Argument Mining on Clinical Trials[1]

Tobias MAYER [a], Elena CABRIO [a], Marco LIPPI [b],
Paolo TORRONI [c] and Serena VILLATA [a]
[a] *Université Côte d'Azur, CNRS, Inria, I3S, France* [2]
[b] *University of Modena and Reggio Emilia, Italy*
[c] *University of Bologna, Italy*

Abstract. Argument-based decision making has been employed to support a variety of reasoning tasks over medical knowledge. These include evidence-based justifications of the effects of treatments, the detection of conflicts in the knowledge base, and the enabling of uncertain and defeasible reasoning in the health-care sector. However, a common limitation of these approaches is that they rely on structured input information. Recent advances in *argument mining* have shown increasingly accurate results in detecting argument components and predicting their relations from unstructured, natural language texts. In this study, we discuss *evidence* and *claim* detection from Randomized Clinical Trials. To this end, we create a new annotated dataset about four different diseases (glaucoma, diabetes, hepatitis B, and hypertension), containing 976 argument components (697 containing evidence, 279 claims). Empirical results are promising, and show the portability of the proposed approach over different branches of medicine.

Keywords. Argument component detection, clinical trials, dataset

1. Introduction

Argumentation-based decision making is becoming increasingly prominent in healthcare applications. Several formal frameworks have been proposed to tackle the issues of reasoning upon clinical evidence and detecting possible conflicts in medical knowledge-bases [13,4,19,26]. Different kinds of data can be explored in this context (e.g., clinical trials, clinical guidelines, Electronic Health Records) combined with the patient and clinician preferences and the specific constraints raised by the particular medical branch taken into account. The general aim of such approaches is to support clinicians and practitioners in taking informed decisions. However, the main limitation of these approaches is that they assume the availability of structured information, e.g., in the form of databases or knowledge-bases.

Unfortunately, such a structured input is not always available. For instance, clinical trials are documents written in natural language, comparing the relative merits of treatments. Hence, there is a well-motivated need to investigate methods able to extract struc-

[1] Submitted to IAT.

[2] This work is partly funded by the French government labelled PIA program under its IDEX UCA JEDI project (ANR-15-IDEX-0001).

tured information from unstructured text, in order to support argument-based decision making frameworks.

In this study, we propose an automated approach to the extraction of argumentative information from clinical data. More precisely, we focus on the definition of *an argument mining system for extracting arguments from clinical trials*.
Argument mining [17,25] is a recent research area aiming at extracting natural language arguments and their relations from unstructured, natural text, with the final goal of providing machine-processable structured data for computational models of argument. In this context, we address the following questions:

- How to distinguish argumentative from non-argumentative components in natural language clinical trials?
- How to classify the detected argumentative components into evidence and claims?

To that end, we propose an argument mining system which addresses both boundary identification for the argument components, i.e., distinguishing in such a way argumentative segments from non-argumentative ones, and then classifies the identified segments into *evidence* (i.e., the premises of the argument) and *claims* (i.e., the conclusion of the argument). As many approaches to detect evidence and claims in texts have been proposed in the argument mining community [15,18,28,5], we decided to rely on an existing system and to tailor it to cope with the clinical data scenario. More precisely, we provide a refined version of MARGOT [18] so that the system is able to detect evidence and claims from clinical data. As for the textual clinical data for the experiments conducted in our approach, we selected Randomized Clinical Trials (RCT), and more precisely, PubMed[3] RCT abstracts about four different diseases, namely, glaucoma, diabetes, hepatitis B, and hypertension. Results are promising and show the portability of the proposed approach to different medical branches.

Our contribution is three-fold. First, to the best of our knowledge, we present the first successful attempt to argumentation mining, and more precisely, argument component detection, on clinical trials. Our approach can be seen as the first building block of formal decision making like the evidence-based approach proposed by Hunter and Williams [13], such that the automatically detected evidence can be provided to the framework to be aggregated to highlight the positive and negative effects of treatments. Second, we create a new manually annotated dataset (together with the related annotation guidelines) containing 976 argument components, classified into 697 evidences and 279 claims. Third, based on said dataset, we present a quantitative and qualitative study on the performance of Tree Kernels in this task.

