Touché: Argument Retrieval
8 September 2022

Argumentative Ranking: Case Studies and Challenges

Andrea Galassi

Introduction: About myself

- Post-doc researcher at **University of Bologna**
- PhD in Computer Science and Engineering
- Experience in deep learning architectures for Natural Language Processing
- Part of **Language Technologies Lab** of prof. **Paolo Torroni**
- Strong collaboration with **Marco Lippi** (Unimore)

Introduction: Language Technologies Lab

Paolo Torroni

Andrea Galassi

Giulia Grundler

Francesco Antici

Marco Lippi

Federico Ruggeri

Elena Palmieri

Eleonora Mancini

https://site.unibo.it/nlp/en

What is this about

- Use of Argumentative Ranking to solve problems
- 3 case studies in 3 different domains
  - News
  - Medical Scientific Literature
  - Legal Domain

Focused Retrieval (FR)

- Given a query, a search engine does not return entire document(s) but only **document sections focused on the query**
- Provide users with direct access to relevant information in retrieved documents.
- Structured as 2-step pipeline:
    1. Identify the relevant documents
    2. Identify the relevant information inside of them

"Passage retrieval and other XML-retrieval tasks" (Trotman and Geva, 2006)
"Evaluating focused retrieval tasks" (Pehcevski and Thom, 2007)

Argument(ation) Mining (AM)

- Area of NLP processing aimed at extracting arguments from text
- Arguments can consists of
  - Claims : (debatable) statements about a certain area of interest
  - Evidences/Premises : supporting facts and notions
  - Their relationship (support/attack)

Focused Retrieval on Arguments

- Use AM as part of the FR pipeline
- Retrieval of the argumentative content of relevant document
  1. Retrieve relevant documents
  2. Apply Argument Mining to extract information
- **Present only the part of the documents that are argumentative**
- Useful for quickly retrieving arguments about a controversial topic or proofs about a certain fact

"Overview of the INEX 2013 social book search track" (Koolen et al, 2013)
"On the Retrieval of Wikipedia Articles Containing Claims on Controversial Topics" (Roitman et al, 2016)

A step forward

- We can do more!
- We can exploit AM to drive the document selection!
- Assumption: **the most interesting documents are the ones that have more argumentative content**
  - Better quality
  - Easier to extract argumentative information

"On the Retrieval of Wikipedia Articles Containing Claims on Controversial Topics" (Roitman et al, 2016)
"Argumentative Ranking" (Lippi and Sarti and Torroni, 2016)

Argumentative Ranking: idea

- Retrieval of documents based on their argumentative content
  1. Apply Argument Mining
  2. Select documents

- Filter out documents that have little argumentative content
- Narrows down the set of documents the one needs to access to obtain a satisfactory overview of the topic

"On the Retrieval of Wikipedia Articles Containing Claims on Controversial Topics" (Roitman et al, 2016)
"Argumentative Ranking" (Lippi and Sarti and Torroni, 2016)

Argumentative Ranking: possible scores

- Having AM method that assigns to each sentence a claim/premise score
- Given a document, we can measure
  - # sentences containing claims/evidences
  - % of sentences containing claims/evidences
  - Average claim/sentence score
  - Sum of scores of sentences containing claims/evidences
  - Average score of sentences containing claims/evidences

"Argumentative Ranking" (Lippi and Sarti and Torroni, 2016)

Argumentative Ranking: choosing scores

- Choice of scores depends on the application
- What do we prefer?
  - Very short document where almost all the sentences are argumentative?
  - Long document that contains many claims but also several non-argumentative sentences?
  - Few sentences that are very probably argumentative?
  - Many sentences with a low probability of being argumentative?

