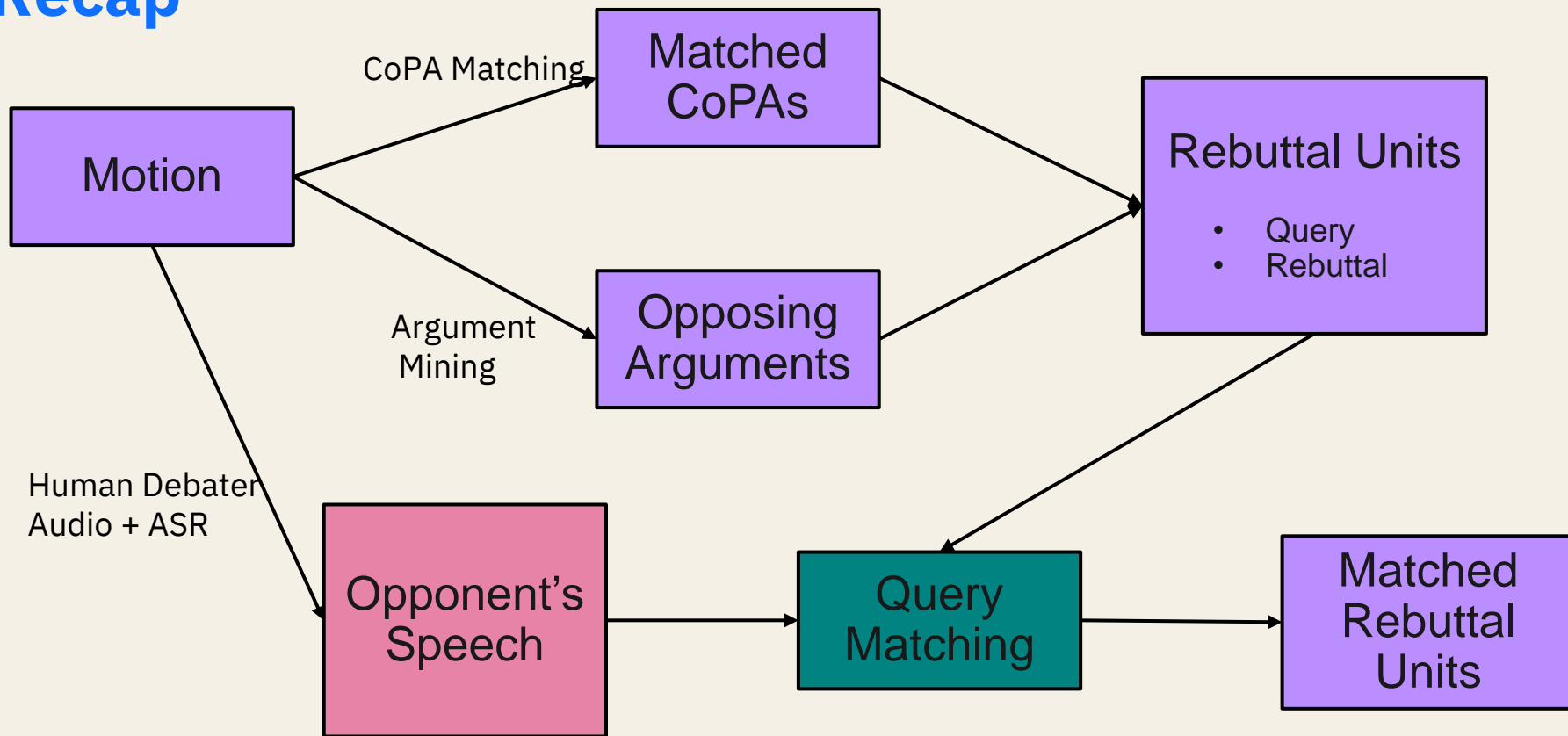
Other matching methods, sentence-by-sentence:

- Keyword matching
- Siamese network
- Word2vec-based sentence matching
- BERT-based sentence matching

Matching Leads to Speeches (CoPA-based queries)



Recap



Related Work: Matching Rebuttal to Lead

Henning Wachsmuth, Shahbaz Syed, Benno Stein, Retrieval of the Best Counter argument without Prior Topic Knowledge, ACL 2018.



This House would have a second Brexit referendum before leaving the EU

POINT

The government is here to serve the people. The concept behind a referendum is public sovereignty, the people themselves are the ones who get to decide. Having had one referendum it has already been agreed that this is an issue where it should not be up to our representatives in Parliament to decide, rather it is an issue for direct democracy.

COUNTERPOINT

The people have set the general principle. It is up to the Government, overseen by Parliament to implement the will of the people as shown by the referendum. We have a parliament precisely so that we have a body of people who have the time to go through a process and consider changes in detail. This is not something that a direct democracy can manage.

Related Work: Matching Rebuttal to Lead

Henning Wachsmuth, Shahbaz Syed, Benno Stein, Retrieval of the Best Counter argument without Prior Topic Knowledge, ACL 2018.

POINT

The government is here to serve the people. The concept behind a referendum is public sovereignty, the people themselves are the ones who get to decide. Having had one referendum it has already been agreed that this is an issue where it should not be up to our representatives in Parliament to decide, rather it is an issue for direct democracy.



COUNTERPOINT

The people have set the general principle. It is up to the Government, overseen by Parliament to implement the will of the people as shown by the referendum. We have a parliament precisely so that we have a body of people who have the time to go through a process and consider changes in detail. This is not something that a direct democracy can manage.

COUNTERPOINT

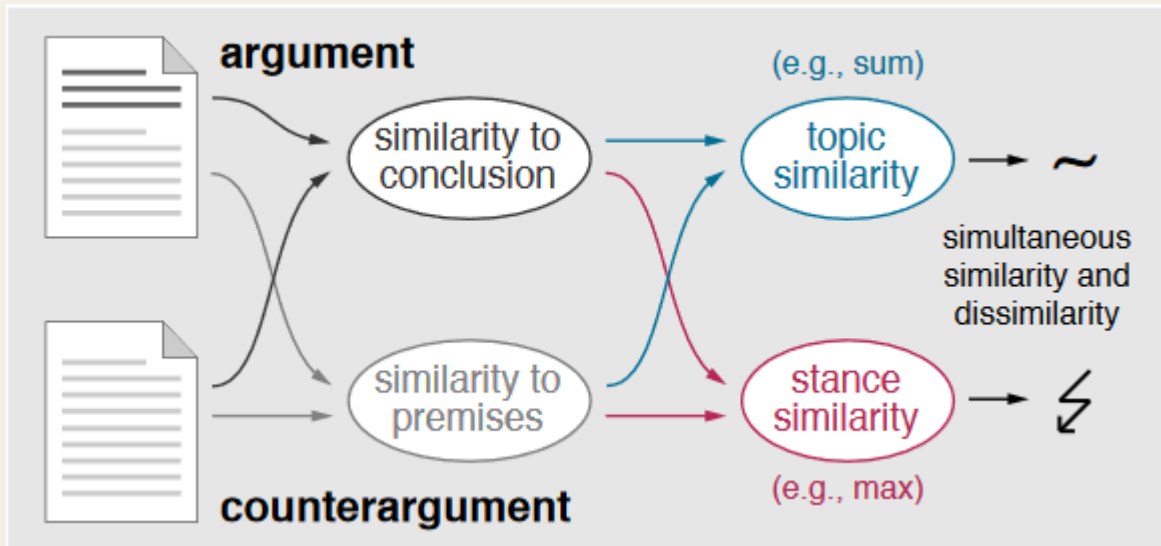
Circumstances are constantly changing but that is not a reason to be constantly having new votes. Should there be a vote to remain in next time and the EU only allows this with concessions from the UK should the public have to vote a third time?

COUNTERPOINT

It is clearly inaccurate that parliamentary sovereignty means that there is no need for consulting the people. If that were the case there would have been no need for the first referendum. Moreover, there would be no need to pay any attention to the results of that referendum.

Related Work: Matching Rebuttal to Lead

Henning Wachsmuth, Shahbaz Syed, Benno Stein, **Retrieval of the Best Counter argument without Prior Topic Knowledge**, ACL 2018.



- Maximize topic similarity and stance dissimilarity.
- Similarity measures based on:
 - Distance between sum of word embeddings vectors.
 - Distance between term frequencies.

Related Work: Neural Generation

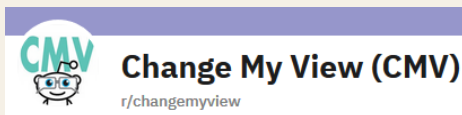
Xinyu Hua and Lu Wang, **Neural Argument Generation Augmented with Externally Retrieved Evidence**, ACL 2018.

Xinyu Hua, Zhe Lu, Lu Wang, **Argument Generation with Retrieval, Planning, and Realization**, ACL 2019.