The paper is organized as follows. Section 2 describes the argument mining research area and its tasks, and describes the main features of the MARGOT system. In Section 3, we present the dataset we annotated with the argument components (evidence, claims), and in Section 4 we describe the methods we used for the argument components detection and classification, and we discuss the obtained results. We conclude with a discussion of the related literature, and we highlight some perspectives for future research.

[3] https://www.ncbi.nlm.nih.gov/pubmed/

2. Preliminaries

In this section, we provide some insights about the argument mining framework and its related tasks, and we describe the MARGOT system.

Argument Mining. *Argument(ation) mining* has been defined as "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" [12]. Two stages are crucial in argument mining:

(1) Arguments' extraction: The first stage is the identification of arguments within the input natural language text. This step may be further split in two different stages such as the detection of argument components (e.g., claim, evidence) and the identification of their textual boundaries. Many approaches have recently been proposed to address this task, that adopt different methods like Support Vector Machines (SVM) [23,18,28,22], Naïve Bayes classifiers [7], Logistic Regression [15].

(2) Relations' prediction: The second stage consists of predicting what are the relations (e.g., *attack* and *support*) holding between the arguments identified in the first stage. Different methods have been employed to address this task, from standard SVMs to Textual Entailment [2]. This stage is also in charge of predicting, in structured argumentation, the internal relations between the argument components, i.e., the connection between the evidence and the claim [28].

Both the tasks require high-quality annotated corpora to train and to evaluate the performances of automated approaches. In this paper, we do not address the second task of the argument mining pipeline, i.e., relation prediction among the identified arguments, leaving this task for future research.

MARGOT. Argument mining being a highly challenging task under many aspects, research has been carried out mostly by targeting individual domains or genres. Many proposals made a choice to exploit the structure of input documents, such as persuasive essays [28] or legal texts [23], making ample use of heavily engineered features. That brought significant progress in such domains, although the focus of these studies is inherently limited. In an effort to overcome genre-dependency, other approaches have been proposed that do not rely on domain-tailored features, but only on features that, at least in principle, could provide meaningful clues across multiple domains, such as occurrence of words and sentence structure. One such approach [16] makes use of tree kernels [21] and underlies the first online argument mining server [18].[4] MARGOT was designed to make argument mining easily accessible outside of the argument mining research community, and was trained on a corpus consisting of 547 Wikipedia articles [1,27]. It was then evaluated on datasets coming from diverse genres such as persuasive essays and social media discussion threads, with encouraging results [18]. MARGOT addresses the first stage of the argument mining pipeline, in particular argument component detection. It carries out both claim and evidence detection, thanks to a SVM classifier that uses bag-of-words and constituency trees with subset tree kernels [3], and is able to predict component boundaries thanks to an SVM-HMM [31].

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

Dataset	*Topic*	*#abstracts*	*#evidences*	*#claims*	*#non arguments*
Training set	glaucoma	79	314	133	535
Test set	glaucoma	30	134	50	208
Test set	diabetes	20	84	41	112
Test set	hepatitis	20	105	22	121
Test set	hypertension	20	60	33	126