"Argumentative Ranking" (Lippi and Sarti and Torroni, 2016)

Case Study:
News Articles

Argumentative Ranking of News Articles

- Objective: argumentative ranking can influence (improve?) retrieval of news?
- Focus on detection of claims
- AM method based on Tree Kernels and SVMs
  - Text represented through constituency trees
- Qualitative study: Argumentative Ranking vs Google Ranking

"Argumentative Ranking" (Lippi and Sarti and Torroni, 2016)

Dataset

- 30 controversial topics
- For each topic, collect pages from the New York Times website
- Discard topics with less than 20 articles (11 topics)
- About 3,000 articles in total

"Argumentative Ranking" (Lippi and Sarti and Torroni, 2016)

Scores and Ranking

- 5 separate partial rankings, one for each score related to the claims
- For each partial ranking, assign
  - 25 points to the 1st document
  - 20 to the 2nd
  - 16 to the 3rd
  - 13 to the 4th,
  - 11, 10, . . . , 1 point to the 5th, 6th, . . . , 15th document
  - 0 to the following ones
- Sum the points of each partial ranking to obtain the final score and ranking

"Argumentative Ranking" (Lippi and Sarti and Torroni, 2016)

Results: gambling

- AR top 1 result: “Majority Back Referendum to Add Casinos, Poll Finds”
  - Does not appear in Google Ranking
  - The article contains both arguments in favor and against expanding casino
- Google tends to include more news, chronicle and event-related articles

	Argumentative Ranking	$\%_C$	$S(D_i)$	Google Ranking	$\%_C$	$S(D_i)$
1.	Majority Back Referendum to Add Casinos...	0.32	94	Rein In Online Fantasy Sports Gambling	0.42	82
2.	Rein In Online Fantasy Sports Gambling	0.42	82	The Trouble With Fantasy Sports Gambling	N/A	N/A
3.	Nevada Says It Will Treat Daily Fantasy...	0.23	51	17 People in Three States Are Held in...	N/A	N/A
4.	Cash Drops and Keystrokes: The Dark...	0.13	51	The Dark World of Fantasy Sports and...	N/A	N/A
5.	Will Other Leagues Join N.B.A.? Don't Bet...	0.19	45	Cash Drops and Keystrokes: The Dark...	0.13	51
6.	N.F.L.'s Unsteady Stance on a Tricky...	0.19	39	Nevada Says It Will Treat Daily Fantasy...	0.23	51
7.	As Casino Vote Nears, Bishops Warn of...	0.38	37	Daily Fantasy Sports and the Hidden Cost...	0.14	12
8.	Seeking to Ban Online Betting, G.O.P...	0.20	36	The Perfect Predictability of Gambling...	0.07	0
9.	An Ad Blitz for Fantasy Sports Games, but...	0.14	27	Whitney Wortman and William Gambling	N/A	N/A
10.	In Sharp Pivot for N.B.A., Commissioner...	0.25	25	An Ad Blitz for Fantasy Sports Games, but...	0.14	27

“Argumentative Ranking” (Lippi and Sarti and Torroni, 2016)

Results: wind power

- AR top 1 result: “Salvation gets cheap”
  - Does not appear in Google Ranking
  - The article contains many arguments about the use of renewable energies and pollution

	Argumentative Ranking	$\%_C$	$S(D_i)$	Google Ranking	$\%_C$	$S(D_i)$
1.	Salvation Gets Cheap	0.29	62	Wind Power Spreads Through Turbines...	N/A	N/A
2.	State of the Union Address - 2012 Transcript	0.06	50	Europe Looks Offshore for Wind Power	0.19	15
3.	Wind Power Is Poised to Spread to All States	0.46	50	Wind Power Is Poised to Spread to All States	0.46	50
4.	Tesla Ventures Into Solar Power Storage for...	0.15	46	Procter & Gamble to Run Its Factories...	0.10	0
5.	Glut of Coal-Fired Plants Casts Doubts on...	0.16	43	The Falling Cost of Wind Power	0.10	0
6.	Natural Gas: Abundance of Supply and Debate	0.22	41	Solar and Wind Energy Start to Win on...	0.17	36
7.	Texas Is Wired for Wind Power, and More...	0.16	37	Texas Is Wired for Wind Power, and More...	0.16	37
8.	Solar and Wind Energy Start to Win on...	0.17	36	Tax Credit for Wind Power	N/A	N/A
9.	A Price Tag on Carbon as a Climate Rescue...	0.11	36	A Texas Utility Offers a Nighttime Special...	0.12	0
10.	China Wins in Wind Power, by Its Own Rules	0.20	29	HP to Power Texas Data Centers With...	0.00	0

