Assisted authoring for avoiding inadequate claims in scientific reporting

Anna Koroleva

Thèse de doctorat

NNT : 2020UPASS021

Assisted authoring for avoiding inadequate claims in scientific reporting

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580 Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat: Informatique
Unité de recherche : Université Paris-Saclay, CNRS, LIMSI, 91400, Orsay, France
Référent : Faculté des sciences d'Orsay

Thèse présentée et soutenue à Amsterdam, le 22 janvier 2020, par

Anna KOROLEVA

Sophia Ananiadou Professeur, University of Manchester	Présidente, Rapporteur
Paolo Rosso Professeur, Universitat Politècnica de València	Rapporteur
Ameen Abu-Hanna Professeur, Universiteit van Amsterdam	Examinateur
Evangelos Kanoulas Professeur, Universiteit van Amsterdam	Examinateur
Olivier Ferret Ingénieur de Recherche, Université Paris-Saclay	Examinateur
Nicolas Sabouret Professeur, Université Paris-Saclay	Examinateur
Patrick Paroubek Ingénieur de Recherche, Université Paris-Saclay	Directeur de thèse
Patrick M.M. Bossuyt Professeur, Universiteit van Amsterdam	Co-Directeur de thèse

Assisted authoring for avoiding inadequate claims in scientific reporting

Anna Koroleva

ISBN: 978-90-830376-4-6

Cover design: James Temple-Smith & Anna Koroleva
Layout: Anna Koroleva
Printed by: Print Service Ede - www.printservice-ede.nl

This project was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

Assisted authoring for avoiding inadequate claims in scientific reporting

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. ir. K.I.J. Maex

ten overstaan van een door het College voor Promoties ingestelde commissie,

in het openbaar te verdedigen in de Agnietenkapel

op woensdag 22 januari 2020, te 16.00 uur

door Anna Koroleva

geboren te Moskou

Promotiecommissie:

Promotor:	prof. dr. P.M.M. Bossuyt	AMC-UvA
Promotor:	dr. P. Paroubek	Université Paris-Saclay
Overige leden:	prof. dr. A. Abu-Hanna	AMC-UvA
	prof. dr. E. Kanoulas	Universiteit van Amsterdam
	prof. dr. S. Ananiadou	University of Manchester
	prof. dr. P. Rosso	Universitat Politècnica de València
	prof. dr. N. Sabouret	Université Paris-Saclay
	dr. O. Ferret	Université Paris-Saclay

Faculteit der Geneeskunde

Dit proefschrift is tot stand gekomen in het kader van het European project "MIROR" GA No. 676207, met als doel het behalen van een gezamenlijk doctoraat.
Het proefschrift is voorbereid in de Faculteit der Geneeskunde van de Universiteit van Amsterdam en in de Centre National de la Recherche Scientific (CNRS) van de Université Paris-Saclay.

This thesis has been written within the framework of the European project "MIROR" GA No. 676207, with the purpose of obtaining a joint doctorate degree.
The thesis was prepared in the Faculty of Medicine at the University of Amsterdam and in the Centre National de la Recherche Scientific (CNRS) at the Université Paris-Saclay.

Contents

Introduction

Many, if not all, of us came across a situation of being prescribed a medication that did not work as it was supposed to and, maybe, even worse, had some unpleasant side effects, leaving us wondering how such a substance could get into clinical use, why our doctor chose to prescribe it, and why there was no warning of side effects.

Absence of effects, or even negative effects, of drugs can occur simply because persons differ, and even established and approved drugs may not work, or do not help, everyone. But what if there are other mechanisms as well that can lead to the dissemination of less effective drugs, to doctors prescribing them, and may lead the public into believing in their effectiveness? I had no idea that such mechanisms can exist before I started my PhD and learned about spin in reporting research results.

The notion of "spin" comes from the domain of politics, where it denotes a form of propaganda and consists in passing on a biased message to the public, to create a certain (positive or negative) perception of a person or event (Tye, 1998; Jackson and Jamieson, 2007; Boardman et al., 2017). Political spin is often related to deception and manipulation.

From politics, the term "spin" came into the domain of scientific research, where it refers to presenting research results in a more positive (or, rarely, more negative) way that the obtained evidence justifies. In 2010, Boutron et al. (2010) introduced and defined the term "spin" for research domain, in particular, for randomized controlled trials (RCTs) - clinical trials studying a new intervention by comparing it to an established intervention or to a placebo. In RCTs with non-significant primary outcome, spin is defined as "the use of specific reporting strategies, from whatever motive, to highlight that the experimental treatment is beneficial, despite a statistically nonsignificant difference for the primary outcome, or to distract the reader from statistically nonsignificant results" (Boutron et al., 2010). A variety of terms was used to denote this phenomenon, such as distorted / misleading presentation, mis-/overinterpretation of research results; "spin" is now the most commonly accepted term for this phenomenon.

If one looks at the examples of spin provided in the relevant articles, it may seem that the difference between "spinned" and fair presentation of results is very slight; as a result, people often wonder if spin is really a problem. To answer this question, Boutron et al. (2014)

compared how clinicians perceive the treatments presented in abstracts with spin and in the same abstract rewritten without spin. The authors showed that the clinicians overestimated the effects of the treatment after reading a "spinned" abstract. Given that the abstract is often the only part of an article available to wide public free of charge, readers often cannot read the full text of an article to assess the correctness of conclusions in the abstract. Hence, clinical decisions can be made on the basis of the information in abstracts; and so spin in abstracts can have a highly negative impact on clinical decision-making. Interventions with unproved efficacy or safety can get into production and clinical use (and maybe this is one of the reasons we have that negative experience with taking medications?).

Spin in research articles can have a wider impact: it was shown to be related to spin in press releases and health news (Haneef et al., 2015; Yavchitz et al., 2012), and thus can impact the beliefs and expectations of the public regarding new treatments.

The phenomenon of spin has started to attract attention not only of the research community, but also of general public during the recent years[1]. Still, spin often remains unnoticed even by editors and peer reviewers: recent studies (2016 - 2019) showed that, in RCTs with non-significant primary outcome, the percentage of abstracts with spin is high in a variety of domains, such as surgical research (40%) (Fleming, 2016), cardiovascular diseases (57%) (Khan et al., 2019), cancer (47%) (Vera-Badillo et al., 2016), obesity (46.7%) (Austin et al., 2018), otolaryngology (70%) (Cooper et al., 2018), anaesthesiology (32.2%) (Kinder et al., 2018), and wound care (71%) (Lockyer et al., 2013).

Spin is a type of research waste —a problem consisting in spending billions of euros per year on low-quality studies that have flaws in their design, are poorly reported or never published (Ioannidis, 2005). In 2014, Macleod et al. (2014) estimated up to 85% of money spent on clinical research to be wasted yearly. In 2018, Glasziou and Chalmers (2018) stated that although some progress has been made, the problem is still far from being solved.