- Train biLSTM encoder-decoder. First decode “talking points”, then attend them (and the claim) to decode counter argument.
- Training Data comes from CMV sub-reddit.
- Additional input data:
 - Retrieve Wikipedia articles related to the claim
 - Extract sentences that are likely to be relevant evidence

Related Work: What to rebut?

Yohan Jo, Seojin Bang, Emaad Manzoor, Eduard Hovy, Chris Reed,
Detecting Attackable Sentences in Arguments, EMNLP 2020.



I believe that Communism is not as bad as everyone says. CMV

I am a Canadian, and as America is my close neighbor, have taken notice to their restriction on Communism, with (for example) Cuba and North Korea currently, as well as the U.S.S.R. I feel that communism is not as bad as America make it seem. A society where everyone is equal seems great to me, as it removes some of the basic faults in society, such as poverty, homelessness, joblessness, as well as touching on moral values such as greed, and envy. I believe a proper Communist society (I.E. one that is not a dictatorship like Joseph Stalin or Fidel Castro) would be beneficial. I could easily be very wrong with my view, but i feel that what i see could definitely work, so Change My View.

⚠ Well, i feel that you have given the most in-depth and convincing argument against what i previously thought. I had a basic understanding on how the higher-class people (ones who have higher standards of living through dedication, intelligence, etc.) would be brought down by communism, but i can see now how communism, no matter how proper, is not beneficial.

A society where everyone is equal seems great to me

That's one of the big problems with communism - what is equality? Is everyone equal? Although I believe every human deserves equality in respect and dignity I don't believe everyone is equal. Will women receive the same opportunities in life as a man? Does a man with locked in-syndrome have the same opportunities as an Olympic athlete? Is it fair you may have a prettier wife than me? Life is unfair and financial equality is not the final answer.

it removes some of the basic faults in society, such as poverty, homelessness, joblessness, as well as touching on moral values such as greed, and envy

Yes there are problems within society but this doesn't mean there is a fault with society. Joblessness is just the growing pains of a changing society. With globalization and increase in women in work, there is much more people looking work than there was 60 years ago. We have always judged men on their ability to work now this has passed on to women and teenagers.

Related Work: What to rebut?

*Yohan Jo, Seojin Bang, Emaad Manzoor, Eduard Hovy, Chris Reed, **Detecting Attackable Sentences in Arguments**, EMNLP 2020.*

- Label sentences as: *successfully attacked*, *unsuccessfully attacked*, *not attacked*.
- Fine tune BERT to identify either *attacked* sentences or *successfully attacked* sentences.

P@1 for...	Random	Longest	BERT	Human
attacked	35.9	42.9	49.6	51.7
Suc. attacked	18.9	22.3	28.3	27.8

Summary

- Speech Rebuttal Task – different from dialog systems
- Constructed rebuttal units
 - CoPA-based
 - Argument-mining based
- Matching units to speeches – data collection, annotation and algorithms
- **Rebuttal is challenging and has much room for further research!**

Advances in Debating Technologies 6: Core NLP Capabilities

Liat Ein-Dor

IBM Research AI

NLP tools in Project Debater - Wikification

Linking words and phrases to Wikipedia articles.



Example: Studies show that physical discipline can cause psychological harm or make the child more aggressive in the longer term.

Aggression

From Wikipedia, the free encyclopedia



Wikification

Usage: Queries for argument mining

Query: Study lexicon → that → **Topic** → Sentiment lexicon

Motion: *Physical punishment* should be banned

Corporal punishment

From Wikipedia, the free encyclopedia
(Redirected from Physical discipline)

Sentence: Studies show that physical discipline can cause psychological harm or make the child more aggressive in the longer term.

Main challenge: Wikification of 10 billion sentences - not feasible for most existing tools

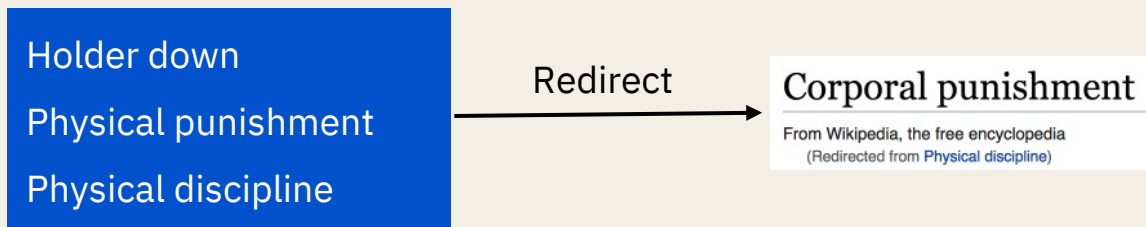


Run-Time Oriented Wikification

Shnayderman *et al.*, Fast end-to-end wikification. *arXiv preprint arXiv:1908.06785* (2019)

Main Idea: Rely on the wisdom of the crowd through “redirect” pages of Wikipedia

Redirect: a page that automatically sends to another page.



Wikification of entire corpus
(10B sentences)

Runtime

TagMe	22 days
RedW	40 hours



Organization of arguments into thematic paragraphs

Motion: Trans fats usage in food should be banned

“A new study has found that the ingestion of trans-fats increase the risk of suffering depression.”

“A study found that junk food high in trans fats can damage sperm in otherwise healthy, young men.”

“Trans fats can increase the physical brain damage and manic behavior associated with amphetamine abuse.”

“...if trans-fat made up too much of an infant's intake of fat, it may affect brain and eye development.”

“Recent research from the FDA has found that trans-fats are linked to infertility in women.”

“Trans fats can have a bad influence on children’s health.”



Organization of arguments into thematic paragraphs

Topic: Trans fats usage in food should be banned

“A new study has found that the ingestion of trans-fats increase the risk of suffering depression.”

“A study found that junk food high in trans fats can damage sperm in otherwise healthy, young men.”

“Trans fats can increase the physical brain damage and manic behavior associated with amphetamine abuse.”

“...if trans-fat made up too much of an infant's intake of fat, it may affect brain and eye development.”

“Recent research from the FDA has found that trans-fats are linked to infertility in women.”

“Trans fats can have a bad influence on children’s health.”

Mental Health

Fertility

Children’s Health



Organization of arguments into thematic paragraphs

Approach:

1. Cluster based on similarity function (iclust, slonim et al PNAS 2005).
2. Measure similarity between sentences

The infertility cluster

Text A: A study found that junk food high in trans fats can damage **sperm** in otherwise healthy, young men.

Text B: Recent research from the FDA has found that trans-fats are linked to **infertility** in women.

Measuring sentence similarity:

- Based on similarity between their Wikipedia concepts .
- Create a benchmark of pairs of terms annotated for their level of relatedness.
- Use the benchmark to train a supervised model for automatic prediction of relatedness between pairs of terms and concepts.



Relatedness Benchmarks

Concept-Relatedness Benchmark

Ein-Dor *et al.*, Semantic relatedness of Wikipedia concepts—benchmark data and a working solution, LREC 2018

- First human annotated dataset of Wikipedia concepts
- 19,276 pairs of concepts
- Used to develop a tool for measuring level of relatedness between pairs of concepts.
- The supervised approach significantly outperforms SOTA methods.

Term-Relatedness Benchmark

Levy *et al.*, Tr9856: A multi-word term relatedness benchmark, ACL 2015

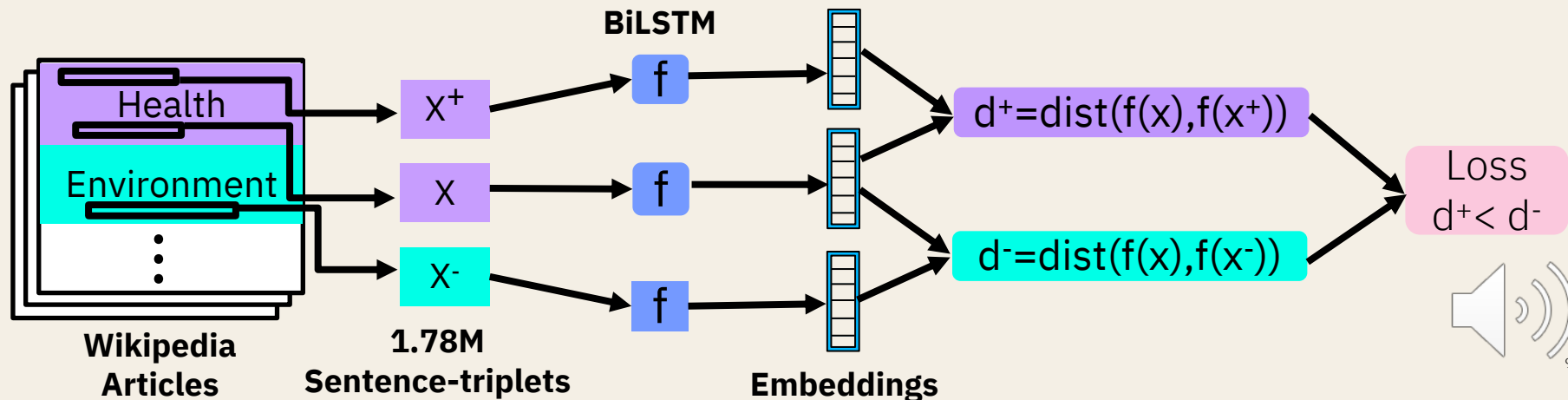
- Largest term-relatedness benchmark (~10K pairs)
- First to consider multi-word terms
- A term-relatedness model trained on this benchmarks outperformed



Thematic Clustering – the Representation Learning Approach

Ein-Dor, *et al.* Learning thematic similarity metric from article sections using triplet networks, ACL 2018

- Learn sentence embedding dedicated to thematic clustering using DNN
- Create a weakly labeled dataset of *thematic* similarity
 - Leverage partition of Wikipedia articles into sections
 - Underlying assumption: sentences from the same section are more thematically related than sentences from different sections.