“Argumentative Ranking” (Lippi and Sarti and Torroni, 2016)

Conclusion

- Argumentative Ranking retrieves **highly argumentative articles** that would be **discarded by the search engine**
- Argumentative Ranking seems promising!
- Could be helpful for journalists, fact checkers, and more…

- Not particularly surprising though: using argumentation as a selection criteria we found documents that are more argumentative

"Argumentative Ranking" (Lippi and Sarti and Torroni, 2016)

Case Study:
Medical Scientific Literature

Context and Motivation

- The amount of scientific literature produced every year is overwhelming
- **The retrieval of relevant publications of good quality** is a big problem
- Difficult to build a dataset for a supervised approach
- Assumption: the **quality of a publication** may be correlated with its **argumentative content**
- Use of Argumentative Ranking to screen scientific literature
- Project AMICA: Argument Mining In Covid-19 Articles

"AMICA: An Argumentative Search Engine for COVID-19 Literature" (Lippi et al. 2022)
"Argument Mining as Rapid Screening Tool of COVID-19 Literature Quality: Preliminary Evidence" (Brambilla et al. 2022)

AMICA: Argument Mining In Covid-19 Articles

- Argumentative Search Engine for COVID-19 Literature
- Mixes **information retrieval** and **AM**
- Rank documents both according to their argumentative content and their pertinence to the query
- Collaboration with domain experts from Italian National Institute of Health (ISS)

"AMICA: An Argumentative Search Engine for COVID-19 Literature" (Lippi et al. 2022)
"Argument Mining as Rapid Screening Tool of COVID-19 Literature Quality: Preliminary Evidence" (Brambilla et al. 2022)

MARGOT: Mining ARGuments frOm Text

http://margot.disi.unibo.it/

- General purpose and software for AM, freely available
- Given a textual document, **labels its sentences as claims and/or evidences**
  - Each sentence is assigned two scores, one for each class (CS and ES)
- Based on Tree Kernels and SVMs
  - Sentences represented through Constituency Trees and TF-IDF
- Trained on (labelled) Wikipedia articles

"MARGOT: A web server for argumentation mining" (Lippi and Torroni, 2016)