Table 1. Statistics on the dataset

3. Dataset on clinical trials

To experiment with the proposed approach to extract argumentative information from clinical data, we built a new annotated corpus of Randomized Clinical Trial abstracts, with annotations for the different argument components (evidences and claims). For building our corpus, we selected the same abstracts used in the corpus of RCT abstracts of Trenta *et al.* [30], which are annotated with PICO[5] elements. RCTs are a common type of experimental studies in the medical domain for evidence-based decision making. In general, to test the effect of a drug or treatment, the test subjects are divided into groups: one receiving the hypothesized treatment (intervention arm) and the other an established treatment (control arm). The results are then compared after certain time intervals. The structure of RCTs should follow the CONSORT[6] policies. Therefore, each reported study has a similar structure. More specifically, the abstract is structured with multiple labels: *background*, *objective*, *methods*, *results* or *conclusion*. The publication policies ensure a minimum consensus of provided information, which makes the studies comparable and ideal for building a corpus. Trenta *et al.* [30] retrieved the RCT abstracts directly from PubMed[7] using three search strategies: (Strategy 1) titles or abstracts containing the word "Glaucoma" and that specified that the studies were randomized clinical trials; (Strategy 2) titles containing at least one element of a list of prescription drugs recognized as those used typically in the treatment of glaucoma or ocular hypertension and that specified that the studies were randomized clinical trials; and (Strategy 3) titles containing at least one element of a list of surgery procedures, identified as those typically used in the treatment of glaucoma or ocular hypertension, and that specified that the studies were randomized clinical trials. Given that in such work the authors' goal is different from ours – they extract elements important for Evidence Based Medicine, i.e. the above mentioned PICO elements: patient group, intervention and control arms, outcome measure description and its measurements in the two arms – we carried out a new annotation process on the same data for the argument mining task (we deleted their annotation tags). Moreover, given that we want to show the system portability to RCT abstracts on different diseases, we have extracted from PubMed 60 additional abstracts on diabetes, hepatitis and hypertension, following Strategy 1 in [30]. Table 1 reports on the statistics of the dataset.

[5] PICO is a framework to answer health care related questions in evidence-based practice. Elements comprise patients/population (P), intervention (I), control/comparison (C) and outcome (O) information.

[6] `http://www.consort-statement.org/`

[7] `https://www.ncbi.nlm.nih.gov/pubmed/`

3.1. Annotation of arguments components in RCT abstracts

We offer an overview of the guidelines we defined for the annotation of argument components in RCT abstracts[8]. We started from the guidelines defined in [28] for argument mining annotation on persuasive essays, and we adapted them to our scenario.

Claims. In the context of RCT abstracts, a *claim* is a concluding statement made by the author about the outcome of the study. It generally describes the relation of a new treatment (intervention arm) with respect to existing treatments (control arm) and is derived from the described results.

Example 1 To compare outcomes of selective laser trabeculoplasty (SLT) with drug therapy for glaucoma patients in a prospective randomized clinical trial. Sixty-nine patients (127 eyes) with open-angle glaucoma or ocular hypertension were randomized to SLT or medical therapy. Target intraocular pressure (IOP) was determined using the Collaborative Initial Glaucoma Treatment Study formula. Patients were treated with SLT (100 applications 360 degrees) or medical therapy (prostaglandin analog). Six visits over 1 year followed initial treatment. If target IOP range was not attained with SLT, additional SLT was the next step, or in the medical arm additional medications were added. Primary outcome: IOP; secondary: number of steps. Sixty-nine patients were treated. Data collection terminated with 54 patients reaching 9 to 12-months follow-up. Twenty-nine patients were in the SLT group, 25 patients in the medical group. Baseline mean IOP for all eyes was 24.5 mm Hg in the SLT group, 24.7 mm Hg in the medical group. Mean IOP (both eyes) at last follow-up was 18.2 mm Hg (6.3 mm Hg reduction) in the SLT arm, 17.7 mm Hg (7.0 mm Hg reduction) in the medical arm. By last follow-up, 11% of eyes received additional SLT, 27% required additional medication. There was not a statistically significant difference between the SLT and medication groups. IOP reduction was similar in both arms after 9 to 12-months follow-up. More treatment steps were necessary to maintain target IOP in the medication group, although there was not a statistically significant difference between groups. **[These results support the option of SLT as a safe and effective initial therapy in open-angle glaucoma or ocular hypertension]$_1$**.

