Towards Natural Language Explanatory Argument Generation: Achieved Results (a few) and Open Challenges (many)

Serena Villata

Toulouse — March 7th, 2022

Artificial Argumentation for Humans

Keywords: Argumentation, Natural Language Processing

- **Argumentation-enhanced intelligent machines** require argumentation technologies to
  - support the **interactive explanation** of the outcome of the deliberation process (**why** the machine deliberated in a certain way) taking into account the user **feedback** through natural language argumentative explanations,
  - **mine**, **analyse**, **summarise**, and **generate** natural language argument structures from different settings (e.g., clinical trials, political debates, legal cases).

High quality explanations for AI deliberations

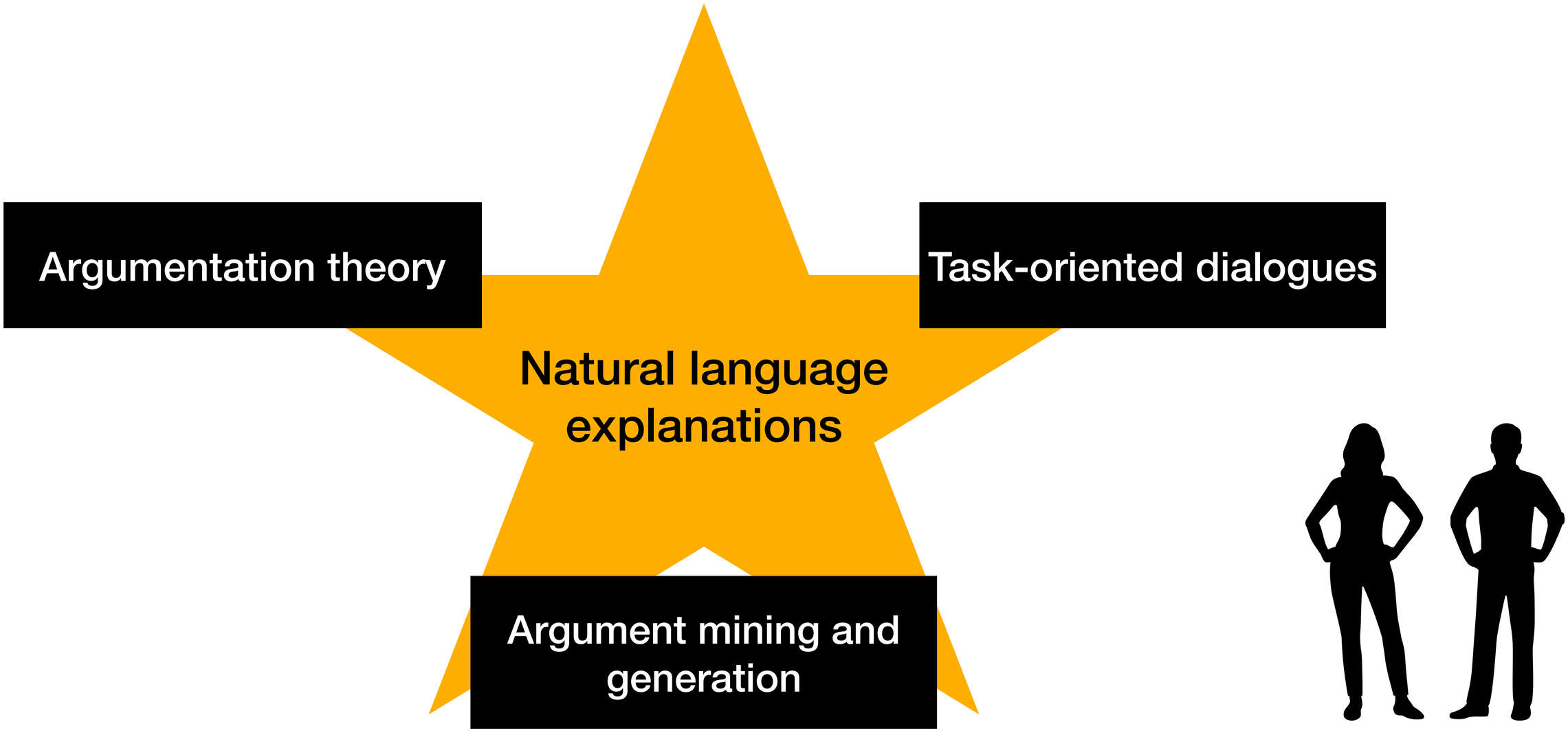
Challenges

- proper level of **generality/specificity** of the explanations
- reference to **specific elements** that have contributed to the deliberation
- analytic statements (e.g., arguments)
- 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