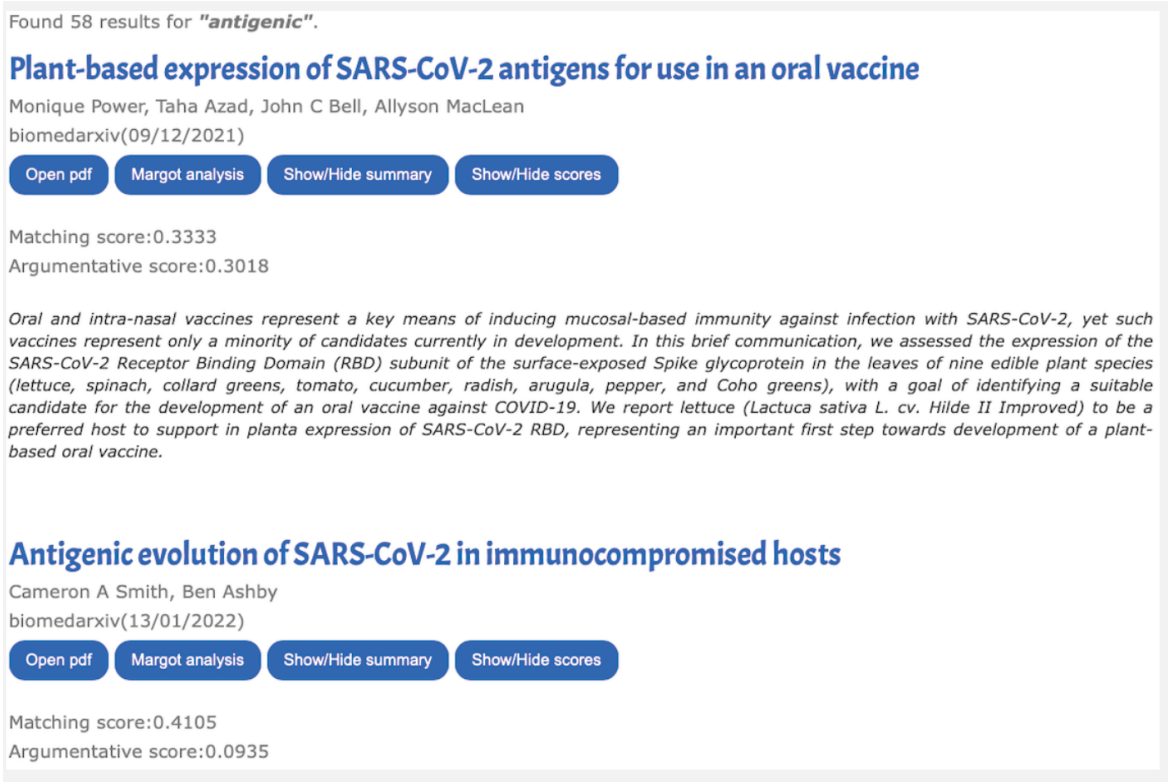
AMICA: Scores

- **Argumentative Score (AS)** of a sentence: maximum between CS and ES
- **Average Argumentative Score (AAS)** of a document: average of AS
- **Argument Ratio (AR)** of a document: the percentage (ratio) of sentences containing either a claim or an evidence
- Similarity Score (SS) of a document: similarity between query and text
  - Similarity between BoW embeddings
- Final score: **product between AAS and SS**

"AMICA: An Argumentative Search Engine for COVID-19 Literature" (Lippi et al. 2022)
"Argument Mining as Rapid Screening Tool of COVID-19 Literature Quality: Preliminary Evidence" (Brambilla et al. 2022)