Example 1 shows a typical example of RCT abstract, where a *claim* concluding the study is highlighted (in the following examples, *claims* are written in bold and are surrounded by square brackets marked with a subscript). It compares the intervention with the control arm on a general level, and claims that the tested option is safe and effective. At large, the comparison could also be that there has been found no significant difference between the two arms. This must not be confused with a concrete numerical comparison of a certain measure or any reporting of the outcomes. Those two types of comparisons differ a lot in their role as argumentative components. Whereas the reporting of experimental observations is a fact/evidence, the general comparison in itself has no factual value and is therefore a statement/claim, which needs supporting evidence in order to be credible. This can also be seen in the same example, where the reporting of the IOP reduction

[8]Full guidelines and the dataset are available at `https://www.dropbox.com/sh/9ms8v2blq8zstep/AAC0lGJ3PxF1L7W8ESv8R0hVa?dl=0`.

(dashed underline) might be confused with a *claim*, but given that it is a reporting of the outcome it should be annotated as *evidence*.

Example 2 Brimonidin tartrate is a highly selective alpha 2-agonist. [...][9] **[In general, the tolerance to medication was acceptable]$_1$.[Brimonidine is safe and effective in lowering IOP in glaucomatous eyes]$_2$. [Brimonidine provides a sustained long-term ocular hypotensive effect, is well tolerated, and has a low rate of allergic response]$_3$.**

Example 2 shows another type of *claim*, the assertion that one object of study, usually the intervention arm, has a specific property. Notice that stated qualities may have to be divided into multiple *claims* on a sentence level, see for instance **claim$_2$** and **claim$_3$**.

Major claims as a stance of the author, as they are defined in [28], are not present in RCT abstracts. Here *major claims* are more a general/concluding *claim*, which is supported by other, more specific, claims. In the following example, *major claims* are marked as underlined claims. In Example 3 the *claims* are specific statements about the conducted studies. The *major claims* are more general statements, for which the other *claims* serve as *evidence*. The concluding statements do not have to occur at the end of the abstract, and may also occur at the beginning of the text as an introductory *claim*.

Example 3 To assess the efficacy and safety of fixed-combination latanoprost-timolol (FCLT) vs latanoprost or timolol monotherapy. [...] **[Fixed-combination latanoprost-timolol therapy is as safe and effective in lowering IOP in patients with either ocular hypertension or glaucoma as monotherapy with latanoprost or timolol]$_1$. [Combination therapy can be used to treat patients for whom monotherapy does not provide sufficient IOP reduction]$_1$.**

Evidence. An *evidence* in a RCT abstract is an observation or measurement in the study (ground truth), which supports or attacks another argument component, usually a *claim*. Those observations comprise side effects and the measured outcome of the intervention and control arm. They are observed facts, and therefore credible without further justifications, since this is the ground truth the argumentation is based on. In the examples below, *evidence* are in italic, underlined and surrounded by square brackets with subscripts.

Example 4 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]$_1$.**

[9] The abstract has been shortened for space reasons.

Example 4 shows different reports of the experimental outcomes as *evidence*. Those can be results without concrete measurement values, see evidence 1, or exact measured values, see evidence 2 and 3. Like for *claims*, different measures are annotated as multiple *evidence*. The reporting of side effects and negative observations are also considered as *evidence*, while irrelevant outcomes, e.g. 'statistically not significant', are not.

An expert in the medical domain (a pharmacist) has validated the proposed guidelines before starting the annotation process. Three annotators with background in computational linguistics[10] have then annotated the data (50 RCT abstracts each) after a training phase. The reliability of an annotated corpus is guaranteed by the calculation of the Inter Annotator Agreement (IAA) that measures the degree of agreement in performing the annotation task among the involved annotators. Hence, IAA among the three annotators has been calculated on 10 additional abstracts, resulting in a Fleiss' kappa of 0.72 (Fleiss' kappa between 0.61 and 0.80 means substantial agreement) for argumentative components and 0.68 for the more fine-grained claim/evidence distinction, attesting the reliability of the obtained dataset.

4. Experimental setting

We will describe the methods for argument component extraction and classification in Section 4.1, and discuss the results in Sections 4.2 and 4.3.

