Evidence Type Classification in Randomized Controlled Trials

Tobias Mayer, Elena Cabrio, Serena Villata

Evidence Type Classification in Randomized Controlled Trials

Tobias Mayer and **Elena Cabrio** and **Serena Villata**
Université Côte d'Azur, CNRS, I3S, Inria, France
{tmayer,cabrio,villata}@i3s.unice.fr

Abstract

Randomized Controlled Trials (RCT) are a common type of experimental studies in the medical domain for evidence-based decision making. The ability to automatically extract the *arguments* proposed therein can be of valuable support for clinicians and practitioners in their daily evidence-based decision making activities. Given the peculiarity of the medical domain and the required level of detail, standard approaches to argument component detection in *argument(ation) mining* are not fine-grained enough to support such activities. In this paper, we introduce a new sub-task of the argument component identification task: *evidence type classification*. To address it, we propose a supervised approach and we test it on a set of RCT abstracts on different medical topics.

1 Introduction

Evidence-based decision making in medicine has the aim to support clinicians and practitioners to reason upon the arguments in support or against a certain treatment, its effects, and the comparison with other related treatments for the same disease. These approaches (e.g., (Hunter and Williams, 2012; Craven et al., 2012; Longo and Hederman, 2013; Qassas et al., 2015)) consider different kinds of data, e.g., Randomized Controlled Trials or other observational studies, and they usually require transforming the unstructured textual information into structured information as input of the reasoning framework. This paper proposes a preliminary step towards the issue of providing this transformation, starting from RCT, i.e., documents reporting experimental studies in the medical domain. More precisely, the research question we answer in this paper is: *how to distinguish different kinds of evidence in RCT, so that fine-grained evidence-based decision making activities are supported?*

To answer this question, we propose to resort on *Argument Mining* (AM) (Peldszus and Stede, 2013; Lippi and Torroni, 2016a), 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" (Habernal and Gurevych, 2017). Two stages are crucial: *(1)* the detection of argument components (e.g., claim, premises) and the identification of their textual boundaries, and *(2)* the prediction of the relations holding between the arguments. In the AM framework, we propose a new task called *evidence type classification*, as a sub-task of the argument component identification task. The distinction among different kinds of evidence is crucial in evidence-based decision making as different kinds of evidence are associated to different weights in the reasoning process. Such information need to be extracted from raw text.

To the best of our knowledge, this is the first approach in AM targeting evidence type classification in the medical domain. The main contributions of this paper are: (i) we propose four classes of evidence for RCT (i.e., *comparative*, *significance*, *side-effect*, and *other*), and we annotate a new dataset of 169 RCT abstracts with such labels, and (ii) we experiment with supervised classifiers over such dataset obtaining satisfactory results.

2 Evidence type classification

In (Mayer et al., 2018), as a first step towards the extraction of argumentative information from clinical data, we extended an existing corpus (Trenta et al., 2015) on RCT abstracts, with the annotations of the different argument components (evidence, claim, major claim). The structure of RCTs should follow the CONSORT policies to ensure a minimum consensus, which makes the studies comparable and ideal for building a corpus[1]. RCT abstracts were retrieved directly from PubMed[2] by searching for the disease name and specifying that it has to be a RCT. This version of the corpus with coarse labels contains 927 argument components (679 evidence and 248 claims) from 159 abstracts comprising 4 different diseases (glaucoma, hypertension, hepatitis b, diabetes).

In particular, an *evidence* in a RCT 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 Example 1, *evidence* are in italic, underlined and surrounded by square brackets with subscripts, while claims are in bold.

Example 1: 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]$_1$**.