AMICA Interface

http://amica.unimore.it/



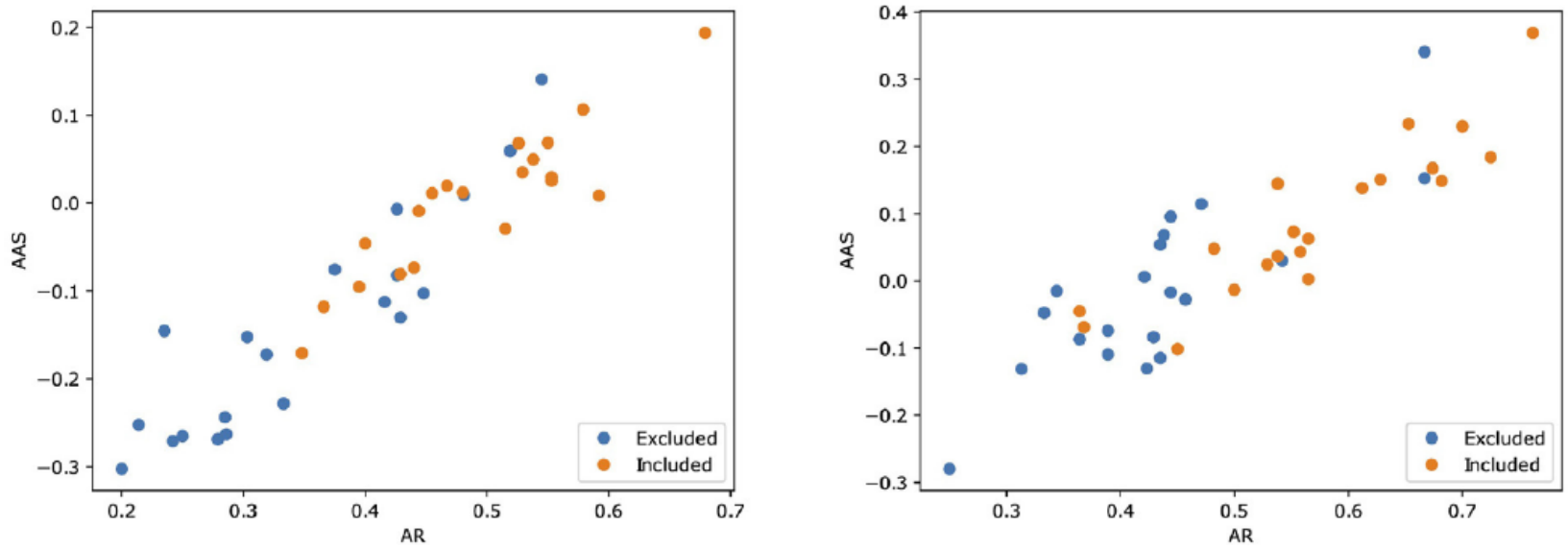
AMICA: Experimental Evaluation

- **Qualitative** comparison with 2 **Cochrane reviews**
  - For each review, a dataset of 40 articles: 20 included, 20 excluded
- **Quantitative** comparison with opinion of **medical experts**
  - Experts assign a quality score from 1 to 5 to 40 papers

"AMICA: An Argumentative Search Engine for COVID-19 Literature" (Lippi et al. 2022)
"Argument Mining as Rapid Screening Tool of COVID-19 Literature Quality: Preliminary Evidence" (Brambilla et al. 2022)

AMICA: Evaluation on Cochrane reviews

- The largest part of papers included in the Cochrane review are also the most argumentative papers for MARGOT



"AMICA: An Argumentative Search Engine for COVID-19 Literature" (Lippi et al. 2022)
"Argument Mining as Rapid Screening Tool of COVID-19 Literature Quality: Preliminary Evidence" (Brambilla et al. 2022)

AMICA: Evaluation on Experts' Opinion

- We measure the Spearman's rank correlation coefficient ρ
  - ρ = 0.463 when considering AAS
  - ρ = 0.562 when considering AR

- **Moderate-to-strong correlation** in both cases
- They are MEDICAL experts, they do not know (or care) about argumentation!

"AMICA: An Argumentative Search Engine for COVID-19 Literature" (Lippi et al. 2022)
"Argument Mining as Rapid Screening Tool of COVID-19 Literature Quality: Preliminary Evidence" (Brambilla et al. 2022)

Conclusion

- Argumentative Ranking seems to be a valid tool to screen scientific literature
- The measure of argumentative content seems related to quality of an article

- BUT they remain two separate concepts
  - E.g., the paper “Cyllage City COVID-19 Outbreak Linked to Zubat Consumption”
- Screening of literature may require **specific criteria not available in the text**
  - E.g., Cochrane reviews considers also: no conflict of interests, certain metadata, etc.

“AMICA: An Argumentative Search Engine for COVID-19 Literature” (Lippi et al. 2022)
“Argument Mining as Rapid Screening Tool of COVID-19 Literature Quality: Preliminary Evidence” (Brambilla et al. 2022)

Case Study:
Law Decisions

Context and Motivation

- Context: database of legal judgments
- Objective: navigate through judgments according to their similarity
- Task: **given a document, find the most similar ones**

- Project ADELE: Analytics for DEcision of LEgal cases

Problem

- What does “similarity between judgments” mean?
- Difficult to build a complete dataset:
  - Many documents are required
  - Requires domain experts
  - The documents are very long => It is very time-consuming
- Unfeasible to build a dataset big enough to rely on supervised techniques

Exploiting Argumentation

- Hypothesis: judgments are similar when the reasoning process is similar
- The reasoning process can be detected through argument mining
- Similar arguments => similar judgments ?