As one of the ways of reducing research waste, an assistance to readers in detecting spin could be useful. Spin can be viewed as a textual phenomenon, related to certain types of expressions that represent inconsistent or incomplete presentation of information in some parts of a text. Thus, identifying spin is a text analysis task, consisting in searching for and analysing certain information in various parts of clinical articles. This task can be performed manually, or with an automated computerized aid. The motivation for using such aid tool is that modern technologies allow computers to understand and analyse text on a near-human level with substantial gain in speed, which means that computer programs can facilitate and speed up the completion of various text analysis tasks. Nowadays, automated and semi-automated aid tools

[1]See some blog posts on the topic: https://blogs.plos.org/absolutely-maybe/2016/03/17/how-to-spot-research-spin-the-case-of-the-not-so-simple-abstract/ (2016), https://www.medicalnewstoday.com/articles/325952.php (2019).

are coming into use in different domains: extraction of trial design elements (Kiritchenko et al., 2010), indexing of medical texts with Medical Subject Headings (MeSH) terms (Mork et al., 2013), risk of bias assessment and evidence synthesis (Marshall et al., 2015, 2017), systematic review process (Ananiadou et al., 2009; Blake and Lucic, 2015; O'Mara-Eves et al., 2015), scientific writing process (Barnes et al., 2015). The majority of the cited tools use methods of Natural Language Processing (NLP) and machine learning (ML). While a complete automation of complex text analysis tasks is still not realistic, semi-automation can be successfully used to extract information relevant for a given task and provide it to human experts, who will make a final conclusion for the task. Similarly to the listed tasks, NLP and ML methods can be leveraged to develop and implement algorithms for detecting potential spin and the supporting information that can help human experts perform assessment of a clinical article for presence of spin.

The aim of the work presented in this thesis was to develop NLP algorithms to identify spin and related information elements. We focus on spin in abstracts of articles of RCT, as RCTs are the main source of data for Evidence-Based Medicine, and abstracts are the most widely available part of articles.

We conducted a study of types of spin and their textual characteristics, using first the existing literature describing spin and the provided examples, and second our observations on a larger corpus of general domain and mental health domain clinical trials. We summarized our observations in the form of annotation scheme and a provisional scheme of spin detection algorithms. We developed a set of baseline rule-based algorithms for a number of key tasks. We explored the possibilities of running a large-scale annotation project for spin annotation and, after it proved unfeasible, we annotated a set of corpora for the most important tasks with the efforts of one annotator. We developed and tested a number of machine learning approaches for the tasks, and chose the best performing approaches for the final implementation of a spin detection pipeline, released as open source code, supplemented with a simple annotation and visualization interface.

The outline of this thesis is as follows:

Part I (chapters 1 - 2) describes our first experiments towards the goal of developing a spin detection pipeline.

Chapter 1 introduces the provisional scheme of a spin detection pipeline and describes state-of-the art, our experiments and possible directions for future work for three spin-related tasks: text classification according to study design (to detect RCT reports); classification of sentences of abstracts to identify sections; and entity extraction (for trial outcomes and population studied).

Chapter 2 reports on our efforts of collecting a corpus of biomedical articles annotated for

spin and spin-related information. It describes the development of an annotation scheme for spin and related information items, annotation guidelines, and arising difficulties, such as the level of knowledge required from annotators, choice of a convenient and easy-to-use annotation tool, and inherent complexity of the task.

Part II (chapters 3 – 6) describes the key algorithms that we built for a spin detection system.

Chapter 3 describes development and evaluation of algorithms for extracting declared (primary) and reported outcomes. For this goal, we annotated a corpus for these two types of entities. We implemented and assessed a number of approaches, including a rule-based baseline approach and a number of machine learning approaches. We employed and compared several deep pre-trained language representation models, including BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), BioBERT (Lee et al., 2019) and SciBERT (Beltagy et al., 2019). We report on the performance of all the models and approaches.

Chapter 4 reports on the development of a semantic similarity assessment algorithm for pairs of trial outcomes. Based on the corpus annotated for the outcome extraction task, we annotated a set of pairs of primary and reported outcomes for their similarity (on a binary scale). We implemented a number of similarity measures, based on strings, tokens and lemmas, and distances between phrases in the WordNet semantic network. We trained and tested a machine learning classifier combining all the measures. Finally, we fine-tuned pre-trained language models (BERT, BioBERT and SciBERT) on our corpus, which proved to be the most successful approach.

Chapter 5 describes our proposed algorithm for detection of outcome switching - a type of spin consisting in unjustified change (omitting or adding) of pre-defined trial outcomes. The algorithm combines outcome extraction and semantic similarity assessment algorithms described in the two previous chapters.

Chapter 6 describes our efforts for annotating a corpus and developing an algorithm for extracting the relation between trial outcomes and their statistical significance levels. For this purpose, we annotated a corpus for pairs of related outcomes and significance levels (in both numerical form - p-value - and qualitative form). We tested a number of machine learning classifiers using manually crafted feature set. Besides, we fine-tuned and tested BERT-based language models, which proved to be superior to the machine classifiers using manually crafted features.

Part III (chapter 7) is devoted to our implementation efforts.

Chapter 7 describes our spin detection prototype system, called DeSpin (Detector of Spin). It outlines the textual features of types of spin addressed, the algorithms and methods used as

well as the best achieved results. The chapter contains a link to an open source release of our code.

Part IV (chapter8) gives an overview of the difficulties encountered in the course of our project.

Chapter 8 describes the most important information element in our pipeline: outcomes. Despite the wide use of this notion, in practice outcomes are highly diverse and lack reporting standards. We describe the observed diversity in the ways of defining and introducing outcomes in clinical study reports.

References

S. Ananiadou, B. Rea, N. Okazaki, R. Procter, and J. Thomas. Supporting systematic reviews using text mining. *Social Science Computer Review - SOC SCI COMPUT REV*, 27:509–523, 10 2009. doi: 10.1177/0894439309332293.

J. Austin, C. Smith, K. Natarajan, M. Som, C. Wayant, and M. Vassar. Evaluation of spin within abstracts in obesity randomized clinical trials: A cross-sectional review: Spin in obesity clinical trials. *Clinical Obesity*, 9:e12292, 12 2018. doi: 10.1111/cob.12292.

C. Barnes, I. Boutron, B. Giraudeau, R. Porcher, D. Altman, and P. Ravaud. Impact of an online writing aid tool for writing a randomized trial report: The cobweb (consort-based web tool) randomized controlled trial. *BMC medicine*, 13:221, 09 2015. doi: 10.1186/s12916-015-0460-y.

I. Beltagy, A. Cohan, and K. Lo. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

C. Blake and A. Lucic. Automatic endpoint detection to support the systematic review process. *J. Biomed. Inform*, 2015.

F. Boardman, N. M. Cavender, and H. Kahane. *Logic and Contemporary Rhetoric: The Use of Reason in Everyday Life. 13 edition.* Cengage Learning, 2017.

I. Boutron, S. Dutton, P. Ravaud, and D. Altman. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*, 2010.

I. Boutron, D. Altman, S. Hopewell, F. Vera-Badillo, I. Tannock, and P. Ravaud. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *Journal of Clinical Oncology*, 2014.

C. M. Cooper, H. M. Gray, A. E. Ross, T. A. Hamilton, J. B. Downs, C. Wayant, and M. Vassar. Evaluation of spin in the abstracts of otolaryngology randomized controlled trials: Spin found in majority of clinical trials. *The Laryngoscope*, 12 2018. doi: 10.1002/lary.27750.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

P. S. Fleming. Evidence of spin in clinical trials in the surgical literature. *Ann Transl Med.*, 4,19(385), Oct 2016. doi: 10.21037/atm.2016.08.23.

P. Glasziou and I. Chalmers. Research waste is still a scandal—an essay by paul glasziou and iain chalmers. *BMJ*, 363, 2018. ISSN 0959-8138. doi: 10.1136/bmj.k4645. URL `https://www.bmj.com/content/363/bmj.k4645`.

