



Towards argument-based explanatory dialogues: from **argument mining** to (explanatory) **argument generation**

Serena Villata

High quality explanations for AI deliberations

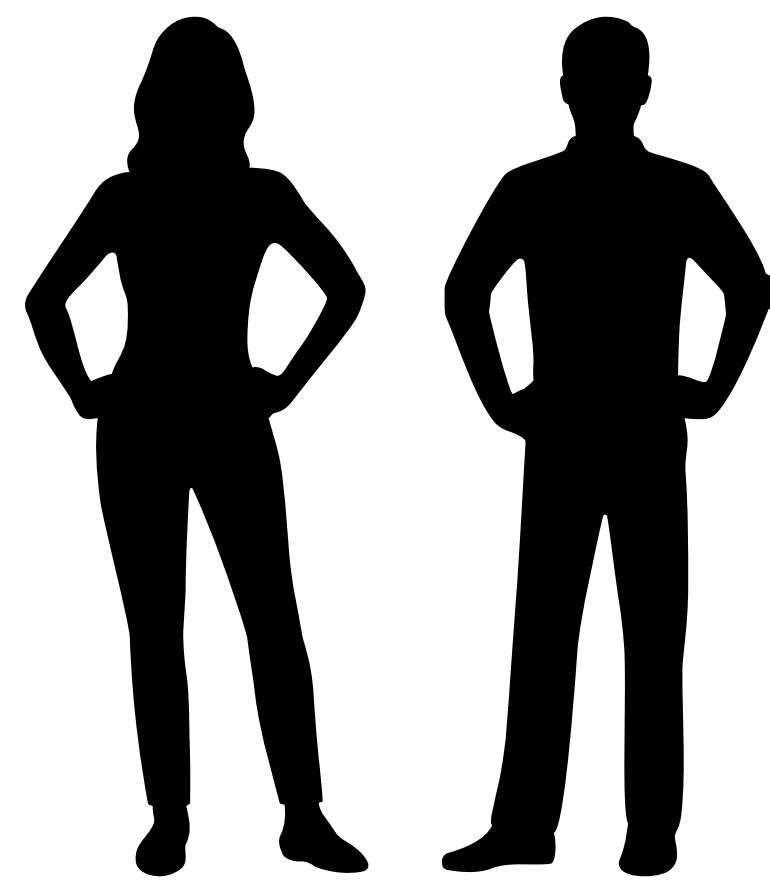
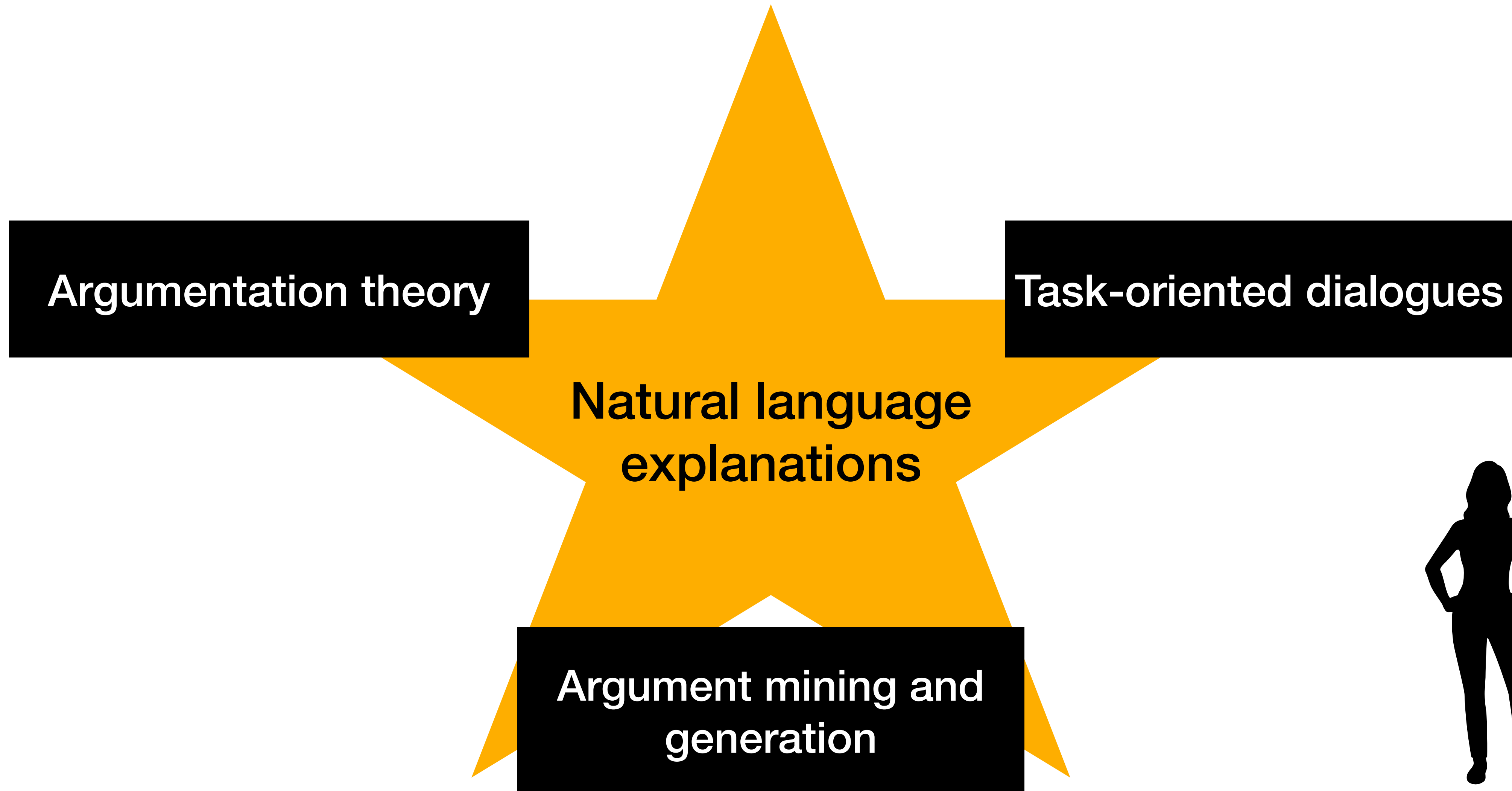
Challenges