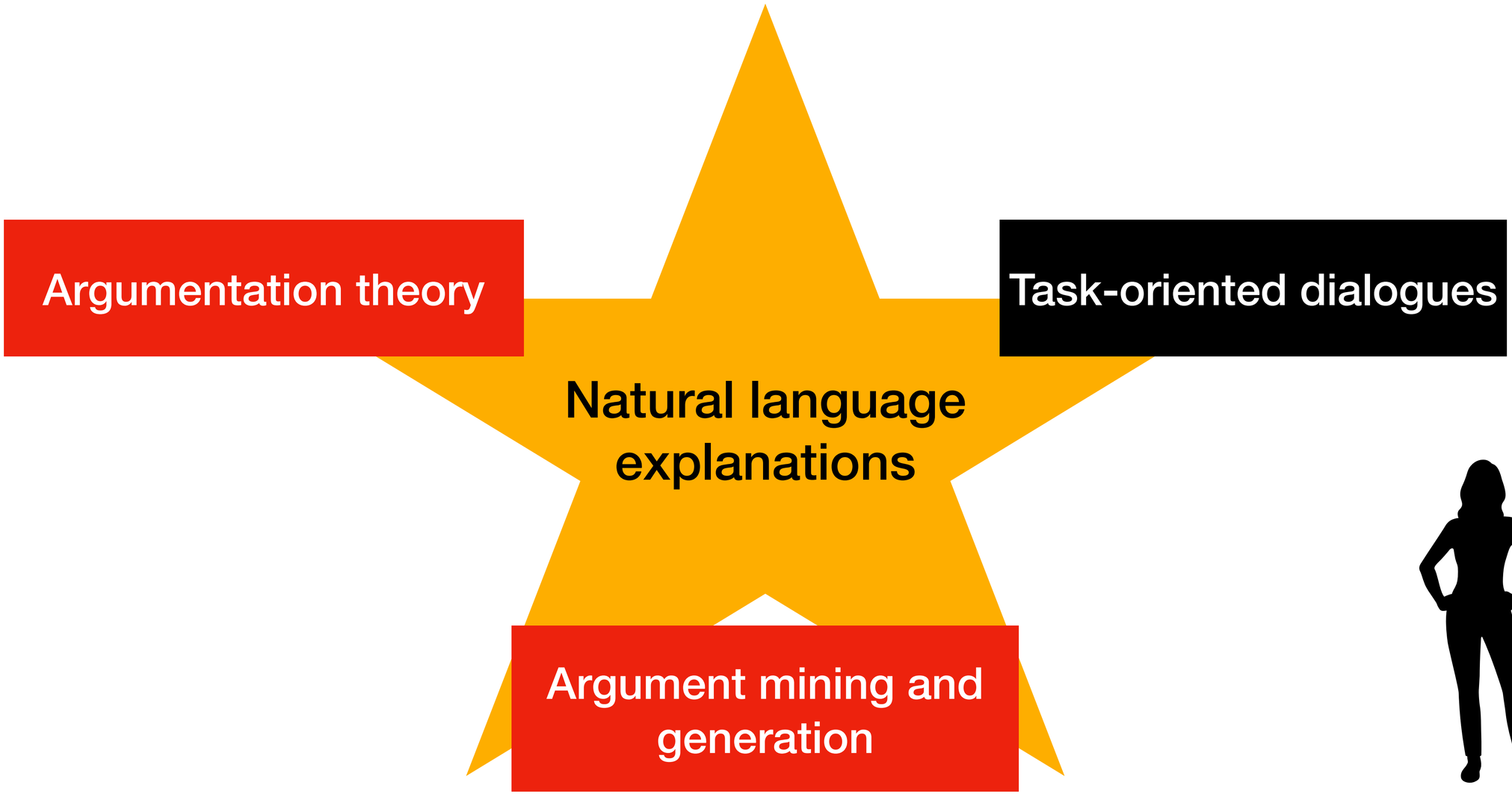
Natural language explanations

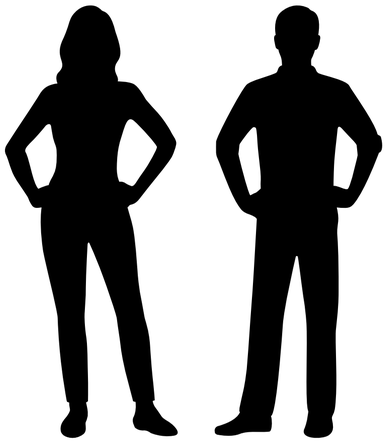
Key features



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

Logic, reasoning and argumentation

Frank Zenker. Logic, Reasoning, Argumentation: Insights from the Wild
Logic and Logical Philosophy · September 2017

Proponent		**Opponent**
1.	Why S?	Because T is true, and T implies S.
2.	Why should I accept T?	Because U is true, and U implies T.
3.	I do accept U.	Do you accept T?
4.	Yes.	Do you accept S?
5.	No.	But you must, because T implies S.

Argument schemes for explanations

Walton, Zenker, Wagermans

F is a finding or given set of facts.

E is a satisfactory explanation of *F*.

No alternative explanation *E'* given so far is as satisfactory as *E*.

Therefore, *E* is plausible, as a hypothesis.

Argument schemes for explanations

Walton, Zenker, Wagermans

Clarification questions (Rao & Daume, 2019; Xu et al., 2019a)

- **CQ1**: [*Absolute merits of explanation*:] How satisfactory is *E* as an explanation of *F*, apart from the alternative explanations available so far in the dialogue?
- **CQ2**: [*Relative merits of explanation*:] How much better an explanation is *E* than the alternative explanations available so far in the dialogue?
- **CQ3**: [*Relative developmental state of dialogue*:] How far has the dialogue progressed? If the dialogue is an inquiry, how thorough has the search been in the investigation of the case?
- **CQ4**: [*Comparative merit of continuing the dialogue*:] Would it be better to continue the dialogue further, instead of drawing a conclusion at this point?

Argument schemes for explanations

Josephson & Josephson

D is a collection of data (facts, observations, …).

H explains *D* (would, if true, explain D).

No other hypothesis can explain *D* as well as *H* does.

Therefore, *H* is probably true.

Argument schemes for explanations

Josephson & Josephson

Clarification questions (Rao & Daume, 2019; Xu et al., 2019a)

- **Threshold** How decisively does *H* surpass the alternatives?
- **Internal merit** How good is *H* by itself, independently of considering alternatives?
- **Data reliability** How trustworthy are data, respectively the processes by which data were obtained?
- **Exhaustiveness** How much confidence is there that all plausible explanations have been considered?
- **Cost and Benefits** What pragmatic considerations matter, including the costs of being wrong, and the benefits of being right?
- **Gravity of issue** How strong is the need to reach a conclusion, especially considering the possibility of seeking further evidence before deciding?