R. Haneef, C. Lazarus, P. Ravaud, A. Yavchitz, and I. Boutron. Interpretation of results of studies evaluating an intervention highlighted in google health news: a cross-sectional study of news. *PLoS ONE*, 2015.

J. Ioannidis. Why most published research findings are false. *PLoS medicine*, 2:e124, 09 2005. doi: 10.1371/journal.pmed.0020124.

B. Jackson and K. H. Jamieson. *unSpun: Finding Facts in a World of Disinformation*. Random House, 2007.

M. Khan, N. Lateef, T. Siddiqi, K. Abdur Rehman, S. Alnaimat, S. Khan, H. Riaz, M. Hassan Murad, J. Mandrola, R. Doukky, and R. Krasuski. Level and prevalence of spin in published cardiovascular randomized clinical trial reports with statistically nonsignificant primary outcomes: A systematic review. *JAMA Network Open*, 2:e192622, 05 2019. doi: 10.1001/jamanetworkopen.2019.2622.

N. Kinder, M. Weaver, C. Wayant, and M. Vassar. Presence of 'spin' in the abstracts and titles of anaesthesiology randomised controlled trials. *British Journal of Anaesthesia*, 122, 11 2018. doi: 10.1016/j.bja.2018.10.023.

S. Kiritchenko, B. D. Bruijn, S. Carini, J. Martin, and I. Sim. Exact: automatic extraction of clinical trial characteristics from journal publications. *BMC Med Inform Decis Mak.*, 2010.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*, 2019.

S. Lockyer, R. W. Hodgson, J. C. Dumville, and N. Cullum. "Spin" in wound care research: the reporting and interpretation of randomized controlled trials with statistically non-significant primary outcome results or unspecified primary outcomes. In *Trials*, 2013.

M. Macleod, S. Michie, I. Roberts, U. Dirnagl, I. Chalmers, J. Ioannidis, R. Al-Shahi Salman, A. Chan, and P. Glasziou. Biomedical research: Increasing value, reducing waste. *Lancet*, 383(9912):101–104, 2014. ISSN 0140-6736. doi: 10.1016/S0140-6736(13)62329-6.

I. Marshall, J. Kuiper, E. Banner, and B. C. Wallace. Automating biomedical evidence synthesis: RobotReviewer. In *Proceedings of ACL 2017, System Demonstrations*, pages 7–12, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P17-4002`.

I. J. Marshall, J. Kuiper, and B. C. Wallace. Robotreviewer: Evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association : JAMIA*, 23, 06 2015. doi: 10.1093/jamia/ocv044.

J. Mork, A. Jimeno-Yepes, and A. Aronson. The nlm medical text indexer system for indexing biomedical literature. *CEUR Workshop Proceedings*, 1094, 01 2013.

A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews*, 4(1):5, Jan 2015. ISSN 2046-4053. doi: 10.1186/2046-4053-4-5. URL `https://doi.org/10.1186/2046-4053-4-5`.

L. Tye. *The Father of Spin: Edward L. Bernays and the Birth of Public Relations*. Crown Publishers, 1998. ISBN 978-0-517-70435-6.

F. E. Vera-Badillo, M. Napoleone, M. K. Krzyzanowska, S. M. Alibhai, A.-W. Chan, A. Ocana, B. Seruga, A. J. Templeton, E. Amir, and I. F. Tannock. Bias in reporting of randomised clinical trials in oncology. *European Journal of Cancer*, 61:29 – 35, 2016. ISSN 0959-8049. doi: 10.1016/j.ejca.2016.03.066. URL `http://www.sciencedirect.com/science/article/pii/S0959804916320287`.

A. Yavchitz, I. Boutron, A. Bafeta, I. Marroun, P. Charles, J. Mantz, and P. Ravaud. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS Med*, 2012.

Part I

Algorithm scheme and annotation

Chapter 1

Automatic detection of inadequate claims in biomedical articles: first steps. Anna Koroleva, Patrick Paroubek. Proceedings of Workshop on Curative Power of MEdical Data, Constanta, Romania, September 12-13, 2017.

Context

We analysed the definitions of spin, its types and subtypes in the existing literature to identify the information elements relevant for spin detection and to outline a provisional scheme for a spin detection algorithm.

After defining the algorithm scheme, we addressed three tasks that are the core tasks for spin detection: text classification according to clinical trial design (to detect randomized controlled trials), classification of sentences in abstracts of scientific articles (to detect the different sections of the abstract), and entity extraction (to detect outcomes and the population studied). We reviewed the state of the art for each task, conducted our first experiments for these tasks in the context of spin detection, and defined possible directions for future work.

Authors' contributions

The work reported in this chapter was conducted by AK under supervision of PP. AK was responsible for data collection, experiment and analysis. AK drafted the manuscript. PP revised the draft critically for important intellectual content.

Abstract

In this article we present the first steps in developing an NLP algorithm for automatic detection of inadequate reporting of research results (known as spin) in biomedical articles. Inadequate reporting consists in presenting the experimental treatment as having a greater beneficial effect than it was shown by the research results. We propose a scheme for an algorithm that would automatically identify important claims in the articles abstracts, extract possible supporting information from the article and check the adequacy of the claims. We present the state of the art and our first experiments for three tasks related to spin detection: classification of articles according to the type of reported clinical trial; classification of sentences in the abstracts aimed at identifying mentions of the Results and Conclusions of the experiment; and extraction of some trial characteristics. For each task, we outline possible directions of further work.

Keywords: Inadequate Reporting, Spin, Biomedical Articles, Text Classification, Entity Extraction.

Introduction

Inadequate claims, or inadequate reporting, are more commonly referred to as 'spin'. Spin in scientific research is a way of distorting the presentation of research results by claiming that they are more positive than what is normally justifiable from the evidences that the experiment yielded. In our project we deal with spin in articles reporting clinical trials which aim at testing a new (experimental) intervention by comparing it against a standard (control) treatment. Spin in medical articles is defined as stating the beneficial effect of the experimental treatment in terms of efficacy or safety to be greater than it is shown by the research results (Boutron et al., 2010, 2014; Haneef et al., 2015; Yavchitz et al., 2016). Two examples of conclusions with spin and the same conclusions rewritten by experts to remove spin are given in Table 1.1.

Spin in the medical field presents an alarming problem as it was proven to change clinicians' interpretation of the efficacy of the experimental treatment, i.e. it makes clinicians overestimate the treatment's beneficial effect (Boutron et al., 2014). Thus, it has a negative impact on the clinical decision-making. The presence of spin also provokes distorted presentation of research findings in press releases and health news (Haneef et al., 2015; Yavchitz et al., 2012).

Spin occurs in articles reporting various types of trials (non-randomized controlled trials, randomized controlled trials, diagnostic accuracy studies) (Boutron et al., 2010, 2014; Lazarus et al., 2015; Yavchitz et al., 2016). We focus on the randomized controlled trials (RCTs) that are the primary source of data for evidence-based medicine (EBM). We concentrate now on spin in abstracts.

The principal objective of our project is to develop an algorithm for automatic spin detection that would assist scientific authors, readers and peer-reviewers in identifying possible instances

Original (anonymized) conclusion	Rewritten conclusion
Treatment A + CAF was well tolerated and is suggested to have efficacy in patients who had not received prior therapy.	Treatment A + CAF was not more effective than CAF + placebo in patients with advanced or recurrent breast cancer.
This study demonstrated improved PFS and response for the treatment A compared with comparator B alone.	The treatment A was not more effective than comparator B on overall survival in patients with metastatic breast cancer.