Summary

- Light-weight end-to-end Wikification tool
- Large term-relatedness dataset with multi-word terms (~10K pairs)
- Large concept-relatedness dataset (>19K pairs)
 - Tool for measuring relatedness between concepts
- DL solution for learning thematic similarity between sentences
- Thematic clustering dataset based on Wikipedia sections



Related Work: Sentence-BERT

Reimers, N and Gurevych, I, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, EMNLP 2019

- Use Siamese and Triplet variations of BERT to compute semantically meaningful sentence embeddings.
- For thematic similarity, fine-tune on the thematic similarity triplet dataset.
- Sentence-BERT outperforms other models in the triplet classification task.
- Can potentially be useful for thematic clustering.



Related Work: Alternative Thematic Clustering Approaches

- Ajjour et al., Modeling Frames in Argumentation, EMNLP 2019
- Reimers et al., Classification and Clustering of Arguments with Contextualized Word Embeddings, ACL 2019
 - Create benchmark of argument pairs with similarity scores
 - Use benchmark to learn argument similarity by fine-tuning BERT
- Misra et al., Measuring the Similarity of Sentential Arguments in Dialog, SIGDIAL 2016
 - Detect argument pairs that are assumed to share the same argument facet, using manually annotated pairs.



Advances in Debating Technologies 7: From Arguments to Narrative

Roy Bar-Haim

IBM Research AI

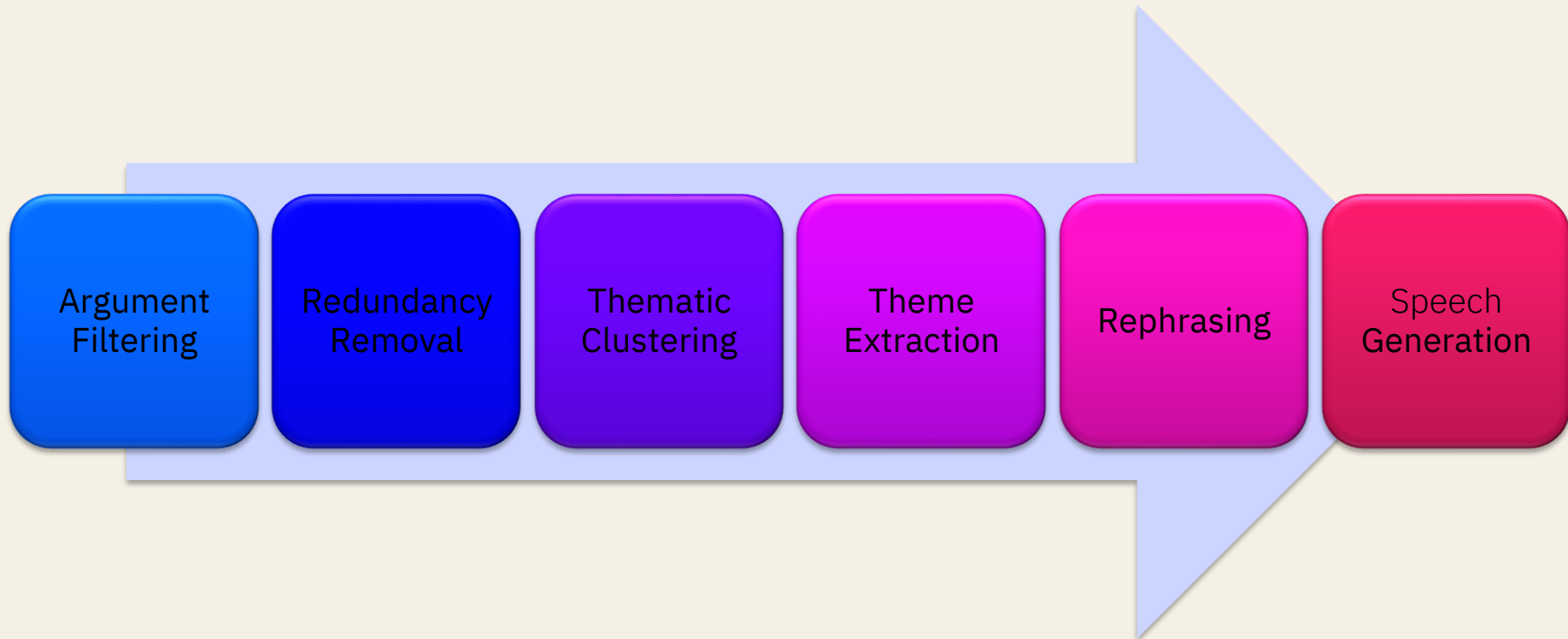
We Now Have All the Ingredients for a Speech

- Ranked lists of Claims and Evidence mined from the Lexis-Nexis corpus, supporting our stance
- Rebuttal Units
- Principled argumentative texts (CoPA-based)
- Boilerplate texts
- Debate topic definition (from Wikipedia)
- And more...

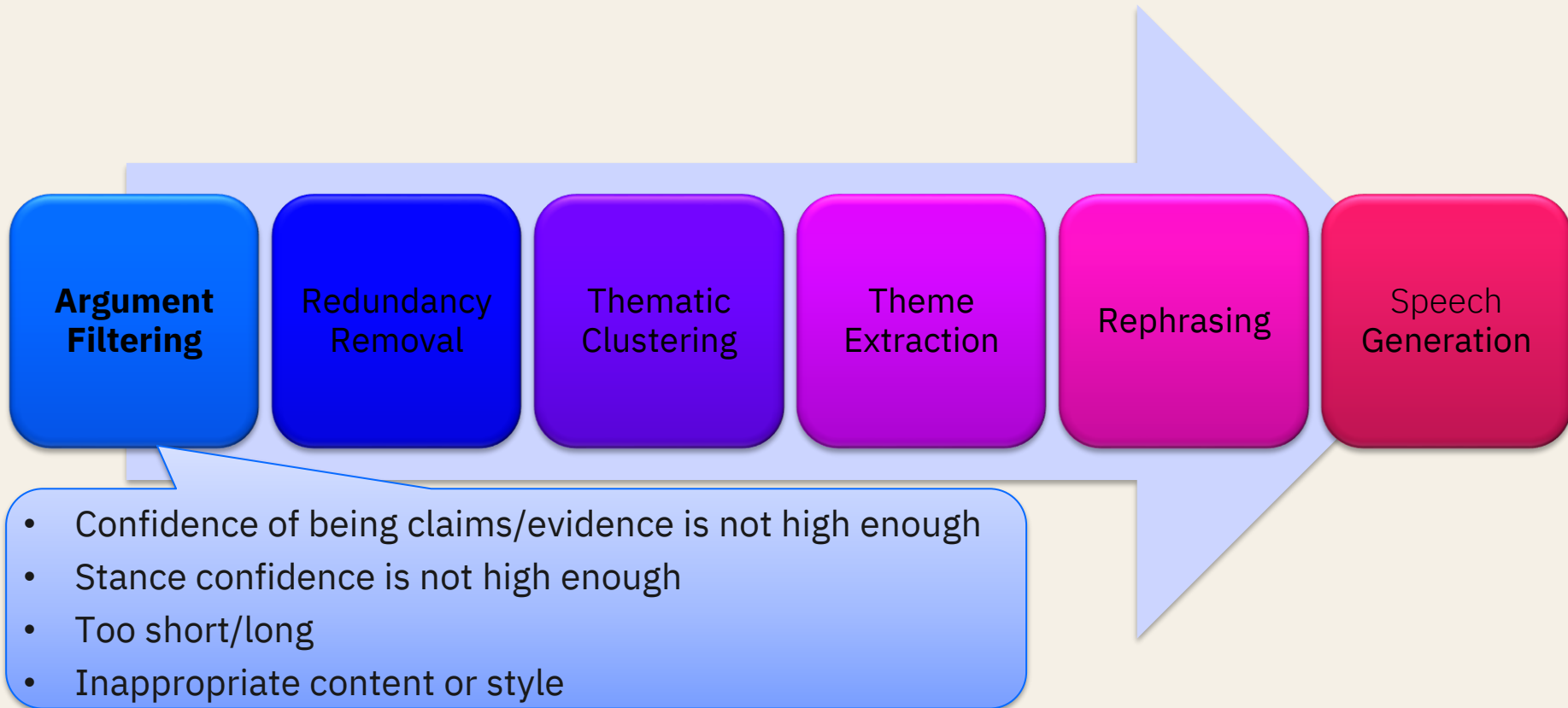


We just need a speech recipe...