- proper level of generality/specificity of the explanations
- reference to specific elements that have contributed to the deliberation
- analytic statements
- use of additional knowledge (common-sense knowledge, domain ontologies, knowledge bases, knowledge graphs, ...)
- use of examples (e.g., from the data the prediction is produced on)
- evidence supporting negative hypotheses

Formulate the explanation in a clearly interpretable, and possibly convincing, way

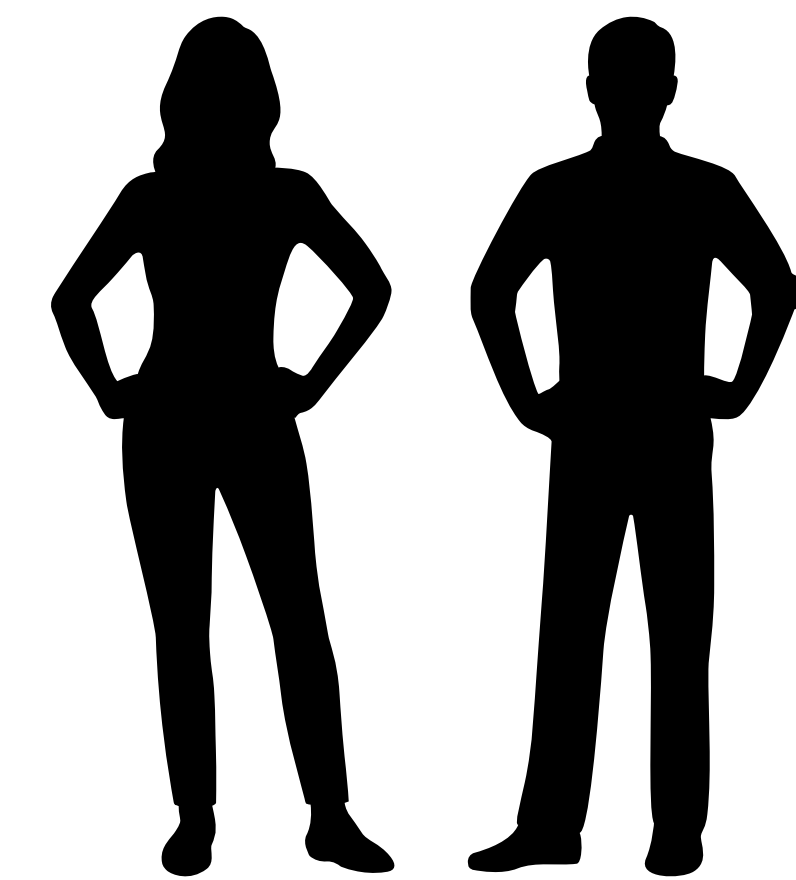
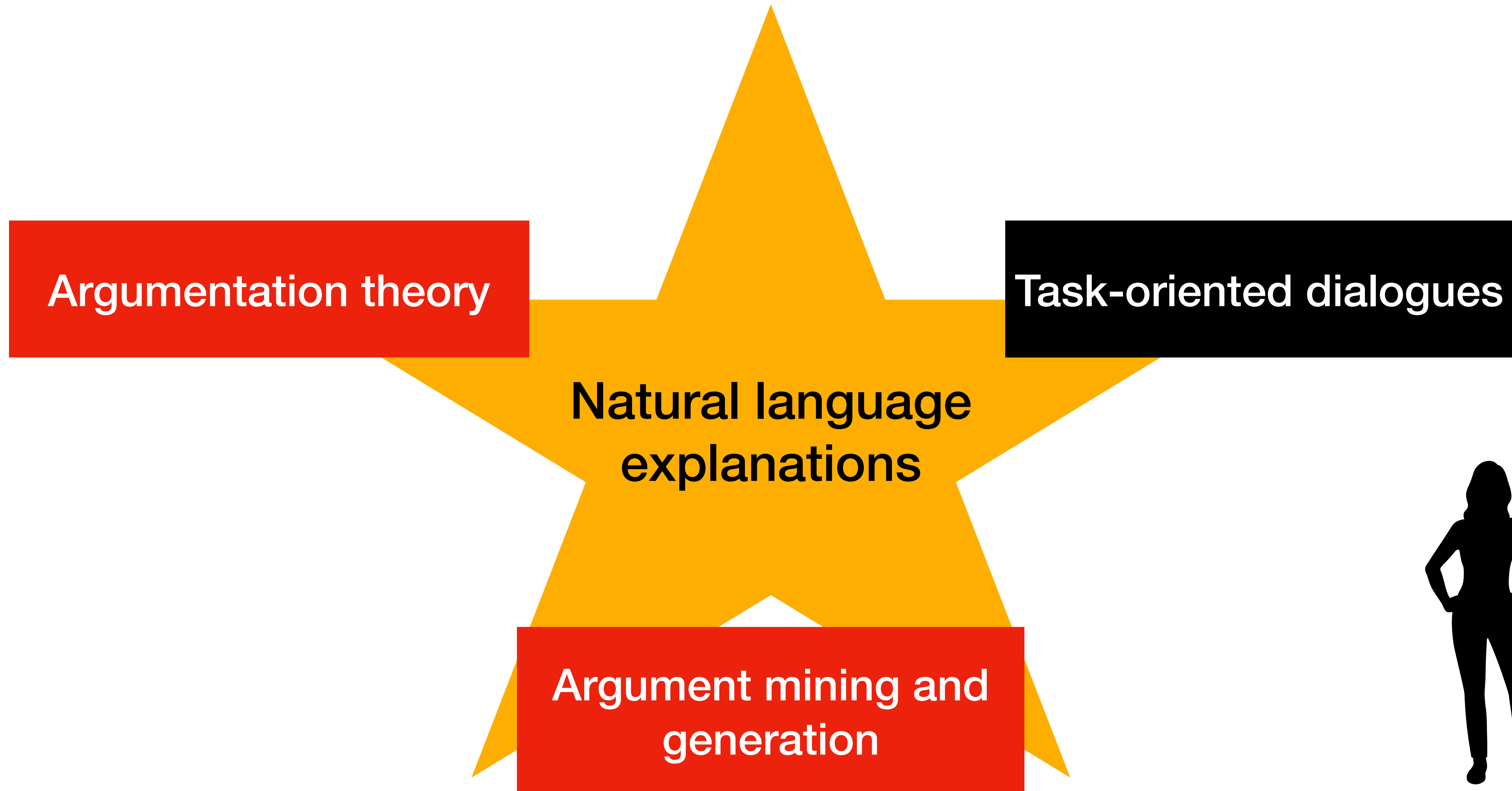
Natural language explanations

