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)
Evidence-based medicine (EBM):

- optimize decision making with evidence from well-conducted research
- meta-analysis and systematic reviews on Randomized Controlled Trials (RCT)
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
"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
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 PICO² annotations
- Extending the collection with more RCT abstracts from PubMed *(glaucoma, diabetes, hepatitis and hypertension)*

²Annotation framework for: Population, Intervention, Control and Outcome
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
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
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

<table>
<thead>
<tr>
<th>Topic</th>
<th>#abstracts</th>
<th>#evidence</th>
<th>#claims</th>
<th>#non arguments</th>
</tr>
</thead>
<tbody>
<tr>
<td>glaucoma</td>
<td>119</td>
<td>448</td>
<td>153</td>
<td>743</td>
</tr>
<tr>
<td>diabetes</td>
<td>20</td>
<td>84</td>
<td>41</td>
<td>112</td>
</tr>
<tr>
<td>hepatitis</td>
<td>20</td>
<td>105</td>
<td>22</td>
<td>121</td>
</tr>
<tr>
<td>hypertension</td>
<td>20</td>
<td>60</td>
<td>33</td>
<td>126</td>
</tr>
</tbody>
</table>

**Inter-annotator agreement**:  
- argumentative components: 0.72  
- claim/evidence distinction: 0.68

---

4agreement is given in Fleiss’ kappa for three annotators
Experimental Settings
METHODOLOGY

MARGOT\textsuperscript{5}:

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

\textsuperscript{5}MARGOT: Mining Arguments from Text. \url{http://margot.disi.unibo.it}
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
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
Our results suggest that ologen may be a useful alternative to MMC in JOAG.

This suggests that brimonidine may be the more reliable choice for first-line therapy of newly diagnosed open-angle glaucoma.
## Methodology

### Data partitioning:

<table>
<thead>
<tr>
<th>topic</th>
<th>training</th>
<th>testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>glaucoma</td>
<td>79 abstracts</td>
<td>30 abstracts</td>
</tr>
<tr>
<td>hepatitis(HB)</td>
<td></td>
<td>20 abstracts</td>
</tr>
<tr>
<td>diabetes(DM)</td>
<td></td>
<td>20 abstracts</td>
</tr>
<tr>
<td>hypertension(HT)</td>
<td></td>
<td>20 abstracts</td>
</tr>
<tr>
<td>mixed</td>
<td></td>
<td>90 abstracts</td>
</tr>
</tbody>
</table>
Results
### Results: Evidence Detection

<table>
<thead>
<tr>
<th>Evidence</th>
<th>Glaucoma</th>
<th>DM</th>
<th>HB</th>
<th>HT</th>
<th>Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoW</td>
<td>0.84</td>
<td>0.79</td>
<td>0.74</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>SSTK</td>
<td>0.86</td>
<td>0.79</td>
<td>0.75</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>SSTK + BoW</td>
<td>0.86</td>
<td>0.79</td>
<td>0.75</td>
<td>0.80</td>
<td>0.80</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Claim</th>
<th>Glaucoma</th>
<th>DM</th>
<th>HB</th>
<th>HT</th>
<th>Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoW</td>
<td>0.75</td>
<td>0.68</td>
<td>0.62</td>
<td>0.64</td>
<td>0.65</td>
</tr>
<tr>
<td>SSTK</td>
<td>0.79</td>
<td>0.73</td>
<td>0.66</td>
<td>0.70</td>
<td>0.72</td>
</tr>
<tr>
<td>SSTK + BoW</td>
<td>0.79</td>
<td>0.74</td>
<td>0.66</td>
<td>0.70</td>
<td>0.72</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Arg. Comp.</th>
<th>Glaucoma</th>
<th>DM</th>
<th>HB</th>
<th>HT</th>
<th>Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoW</td>
<td>0.82</td>
<td>0.74</td>
<td>0.70</td>
<td>0.72</td>
<td>0.74</td>
</tr>
<tr>
<td>SSTK</td>
<td>0.86</td>
<td>0.76</td>
<td>0.71</td>
<td>0.74</td>
<td>0.78</td>
</tr>
<tr>
<td>SSTK + BoW</td>
<td>0.86</td>
<td>0.76</td>
<td>0.71</td>
<td>0.74</td>
<td>0.78</td>
</tr>
</tbody>
</table>

- 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, generalizability 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."
Results for multi-class classification using SVMs:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>glaucoma</th>
<th>combined.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold standard</td>
<td>RANDOM</td>
<td>0.33</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>MAJORITY</td>
<td>0.27</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>N-GRAMS</td>
<td>0.80</td>
<td>0.74</td>
</tr>
<tr>
<td>whole pipeline</td>
<td>RANDOM</td>
<td>0.38</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>MAJORITY</td>
<td>0.38</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>N-GRAMS</td>
<td>0.71</td>
<td>0.66</td>
</tr>
</tbody>
</table>

**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!
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."
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."
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."
Negated sentences often mis-classified:

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