Debate Construction - Overview



Debate Construction - Overview



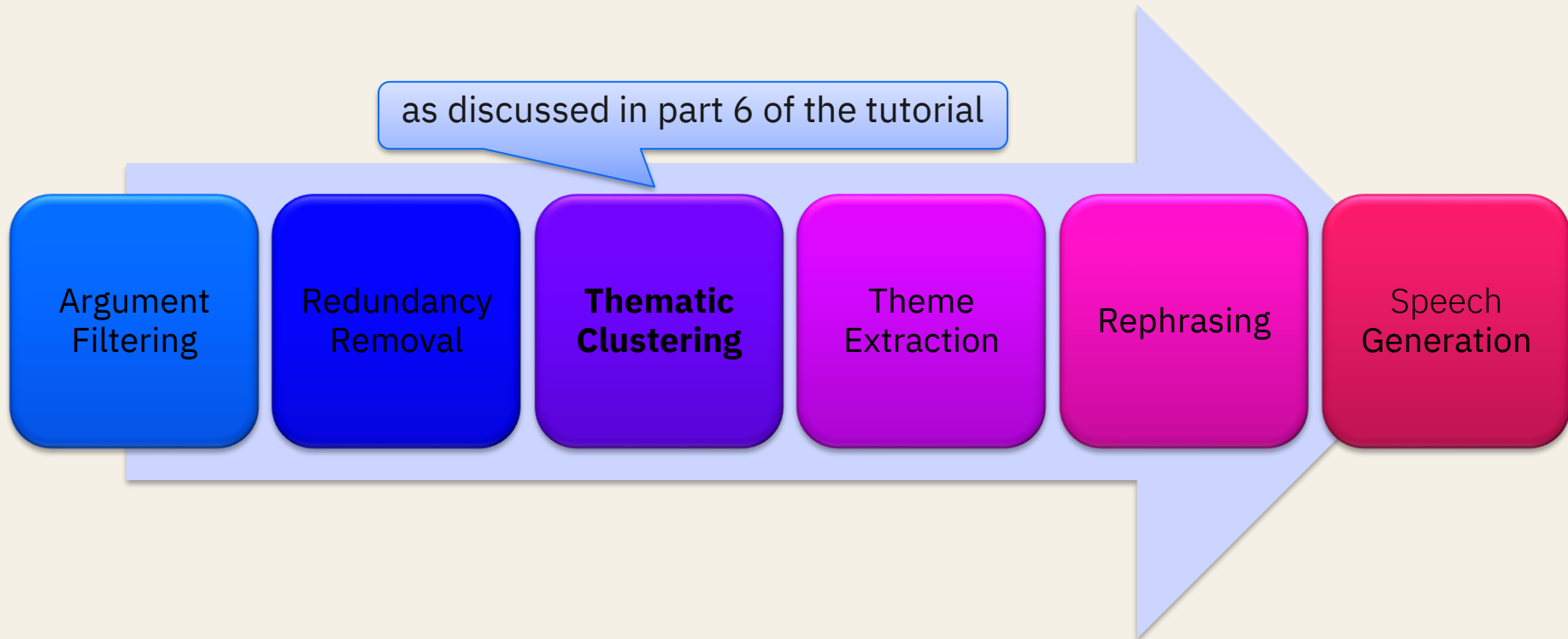
Debate Construction - Overview

- Agglomerative clustering of arguments using weighted word2vec similarity
- Keep one representative per cluster



Debate Construction - Overview

as discussed in part 6 of the tutorial

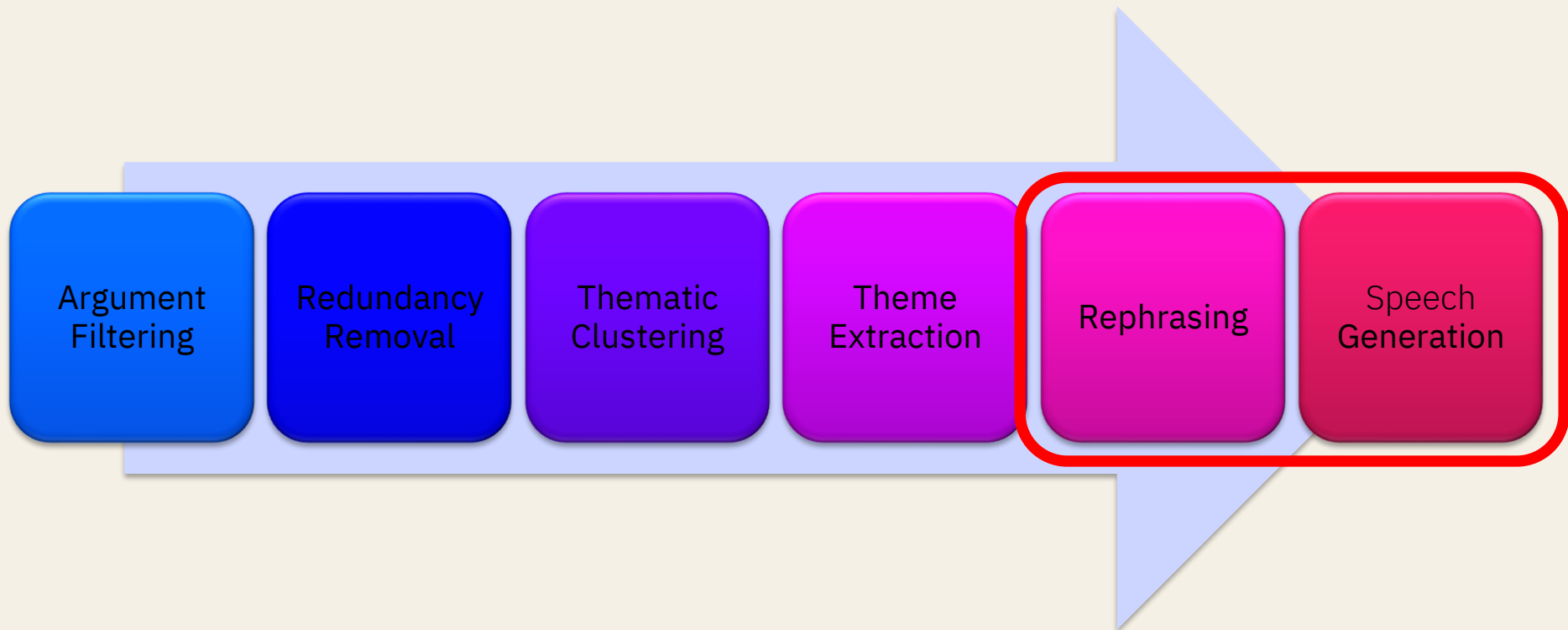


Debate Construction - Overview



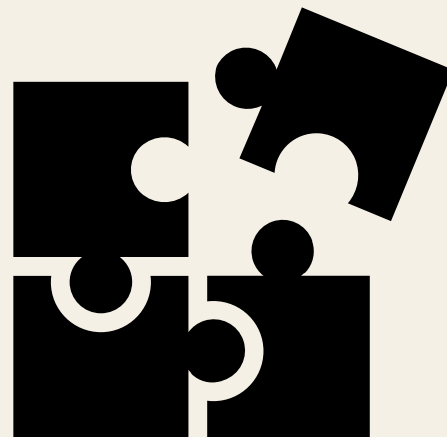
- Assign a theme (Wikipedia concept) for each cluster
- Select a theme that is both statistically enriched and frequent in the cluster
- Aim to extract a simple “theme claim” linking the theme to the topic. Will be used to introduce the theme

Debate Construction - Overview



Rephrasing

- **The challenge:** transform a collection of arguments extracted from many different documents to a coherent paragraph
- Overall – 62 different rephrasers!
 - Argument level / paragraph level
- Examples for Basic rephrasers
 - Anaphora resolution
 - Replacing a repeating noun phrase with an anaphor
 - Combining several short arguments into one sentence



Rephrasing Additional Examples 1/2

Bryan Caplan argues that higher education “is a big waste of time and money” and that “students spend thousands of hours studying subjects irrelevant to the modern labor market.”

Rephrasing Additional Examples 1/2

Who is Bryan Caplan?