Example 1 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). Different measures are annotated as multiple evidence. The reporting of side effects and negative observations are also considered as evidence. Traditionally evidence-based medicine (EBM) focuses mainly on the study design and risk of bias, when it comes to determining the quality of the evidence. As stated by (Bellomo and Bagshaw, 2006) there are also other aspects of the trial quality, which impinge upon the truthfulness of the findings. As a step forward, in this work we extend the corpus annotation, specifying four classes of *evidence*, which are most prominent in our data and assist in assessing these complex quality dimensions, like reproducibility, generalizability or the estimate of effect:

comparative: when there is some kind of comparison between the control and intervention arms (Table 1, example 2). Supporting the search for similarities in outcomes of different studies, which is an important measure for the reproducibility.

significance: for any sentence stating that the results are statistically significant (Table 1, example 3). Many comparative sentences also contain statistical information. However, this class can be seen more as a measure for the strength of beneficial or potentially harmful outcomes.

side-effect: captures all evidence reporting any side-effect or adverse drug effect to see if potential harms outweigh the benefits of an intervention (Table 1, example 4).

other: all the evidence that do not fall under the other categories, like non-comparative observations, risk factors or limitations of the study (too rare occurrences to form new classes). Especially the latter can be relevant for the generalizability of the outcome of a study (Table 1, example 5).

Table 2 shows the statistics of the obtained dataset. Three annotators have annotated the data after a training phase. Inter Annotator Agreement has been calculated on 10 abstracts comprising 47 evidence, resulting in a Fleiss' kappa of 0.88.

3 Proposed methods

In (Mayer et al., 2018), we addressed the argument component detection 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 classifier to detect the argumentative sentences. We retrained an existing system, i.e. MARGOT (Lippi and Torroni, 2016b), to detect evidence and claims from clinical data. The methods we used are SubSet Tree Kernels (SSTK) (Collins and Duffy, 2002),

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

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

1.	Brimonidine provides a sustained long-term ocular hypotensive effect, is well tolerated, and has a low rate of allergic response.
2.	The overall success rates were 87% for the 350-mm2 group and 70% for the 500-mm2 group ($P = 0.05$).
3.	All regimens produced clinically relevant and statistically significant ($P < .05$) intraocular pressure reductions from baseline.
4.	Allergy was seen in 9 % of subjects treated with brimonidine.
5.	Risk of all three outcomes was higher for participants with chronic kidney disease or frailty.

Table 1: Sample of each class represented in the corpus (*claim, comparative, significance, side-effect, other*).

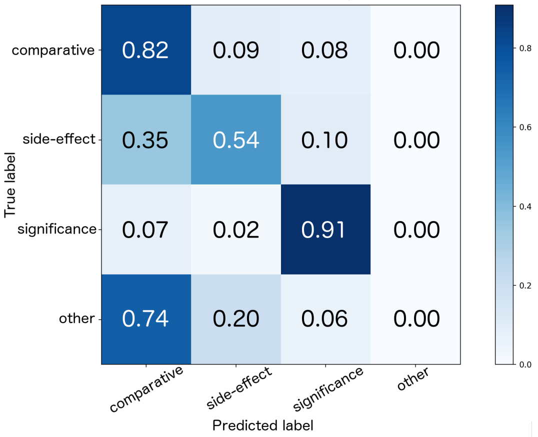

Figure 1: Normalized confusion matrix of the combined test set.

which offer a reasonable compromise between expressiveness and efficiency (Lippi and Torroni, 2016b). In SSTK, a fragment can be any subtree of the original tree, which terminates either at the level of pre-terminal symbols or at the leaves. 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. All the pre-processing steps were performed with Stanford CoreNLP (version 3.5.0). We conducted experiments with different classifiers and feature combinations. 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. 5-fold cross validation was performed optimizing for the F_1-score. The model was evaluated on the test set in Table 2 obtaining 0.80 and 0.65 F_1-score for evidence and claim detection respectively.

As a step forward - after the distinction between argumentative (claims and evidence) and non-argumentative sentences - we address the task of distinguishing the different types of evidence (see Section 2). We cast it as a multi-class classification problem. For that we use Support Vector Machines (SVMs)[3] with a linear kernel and different strategies to transform the multi-class into a binary classification problem: *(i)* ONEVSREST, and *(ii)* ONEVSONE. The first strategy trains one classifier for each class, where the negative examples are all the other classes combined, outputting a confidence score later used for the final decision. The second one trains a classifier for each class pair and only uses the correspondent subset of the data for that. As features, we selected lexical ones, like TF-IDF values for BoW, n-grams and the MedDRA[4] dictionary for adverse drug effects. As for the argument component classification, the model was evaluated on different test sets with respect to the weighted average F_1-score for multi-class classification. The models were compared against a random baseline, based on the class distribution in the training set and a majority vote classifier, which always assigns the label of the class with the highest contingent in the training set. The first dataset consisting only of the glaucoma data, and the second one comprising all the other maladies as well (see Table 2).