Table 1.1: Examples of conclusions with spin and the same conclusions rewritten without spin

of spin. For this purpose we plan to use Natural Language Processing techniques to detect important claims in scientific articles, extract possible supporting information for them and evaluate the adequacy of the claims.

The structure of this paper is the following: in section 2 we present existing types of spin and the supporting information relevant for various types; in section 3 we present the proposed scheme of our algorithm; in section $4 - 6$ we address some of the subtasks of spin detection: we present the related research, our current work and obtained results, and we provide an outline of our future work.

Types of spin

Spin in medical articles can be classified into the following types (Boutron et al., 2010; Lazarus et al., 2015; Yavchitz et al., 2016):

1. misleading reporting of study results: selective reporting (not reporting the primary outcome; focus on statistically significant secondary outcomes or subgroups of patients); misleading reporting of study design; not reporting adverse events; linguistic spin; no consideration of limitations; selective citation of other studies.

2. inadequate interpretation of the results: claiming a beneficial or equivalent effect of the intervention for statistically non-significant results or with no comparison test performed; claiming the treatment's safety for statistically non-significant safety outcomes; interpretation of the results according to statistical significance instead of clinical relevance; claiming a causal effect between the intervention assessed and the outcome of interest despite a non-randomized design

3. inadequate extrapolation: inadequate extrapolation from the population, interventions

or outcome actually assessed in the study to a larger population, different interventions or outcomes; inadequate implications for clinical practice.

Basing on this classification, we can highlight the following categories of supporting information for spin (information that could prove the conclusions): study design; outcomes (primary and secondary); statistical significance of results; patient population studied; adverse events; limitations of a trial; interventions studied.

Algorithm description

Out future algorithm is intended to assist both authors and readers. The default input of the algorithm is a full-text article with title and abstract. When used by an author, it may benefit from additional information, e.g. division of the text into structural parts (title, abstract, body text) or information about the trial (design, interventions, etc.) provided by the author, by default we suppose that no such information is available; thus, our algorithm ought to be able to find or infer the required pieces of information.

We propose the following provisionary procedure for spin detection:

1. File preprocessing: if the source file is not in a raw text format (e.g. a .doc of .pdf document), then convert it.

2. Divide the text into structural parts: title, abstract, body text.

3. Automatically identify whether the text is an article reporting an RCT. If not, it will not be considered by the algorithm.

4. Automatically classify sentences in the abstract to identify those containing mentions of RCT results and conclusions. These sentences are supposed to contain important claims that are to be checked for the presence of spin.

5. Identify the tonality of reported results in the abstract: positive/ neutral / negative / mixed. If no positive or mixed results are reported, the abstract is considered not to contain spin.

6. If positive or mixed results are reported, the next stage is information extraction, which concerns:

    - Entity extraction. For the moment we are focusing on the types of spin related to misreporting of outcomes and patient population, thus, our primary goal is to extract information about pre-defined outcomes, patient population, and statistical

significance of results. Detecting other types of spin would also require extracting other information such as interventions examined, or observed adverse events.

- Relation extraction: finding relations between entities extracted at the previous stage, e.g. the link between outcomes and their significance levels, which will be used to identify the cases where non-significant results are presented as positive.
- Exploring specific linguistic features: looking for specific constructions that can represent a certain type of spin, e.g. similarity statements in the abstract results and conclusions, advice to use the experimental treatment; other linguistic features that may be related to spin (e.g. "hedging" —expressions of uncertainty).

7. Look for specific spin markers, e.g.:

   - Is the primary outcome reported in the abstract? If positive results for the primary outcome are reported, are they statistically significant?
   - Is the patient population mentioned in the results/conclusions of the abstract the same that the population initially studied?
   - If there is a similarity statement for the two treatments compared, was the trial of the non-inferiority/equivalence type?

Text classification according to study design

Related work

Identification of RCTs among different types of medical texts has received sufficient attention since finding RCTs relevant to a given topic is required for systematic reviews and other tasks in the domain of EBM. In some databases such as Medline, texts are manually annotated with several types of metadata, including Medical Subject Headings (MeSH) terms and publication types (e.g. "randomized controlled trial", "observational study", etc.). However, the manual annotation is not always complete and precise; thus, several articles addressed the problem of creating search strategies for identifying RCTs in Medline (Glanville et al., 2006; Higgins and Green, 2008; McKibbon et al., 2009). These works explore both annotation metadata and terms present in the articles. Although not complete, the annotation metadata has been proven to be the most useful feature for identifying RCTs (Glanville et al., 2006).

Cohen et al. (2015) addressed the task of creating a binary classifier aimed at identifying RCTs in Medline, using the textual features of the title and abstract, bibliographic features and annotation metadata such as MeSH terms. Manually annotated publication types served

as a gold standard for classification. The whole corpus consisted of over 5 million articles; a 7.5% sample was used for training and cross-validation. The classifier performed well with reported accuracy ≥ 0.984 and F-score ≥ 0.807

Experiments

Our primary aim is to identify RCTs, but we also examine the possibility to distinguish non-randomized clinical trials as their automatic detection may be useful for future works on spin identification. Thus, our classification model has three categories: RCT, clinical study (which means here a non-RCT), and other.

Our corpus is a set of PMC[1] articles collected in the course of some previous experiments. The initial corpus consists of 119,339 texts; using the Medline metadata we obtained the publication type for 65,396 articles: 3,938 had the type "Randomized controlled trial", 1,139 had the type "Clinical Trial" (excluding the RCTs) and 60,319 were of other types. A disadvantage of our corpus is imbalance between the numbers of articles belonging to different types. However, we were interested in exploring features of the full-text articles and not only of titles and abstract. Retrieving full-text articles is a complex and time-consuming task. Thus, we decided to evaluate the quality that we can achieve with this corpus which was already available.

We compared different sets of features. They can be divided into the following types: information about the structure of the text (division into title, abstract and body text), textual features (n-grams and their position in the text, i.e., whether an n-gram occurred in the title, abstract or body text; relative position of an n-gram in the body text), metadata (authors' names, journal that published the paper). As our future algorithm is to be used for papers yet unpublished, one of our points of interest was the performance of classifier without the use of the metadata.

We compared performance of several classifiers implemented in Weka software (Hall et al., 2008). The best performance was shown by SMO classifier using textual features of the whole text of articles (title, abstract and body text), taking into account information about the division of the text into the three structural parts, but excluding metadata. The overall performance was the following: precision = 0.955, recall = 0.966, F-measure = 0.958. However, as our corpus is highly imbalanced, we were more interested in the quality of classification for the two target classes: RCT and clinical study classes. For RCT, the classifier shows relatively good performance: precision = 0.889, recall = 0.805 and F-measure = 0.845. For the class "clinical study" the performance is low: precision = 0.318, recall = 0.042 and F-measure = 0.074. These results may stem from the fact that the corpus is highly imbalanced.

[1]PMC (PubMed Central) is a database of full-text articles in the domains of biomedicine and life sciences. Official site: https://www.ncbi.nlm.nih.gov/pmc/

Future work

One of the directions for future work is exploring the feasibility of adding syntactic features to the classification model, e.g. the pairs and triples of the type (Word, Word) or (Word, Syntactic Group) and (Word, Relation, Word) or (Word, Relation, Syntactic Group), some of which may be associated with a certain class of texts. We will evaluate the performance of the classifier with these features added. Another possible way to improve the classification quality is enlarging the training corpus.