Bryan Caplan argues that higher education “is a big waste of time and money” and that “students spend thousands of hours studying subjects irrelevant to the modern labor market.”

Rephrasing Additional Examples 1/2

Extracted from the same document

Bryan Caplan, an economics professor at George Mason University, argues that higher education “is a big waste of time and money” and that “students spend thousands of hours studying subjects irrelevant to the modern labor market.”

Rephrasing Additional Examples 2/2

“Scientists have abundant evidence that birth control has significant health benefits for women and their families, is documented to significantly reduce health costs, and is the most commonly taken drug in America by young and middle-aged women,” **U.S. Department of Health and Human Services Secretary Kathleen Sebelius said in a 2012 statement.**

confusing structure for spoken language

Rephrasing Additional Examples 2/2

Moved to the beginning

U.S. Department of Health and Human Services Secretary Kathleen Sebelius said in a 2012 statement that “Scientists have abundant evidence that birth control has significant health benefits for women and their families, is documented to significantly reduce health costs, and is the most commonly taken drug in America by young and middle-aged women”.

Template-Based Speech Generation

Motion: We should not abandon the Paris Agreement.

1. Greeting
2. Definition of debate topic (Paris Agreement)
3. CoPA-based opening (for “Environment”)
4. CoPA-based arguments (principled arguments against harming the environment)
5. Arguments intro – themes and theme claims



There are several issues I would like to address. They explain why we should not abandon the Paris Agreement. I will start by explaining why **the Paris Climate Accord represents the best way to address climate change [...]**.

Template-Based Speech Generation

We should not abandon the Paris Agreement.

6. {Paragraph from mined arguments cluster} × 3
7. Arguments summary
8. Pre-closing
9. CoPA-based proactive argument
10. Closing

My co-debater today will likely say that immediate human interests come first. I would like them to supply evidence showing how ruining the environment would benefit anyone in the long run.

Rebuttal Paragraph

*First, **I will address some of the points made by John Smith.** I think that one of the claims made by John was that the Paris agreement has failed. The Paris agreement certainly has flaws in its implementation, but those flaws are not essential to its mission or construction. Making changes in organizational structure and leadership are a more efficient way to fulfill the same goal.*

Keeping a Live Debate Lively

Keeping a Live Debate Lively



Giving Project Debater “AI Personality”

“While I can not experience poverty directly, and have no complaints concerning my own standards of living, I still have the following to share”.

Scripted comment for the theme **Poverty**

Concluding Words (CoPA-Basd)

Always good to end with a quote...

*Mankind has inflicted extensive damage on the environment. Yet, the fight is not over. We could save natural resources, flora and fauna if we put all of our efforts into it. Let's not give up, or we will also be remembered eventually for our brief existence on this planet. **In the words of Mahatma Gandhi: "this planet can provide for human need, but not for human greed."***

Advances in Debating Technologies

Part 8: Summary and Moving Forward

Noam Slonim

IBM Research AI



Research AI



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

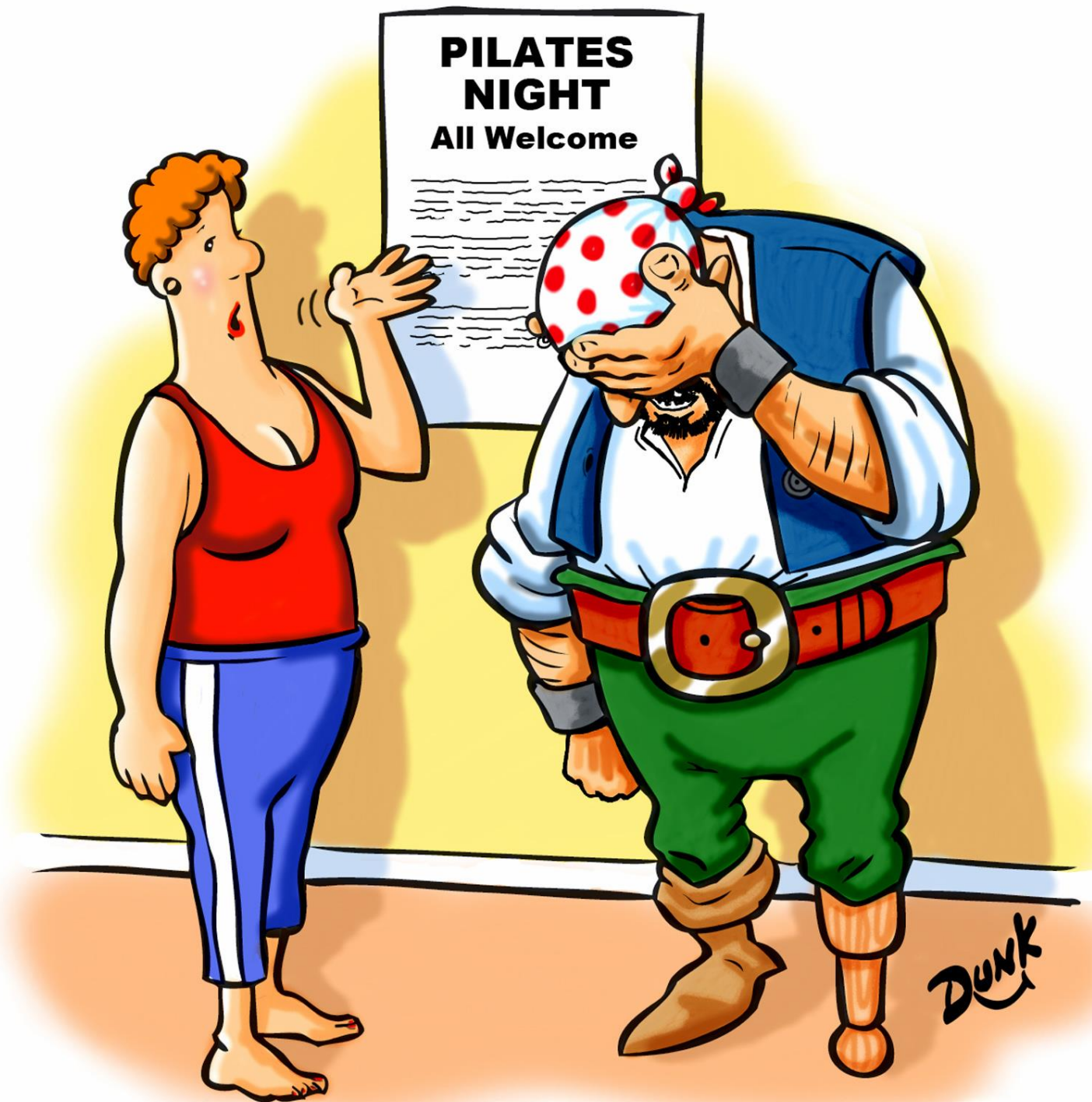
- identify key claims hidden in long continuous spoken language
- Compare to personal assistants - simple short commands