The ANTIDOTE Project
CHIST-ERA Call 2019 — XAI

ArgumeNtaTIon-Driven explainable artificial intelligence fOr digiTal mEdicine

UNIVERSITÉ CÔTE D'AZUR

FONDAZIONE BRUNO KESSLER

NOVA UNIVERSIDADE NOVA DE LISBOA

KU LEUVEN

UPV EHU

Motivations

Towards argument-based explanations

- in neural architectures the correlation between internal states of the network (e.g., weights assumed by single nodes) and the justification of the network classification outcome is not well studied;

- high quality explanations are crucially based on argumentation mechanisms (e.g., provide supporting examples and rejected alternatives);

- in real settings, providing explanations is inherently an interactive process involving the system and the user.

  General objective: providing a unified computational framework for jointly learning clinical predictions and the associated argumentative justifications, fostering a natural interaction with clinicians through explanatory dialogues.

(Ideal) Use case medical scenario

A dialogue between a student and a teacher

- We describe the case of a 21-year-old male, without known allergic drug reactions, smoker and social drinker during the weekends. He had been referring for a month and a half, after a fortuitous fall on a terrain with vegetation, pain and inflammatory signs in the front of the right leg, so he had received treatment with non-steroidal anti-inflammatory drugs (NSAIDs) and ciprofloxacin 750mg every 12 hours improving partially. During the following weeks, after starting his usual physical activity, playing soccer, the symptoms got worse, becoming more intense than at the beginning. The appearance of a swelling on the face forefoot of his right leg, severe pain and inability to dorsiflex the right foot and first toe were the reasons why he came to our hospital.

- Results of the blood test:
  1. Evident increase in the white series (14,000 leukocytes with 74% neutrophil),

  2. C-reactive protein (PCR) of 3.82mg dl, 3. High D-dimer (550 ng / dl).

- Results of the Doppler ultrasound of the right lower limb: Permeability of the deep venous axis, warm-fibular trunk, arch of the internal saphenous and of the distal trunks.

- Results from the microbiological culture: appearance of Streptococcus intermedius.

(Ideal) Use case medical scenario

A dialogue between a student and a teacher

- **Teacher**: Up to you, which are the possible diagnoses compatible with this clinical case?
- **Student**: According to the symptoms referred by the patient initially, possible compatible diagnosis could be:
  1. Deep vein thrombosis (ICD10 I80.2);
  2. Necrotizing cellulitis (ICD10 -)

  3. Erysipelas (ICD10 A46);
  4. Necrotizing fasciitis type 2 (ICD10 M72.6);
  5. Streptococcal gangrene (ICD10 B95.5);
  6. Clostridic myonecrosis (or caseous gangrene) (ICD10 A48.0);
  7. Mucormycosis (ICD10 B46.5);
  8. Pyomyositis (ICD10 M60.003);
  9. Mixed cellulite of polymicrobial origin (ICD10 -).