Key features



Natural language explanations

Key features



Explanatory dialogues

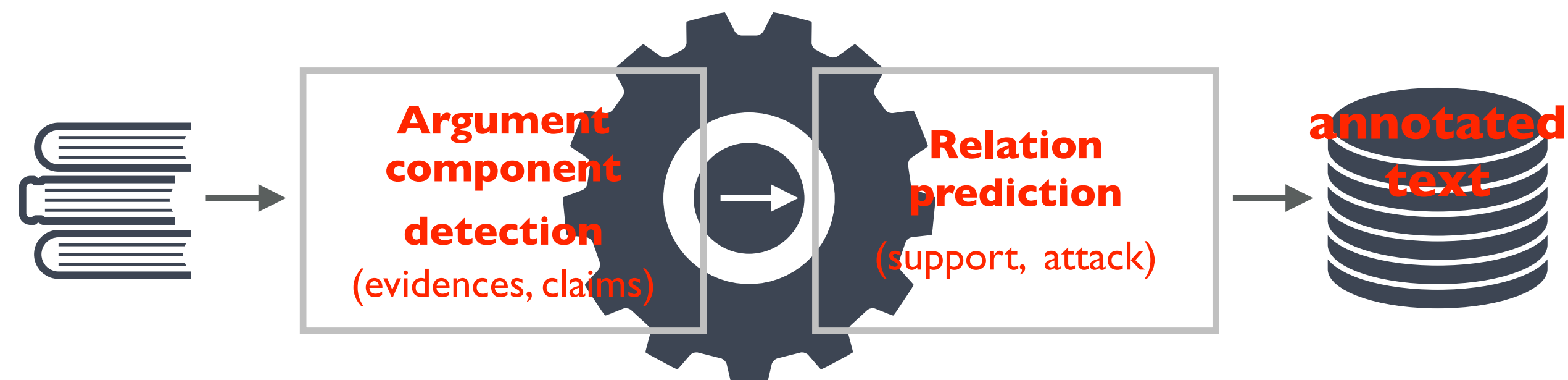
Argumentation theory

- Argumentation as reasoning-in-interaction
- Arguments need not only be rational, but “**manifestly**” **rational** (Johnson (2000))
- Arguers can see for themselves the rationale behind inferential steps taken
- In explanations
 - an agent accepts the conclusion but queries premises “OK that the diagnosis you proposed is D, but why?”
 - pragmatic goal is understanding, typically reached via causal reasoning

Explanatory argumentative dialogues

From argument mining to generation through extractive summaries

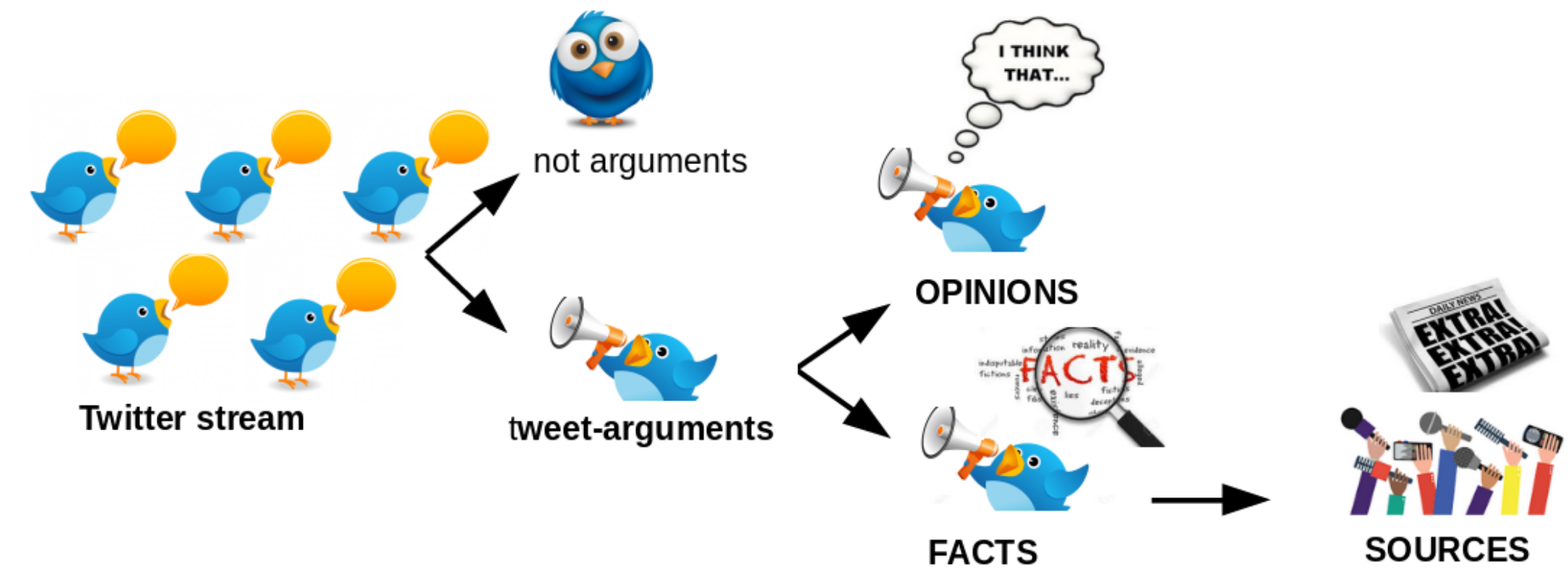
- The **task** of analysing discourse on the pragmatics level and applying a certain argumentation theory to model and automatically analyze the data at hand.
- Providing structured data for computational models of argument.
- Large resources of natural language texts: user-generated arguments on blogs, product reviews, newspapers,...
- Computational linguistics and machine learning advances.
- Argument mining IS NOT opinion mining.



Argument Mining

Argument mining

Twitter (LREC16, EMNLP17)



Tasks: argument detection (binary classification), factual vs. opinion classification, source identification.

Data: DART [Bosc et al., LREC2016], thread *#Grexit* (987 tweets) + 900 tweets from *#Brexit*.

2 annotators, IAA: $\kappa=0.767$ (1st task, 100 tweets), $\kappa=0.727$ (2nd task, 80), Dice=0.84 (3rd task, whole dataset)).