4 Results and Discussion

We run two sets of experiments. In the first one, we test the evidence type classifier on the gold standard annotations of the evidence. In the second one, we test the whole pipeline: the evidence type classifier is run on the output of the argument component classifier described in the previous section. In both cases, the best feature combination was a mix of BoW and bi-grams. The dictionary of adverse drug effects did not increase the performance. Together with the fact that the data contains just a small group of reoccurring side-effects, this suggests that the expected discriminative information from the dictionary is captured within the uni- and bi-gram features. This might change for bigger datasets with a broader range of adverse effects. Results of the best feature combinations and the random baseline are reported in Table 3. For the evidence type classifier on gold stan-

[3] scikit-learn, version 0.19.1

[4] https://www.meddra.org/

Dataset	*Topic*	*#abstract*	*#comp.*	*#sign.*	*#side-eff.*	*#other*
Training set	glaucoma	79	151	83	65	10
Test set	glaucoma, diabetes, hepatitis, hypertension	90 (resp. 30, 20, 20, 20)	160	98	79	33

Table 2: Statistics on the dataset showing the class distributions.

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 3: Results (weighted average F_1-score).

dard annotations, the observed results regarding the different multi-class strategies did not differ significantly. A F_1-score of 0.80 and 0.74 respectively for the glaucoma and combined test set was achieved. Reviewing the best n-grams, they contain very specific medical terminology, explaining the performance difference between the two test sets. For the future, another pre-processing step with better abstraction capability, e.g., substituting concrete medical related terms with more general tags, could provide benefits for the trained model on the out-of-domain task. The F_1-score of the whole pipeline is 0.71 for the glaucoma and 0.66 for the combined test set. As expected, the errors of the argument component classifier have an impact on the performances of the second step, but that corresponds to a more realistic scenario.

Error analysis. As shown in Figure 1, *side-effect* were often misclassified as *comparative*. Certain types of *side-effect* comprise comparisons of side-effects between the two groups including statements of their non-existence. The structure and wording of those sentences are very similar to correct *comparative* examples and only differ in the comparison criteria (side-effect vs. other measurement), see Examples 2 and 3. Furthermore, *comparative* and *significance* labels were often confused. As explained above, comparisons can also state information about the statistical significance and could therefore belong to both classes, see Example 4. For future work, we plan to adopt a multi-label approach to overcome this problem.

Example 2: Headache, fatigue, and drowsiness were similar in the 2 groups.

Example 3: The number of adverse events did not differ between treatment groups, with a mean (SD) of 0.21 (0.65) for the standard group and 0.32 (0.75) for the intensive group (P=0.44).

Example 4: The clinical success rate was 86.2% in the brimonidine group and 81.8% in the timolol group, making no statistically significant difference between them (p=0.817).

5 Concluding remarks

We have presented a first step towards mining fine-grained evidence from RCTs, contributing in *i)* the definition of the AM sub-task of evidence type classification for medical data, *ii)* a new dataset of RCT annotated with claims and four kinds of evidence, and *iii)* a supervised classifier to address this task.

A similar task is comparative structure identification in clinical trials. It relies on under-specified syntactic analysis and domain knowledge (Fiszman et al., 2007). (Gupta et al., 2017) applied syntactic structure and dependency parsers to extract comparison structures from biomedical texts. (Trenta et al., 2015) built an information extraction system based on a maximum entropy classifier with basic linguistic features for the tasks of extracting the patient group, the intervention and control arm, and the outcome measure description. Differently from us, they extract information to fill in evidence tables, ignoring the linguistic phrases to reconstruct the whole argumentation. (Dernoncourt et al., 2017) developed a neural network with word embeddings to assign PubMed RCT abstract labels to sentences showing that considering sequential information to jointly predict sentence labels improves the results. However, their task differs from ours as they predict the abstracts structure, which depends on contextual information. Concerning the evidence classification, (Rinott et al., 2015) tackled this problem on Wikipedia based data, dividing the evidence into *study, anecdotal and expert* evidence. This taxonomy is not applicable for the here presented type of data. Beside the extraction of evidence, another relevant task is their qualitative evaluation. The traditional quality-based hierarchy for medical evidence grades them based on the employed research method, e.g., the applied statistical principles (Schünemann et al., 2008). Top ranked methods comprise systematic reviews and meta-analyses of RCTs (Manchikanti et al., 2009). While they focus on collecting and using metadata from the studies to draw general conclusions to define, e.g., recommendation guidelines, they do not consider 'why' an author came to certain conclusion. This issue is tackled in our paper.