Abstracts sentence classification

Related work

The problem of identifying sentence types in medical articles abstracts (e.g. general categories such as Introduction, Method, Result, or Conclusion, or more specific types such as Intervention, Participants and Outcome) has been addressed by several studies (Hirohata et al., 2008; Kim et al., 2011; McKnight and Srinivasan, 2003; Yamamoto and Takagi, 2005). Simple bag-of-words approach was explored and showed good performance (McKibbon et al., 2009). Other features used to enhance the classification performance include: structural information (position of a sentence within an abstract) (McKnight and Srinivasan, 2003), semantic information (semantic categories of words and phrases, obtained through MetaMap (Aronson, 2001)), sequential information (features of preceding/following sentences) (Hirohata et al., 2008; Kim et al., 2011). Classifiers used for this task include SVM and CRF. Classifiers are trained on manually annotated corpora (Kim et al., 2011; McKnight and Srinivasan, 2003; Yamamoto and Takagi, 2005) or use structured abstracts as gold standard (Hirohata et al., 2008; McKnight and Srinivasan, 2003; Yamamoto and Takagi, 2005).

Experiments

We seek to classify sentences in the abstracts into three categories: Results, Conclusions and Other. Following the approach adapted in Hirohata et al. (2008); McKnight and Srinivasan (2003); Yamamoto and Takagi (2005), we use the structured abstracts as the gold standard. The structure of abstracts coming from different sources may differ: an abstract may contain general sections such as Background, Methods, Results, Conclusions, or authors may divide it into more specific parts such as Problem, Objective, Importance, which correspond to Background; Participants, Outcomes, Intervention, which correspond to Methods, etc. We chose the three above-mentioned categories for our classification because Results and Conclusions

sections are the most important for our final goal of spin detection and because they are among the basic sections, most often present in structured abstracts.

We explored textual features of the abstracts (n-grams) and structural information (relative position of a sentence in the abstract). With the use of SMO classifier in Weka we achieved the following overall performance: precision = 0.899, recall = 0.899. For the class "Conclusion", precision is 0.915 and recall is 0.844; for the class "Results", precision is 0.896 and recall is 0.888

Future work

Our current results are relatively good and comparable to some of the previously reported approaches (McKnight and Srinivasan, 2003; Yamamoto and Takagi, 2005), but they are still lower than the best results obtained for this task, e.g. Hirohata et al. (2008). Our future work will be aimed at exploring the possibilities to improve the classification quality using semantic and sequential information as it was done by previous works. We will further test the classifier on unstructured abstracts.

Information extraction: outcomes and population

Related work

Extraction of entities that represent clinical study characteristics (patient population, interventions, diseases, outcomes, negative side effects, etc.) receives sufficient attention as it is crucial for automatic text summarization, question-answering systems or tasks related to creation and use of structured databases.

Some of the authors (Bruijn et al., 2008; Kiritchenko et al., 2010) aimed at extracting a large variety of information about a trial, such as experimental and control treatment, patients eligibility criteria, dosage, duration and frequency of treatment administration, sample sizes, primary and secondary outcomes, financing, etc. Some other works are focused on a limited set of entities relevant to a certain task, e.g. treatment names, intervention groups and outcomes (Summerscales et al., 2009, 2011); descriptions and sizes of patient groups, outcomes examined, and numerical data for outcomes (Summerscales et al., 2011); intervention arms (Chung, 2009); patient population including general description, sample sizes, medical condition (Raja et al., 2016; Xu et al., 2007).

We can draw some interesting observations about the approaches and methods used. The majority of the articles is focused on RCTs; and are aimed at extracting the data from abstracts, with only a few taking into consideration the whole text of an article (Bruijn et al., 2008;

Kiritchenko et al., 2010). The most common approach consists of two stages. First, the sentences are filtered, most often with the use of a classifier, to choose those that are likely to contain the target entities (Chung, 2009; Bruijn et al., 2008; Kiritchenko et al., 2010; Raja et al., 2016; Summerscales et al., 2011; Xu et al., 2007); second, the sentences identified at the first stage are searched for entity mentions, which is done by means of rule based approaches (Bruijn et al., 2008; Kiritchenko et al., 2010; Raja et al., 2016; Xu et al., 2007) or CRF-based automatic classifiers (Chung, 2009; Raja et al., 2016). A common approach is thus to combine the rule-based techniques and machine learning.

Some of the works focused on syntactic features in abstracts since they explore extraction of relevant information from specific syntactic constructions (Chung, 2009). Semantic information retrieved with the use of systems such as MetaMap, that links the terms of a text to the terms of medical thesauri, is frequently used (Chung, 2009; Summerscales et al., 2009, 2011). Semantic information is reported to be more useful than information about word shape (Summerscales et al., 2009, 2011).

Experiments

Our first goal is to identify 1) outcomes and 2) patient population, because these two types of information are most often misrepresented in the medical articles abstracts, with pre-specified outcomes and population being changed, replaced, or removed.

One of the possible ways to obtain this information is to extract it from trial registries (online databases containing trial data, with each registered trial assigned a unique identifier). Trial registration becomes more and more common, and the registration number is likely to be reported in an article. Registration numbers follow some fixed patterns, including usually a registry identifier and a trial identifier, e.g. NCT00000001 would be a trial registered at the ClinicalTrials.gov registry under the number 00000001. Given the registration number, it is possible to automatically access the webpage of the trial and download the data, which usually includes the outcomes and patient information. This task belongs rather to the domain of Document Retrieval and structured information parsing than to NLP, so we will not go into further details here, though we will likely use data obtained this way in our future work.

We will consider now the NLP task of extracting outcomes and population information from the articles texts.

Later in the course of our project we will collect a corpus for spin detection and annotate it for the types of spin and probable supporting information. We plan to implement machine learning strategies for the task of entity extraction after annotating the corpus; at the current stage we use a rule-based approach to extract a set of manually identified linguistic constructions. We suppose to use these rules as a baseline and for pre-annotating the corpus to assist

human annotators. Our rules are implemented as finite-state automata in Unitex (Paumier, 2016), following the successful reports of previous experience along this approach in Friburger and Maurel (2004); Maurel et al. (2011).

Below we describe the constructions targeted by our rules.

Outcomes

Unlike previous studies, we are not aiming now at extracting outcomes from the phrases reporting results such as an example from (Summerscales et al., 2009):

(1) ***Mortality*** *was higher in the quinine than in the artemether.*

Some of the most common and alarming types of spin are related to not reporting or inadequate reporting of the primary outcome; thus, our main task is not only to identify the outcomes, but to distinguish between primary and secondary ones. We seek thus to detect the phrases stating explicitly the type of an outcome, e.g.:

(2) *The primary outcome was* ***mortality rate****.*

As such phrases may be absent in the article, we consider more general descriptions of objectives and measures assessed to be potentially useful for our task, e.g.:

(3) *Our goal was* ***to compare mortality rate between patients using treatment A and placebo****.*

(4) ***Mortality rate*** *was measures/assessed/...*

Patient population

The most common types of spin concerning patient population include reporting the results for a subgroup instead of the whole population studied (e.g. for a certain gender, age or nationality) or presenting a population broader than the one studied (e.g. generalizing the result achieved for a population with a specified age range to the whole population with the condition examined). Thus, our main goal is to find the descriptions of patients including some basic information such as their age and gender and some more specific information regarding their medical condition. We do not aim at extracting sizes for the whole population or treatment groups as patients may leave a trial for some reasons, so changes in the number of participants may occur and are not to be checked by a spin detection algorithm. We do not aim now at extracting the detailed description of inclusion and exclusion criteria for trial participants as this information is complex and difficult to extract and analyze. We plan to explore the possibility to identify population-related types of spin basing on simple descriptions such as "children aged 8-12 suffering from pneumonia".