4.1. Methods

Argument component detection is typically addressed as a supervised text classification problem: given a collection of sentences, each labeled with the presence/absence of an argument component, the goal is to train a machine learning classifier to detect the argumentative sentences. Formally, given a dataset $\mathscr{D} = \{(x_j, y_j)\}_{j=1}^{N}$, where x_j is a sentence and y_j is the corresponding label (whether the sentence contains an argument or not), the goal is to learn a discrimination function $f : X \rightarrow Y$ to infer the label from the input text. Such a task can be addressed by a variety of machine learning algorithms [17].

A very common approach in natural language processing is to employ a bag-of-words (BoW) to represent sentences. This solution exploits lexical information, since each word in the vocabulary is a feature for the classifier, that is typically a Support Vector Machine. The method can be generalized to *n*-grams rather than just words. Despite its simplicity, this approach is often a strong baseline in argument mining [18,17]. The methodology implemented in MARGOT consists instead in a kernel machine that exploits a Tree Kernel (TK) to measure similarity between examples, namely between constituency parse trees. The key idea behind this approach is that the *structure* of a sentence is typically highly informative of the presence of an argument, or part thereof, within the sentence itself [16]. TKs aim to compare two trees by considering common *fragments*. Different definitions of fragments induce different TK functions. In this paper, as in the original MARGOT implementation, we employ the SubSet Tree Kernel (SSTK) [3], which offers a reasonable compromise between expressiveness and efficiency [18]. In

[10] In [10] researchers from a variety of backgrounds (biology, computer science, computer- supported argumentation pedagogy, and BioNLP) have been selected to annotate medical data for an argument mining research task, showing that they performed equally well despite their backgrounds.

SSTK, a fragment can be any sub-tree of the original tree, which terminates either at the level of pre-terminal symbols or at the leaves. The kernel between two trees T_x and T_z is evaluated as:

$$K(T_x, T_z) = \sum_{n_x \in N_{T_x}} \sum_{n_z \in N_{T_z}} \Delta(n_x, n_z) \tag{1}$$

where N_{T_x} (respectively, N_{T_z}) is the set of nodes of tree T_x (respectively, T_z), and $\Delta(n_x, n_z)$ measures the score nodes n_x and n_z, depending on the chosen definition of fragments. Given the (tree) kernel function K, the discrimination function f is defined as:

$$f(T_x) = \sum_{i=1}^{N} \alpha_i y_i K(T_{x_i}, T_x) \tag{2}$$

where N is the number of support vectors, and α_i is the (learned) coefficient of the i-th support vector. In our case, the problem is formulated as a binary classification task (i.e., a sentence contains an argument component, or it doesn't), therefore $y_i \in Y = \{\pm 1\}$.

A very interesting characteristic of TKs is that the similarity measure implicitly allows to define a rich and expressive feature space, that basically consists of all the possible fragments that can be encountered in the parse tree.

4.2. Experimental Setup

Data was pre-processed (tokenisation and stemming), and the constituency parse tree for each sentence was computed. Furthermore, the bag-of-words (BoW) features with Term Frequency and Inverse Document Frequency (TF-IDF) values were also computed. TF-IDF assigns higher weights to more distinctive words, thus lowering the impact of terms that are common among all documents (a document in our case is a sentence). All the pre-processing steps were performed with Stanford CoreNLP, version 3.5.0.

We conducted experiments with three different classifiers: (i) SSTK exploiting constituency parse trees, (ii) SVM with BoW features weighted by TF-IDF, (iii) a kernel machine combining the two approaches. Two datasets were prepared to train two binary classifiers for each approach: one for claim detection, and one for evidence detection. Both training sets only differ in the labels, which were assigned to each sentence.

For tuning the hyper-parameter C (SVM regularization parameter) and the decay factor for the tree kernel, we performed a grid search using 5-fold cross validation optimizing for the F_1-score. Substituting the SSTK with a Partial Tree Kernel (PTK) [20] did not improve the results.

4.3. Results and discussion

Evaluation of the models was conducted on multiple datasets, as described in Section 3. We computed the F_1-score for the tasks of (1) evidence, (2) claim and (3) argumentative component (evidence or claim) detection. Results are shown in Table 2.