For future work, we plan to weight the argument strength based on the different evidence types (similar to the categories proposed in (Wachsmuth et al., 2017) and (Gurevych and Stab, 2017)). A scale for side-effects could be based on a weighted taxonomy of adverse drug effects. Furthermore, we plan to mine the full RCT reports, to get relevant information on the limitations of the study and risk factors, currently annotated with the *other* label since they rarely appear in the abstracts.

Acknowledgments

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

References

Rinaldo Bellomo and Sean M. Bagshaw. 2006. Evidence-based medicine: Classifying the evidence from clinical trials – the need to consider other dimensions. *Critical Care*, 10(5):232.

Michael Collins and Nigel Duffy. 2002. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *ACL*, pages 263–270. ACL.

Robert Craven, Francesca Toni, Cristian Cadar, Adrian Hadad, and Matthew Williams. 2012. Efficient argumentation for medical decision-making. In *KR*. AAAI Press.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. Neural networks for joint sentence classification in medical paper abstracts. In *EACL*, pages 694–700.

Marcelo Fiszman, Dina Demner-Fushman, François-Michel Lang, Philip Goetz, and Thomas C. Rindflesch. 2007. Interpreting comparative constructions in biomedical text. In *BioNLP@ACL*, pages 137–144.

Samir Gupta, A. S. M. Ashique Mahmood, Karen Ross, Cathy H. Wu, and K. Vijay-Shanker. 2017. Identifying comparative structures in biomedical text. In *BioNLP 2*, pages 206–215.

Iryna Gurevych and Christian Stab. 2017. Recognizing insufficiently supported arguments in argumentative essays. In (Lapata et al., 2017), pages 980–990.

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

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

Mirella Lapata, Phil Blunsom, and Alexander Koller, editors. 2017. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*. Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2016a. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn.*, 16(2):10.

Marco Lippi and Paolo Torroni. 2016b. MARGOT: A web server for argumentation mining. *Expert Systems with Applications*, 65:292–303.

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

Laxmaiah Manchikanti, Sukdeb Datta, Howard Smith, and Joshua A Hirsch. 2009. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: part 6. systematic reviews and meta-analyses of observational studies. *Pain physician*, 12 5:819–50.

Tobias Mayer, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata. 2018. Argument mining on clinical trials. In *Proceedings of COMMA'18*.

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

Malik Al Qassas, Daniela Fogli, Massimiliano Giacomin, and Giovanni Guida. 2015. Analysis of clinical discussions based on argumentation schemes. *Procedia Computer Science*, 64:282–289.

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

Holger Schünemann, Andrew D Oxman, Jan Brozek, Paul Glasziou, Roman Jaeschke, Gunn Vist, John Williams, Regina Kunz, Jonathan Craig, Victor M Montori, Patrick Bossuyt, and Gordon Guyatt. 2008. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ (Clinical research ed.)*, 336:1106–10.

Antonio Trenta, Anthony Hunter, and Sebastian Riedel. 2015. Extraction of evidence tables from abstracts of randomized clinical trials using a maximum entropy classifier and global constraints. *CoRR*, abs/1509.05209.

Henning Wachsmuth, Benno Stein, Graeme Hirst, Vinodkumar Prabhakaran, Yonatan Bilu, Yufang Hou, Nona Naderi, and Tim Alberdingk Thijm. 2017. Computational argumentation quality assessment in natural language. In (Lapata et al., 2017), pages 176–187.