FACT: *The Guardian*: Greek crisis: European leaders scramble for response to referendum no vote. <http://t.co/cUNiyLGfg3>

OPINION: *Trump is going to sell us back to England. #Brexit #RNCinCLE*

Method and results:

Task	Method	Features	Results
argument detection	LR	lex., Twitter, synt., sem., sent.	0.78
factual/opinion classification	LR	lex., Twitter, synt., sem., sent.	0.80
source identification	Matching + heuristics		0.67

Mining argumentative structures from clinical trials

AI in Medicine 2021, ECAI20, COMMA2020, IJCAI19

Task: argument component detection (evidence, claims) and relation prediction (attack, support).

Data: 4073 argument components (2808 evidence, 1265 claims). IAA: 3 ann., 10 abs., Fleiss' $\kappa = 0.72$ (arg. comp.) and $\kappa = 0.68$ (c/e) – 2601 argument relations (2259 supports, 342 attacks). IAA: 3 ann., 30 abs., Fleiss' $\kappa = 0.62$.

Topics: neoplasm, glaucoma, hepatitis, diabetes, hypertension.

[*The diurnal intraocular pressure reduction was significant in both groups ($P < 0.001$)*]₁. [*The mean intraocular pressure reduction from baseline was 32% for the latanoprost plus timolol group and 20% for the dorzolamide plus timolol group*]₂. [*The least square estimate of the mean diurnal intraocular pressure reduction after 3 months was -7.06 mm Hg in the latanoprost plus timolol group and -4.44 mm Hg in the dorzolamide plus timolol group ($P < 0.001$)*]₃. This study clearly showed that **[the additive diurnal intraocular pressure-lowering effect of latanoprost is superior to that of dorzolamide in patients treated with timolol]**₁.

Method: Gated Recurrent Unit + Conditional Random Fields, sciBERT.

Results : evidence (F1: **0.92**), claim (F1: **0.88**), arg. comp. (F1: **0.87**) – relation classification F1: **.68**.

[Review](#) > [Infez Med.](#) 2020 Ahead of print Jun 1;28(2):198-211.

Update on treatment of COVID-19: ongoing studies between promising and disappointing results

[Silvano Esposito](#)¹, [Silvana Noviello](#)¹, [Pasquale Pagliano](#)¹

Affiliations + expand

PMID: 32335561

[Free article](#)

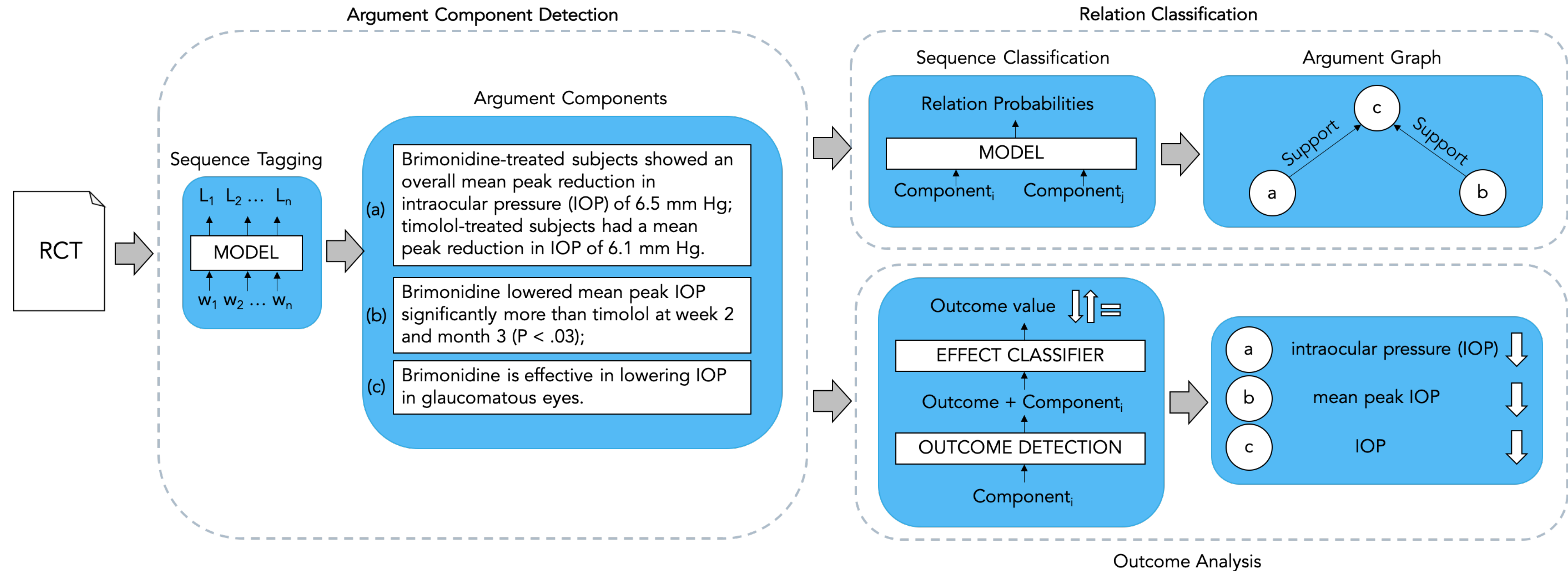
Abstract

The COVID-19 pandemic represents the greatest global public health crisis since the pandemic influenza outbreak of 1918. We are facing a new virus, so several antiviral agents previously used to treat other coronavirus infections such as SARS and MERS are being considered as the first potential candidates to treat COVID-19. Thus, several agents have been used by the beginning of the current outbreak in China first and all over the world successively, as reported in several different guidelines and therapeutic recommendations. At the same time, a great number of clinical trials have been launched to investigate the potential efficacy therapies for COVID-19 highlighting the urgent need to get as quickly as possible high-quality evidence. Through PubMed, we explored the relevant articles published on treatment of COVID-19 and on trials ongoing up to April 15, 2020.

