Argument Mining on Clinical Trials

Tobias MAYER
Supervisor: Serena Villata
Co-director: Céline Poudat

Université Côte d'Azur, CNRS, Inria, I3S
Base, Corpus, Langage (BCL)

UNIVERSITÉ
CÔTE D'AZUR

Motivation

Evidence-based medicine (EBM):

- optimize decision making with evidence from well-conducted research
- meta-analysis and systematic reviews on Randomized Controlled Trials (RCT)

Motivation

Evidence-based medicine (EBM):

- optimize decision making with evidence from well-conducted research
- meta-analysis and systematic reviews on Randomized Controlled Trials (RCT)

How to assist with automatic processing?

Our approach

Argument mining system for clinical trials:

- automated approach to extract argumentative information from trials
- detection of claims and evidence
- domain unspecific applicability

Argument Mining

"The general task of analyzing discourse on the pragmatics level and applying a certain argumentation theory to model and automatically analyze the data at hand"

[Habernal and Gurevych, 2017]

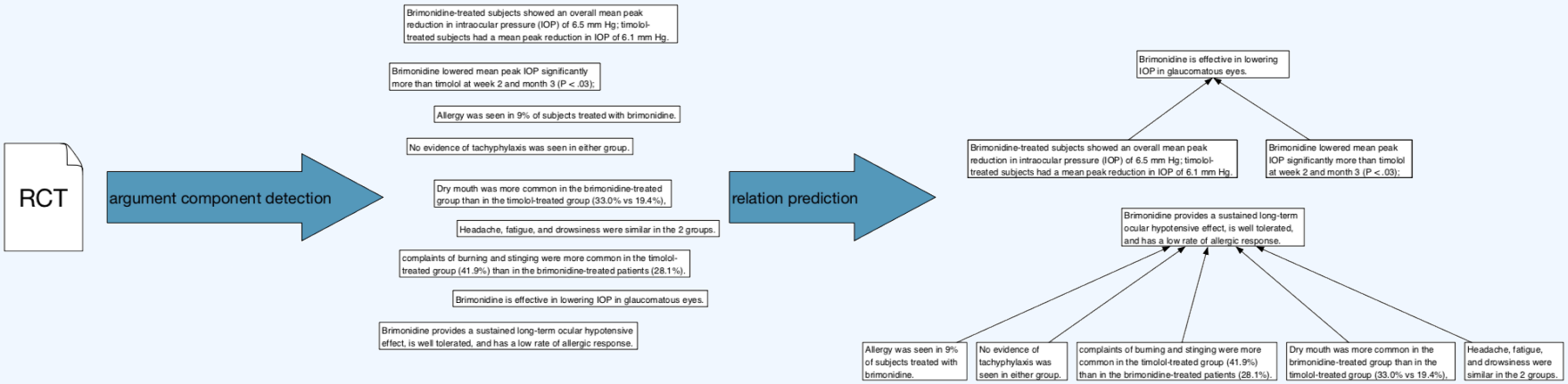
Argument Mining

Argument extraction:

- distinguish argumentative from non-argumentative components
- classify the components into evidence and claims

Relations prediction:

- intra-argument relation prediction
- inter-argument relation prediction



Dataset

Dataset

Randomized Controlled Trials (RCT):

- common type of **experimental studies** in the medical domain
- **comparison** between intervention and control arm
- used for **evidence-based medical decision making** (systematic reviews and meta-analysis)
- **PubMed**: freely available citation database from the United States National Library of Medicine (NLM)
- structure should follow CONSORT[1] policies

[1] http://www.consort-statement.org/

Dataset

Data collection:

- Annotate existing **collection of RCT abstracts** on glaucoma treatments with argumentative labels
- Delete existing PICO[2] annotations
- Extending the collection with more RCT abstracts from PubMed **(glaucoma, diabetes, hepatitis and hypertension)**

[2] Annotation framework for: **P**opulation, **I**ntervention, **C**ontrol and **O**utcome

Annotation

Claim

- concluding statement made by the author about the outcome of the study:
  - *"Brimonidine is well tolerated and has a low rate of allergic response."*
- general description of the relation between intervention and control arm:
  - *"Trabeculectomy was more effective than viscocanalostomy in lowering IOP in glaucomatous eyes of white patients."*
- should logically follow from the described results

Annotation

Evidence/Premise

- observation in the study (side-effect or other measured outcome):
  - *"Allergy was seen in 9% of subjects treated with brimonidine."*
  - *"Brimonidine lowered mean peak IOP significantly more than timolol at week 2 ($P < .03$)."*
- credible without further evidence (ground truth)
- supports or attacks another argument component