The model behavior is different for claim detection and for evidence detection. As for claim detection, the best performance is still on the glaucoma set with 0.79 F_1 score and 0.75, respectively. But here the difference between the SSTK and BoW model is significantly higher. This suggests that claims have a distinctive syntactic structure which

		Glaucoma	Diabetes	Hepatitis	Hypertension	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
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
Argumentative Component	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

Table 2. Results for the glaucoma, diabetes, hepatitis, hypertension and mixed test set on the task of evidence, claim and argumentative component detection. Results are given in F_1 score.

can be learned, and that is useful to distinguish them from non-argumentative sentences and evidences. This is true also when the test set comprises the same topic as the training set and thus the lexical approach should have a natural advantage. Furthermore, when comparing the results from the glaucoma test set with the other test sets, the performance of the SSTK does not decrease as strongly as the one of the BoW model, e.g. -0.07 vs. -0.10 F_1 score on the mixed dataset (joint test set over all domains). Differently from the case of evidence detection, here the BoW relies more on specialized medical terminology, which differs with the individual test set domain, whereas TKs generalize better on out-of-domain data. The combined model delivers similar results compared to the pure SSTK model. Again, as for evidence detection, this suggests that the lexical information representing the characteristics of claims is also contained in the syntactic representation. For evidence detection, all models performed best on the glaucoma test set. This is intuitive, since in that case training and test domains coincide. Comparing the different models on this test set, the SSTK performed slightly better with 0.86 F_1 score, but still the difference to the BoW baseline (0.84 F_1 score) is only marginal. This difference becomes even smaller on the hepatitis dataset and vanishes completely for diabetes and hypertension. Thus, we can conclude that evidence is not that highly distinctive with respect to syntactic structure from non-argumentative sentences or claims, while it can be easily identified by lexical information. Moreover, this lexical information is domain independent, otherwise the performance of the BoW model would strongly decrease with respect to the SSTK on the other test sets. Therefore, the distinctive vocabulary is likely to be related to the domain of statistical evaluation, rather than to medical terminology, as one could expect. These observations will need further investigation. Interestingly, the combination of the syntactic (SSTK) and lexical (BoW) approach did not increase the results, meaning that those two models share the equal amount of information representing the characteristics of evidence, and that the two models generalize equally well.

The results of the third classification problem, the detection of argumentative sentences, reflect the above described findings. The best performance for each model was obtained on the glaucoma set. The SSTK model outperforms the BoW baseline, but a combination of TKs and BoW does not increase the results. Again, the TK generalizes better over the different test set domains. Comparing the outcomes of the experiments for claim and argumentative component detection tasks, the best models for the glaucoma and mixed test set perform in a comparable range. In theory the performance for the combined task should be lower, since the claim detection has a significant lower F_1 score

1.	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.
2.	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 .
3.	Among 426 participants (53% male, median age 35 years, median CD4 count 19 cells/L), 31 developed hepatotoxicity (7.3%).
4.	Overall, there were no significant differences in pregnancy-induced hypertension across supplement groups.
5.	No patients developed additional resistance mutations throughout the study period.

Table 3. Sample classification errors.

than the evidence detection. This can be explained when looking at the errors made by the claim classifier. A sizable amount of false positives, sentences which are classified as containing a claim, but actually do not, were sentences containing evidence. When merging together evidence and claims into argumentative components, those false positives become true positives, increasing the overall results.

Error analysis on claim detection indicates that a significant amount of false positives are sentences describing the objective of the RCT (Table 3, Example 1). This might be due to the comparative nature of the sentences, since comparative statements are common among claims. Many false negatives have complex syntactic structures (Example 2) where either the whole sentence is a claim, or it contains multiple fragments with claims. Those complex structures might have been missing from the training set.