Collaborations:
INSERM, CHU Nice

Mining argumentative structures from clinical trials

AI in Medicine 2021, ECAI20, COMMA2020, IJCAI19



Argumentative Clinical Trial Analysis

The image displays the ACTA web application interface. At the top, the logo 'ACTA Argumentative Clinical Trial Analysis' is visible. A navigation menu includes 'Home', 'About', 'Contacts', and 'Services'. The main content area features a network diagram with a central node 'Claim ID:5' connected to 'Claim ID:1' and four evidence nodes: 'Evidence ID:0', 'Evidence ID:2', 'Evidence ID:3', and 'Evidence ID:4'. Below the diagram, two panels show abstract text from clinical trials. The left panel shows the text with various components highlighted in different colors (green, red, purple, blue, orange) to represent argumentative elements. The right panel shows the same text with a different highlighting scheme. At the bottom of each panel, there are buttons for 'Highlight Argumentative Components' and 'Highlight PICO Elements'.

ACTA Argumentative Clinical Trial Analysis

Home About Contacts Services

22340282: Topical photodynamic therapy (PDT) with aminolevulinic acid (ALA) and 5% [...]

21871978: The postoperative clinical superiority of the

Claim ID:1

Claim ID:5

Evidence ID:0

Evidence ID:2

Evidence ID:3

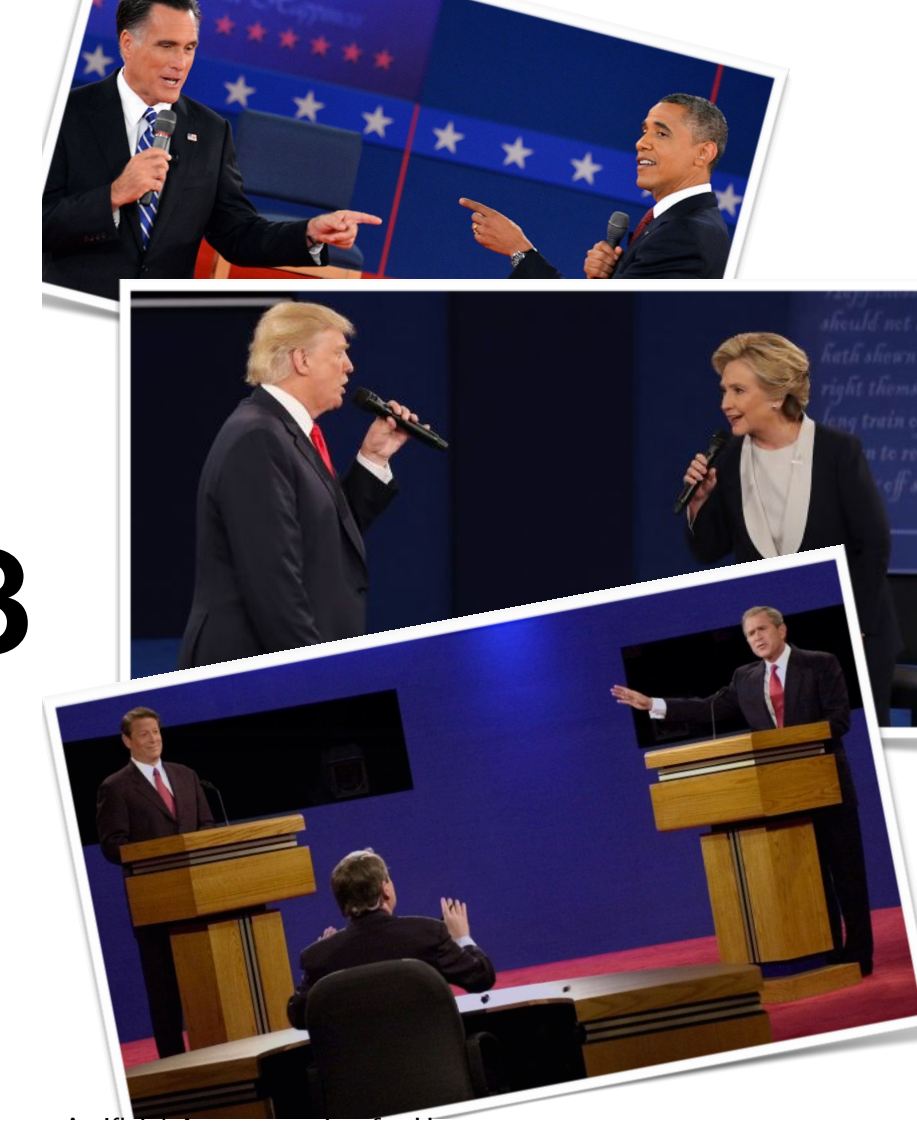
Evidence ID:4

Abstract: One attempt to improve long-term survival in patients with advanced ovarian cancer was thought to be the addition of more non-cross-resistant drugs to platinum-paclitaxel combination regimens. Gemcitabine was among the candidates for a third drug. We performed a prospective, randomized, phase III, intergroup trial to compare carboplatin plus paclitaxel (TC; area under the curve [AUC] 5 and 175 mg/m², respectively) with the same combination and additional gemcitabine 800 mg/m² on days 1 and 8 (TCG) in previously untreated patients with advanced epithelial ovarian cancer. TC was administered intravenously (IV) on day 1 every 21 days for a planned minimum of six courses. Gemcitabine was administered by IV on days 1 and 8 of each cycle in the TCG arm. Between 2002 and 2004, 1,742 patients were randomly assigned; 882 and 860 patients received TC and TCG, respectively. Grades 3 to 4 hematologic toxicity and fatigue occurred more frequently in the TCG arm. Accordingly, quality-of-life analysis during chemotherapy showed a disadvantage in the TCG arm. Although objective response was slightly higher in the TCG arm, this did not translate into improved progression-free survival (PFS) or overall survival (OS). Median PFS was 17.8 months for the TCG arm and 19.3 months for the TC arm (hazard ratio [HR], 1.18; 95% CI, 1.06 to 1.32; P = .0044). Median OS was 49.5 for the TCG arm and 51.5 months for the TC arm (HR, 1.05; 95% CI, 0.91 to 1.20; P = .5106). The addition of gemcitabine to carboplatin plus paclitaxel increased treatment burden, reduced PFS time, and did not improve OS in patients with advanced epithelial ovarian cancer. Therefore, we recommend no additional clinical use of TCG in this population.

Highlight Argumentative Components Highlight PICO Elements

Mining political arguments

COLING20, IJCAI19 demo, ACL19 short, AAAI18



39 political debates
from the last 50 years
of US presidential
campaigns (29k
argument components)

↓
Argument Mining
for fallacies detection

Task: argument component detection (claim, premises) and relation classification (attack, support).