Annotation Example

To compare the intraocular pressure-lowering effect of latanoprost with that of dorzolamide when added to timolol. [...] [*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 ($P < 0.001$)*]$_3$. Drugs administered in both treatment groups were well tolerated. 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]$_1$**.[3]

[3] **claims** are written in bold, *evidence* are underlined

Corpus Statistics

Topic	*#abstracts*	*#evidence*	*#claims*	*#non arguments*
glaucoma	119	448	153	743
diabetes	20	84	41	112
hepatitis	20	105	22	121
hypertension	20	60	33	126

Inter-annotator agreement[4]:

- argumentative components: 0.72
- claim/evidence distinction: 0.68

[4] agreement is given in Fleiss' kappa for three annotators

Experimental Settings

MARGOT[5]:

- argument mining approach to **overcome genre-dependency**
- addresses **argument component detection**
- **cross-domain features** (word occurrences, sentence structure)
- trained on Wikipedia articles

[5] MARGOT: Mining Arguments from Text. http://margot.disi.unibo.it

Methodology

Model:

- SVM classifier for detection of claim/evidence
- SVM+HMM for detection of component boundaries

Features:

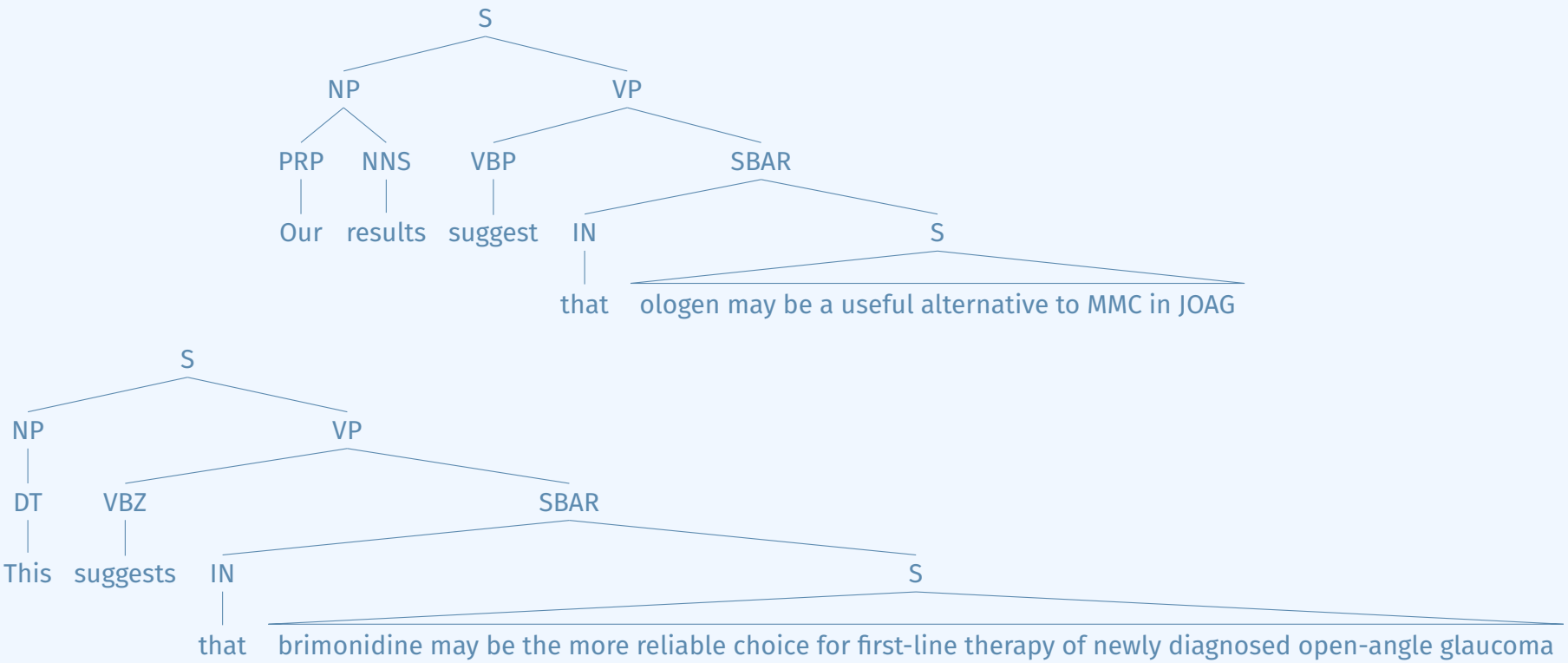
- SubSet Tree Kernels (SSTK)
- bag-of-words with TF-IDF values

Tree Kernel

What is a tree kernel?