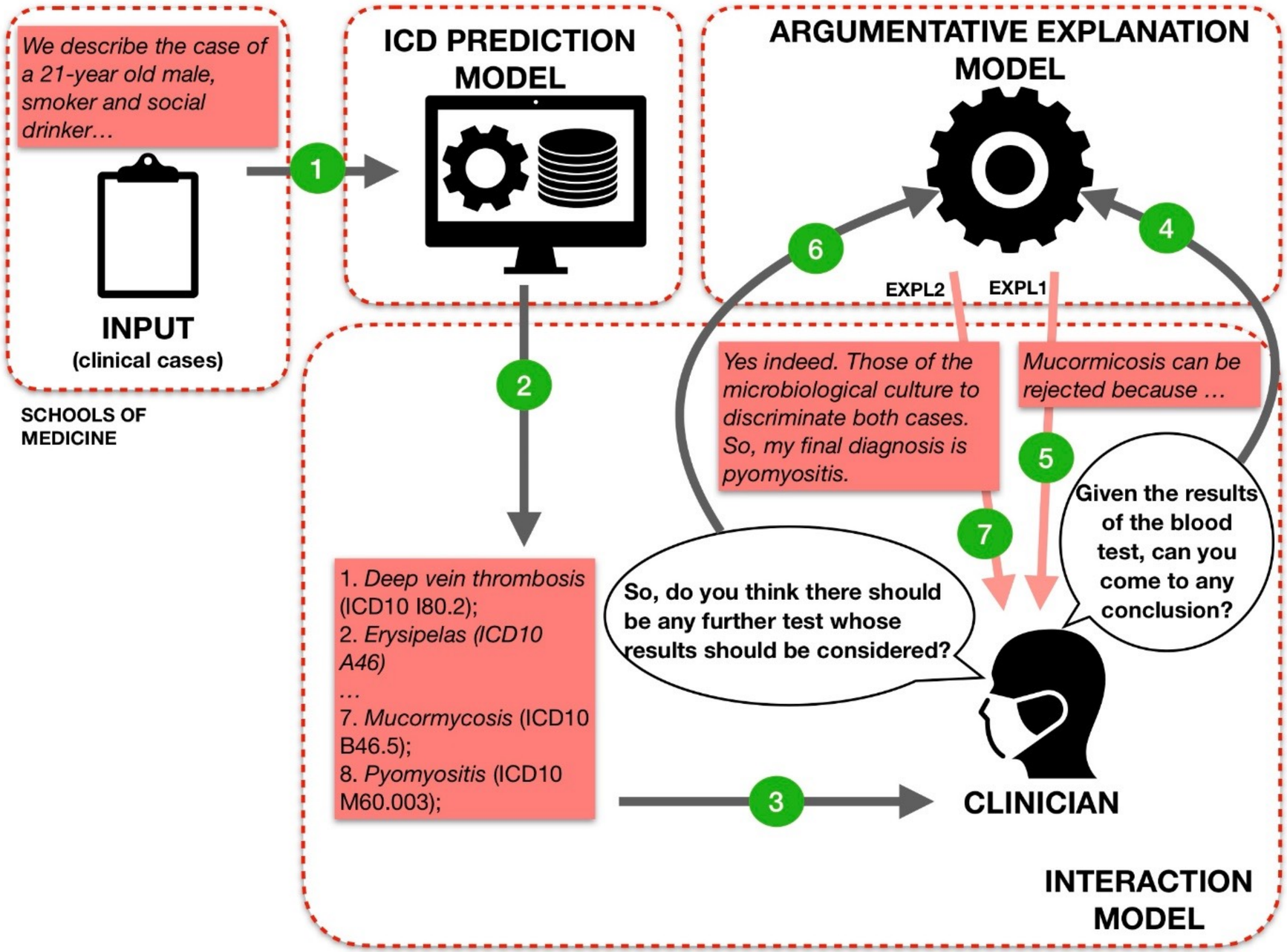
  But due to the fact that the evolution has been slow over time (some weeks after referring the first symptoms), I think that I can exclude necrotizing cellulitis, erysipelas, necrotizing fasciitis type 2, streptococcal gangrene, and clostridic myonecrosis (or caseous gangrene). Did the patient in the very beginning present any considerable size injury on the skin?
- **Teacher**: No, it did not present any relevant injury.

(Ideal) Use case medical scenario

A dialogue between a student and a teacher

- **Student:** Then, mucormycosis can also be rejected because it is caused by fungi of the Mucorales family present in the soil of vegetated areas that are generally introduced in the form of spores in the dermis when there is an injury on the skin. It usually produces disease in immune patients uncommitted or with underlying diseases, and this is not the case since the patient is a young boy with no previous health problems.
- **Teacher:** Given the results of the blood test, can you come to any conclusion?
- **Student**: Not yet, I have a question, did the patient have fever?
- **Teacher:** No, the patient denied having had a fever at any time.
- **Student:** Ok. If I consider also the results of the ultrasound, I can exclude deep vein thrombosis. The functional impotence that the patient presents and the pain in a so defined area orientates me to think about pyomyositis or suppurative myositis as another possibility. Although Mixed cellulite of polymicrobial origin cannot be excluded yet.
- **Teacher:** So, do you think there should be any further test whose results should be considered?
- **Student**: Yes indeed. Those of the microbiological culture to discriminate both cases. Considering those, my final diagnosis is pyomyositis.
- **Teacher:** The diagnose is indeed correct.