Data: 29521 argument components (16087 claims and 13434 premises) and 25012 relations (3723 attacks and 21289 supports). IAA: 3 ann., moderate/faire agreement.

Method: LSTM + Fine tuned BERT

Results: evidence (F1: **0.72**), claim (F1: **0.69**), argument components (F1: **0.84**), relation classification (F1: **0.68**)

Collaborations:
Univ. of
Luxembourg

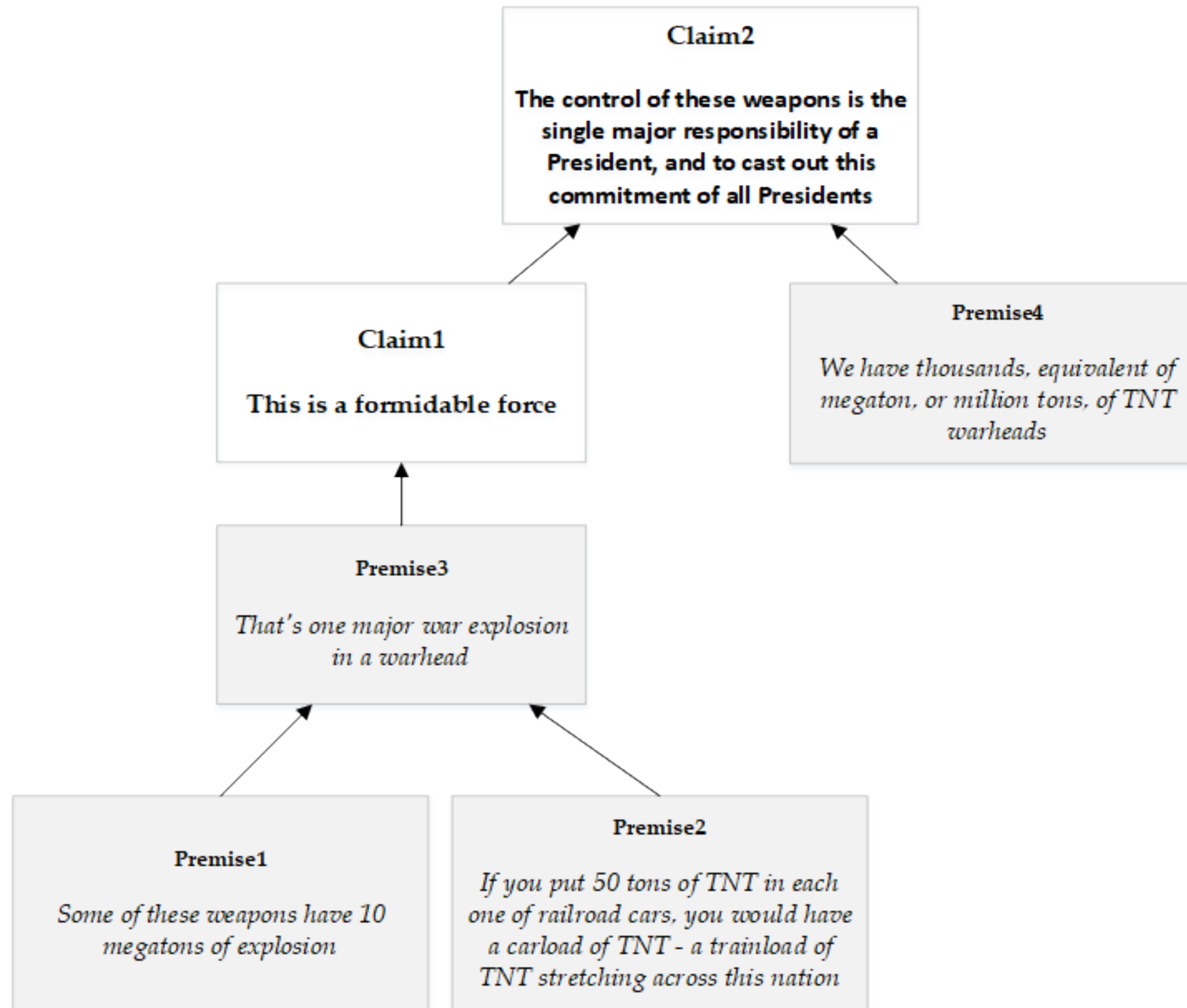
Mining political arguments

COLING20, IJCAI19 demo, ACL19 short, AAAI18



39 political debates
from the last 50 years
of US presidential
campaigns (29k
argument components)