We have constructed 9 automata for outcomes and 5 for patient descriptions. Descriptions of primary outcomes are found in 51% of the texts, with more general constructions describing

objectives and measures assessed occur in 91.5% and 94% respectively. Patient descriptions are found in 99.9% of the texts.

Future work

Our next tasks include corpus collection and annotation for further implementation of machine learning techniques. Besides, we will explore approaches for 1) checking the presence of the primary outcome in the abstract results/conclusions; 2) checking if the population mentioned in results/conclusions corresponds to the population studied. These tasks are related but not identical to the task of textual entailment (Kouylekov and Magnini, 2005), which seeks to detect if the meaning of one text can be inferred from another text.

For the task of comparing outcomes, there are two possible directions for achieving our goal. The first way is to extract the outcome from the relevant sentences (such as example (1) above) and compare them to the outcomes extracted from explicit descriptions. A problem that can undermine this approach is the difficulty of extracting outcomes from results and conclusions sentences (Summerscales et al., 2011). The second way is to check the presence of explicitly described outcome in the relevant sentences (as a string, set of words, set of semantically related terms, etc.).

For comparing population descriptions, only the first approach is feasible as the absence of mentions of a population in the results/conclusions does not represent spin.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

We would like to thank Professor Isabelle Boutron from University Paris Descartes for her insight and expertise in the field of our research topic.

References

A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Proceedings. AMIA Symposium*, pages 17–21, 2001.

I. Boutron, S. Dutton, P. Ravaud, and D. Altman. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*, 2010.

I. Boutron, D. Altman, S. Hopewell, F. Vera-Badillo, I. Tannock, and P. Ravaud. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *Journal of Clinical Oncology*, 2014.

B. D. Bruijn, S. Carini, S. Kiritchenko, J. Martin, and I. Sim. Automated information extraction of key trial design elements from clinical trial publications. In *Proceedings of the AMIA Annual Symposium*, 2008.

G. Y.-C. Chung. Towards identifying intervention arms in randomized controlled trials: Extracting coordinating constructions. *Journal of Biomedical Informatics*, 42(5):790 – 800, 2009. ISSN 1532-0464. doi: 10.1016/j.jbi.2008.12.011. URL `http://www.sciencedirect.com/science/article/pii/S1532046408001573`. Biomedical Natural Language Processing.

A. M. Cohen, N. R. Smalheiser, M. S. McDonagh, C. T. Yu, C. E. Adams, J. M. Davis, and P. S. Yu. Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine. In *JAMIA*, 2015.

N. Friburger and D. Maurel. Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, 313(1):93 – 104, 2004. ISSN 0304-3975. doi: 10.1016/j.tcs.2003.10.007. URL `http://www.sciencedirect.com/science/article/pii/S0304397503005371`. Implementation and Application of Automata.

J. M. Glanville, C. Lefebvre, J. N. V. Miles, and J. Camosso-Stefinovic. How to identify randomized controlled trials in medline: ten years on. *Journal of the Medical Library Association : JMLA*, 94 2:130–6, 2006.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11:10–18, 11 2008.

R. Haneef, C. Lazarus, P. Ravaud, A. Yavchitz, and I. Boutron. Interpretation of results of studies evaluating an intervention highlighted in google health news: a cross-sectional study of news. *PLoS ONE*, 2015.

J. P. Higgins and S. Green, editors. *Cochrane handbook for systematic reviews of interventions*. Wiley & Sons Ltd., West Sussex, 2008.

K. Hirohata, N. Okazaki, S. Ananiadou, and M. Ishizuka. Identifying sections in scientific abstracts using conditional random fields. In *Third International Joint Conference on Natural Language Processing: Volume-I*, 2008. URL `https://www.aclweb.org/anthology/I08-1050`.

S. Kim, D. Martinez, L. Cavedon, and L. Yencken. Automatic classification of sentences to support evidence based medicine. *BMC bioinformatics*, 12 Suppl 2:S5, 03 2011. doi: 10.1186/1471-2105-12-S2-S5.

S. Kiritchenko, B. D. Bruijn, S. Carini, J. Martin, and I. Sim. Exact: automatic extraction of clinical trial characteristics from journal publications. *BMC Med Inform Decis Mak.*, 2010.

M. Kouylekov and B. Magnini. Tree edit distance for textual entailment. 01 2005. doi: 10.1075/cilt.292.22kou.

C. Lazarus, R. Haneef, P. Ravaud, and I. Boutron. Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. *BMC Med Res Methodol.*, 2015.

D. Maurel, N. Friburger, J.-Y. Antoine, I. Eshkol, and D. Nouvel. Cascades de transducteurs autour de la reconnaissance des entités nommées. *TAL*, 52, 01 2011.

K. A. McKibbon, N. L. Wilczynski, and R. B. Haynes. Retrieving randomized controlled trials from medline: a comparison of 38 published search filters. *Health information and libraries journal*, 26 3:187–202, 2009.

L. McKnight and P. Srinivasan. Categorization of sentence types in medical abstracts. *AMIA Annual Symposium*, pages 440–4, 2003.

S. Paumier. Unitex 3.1 user manual. `http://unitexgramlab.org/releases/3.1/man/Unitex-GramLab-3.1-usermanual-en.pdf`, 2016.

K. Raja, N. Dasot, B. Tech, P. Goyal, and S. R. Jonnalagadda. Towards evidence-b ased precision medicine : Extracting population information from biomedical text u sing binary classifiers and syntactic patterns. In *AMIA Jt Summits Transl Sci Proc*, 2016.

R. Summerscales, S. Argamon, J. Hupert, and A. Schwartz. Identifying treatments, groups, and outcomes in medical abstracts. In *Proceedings of the Sixth Midwest Computational Linguistics Colloquium (MCLC)*, 2009.

R. L. Summerscales, S. E. Argamon, S. Bai, J. Hupert, and A. Schwartz. Automatic summarization of results from clinical trials. *2011 IEEE International Conference on Bioinformatics and Biomedicine*, pages 372–377, 2011.

R. Xu, Y. Garten, K. S Supekar, A. Das, R. Altman, and A. M Garber. Extracting subject demographic information from abstracts of randomized clinical trial reports. *Studies in health technology and informatics*, 129:550–4, 02 2007. doi: 10.3233/978-1-58603-774-1-550.

Y. Yamamoto and T. Takagi. A sentence classification system for multi biomedical literature summarization. *21st International Conference on Data Engineering Workshops (ICDEW'05)*, pages 1163–1163, 2005.

A. Yavchitz, I. Boutron, A. Bafeta, I. Marroun, P. Charles, J. Mantz, and P. Ravaud. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS Med*, 2012.

A. Yavchitz, P. Ravaud, D. G. Altman, D. Moher, A. Hróbjartsson, T. J. Lasserson, and I. Boutron. A new classification of spin in systematic reviews and meta-analyses was developed and ranked according to the severity. *Journal of clinical epidemiology*, 75:56–65, 2016.

Chapter 2

Annotating Spin in Biomedical Scientific Publications: the case of Randomized Controlled Trials (RCTs). Anna Koroleva, Patrick Paroubek. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 7-12, 2018

Context

A key element in developing NLP algorithms is data; in particular, substantial amounts of high-quality annotated data are required for evaluation of any algorithms, including rule-based ones, and for training supervised machine learning algorithms.