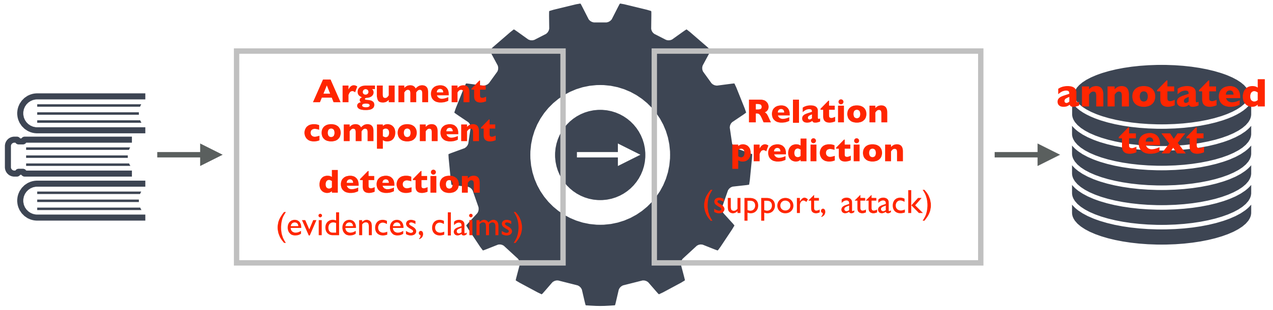
Overall architecture



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.



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 word 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.