First (naïve) attempt

- First unsupervised approach:
    1) Extract arguments from two documents
    2) Measure similarity between pairs of arguments (BERT-like embeddings)
    3) Aggregate the score to find similarity between documents
- Test on a small dataset (about 20 pairs of documents)
- Use of threshold to establish which documents are similar
- Terrible results!!

Hypotesis

- The concept of similarity should be defined differently according to the type of premises

- **Factual premises**: subject of the reasoning (topic?)
- **Legal premises**: type of reasoning (scheme?)
  - Different schemes: legal rules, precedents, interpretations

"Detecting Arguments in CJEU Decisions on Fiscal State Aid" (Grundler et al. 2022)

Clustering of Premises

- Qualitative analysis of the clustering of premises

- Purpose:
  - Investigate if there is enough information inside premises
  - Investigate if the premise type influence the results

- We analyze the most representative features and randomly sample premises

Results: Factual Premises

- Premises divided according the **topic** of the judgment
- Low level features, such as words (e.g., “tax”)
- Pro: can be retrieved easily
- Con: need many documents to cover many topics => **quantity**

Results: Legal Premises

- The clustering seem related to the **scheme** of the premise
- More abstract concept of similarity
- Pro: each document will likely contain many different schema
- Con: difficult to extract, since it’s difficult to define => **quality**

Discussion

- Two dimension of similarity: subject of reasoning and type of reasoning
- We should compare only similarities of the same type

- Does it work?
  - Difficult to prove: challenging to build a dataset (quantity and quality)
- Does this approach really reflect what the end users want?
  - Difficult for experts to define similarity in abstract terms

Conclusion

- The legal domain seems great to study opportunities and limits of argumentative ranking/retrieval/mining/clustering/… :
  - Still an open problem
  - Multiple objectives
  - Necessity of specific and tailored solutions
  - Experts (lawyers, judges, researchers) are interested in getting involved!
- It’s necessary to conduct a **study with human experts** to understand their perspective and gather data

Conclusion

Open Challenges

- The detection of arguments can help the retrieval of documents
- The measure of argumentative content can be partially related to the quality of the documents

- Additional (symbolic) features must be considered for specific application
- Certain domains may require tailored solutions

- The legal domain seems to be a challenging and promising frontier

References

- Trotman and Geva, “Passage retrieval and other XML-retrieval tasks”, SIGIR Workshop on XML Element Retrieval Methodology, 2006
- Pehcevski and Thom, “Evaluating focused retrieval tasks”, SIGIR Workshop on Focused Retrieval, 2007
- Koolen, Kazai, Preminger, and Doucet, “Overview of the INEX 2013 social book search track”, CLEF, 2013
- Roitman, Hummel, Rabinovich, Sznajder, Slonim, and Aharoni, “On the Retrieval of Wikipedia Articles Containing Claims on Controversial Topics”, WWW, 2016
- Lippi, Sarti, and Torroni, “Argumentative Ranking”, IJCAI Workshop on Natural Language Processing Meets Journalism, 2016
- Brambilla, Rosi, Antici, Galassi, Giansanti, Magurano, Ruggeri, Torroni, Cisbani, and Lippi, “Argument Mining as Rapid Screening Tool of COVID-19 Literature Quality: Preliminary Evidence”, Frontiers on Public Health, 2022
- Lippi, Antici, Brambilla, Cisbani, Galassi, Giansanti, Magurano, Rosi, Ruggeri, and Torroni, “AMICA: An Argumentative Search Engine for Covid-19 Literature”, IJCAI, 2022
- Grundler, Santin, Galassi, Galli, Godano, Lagioia, Palmieri, Ruggeri, Sartor, and Torroni “Detecting Arguments in CJEU Decisions on Fiscal State Aid”, ArgMining (accepted), 2022

Thank you for your time!
Any questions?