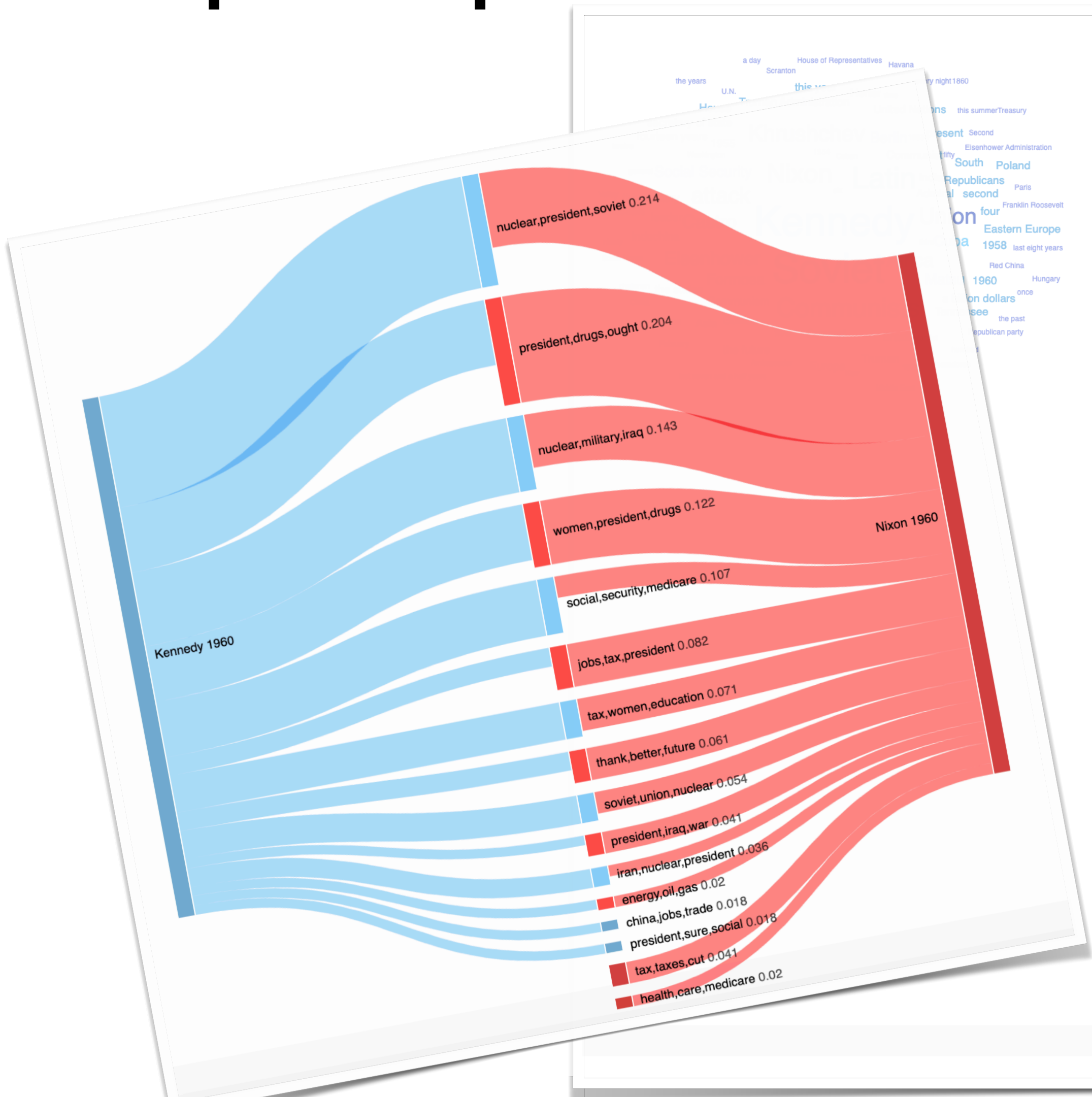
Argument Mining
for fallacies detection



Collaborations:
Univ. of
Luxembourg

DispuTOOL

<https://disputool.uni.lu/>



Filter data

Based on Year:

- 1960
- 1976
- 1980
- 1984
- 1988
- 1992
- 1996
- 2000
- 2004
- 2008
- 2012
- 2016

Based on NER Type:

- LOCATION
- NATIONALITY
- ORGANIZATION
- PERSON
- RELIGION

Based on Speaker:

- Albert A. Gore
- Barack H. Obama
- Donald J. Trump
- George H. W. Bush
- George W. Bush
- Gerald R. Ford
- Geraldine A. Ferraro
- Henry Ross Perot
- Hillary D. R. Clinton
- Jack F. Kemp
- James B. Stockdale
- James D. Quayle
- Jimmy E. Carter
- John B. Anderson
- John F. Kennedy
- John F. Kerry
- John S. McCain
- Johnny(John) R. Edwards
- Joseph I. Lieberman
- Joseph(Joe) R. Biden
- Lloyd M. Bentsen
- Michael S. Dukakis
- Paul D. Ryan
- Richard M. Nixon
- Richard(Dick) B. Cheney
- Robert J. Dole
- Ronald W. Reagan
- Sarah L. Palin
- Walter F. Mondale
- Willard(Mitt) M. Romney
- William(Bill) J. Clinton

APPLY FILTER

21 Oct 1960

Filter

- Highlight Claims.
- Highlight Premises.

Our policies are very different. I think that Senator Kennedy's policies and recommendations for the handling of the Castro regime are probably the most dangerous- dangerously irresponsible recommendations that he's made during the course of this campaign. In effect, what Senator Kennedy recommends is that the United States government should give help to the exiles and to those within Cuba who oppose the Castro regime - provided they are anti-Batista. Now let's just see what this means. We have five treaties with Latin America, including the one setting up the Organization of American States in Bogota in 1948, in which we have agreed not to intervene in the internal affairs of any other American country - and they as well have agreed to do likewise. The charter of the United Nations - its Preamble, Article I and Article II - also provide that there shall be no intervention by one nation in the internal affairs of another. Now I don't know what Senator Kennedy suggests when he says that we should help those who oppose the Castro regime, both in Cuba and without. But I do know this: that if we were to follow that recommendation, that we would lose all of our friends in Latin America, we would probably be condemned in the United Nations, and we would not accomplish our objective. I know something else. It would be an open invitation for Mr. Khrushchev to come in, to come into Latin America and to engage us in what would be a civil war, and possibly even worse than that. This is the major recommendation that he's made. Now, what can we do? Well, we can do what we did with Guatemala. There was a Communist dictator that we inherited from the previous Administration. We quarantined Mr. Arbenz. The result was that the Guatemalan people themselves eventually rose up and they threw him out. We are quarantining Mr. Castro today. We're quarantining him diplomatically by bringing back our Ambassador: economically by cutting off trade, and Senator Kennedy's suggestion that the trade that we cut off is not significant is just one hundred percent wrong. We are cutting off the significant items that the Cuban regime needs in order to survive. By cutting off trade, by cutting off our diplomatic relations as we have, we will quarantine this regime so that the people of Cuba themselves will take care of Mr. Castro. But for us to do what Senator Kennedy has suggested would bring results which I know he would not want, and certainly which the American people would not want.



Richard M. Nixon



John F. Kennedy

Mr. Nixon uh - shows himself i- misinformed. He surely must be aware that most of the equipment and arms and resources for Castro came from the United States, flowed out of Florida and other parts of the United States to Castro in the mountains. There isn't any doubt about that, number one. Number two, I believe that if any economic sanctions against Latin America are going to be successful they have to be multilateral. They have to include the other countries of Latin America. The very minute effect of the action which has been taken this week on Cuba's economy - I believe Castro can replace those markets very easily through Latin America, through Europe, and through Eastern Europe. If the United States had stronger prestige and influence in Latin America it could persuade - as Franklin Roosevelt did in 1940 - the countries of Latin America to join in an economic quarantine of Castro. That's the only way you can bring real economic pressure on the Castro regime - and also the countries of Western Europe, Canada, Japan and the others. Number three, Castro is only the beginning of our difficulties throughout Latin America. The big struggle will be to prevent the influence of Castro spreading to other countries - Mexico, Panama, Bolivia, Colombia. We're going to have to try to provide closer ties, to associate ourselves with the great desire of these people for a better life if we're going to prevent Castro's influence from spreading throughout all of Latin America. His influence is strong enough today to prevent us from joining - getting the other countries of Latin America to join with us in economic quarantine. His influence is growing - mostly because this Administration has ignored Latin America. You yourself said, Mr. Vice President, a month ago, that if we had provided the kind of economic aid five years ago that we are now providing we might never have had Castro. Why didn't we?