For evidence classification, many sentences describing the participants of the studies (Example 3) have been mis-classified as evidences by all approaches. This might be due to their sub-clauses containing statistical descriptions, as many pieces of evidence have too. Similarly, sentences describing the initial condition of the different groups (Example 4) were confused as evidences. The problem here is that those sentences are highly context-dependent: Example 4 could be a valid evidence, if the context was the description of the results and not the description of the initial conditions. There is no way to distinguish those cases without considering a larger context. Other mis-classified evidences are negated sentences like Example 5, reporting the non-existence of an effect.

5. Related Work

Most of the previous work in argument mining focuses on non-scientific domains, e.g., legal documents, social media, on-line debates or persuasive essays. However, automatic argument extraction can support scientific disciplines by simplify time-consuming tasks, such as literature research, reasoning, and knowledge extraction.

Related work aims to define a fine-grained and informative model of justifications in scientific argumentations [29]. Another related topic in this domain is the detection of comparative structures. Comparisons are a crucial part of the scientific exchange and communication. Naive Bayes, SVM and Bayesian networks are used in [24] to automatically detect comparison claims in full-text scientific articles.

Previous work on comparative structure identification in clinical trials relied on under-specified syntactic analysis and domain knowledge [8]. Gupta *et al.* [11] ap-

plied syntactic structure and dependency parsers to extract comparison structures from biomedical texts. Those findings justify our choice of syntax-based tree kernels, since a sizable amount of the argumentative components in our dataset are comparisons.

Comparing RCTs as a common way of gathering evidence in support of medical decision making, Trenta *et al.* [30] built an information extraction system based on a maximum entropy classifier with basic linguistic features for the tasks of extracting the patient group, intervention and control arm, and outcome measure description. Differently from us, they extract components to extend evidence tables, not considering linguistic phrases to reconstruct the whole argumentation. Dernoncourt *et al.* [6] developed an artificial neural network with word embeddings to assign PubMed RCT abstract labels to sentences. They show that considering sequential information to jointly predict sentence labels improves the results. However, their task differs from ours in the sense that they predict the structure of abstracts, which depends on contextual information, since abstract composition should follow the standardized CONSORT policies.

One of the few studies in argument mining focusing on the biomedical domain was presented by Green [9,10], who proposed argumentation schemes and inter-argument relations for the annotation of arguments in research articles. Yet, such annotation schemes are only partially applicable for argument extraction in RCT abstracts. Other NLP approaches in the medical domain perform sequence labeling, focusing on named entity recognition or medical event detection, and apply either conditional random field or artificial neural networks [14], which may also be applicable in our case.

6. Conclusion

This study is the first application of argument mining methods for the automatic extraction of evidence and claims from clinical trials. We annotated a dataset of 169 RCT abstracts, where we identified 976 argument components. We discussed the performance of Tree Kernels, a method that underlies an existing online argumentation mining system, by performing cross-topic evaluation to show the portability of our findings.

Although argument mining in the medical domain has a long way to go, we believe that it has a huge potential for effectively supporting key tasks, such as information seeking for better-informed clinical decision making or support for evidence quality and weight assessment. We also plan to provide a finer-grained classification of evidence (e.g., side effects, biological plausibility, precision of estimate of effect) so that evidence weights may be automatically assessed depending on these parameters. For this more fine-grained annotation, we plan to collaborate with experts in the respective fields to ensure the reliability of those specific annotations from the medical point of view.

References

[1] E. Aharoni, A. Polnarov, T. Lavee, D. Hershcovich, R. Levy, R. Rinott, D. Gutfreund, and N. Slonim. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *ArgMining@ACL*, 2014.

[2] E. Cabrio and S. Villata. A natural language bipolar argumentation approach to support users in online debate interactions. *Argument & Computation*, 4(3):209–230, 2013.

[3] M. Collins and N. Duffy. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *ACL*, pages 263–270. ACL, 2002.

[4] R. Craven, F. Toni, C. Cadar, A. Hadad, and M. Williams. Efficient argumentation for medical decision-making. In G. Brewka, T. Eiter, and S. A. McIlraith, editors, *KR*. AAAI Press, 2012.

[5] J. Daxenberger, S. Eger, I. Habernal, C. Stab, and I. Gurevych. What is the essence of a claim? cross-domain claim identification. In *EMNLP*, pages 2055–2066, 2017.