- **similarity measure** between constituency parse trees
- considers **common fragments** between two trees
- defines a rich feature space
- SSTK provides a **good compromise** between expressiveness and efficiency

Tree Kernel Example



Data partitioning:

topic	training	testing
glaucoma	79 abstracts	30 abstracts
hepatitis(HB)		20 abstracts
diabetes(DM)		20 abstracts
hypertension(HT)		20 abstracts
mixed		90 abstracts

Results

Results: Evidence Detection[6]

		Glaucoma	DM	HB	HT	Mixed
Evidence	BoW	0.84	0.79	0.74	0.80	0.80
	SSTK	0.86	0.79	0.75	0.80	0.80
	SSTK + BoW	0.86	0.79	0.75	0.80	0.80

- SSTK slightly better than BoW, but still comparable
- no differences between SSTK and BoW for out-of-domain topics
- distinctive vocabulary might be related to statistical evaluation rather than medical terminology

[6] results are given in f_1-score

Results: Claim Detection[6]

		Glaucoma	DM	HB	HT	Mixed
Claim	BoW	0.75	0.68	0.62	0.64	0.65
	SSTK	0.79	0.73	0.66	0.70	0.72
	SSTK + BoW	0.79	0.74	0.66	0.70	0.72

- SSTK significantly better than BoW
- distinctive syntactic structure for claims
- SSTK generalizes better than BoW
- combining the models do not increase results
- lexical information also captured in syntactic structure

[6] results are given in f_1-score

Results: Argumentative Component Detection[6]

		Glaucoma	DM	HB	HT	Mixed
Arg. Comp.	BoW	0.82	0.74	0.70	0.72	0.74
	SSTK	0.86	0.76	0.71	0.74	0.78
	SSTK + BoW	0.86	0.76	0.71	0.74	0.78

- TK model performs better
- results similar to evidence detection
- many errors were made between claim and evidence distinction

[6] results are given in f_1-score

Evidence Classification

Motivation

- EBM focuses mainly on study design and risk of bias as quality of evidence
- need for other aspects to measure trial quality (reproducability, generalizabilty or estimate of effect)
- first step towards creating arguments for argumentation framework

Evidence Types

- **comparative:**
  - *"The overall success rates were 87% for the 350-mm2 group and 70% for the 500-mm2 group (P = 0.05)."*
- **significance:**
  - *"All regimens produced clinically relevant and statistically significant (P < .05) intraocular pressure reductions from baseline."*
- **side-effect:**
  - *"Allergy was seen in 9 % of subjects treated with brimonidine."*
- **other:** risk factors, limitations
  - *"Risk of all three outcomes was higher for participants with chronic kidney disease or frailty."*

Evidence Classification

Results for multi-class classification using SVMs:

Dataset	*Method*	*glaucoma*	*combined.*
Gold standard	RANDOM	0.33	0.32
	MAJORITY	0.27	0.26
	N-GRAMS	0.80	0.74
whole pipeline	RANDOM	0.38	0.38
	MAJORITY	0.38	0.39
	N-GRAMS	0.71	0.66

Table: Results (weighted average F_1-score).

Conclusion

- creation of a dataset of RCTs labeled with argumentative components
- application of Argument Mining on clinical trials
- first step to evidence classification

Next Steps

- relation prediction (building argumentation trees)
- annotation of CHU data (French) and corpus building (together with BCL)
- evidence quality assessment
- reproducible support for clinical decision making

Thank you for your attention!

Error Analysis

Description of the objective of a study confused as claims:

- *"The goal of this research is to evaluate efficacy and safety of herbal medicine as compared to allopathic medicine in patients suffering from hepatitis B."*

Error Analysis

Claims sometimes with a very complex syntactic structure:

- *"The authors tested the hypothesis that a valsartan/cilnidipine combination would suppress the home morning blood pressure (BP) surge (HMBPS) more effectively than a valsartan/hydrochlorothiazide combination in patients with morning hypertension , defined as systolic BP (SBP) 135 mm Hg or diastolic BP 85 mm Hg assessed by a self-measuring information and communication technology-based home BP monitoring device more than three times before either combination 's administration."*

Error Analysis

Group descriptions (group sizes or initial medical conditions) mis-classified as evidences:

- *"Among 426 participants (53% male, median age 35 years, median CD4 count 19 cells/μL), 31 developed hepatotoxicity (7.3%)."*
- *"Overall, there were no significant differences in pregnancy-induced hypertension across supplement groups."*

Error Analysis

Negated sentences often mis-classified:

- *"No patients developed additional resistance mutations throughout the study period."*