Modeling human dilemmas

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



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!”*

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.*

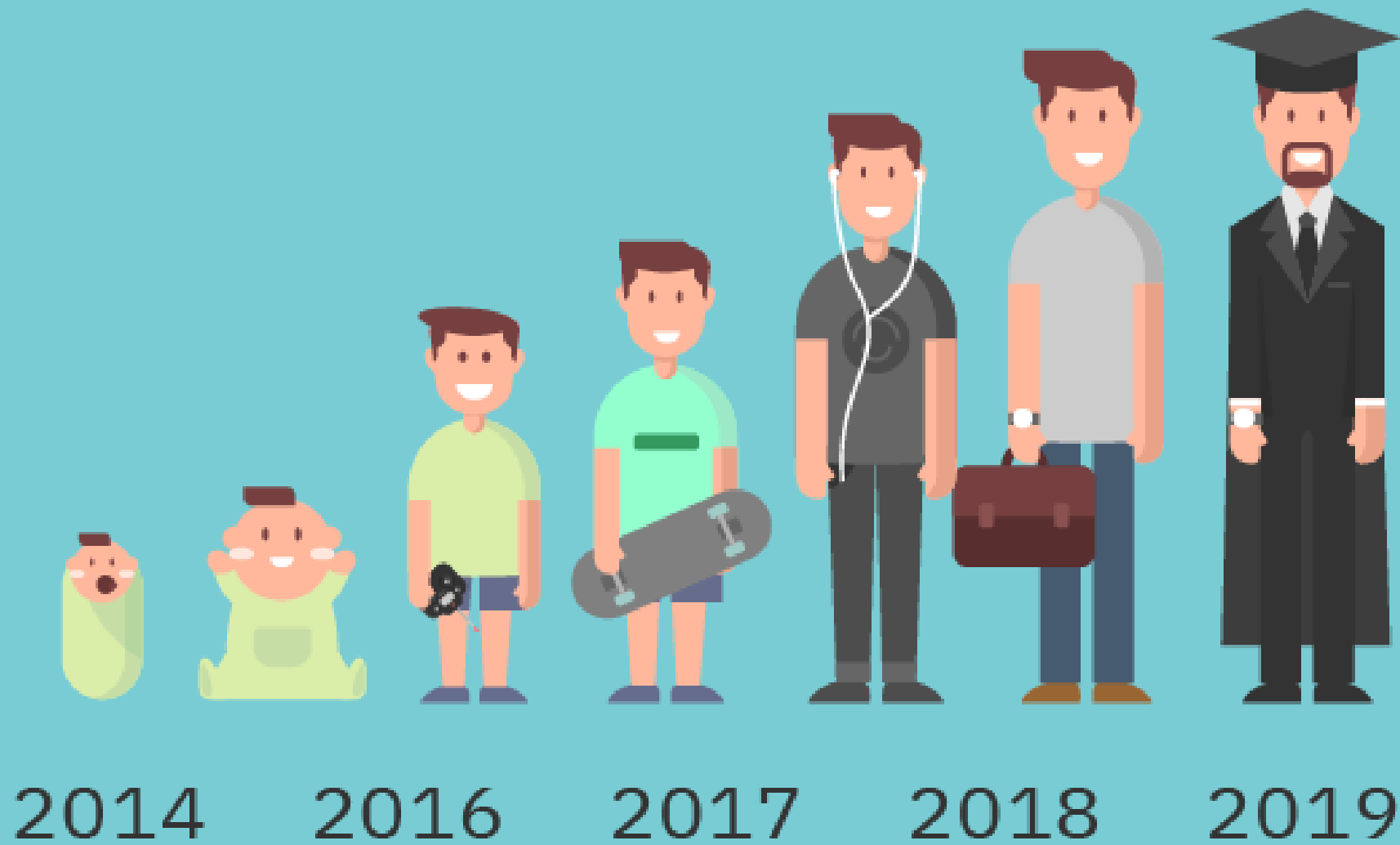


Many things can go wrong... / Examples - cont.

- 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"



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 Moshkovich, 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



How to evaluate an autonomous debating system?

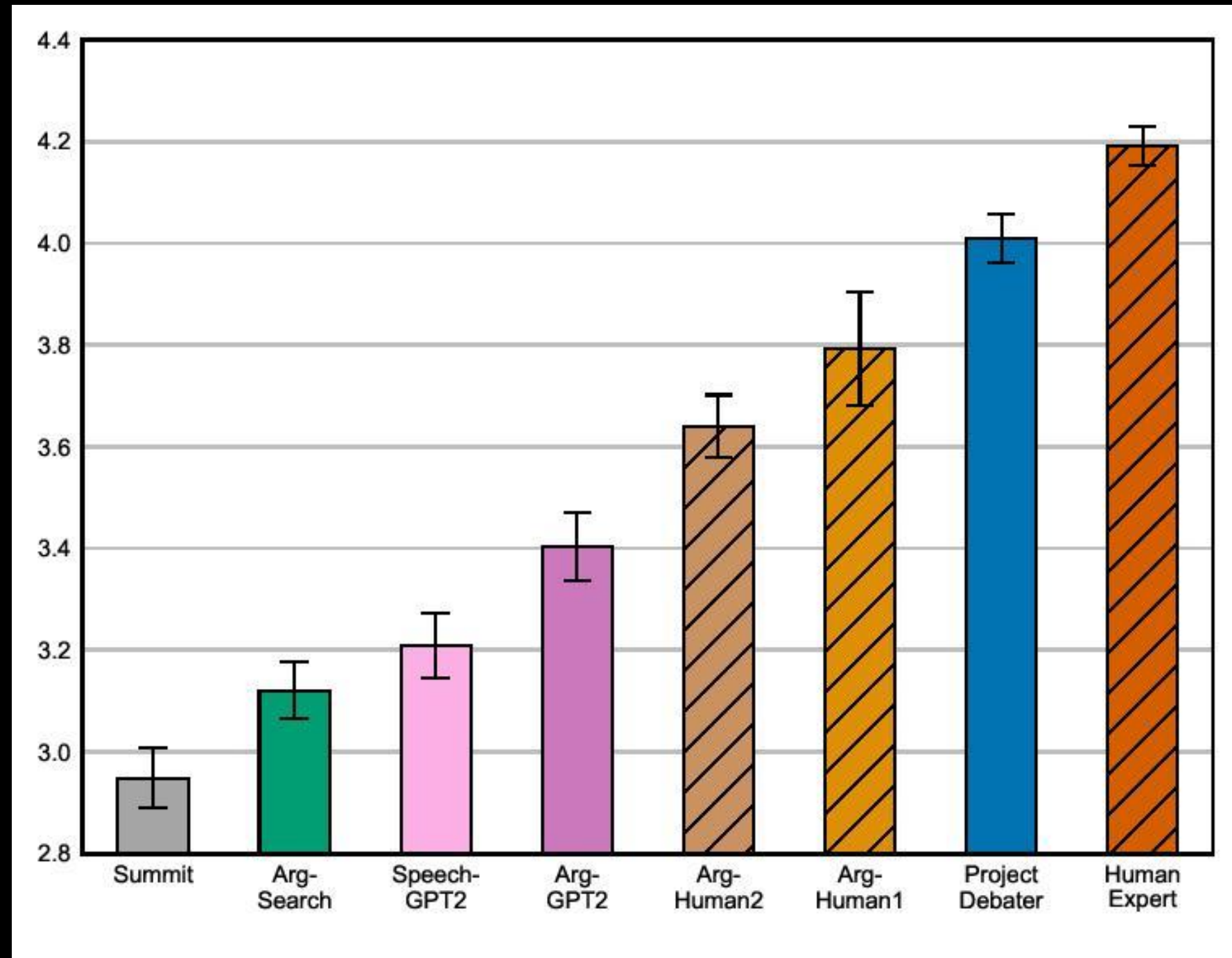
- **Public debate approach – the audience votes before and after, and the side who pulled more votes to their side is declared the ‘winner’**
- **Limitations of this approach –**
 - **Unbalanced pre-debate vote will increase the burden on the leading side**
 - **Voting is subjective and affected by various 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**

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. SUMMIT - 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, AAI 2020)
 7. Arg-Human2 – based on arguments retrieved by Project Debater and manually curated (Ein-Dor et al, AAI 2020)
 8. Human – two speeches delivered by expert human debaters

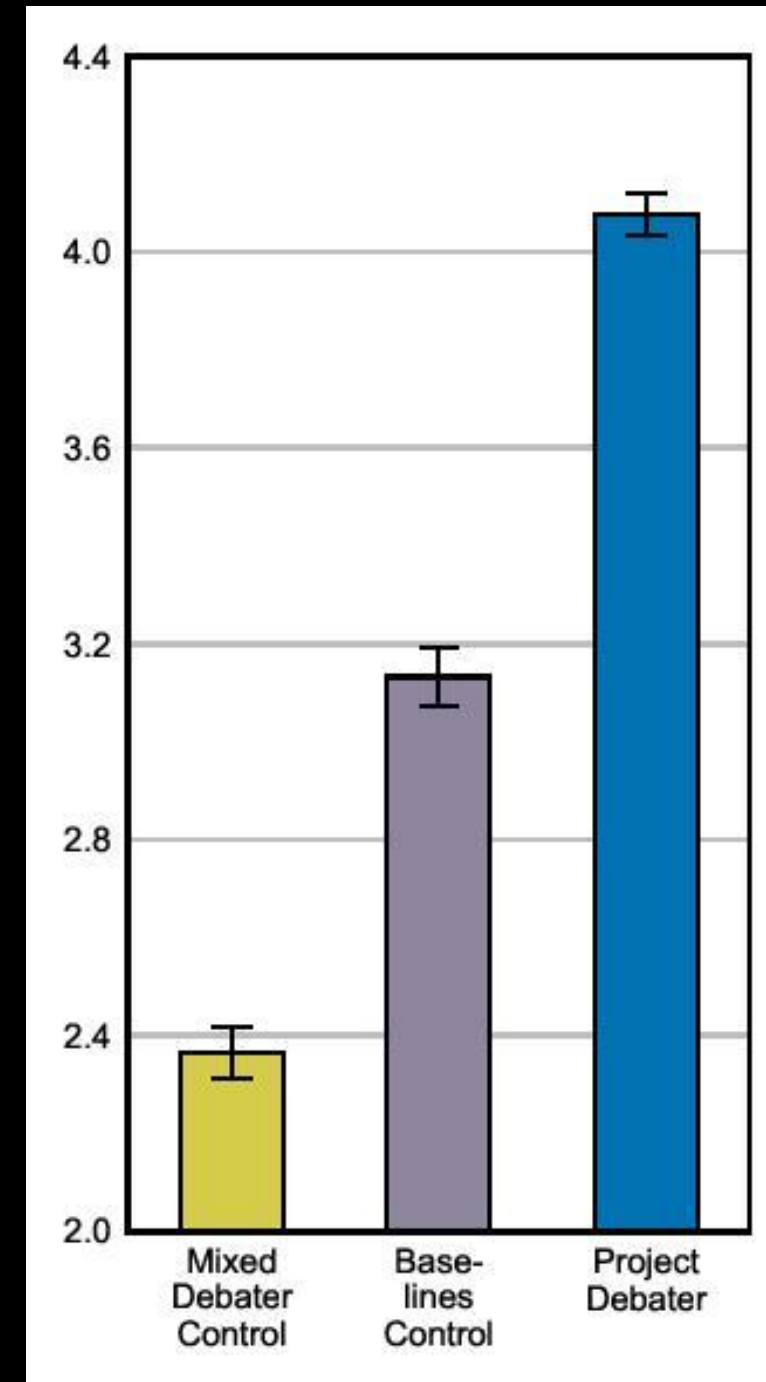
Comparison to baseline systems in producing an opening speech