In the previous chapter we identified tasks that are to be included into a spin detection pipeline. For some of these tasks (e.g. detection of sections in abstracts of scientific articles) annotated datasets exist and are available. For some others, no datasets were available when we began working on automatic spin detection. Previous works on spin analysed a limited number of articles, not aiming at annotating data, and thus did not produce annotated datasets large enough for the development of NLP algorithms.

Hence, a major step in our project was collecting and annotating of corpora for spin de-

tection. This chapter reports on our efforts in designing a spin annotation project that we planned to conduct within our project.

Authors' contributions

The work reported in this chapter was conducted by AK under supervision of PP. AK was responsible for data collection, experiment and analysis. AK drafted the manuscript. PP revised the draft critically for important intellectual content.

Abstract

In this paper we report on the collection in the context of the MIROR project of a corpus of biomedical articles for the task of automatic detection of inadequate claims (spin), which to our knowledge has never been addressed before. We present the manual annotation model and its annotation guidelines and describe the planned machine learning experiments and evaluations.

Keywords: spin, annotation scheme, biomedical articles

Introduction

Merriam Webster dictionary defines spin doctor as "*a person (such as a political aide) responsible for ensuring that others interpret an event from a particular point of view*"[1]. In the context of the MIROR[2] project, we address spin in biomedical scientific publications, where it refers to misleading presentation of scientific results, in particular in articles reporting randomized controlled trials (RCTs), an important type of clinical trial. In our case, spin consists in presenting the examined treatment as having greater beneficial effects than the experiments show (Boutron et al., 2010, 2014; Haneef et al., 2015; Yavchitz et al., 2016). Spin in RCTs affects clinical decision-making (Boutron et al., 2014) and results in distorted presentation of research findings in media (Yavchitz et al., 2012; Haneef et al., 2015). We present here the first steps aiming at proposing an algorithm for automatic spin identification in biomedical abstracts, something which to the best of our knowledge has not been attempted before. We present here the construction and annotation of a corpus of medical publication extracted from PubMed Central[3] (PMC) about RCT and describe the annotation model and guidelines.

[1] https://www.merriam-webster.com/dictionary
[2] http://miror-ejd.eu/
[3] https://www.ncbi.nlm.nih.gov/pmc/

On spin types

From previous research on spin classification (Boutron et al., 2010; Lazarus et al., 2015; Yavchitz et al., 2016), we can outline three main types and their subtypes of spin in RCT reports:

1. misleading reporting of study results:
    - selective reporting of outcomes (omission of the primary outcome; focus on statistically significant results different from the main outcome);
    - occulting adverse events;
    - misleading reporting of study design;
    - linguistic spin (beautifying formulations);
    - discarding limitations;
    - selective citation of other studies
2. inadequate interpretation of the results:
    - claiming a beneficial effect of the intervention despite statistically non-significant results;
    - claiming an equivalent effect of the interventions for statistically non-significant results;
    - claiming that the treatment is safe for statistically non-significant safety outcomes;
    - concluding a beneficial effect despite no comparison test performed;
    - interpretation of the results according to statistical significance instead of clinical relevance;
3. inadequate extrapolation:
    - inadequate extrapolation from the population, interventions or outcome actually assessed in the study to a larger population, different interventions or outcomes;
    - inadequate implications for clinical practice.

Example of spin putting focus on secondary result ("*improved PFS and response for treatment*") instead of the main result, object of the experiment ("*survival rate*"):

In the rest of this paper, we present our linguistic model of spin (section 2), the annotation scheme (section 3), the the annotation guidelines (section 4), conclusions and plans for future work (section 5).

"*This study demonstrates improved PFS and response for the treatment A compared with comparator B, although this* ***did not result in improved survival***".

Figure 2-1: Example of spin (focus on secondary result)

Model of spin

To the best of our knowledge, this is the first attempt at addressing the analysis of spin in the biomedical literature from a Natural Language Processing point of view. Spin detection overlaps partially with previous works in NLP, in particular objectivity/subjectivity identification (Wiebe and Riloff, 2005), sentiment analysis (Pak, 2012), fact checking (Nakashole and Mitchell, 2014) or deception detection (Hancock et al., 2010; Litvinova et al., 2017); a point to note is that these works address texts of general domain while we deal with spin in biomedical texts. We regard spin detection as a task most closely related to deception detection. Deception is defined as a deliberate act of communicating information that the speaker/author believes to be false, with the intention to induce listeners/readers to believe a distorted presentation of the topic. Strictly speaking, spin is not necessarily a form of deception, as the intention is difficult to establish most of the time, e.g., spin in abstracts may be conditioned by limited space; by author's wish to report the results that he/she perceives to be most important; by unclear/absent reporting guidelines; by lack of training etc. However, spin is similar to deception for what concerns its impact and the method required to detect it from textual content only (Mihalcea and Strapparava, 2009).

Spin can be considered as the most serious form of incomplete or incoherent reporting of trial data and results (omission or inconsistent presentation of information). We aim at creating a general model that would be able to represent the information about a trial data and all possible realizations of spin in reporting.

For trial data, we choose to follow the information structure accepted in trial registries (official catalogues for registering clinical trials, containing in a structured form trial data provided by the investigators who carry out a trial).

Trial registries may slightly vary regarding the level of detalisation used for information presentation, so we reviewed several registries (ClinicalTrials.gov[4], ISRCTN[5], plus some national registries) and generalized the categories used. We compiled the following list of data describing a trial:

- Information about interventions: intervention name, dosage, administration schedule, treatment duration;

[4]https://clinicaltrials.gov/

[5]https://www.isrctn.com/ (International Standard Randomised Controlled Trial Number Register)

Start Date ICMJE	September 2006
Primary Completion Date	*Not Provided*
Current Primary Outcome Measures ICMJE **(submitted: June 8, 2007)**	Sleep quality assessed by the Leed Sleep Evaluation Questionnaire, measured after one month of treatment
Original Primary Outcome Measures ICMJE	*Same as current*
Change History	No Changes Posted
Current Secondary Outcome Measures ICMJE **(submitted: June 8, 2007)**	• Sleep quality assessed by the Leed Sleep Evaluation - Questionnaire, measured 15 days after treatment withdrawal • Sleep quality assessed by a Sleep diary completed each day during all the study period by the participants • Sleep efficiency measured by ambulatory actigraphy (2 period) • Evolution of melatonin/6sulfatoxymelatonin ratio (before and after treatment) • Clinical General Impression of the clinician, before and after treatment. • Safety of the treatment (adverse event reporting)

Figure 2-2: Excerpt from an RCT description queried on ClinicalTrials.gov with the keywords: '*RCT insomnia France*'

- Information about participants: age, gender, health condition, health type, nationality/ethnicity, recruitment country/region; information regarding intervention assigned; other information. Can be represented in a form of a list of inclusion and exclusion criteria, that can contain all of the above information;
- Trial methods / trial design: allocation concealment, allocation type, blinding, sample sizes for groups examined, study type, study subtype, trial phase, statistical tests used;
- Trial objectives / outcomes (with their methods of measurement and associated time points);
- Data about registration: registration number, registration time;
- Financing: sponsors;
- Hypothesis, hypothesis type;
- General information: medical domain;
- Summary.

We also introduced some other categories that are not typically present in registries but that are relevant to trial description: limitations and reported statistical measures.