[6] F. Dernoncourt, J. Y. Lee, and P. Szolovits. Neural networks for joint sentence classification in medical paper abstracts. In *EACL*, pages 694–700, 2017.

[7] R. Duthie, K. Budzynska, and C. Reed. Mining ethos in political debate. In *COMMA*, volume 287 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2016.

[8] M. Fiszman, D. Demner-Fushman, F. Lang, P. Goetz, and T. C. Rindflesch. Interpreting comparative constructions in biomedical text. In *BioNLP@ACL*, pages 137–144, 2007.

[9] N. Green. Argumentation for scientific claims in a biomedical research article. In *Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, 2014.

[10] N. L. Green. Annotating evidence-based argumentation in biomedical text. *IEEE BIBM*, pages 922–929, 2015.

[11] S. Gupta, A. S. M. A. Mahmood, K. Ross, C. H. Wu, and K. Vijay-Shanker. Identifying comparative structures in biomedical text. In *BioNLP 2*, pages 206–215, 2017.

[12] I. Habernal and I. Gurevych. Argumentation mining in user-generated web discourse. *Comput. Linguist.*, 43(1):125–179, 2017.

[13] A. Hunter and M. Williams. Aggregating evidence about the positive and negative effects of treatments. *Artificial Intelligence in Medicine*, 56(3):173–190, 2012.

[14] A. Jagannatha and H. Yu. Structured prediction models for RNN based sequence labeling in clinical text. In *EMNLP*, pages 856–865, 2016.

[15] R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni, and N. Slonim. Context dependent claim detection. In *COLING*, 2014.

[16] M. Lippi and P. Torroni. Context-independent claim detection for argument mining. In *IJCAI*, 2015.

[17] M. Lippi and P. Torroni. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn.*, 16(2):10, 2016.

[18] M. Lippi and P. Torroni. MARGOT: A web server for argumentation mining. *Expert Systems with Applications*, 65:292–303, 12 2016.

[19] L. Longo and L. Hederman. Argumentation theory for decision support in health-care: A comparison with machine learning. In *BHI*, pages 168–180, 2013.

[20] A. Moschitti. Efficient convolution kernels for dependency and constituent syntactic trees. In *ECML*, volume 4212 of *LNCS*, pages 318–329. Springer, 2006.

[21] A. Moschitti. State-of-the-art kernels for natural language processing. In *Tutorial Abstracts of ACL 2012*, ACL '12, pages 2–2, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[22] V. Niculae, J. Park, and C. Cardie. Argument mining with structured SVMs and RNNs. In *ACL*, 2017.

[23] R. M. Palau and M. Moens. Argumentation mining. *Artif. Intell. Law*, 19(1):1–22, 2011.

[24] D. H. Park and C. Blake. Identifying comparative claim sentences in full-text scientific articles. In *Workshop on Detecting Structure in Scholarly Discourse*, pages 1–9, 2012.

[25] A. Peldszus and M. Stede. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Inform. Nat. Intell.*, 7(1):1–31, Jan. 2013.

[26] M. A. Qassas, D. Fogli, M. Giacomin, and G. Guida. Analysis of clinical discussions based on argumentation schemes. *Procedia Computer Science*, 64:282–289, 2015.

[27] R. Rinott, L. Dankin, C. A. Perez, M. M. Khapra, E. Aharoni, and N. Slonim. Show me your evidence - an automatic method for context dependent evidence detection. In *EMNLP*, 2015.

[28] C. Stab and I. Gurevych. Parsing argumentation structures in persuasive essays. *Comput. Linguist.*, 43(3):619–659, 2017.

[29] S. Teufel. Scientific argumentation detection as limited-domain intention recognition. In *Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, 2014.

[30] A. Trenta, A. Hunter, and S. Riedel. Extraction of evidence tables from abstracts of randomized clinical trials using a maximum entropy classifier and global constraints. *CoRR*, abs/1509.05209, 2015.

[31] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.