- Average results over ~80 debate topics
- Each opening 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



Moving forward

- IBM Research is in a journey to develop technologies to master human language
- The Debater team mission is to develop **language technologies to enhance decision-making in enterprises**
- Informed decisions require considering pros and cons, typically done via **Reading / Consulting**
- Key example – Debater **Speech by Crowd** and Debater **Key Point Analysis**



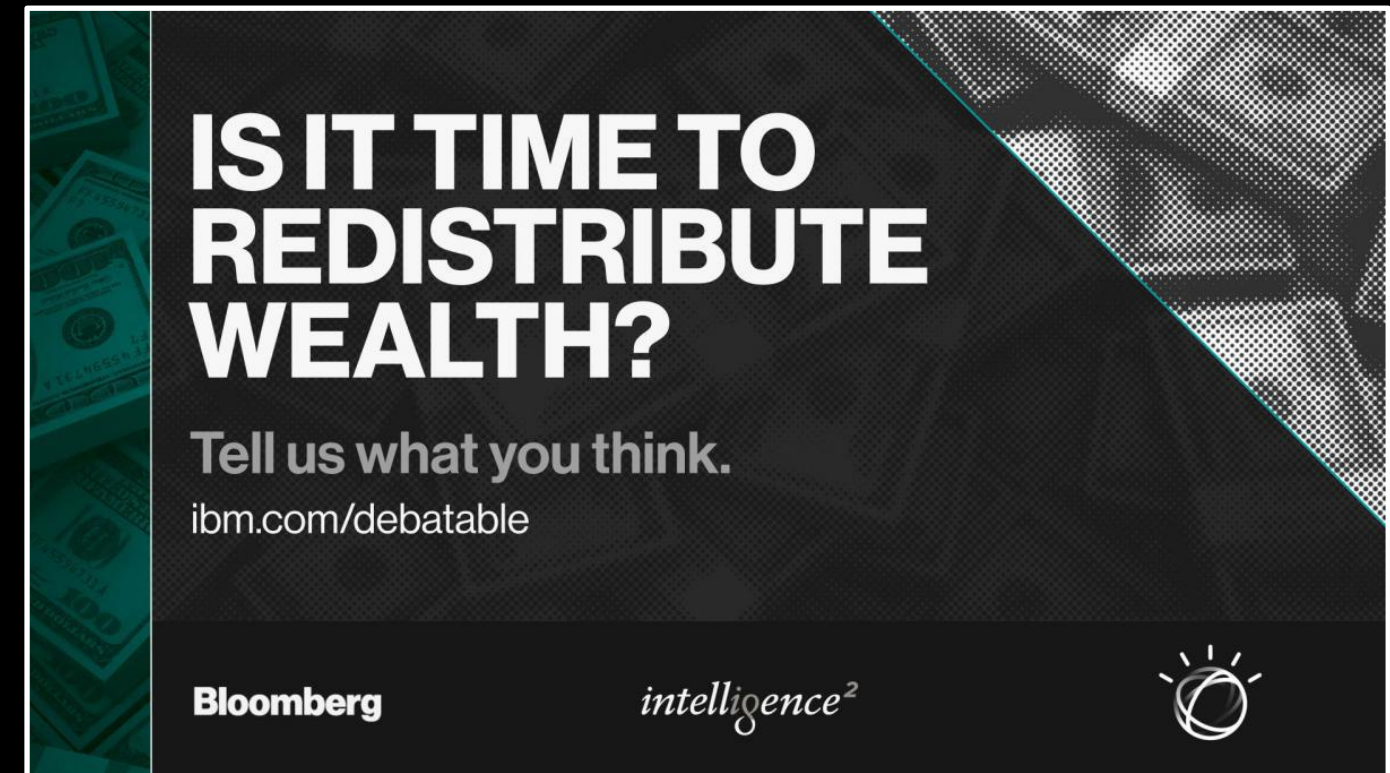
Debater Speech by Crowd

- Can we generate compelling narratives out of arguments contributed by **many** different people?
- Challenges – Pro or con? Redundant? Themes? Argument quality?
- Live demonstrations - e.g., Tel-Aviv, Las Vegas (CES), **Cambridge UK**
- Speech by Crowd Use cases – Company/Customers, Employer/Employees, Government/Citizens...
- **A new communication channel between the decision maker and those impacted by the decision?**




Debater Key Point analysis

- Summarize numerous arguments, survey responses, reviews, etc., into a concise set of key-points and their relative prevalence.
- A **quantitative** summarization technique with an associated well-defined quantitative evaluation.
- Used by Bloomberg TV & IQ2 in “That’s Debatable”
→ **Live Demo** at ibm.com/debatable
- References –
 - **From Arguments to Key Points: Towards Automatic Argument Summarization**, Bar-Haim et al, ACL 2020
 - **Quantitative Argument Summarization and Beyond: Cross-Domain Key Point Analysis**, Bar-Haim, Kantor et al, EMNLP 2020
 - **Every Bite Is an Experience: Key Point Analysis of Business Reviews**; Bar-Haim et al, ACL 2021



**IS IT TIME TO
REDISTRIBUTE
WEALTH?**

Tell us what you think.
ibm.com/debatable

Bloomberg *intelligence²* 



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
 - Workshops & Tutorials
- **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**



INPUT
Typically short texts



Debater Early Access Program



OUTPUT
Several options

- Arguments from corpus / humans
- Survey Responses
- Opinions in Reviews
- More...?

Wikification	Semantic Relatedness	Text Clustering
Claim Detection	Evidence Detection	Pro/Con Analysis
Argument Quality	Theme Extraction	Key-point Analysis
	Narrative Generation	

- 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



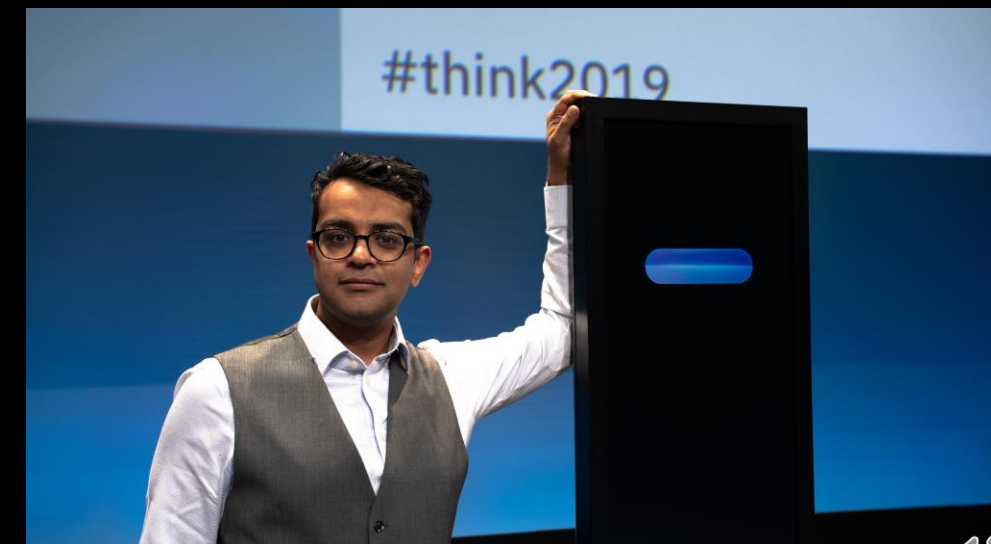
From Checkers to Debate and beyond...

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

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



A new territory for AI
Grand Challenges?





Thank you!



Advances in Debating Technologies: Building AI That Can Debate Humans

Part 9: Demo Session – Using Debating Technologies in Your Application

Elad Venezian

IBM Research AI

Project Debater has transitioned from a technology showcase in the context of a competitive debate into a business asset

designed to collegially assist people in decision making



Package APIs into deployable containers for Cloud



Work with clients on real use cases through an Early Access program



Integrate components into IBM Watson products



Continue the work to advance the research into NLP, NLU, NLG, and computational argumentation

Project Debater for Academic Use

4

Knowledge Application

Project Debater is the first AI to successfully engage with a human in a live debate. In February 2019, Project Debater debated Mr. Harish Natarajan, one of the world's leading professional debaters in an event held in San Francisco and broadcasted live worldwide.