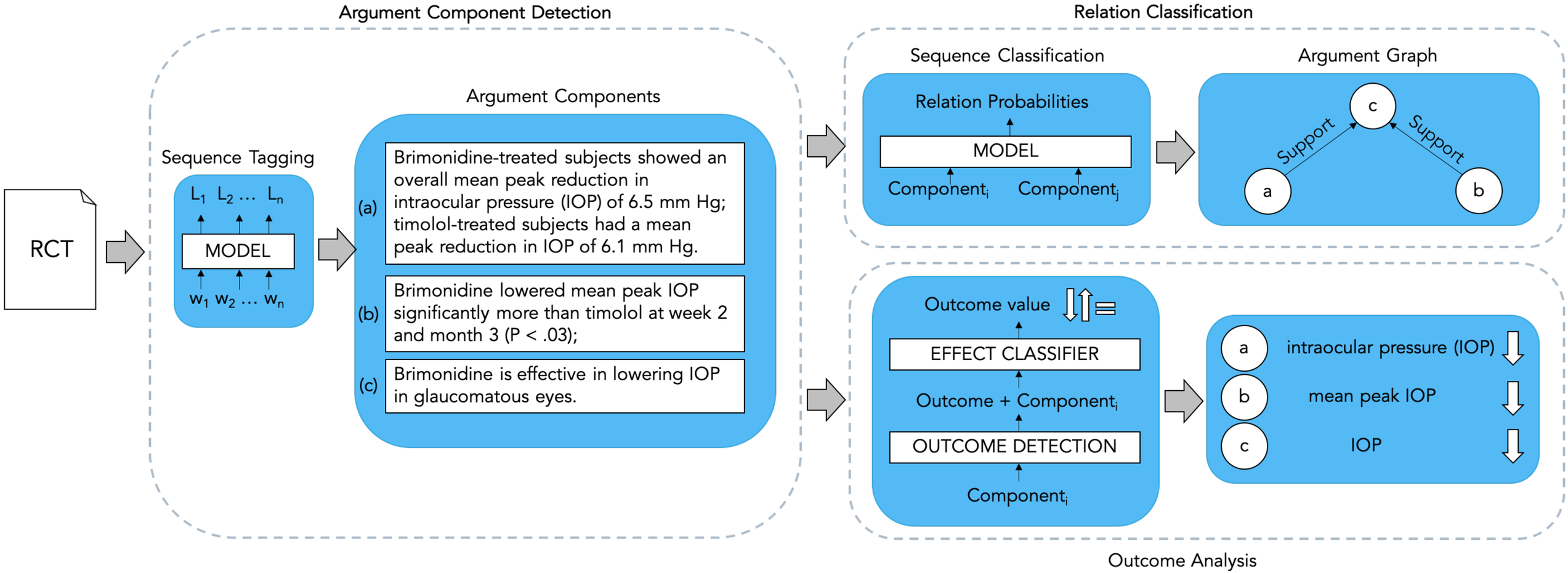
Collaborations:
INSERM, CHU Nice

PhD of Tobias Mayer

In collaboration with E. Cabrio

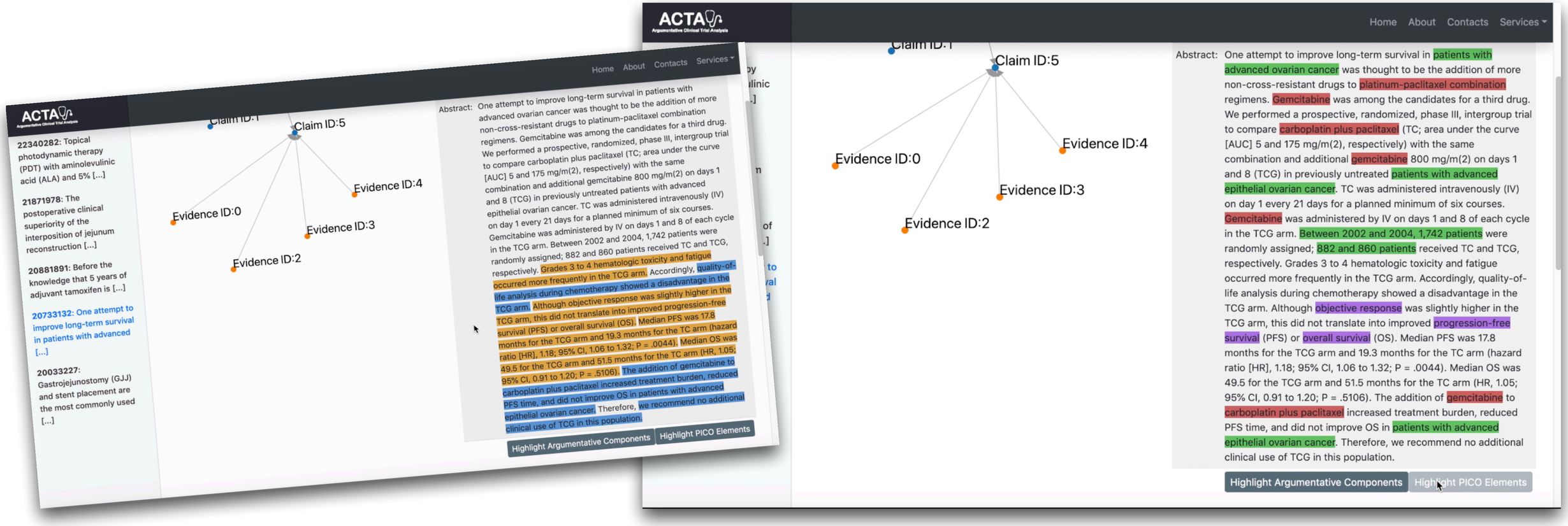
Mining argumentative structures from clinical trials

AI in Medicine 2021, ECAI20, COMMA2020, IJCAI19



ACTA

http://ns.inria.fr/acta/



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

<u>Why?</u>

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 !