Explanatory arguments (and their further use in dialogues)

Argument-based explanation patterns

(Darpa XAI Program Update)

- **analytic statements** in NL that describe the elements and context that support a choice,
 - ➡ the arguments (evidence, claim, warrant if any)
- **visualizations** that highlight portions of the raw data that support a choice,
- cases that invoke **specific examples**, and
 - ➡ hard, you need more than one case to support by examples the choice
- **rejections of alternative choices** that argue against less preferred answers based on analytics, cases, and data.
 - ➡ hard, you need the arguments from the rejected options

Use case example to build the dataset

A 37-year-old woman is brought to the emergency department because of intermittent chest pain for 3 days. The pain is worse with inspiration, and she feels she cannot take deep breaths. She has not had shortness of breath, palpitations, or nausea. She had an upper respiratory tract infection 10 days ago and took an over-the-counter cough suppressant and decongestant and acetaminophen. Her temperature is 37.2°C (98.9°F), pulse is 90/min, and blood pressure is 122/70 mm Hg. The lungs are clear to auscultation. S1 and S2 are normal. A rub is heard during systole. There is no peripheral edema. An ECG shows normal sinus rhythm and diffuse, upwardly concave ST-segment elevation and PR-segment depression in leads II, III, and a VF.

Use case example

Training residents to improve argument-based diagnosis

Which of the following is the most likely diagnosis?

- (A) Acute pericarditis
- (B) Aortic dissection
- (C) Gastroesophageal reflux disease
- (D) Myocardial infarction
- (E) Peptic ulcer disease
- (F) Pulmonary embolism
- (G) Unstable angina pectoris

**ALTERNATIVE
OPTIONS**

Use case example

Training residents to improve argument-based diagnosis

Which of the following is the most likely diagnosis?

(A) Acute pericarditis

(B) Aortic dissection

(C) Gastroesophageal reflux disease

(D) Myocardial infarction

(E) Peptic ulcer disease

(F) Pulmonary embolism

(G) Unstable angina pectoris

**ALTERNATIVE
OPTIONS**

Use case example

Training residents to improve argument-based diagnosis

Which of the following is the most likely diagnosis?

(A) Acute pericarditis

Why?

A friction rub and diffuse low-grade ST-segment elevation equals pericarditis.

Use case example

- Clinical case: a 37-year-old woman is brought to the emergency department because of intermittent chest pain for 3 days. The pain is worse with inspiration, and she feels she cannot take deep breaths. She has not had shortness of breath, palpitations, or nausea. She had an upper respiratory tract infection 10 days ago and took an over-the-counter cough suppressant and decongestant and acetaminophen. Her temperature is 37.2°C (98.9°F), pulse is 90/min, and blood pressure is 122/70 mm Hg. The lungs are clear to auscultation. S1 and S2 are normal. A rub is heard during systole. There is no peripheral edema. An ECG shows normal sinus rhythm and diffuse, upwardly concave ST-segment elevation and PR-segment depression in leads II, III, and a VF.
- Diagnosis: the patient is showing a pericarditis **because** she has a friction rub and diffuse low-grade ST-segment elevation.

First step: **extractive** explanatory argument generation

- Clinical case: *[a 37-year-old woman is brought to the emergency department because of intermittent chest pain for 3 days]. [The pain is worse with inspiration], and she feels [she cannot take deep breaths]. [She has not had shortness of breath, palpitations, or nausea]. [She had an upper respiratory tract infection 10 days ago] and [took an over-the-counter cough suppressant and decongestant and acetaminophen]. [Her temperature is 37.2°C (98.9°F)], [pulse is 90/min], and [blood pressure is 122/70 mm Hg]. [The lungs are clear to auscultation]. [S1 and S2 are normal]. [A rub is heard during systole]. [There is no peripheral edema]. [An ECG shows normal sinus rhythm and diffuse], [upwardly concave ST-segment elevation] and [PR-segment depression in leads II, III, and a VF].*
- Diagnosis: the patient is showing a pericarditis **because** *[a rub is heard during systole]* and the ECG shows *[concave ST-segment elevation]*.

Extractive explanatory argument generation

Argument Mining + Knowledge graphs

- **Diagnosis with explanation by expert:** the patient is showing a pericarditis **because** she has a friction rub and diffuse low-grade ST-segment elevation.
- **Diagnosis with extracted explanatory arguments:** the patient is showing a pericarditis **because** [*a rub is heard during systole*] and the ECG shows [*concave ST-segment elevation*].
- **What we have?**
 - Premises extracted from description of the case, correct diagnosis.
- **What we need further?**
 - Criteria to choose among the premises to pick the right ones, those which justify the diagnosis → knowledge graphs of clinical knowledge
 - What if the explanation is not “contained” in the evidence ?

Explanatory dialogues

Argument mining and generation

- (Counter-)argument generation SoA (e.g., (Park et al., 2019, Hua et al., 2019)): mainly reformulation of arguments mined from Wikipedia and newspaper articles
- Insufficient to generate effective and interactive explanatory arguments
- **Extractive argument generation vs. abstractive argument generation**
- Large-scale unsupervised language models to generate arguments
- **Explanatory arguments meet high quality arguments:**
 - quality (i.e., variability of the explanatory arguments, no repetitiveness)
 - quantity
 - standard evaluation metrics: BLEU and BertScore

Main open challenges

- **(Annotated) Data**
- **World knowledge and specific domain knowledge**
 - To allow for generalisations, instantiations, inferences
- **How to evaluate explanatory dialogues?**
 - quality and quantity of the generated arguments
 - structural simplicity, coherence, minimality
 - what else?
- **Are these explanations actually for humans?** If so, human feedback required!



Serena Villata

CR1 CNRS, HDR

Université Cote d'Azur, CNRS, Inria

Laboratoire I3S (SPARKS-WIMMICS team)



serena.villata@univ-cotedazur.fr



<http://www.i3s.unice.fr/~villata/>



@serena_villata

Thanks !