The underlying technologies that enabled the live event are now available as software services that include core natural language understanding capabilities, argument mining, and narrative generation.

We offer free access to these services as Cloud APIs for non commercial academic use. The early access website is available at [early-access-program](#)

You can login to the website as guest , view the documentation and run online interactive demos of most of the services.

You can then request an API key to access the services from your code for your research use by mailing project.debater@il.ibm.com.

[Start using the Early Access Program >](#)

Try at:

<http://early-access-program.debater.res.ibm.com/>

In order to get an API KEY

project.debater@il.ibm.com



Project Debater Early Access Program services

Based on:

- Shnayderman et al. 2019
- Ein Dor et al. LREC 2018
- Slonim et al. 2002

Core NLU Services

Text wikification,
Concepts relatedness ,
Text clustering,
Common theme extraction

Argument Mining and Analysis

Claim detection,
Evidence detection,
Detecting claim boundaries,
Argument quality,
Pro/con classification

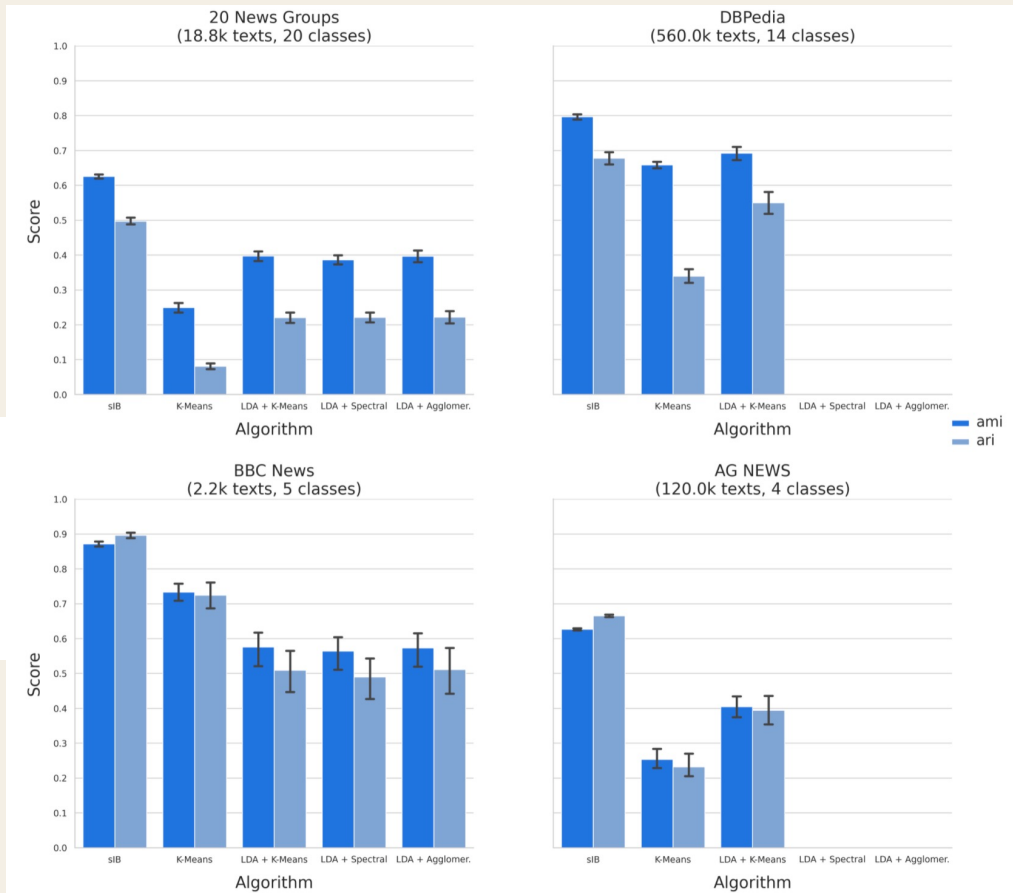
Content Summarization

Narrative generation,
Key point analysis

Text clustering – Sequential Information Bottleneck.

Unsupervised document classification using sequential information maximization (Slonim et al. 2002)

Dataset	sIB	K-Means	LDA + K-Means	LDA + Spectral	LDA + Agglomerative
20 NG	25.65	24.23	77.74	84.92	80.70
DBPedia	159.13	124.78	580.16		
BBC News	1.25	0.74	4.93	4.93	4.60
AG News	20.06	7.90	124.76		



- One of the Early Access Program services.
- Python package open source <https://github.com/IBM/sib>

Project Debater Early Access Program services

Core NLU Services

Text wikification,
Concepts relatedness ,
Text clustering,
Common theme extraction

Argument Mining and Analysis

Claim detection,
Evidence detection,
Detecting claim boundaries,
Argument quality,
Pro/con classification

Content Summarization

Narrative generation,
Key point analysis

Project Debater Early Access Program services

Based on:

- Levy et al. COLING 2014
- Ein-Dor et al. AAAI 2020
- Toledo et al. EMNLP 2019
- Gretz et al. AAAI 2020:
- Bar-Haim et al. EACL 2017
- Toledo et al. EMNLP 2020

Core NLU Services

Text wikification,
Concepts relatedness ,
Text clustering,
Common theme extraction

Argument Mining and Analysis

Claim detection,
Evidence detection,
Detecting claim boundaries,
Argument quality,
Pro/con classification

Content Summarization

Narrative generation,
Key point analysis

Project Debatr Early Access Program services

Core NLU Services

Text wikification,
Concepts relatedness ,
Text clustering,
Common theme extraction

Argument Mining and Analysis

Claim detection,
Evidence detection,
Detecting claim boundaries,
Argument quality,
Pro/con classification

Content Summarization

Narrative generation,
Key point analysis

Project Debater Early Access Program services

Based on:

- Slonim et al. Nature 2021
- Bar-Haim et al. ACL 2020
- Bar-haim et al. EMNLP 2020

Core NLU Services

Text wikification,
Concepts relatedness ,
Text clustering,
Common theme extraction

Argument Mining and Analysis

Claim detection,
Evidence detection,
Detecting claim boundaries,
Argument quality,
Pro/con classification

Content Summarization

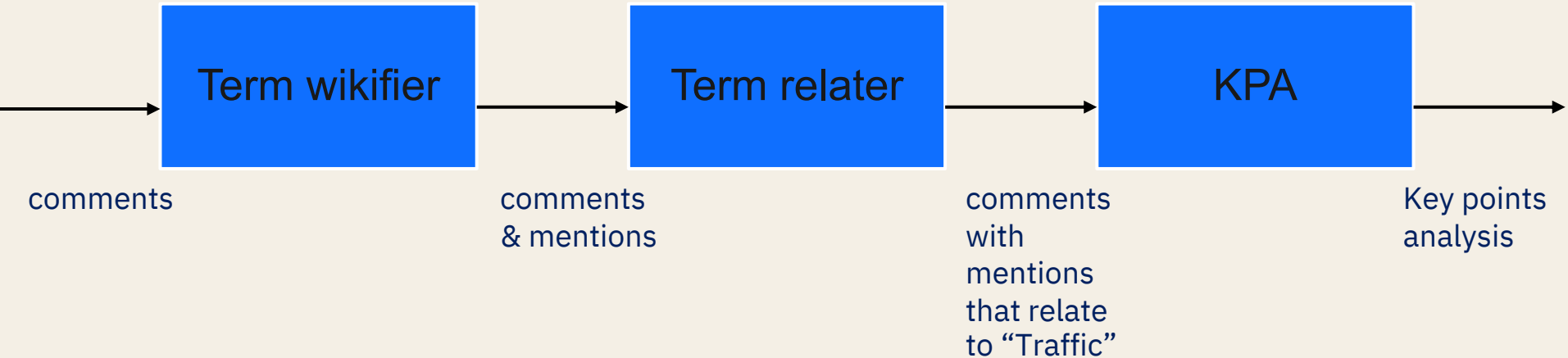
Narrative generation,
Key point analysis

Use case: Using Project Debater services for analyzing and finding insights in a survey data

<https://github.com/IBM/debater-eap-tutorial>

Deep dive into the traffic problem

Validating the match:



Summary

- The Debater Early Access program includes 12 APIs.
- It is free for non-commercial use.
- There is a gitub repository that contains examples of complex problems that can be solved with the Early Access Program, in addition to the example tab on the web page
- Efficient python implementation of SIB clustering algorithm is available at github.
- All Debater public datasets are available, there is a link at the Early Access Program website.

Thank you for listening.