In order to be able to capture instances of spin, we further need to reflect the following phenomena:

1. Incomplete reporting, which can take many forms, but we are most interested in omission of information that is normally supposed to be present in a well-reported abstract, such as:

- clear definition of the primary outcome;
- results for primary outcome;
- results for non-significant secondary outcomes;
- information about adverse events (their absence should be stated explicitly).

Omission of some other types of information (design, methods, statistical tests used, etc.) should not be considered as spin but rather as incomplete reporting acting as 'spin facilitator' hindering fact checking.

2. Incoherent reporting:

   - primary outcome described in the trial registry differs from the primary outcome described in the text;
   - patient population reported in the abstract does not correspond to the population studied in its qualitative characteristics (age, gender, etc.);
   - reported results do not correspond to trial design;
   - the compared treatments are reported to be similar when the design does not allow to conclude on similarity (i.e. the trial is not a 'non-inferiority' of 'equivalence' trial);
   - within-group comparison reported when the trial objective was not to examine changes within groups (i.e. the trial is not a 'before-and-after trial');
   - focus on significant secondary outcomes instead of primary outcome;
   - positive conclusions are made (efficacy stated, treatment recommended for use) when the primary outcome is not significant.

   Incoherence or incompleteness of reporting can be established by checking the completeness of the abstract, discrepancies between abstract and article body or between trial registry entry (if available) and article content. We thus work with two types of documents: articles and registry entries. For articles, the model comprises information about its structure: its division into title, abstract and body text, for registries we rely on their internal structure, in general a tabular form holding short pieces of text or data.

Annotation scheme

We proposed a description of an algorithm of spin detection elsewhere (Koroleva and Paroubek, 2017b). The main steps are the following:

- dividing a given article into title, abstract and body text; finding results and conclusions within the abstract;
- identifying positive evaluation of the studied treatment in results/conclusions of the abstract;
- extracting elements of trial data relevant to spin assessment, such as outcomes, patient population, statistical significance of results;
- extracting relation between elements of trial data, such as an outcome and its statistical significance;
- extracting specific constructions possibly related to spin (see below);
- final assessment of spin: checking if the information in the results and conclusions of the abstract corresponds to the extracted trial data, for example, if the pre-defined outcomes are reported correctly or if the positive evaluation of the treatment is supported by statistically significant results.

We propose here an annotation scheme comprising the information elements relevant for the future algorithm. Our annotation scheme is implemented in XML and includes several levels of information:

1. Document type (article/registry entry).
2. Structural information (for articles). For this annotation level we adopt the existing annotation scheme used in PubMed1, simplified for our needs. Our scheme includes journal name, article title, authors list, abstract, body text, bibliography. Within abstracts, Results and Conclusions sections are marked.
3. Elements describing the trial (what was studied and how: compared interventions, outcomes, population studied, statistical measures used, etc.): we introduce a separate tag for each type of trial data. This decision is motivated by the fact that we need specific sets of attributes for different types of trial data, and we need to introduce particular relations for specific types of trial information. As outcome is the most important type of trial data for spin detection, for outcomes (or trial objectives) we use several tags that are needed to distinguish between different specific constructions:

   The type of an outcome can have three different type attribute values: Prim (primary) / Sec (secondary) / None (undefined). Outcome has also an attribute 'status' which can have two values: Declared when it is explicitly stated in the text to be an outcome (e.g.

The <Prol>primary outcome measure will be< /Prol> <Out type="Prim" Status="Declared">**QoL**< /Out>, assessed with the ALS Assessment Questionnaire...

Figure 2-3: Example of annotation for a primary (attribute type is Prim) outcome (Out) explicitly declared, with the annotation of its linguistic marker (Prol).

Fig 2-3), which is its value by default and Reported, when the outcome is only reported in results or conclusion section without referring explicitly to its nature.

Our <Prol>secondary aim is< /Prol> <Aim type="Sec">**to describe the costs**< /Aim> associated with RESERVE-DSD.

Figure 2-4: The AIM is the objective of the trial.

4. Relations between elements of trial data: relations that link a pair of elements that describe different features of a single concept, e.g. an outcome with its method of measurement or with its time points, or an intervention to its dosage, administration schedule, etc.

5. Particular constructions of interest:

   - Positive evaluations of treatment (positive results regarding the treatment);
   - Statements of similarity between treatments regarding their efficacy or safety;
   - Within-group comparisons (statements of changes that occur within a group receiving the studied treatment, without comparing it to the group receiving the control treatment);
   - Recommendations to use treatment.

These include: i) an analysis which shows that the ethnic difference in performance in this 2006/7 <Subj>**cohort of Year 3 students**< /Subj> was similar in size to that in <Subj>**previous cohorts on the course**< /Subj>.

Figure 2-5: Example of a similarity statement. Subj – trial subjects.

A problem that arises with this type of information consists in deciding which fragment of text should be annotated. Normally these constructions comprise a whole proposition, but we can as well highlight some words that are the most 'representative' of the meaning of each construction. We choose to annotate the smallest possible fragments that are indicators of relevant constructions. The motivation behind this decision is the need to

make the annotation as clear and simple as possible for annotators, and the fact that, having annotation on word level, we can easily expand it to the sentence level.

6. Annotation for spin: annotation level that is meant to capture all the cases of incoherence and incompleteness regarding the types of information enumerated above. This type of annotation resembles most to a well-known task of relation annotation, but here the most important is not to capture relation that holds between two text fragments, but to mark the cases when there is no relation when we expect it to exist. For example, a relation between a declared primary outcome in article text or protocol and a corresponding reported outcome in abstract means no spin, but a declared primary outcome with no related reported outcome is a case of spin. A similarity statement is not spin if the trial was of equivalence type, but it is a spin if there is no text fragment indicating that the trial belongs to equivalence trials.

   To annotate this information, we follow the system accepted in TimeML (Pustejovsky et al., 2003) annotation for relations: we introduce empty tags that contain reference to IDs of fragments that are linked in case of good reporting; in case of incoherence/ incompleteness, the tag contains ID of the present text fragment. These tags have an attribute 'spin' that is set to 'yes' or 'no'.

   Another form of actual spin or of 'spin facilitator' is omitting some information about methods, design or results in the abstract, e.g. not stating clearly the primary outcome. For this type of omission, we do not need to refer to an ID, we only need an empty tag to mark which type of information is omitted in abstract.

Thus, the annotation for spin is done on the lowest level: as a relation between text fragments. We can then calculate the value of 'spin' attribute for the whole text.

Figure 2-6 shows an example of text with spin (the example comes from the appendix of Boutron et al. (2014) and the process of assessment of outcome-related spin.

The first step in annotating this text would be to annotate all outcomes reported in the abstract (IDs 1 – 5) and the declared primary outcome (ID 6). The following steps to fully annotate all types of spin related to primary outcome would be the following:

1. Check and mark if there is a definition of the primary outcome in the abstract. Here it is absent (full text if abstract omitted for the sake of space) – we conclude incomplete reporting.

2. Check and mark if the declared primary outcome is present among the reported outcomes. Here it can be considered to correspond to the outcomes 3 and 5 – we conclude correct reporting.

<Abstract>Abstract
< ... > <Res>Results
< ... > <Out ID="1" Type="None" Status="Reported">**The International Union Against Cancer R0 resection rate**< /Out> was 81.9% after treatment A as compared with 66.7% with surgery alone (P = .036). The surgery-only group had more <Out ID="2" Type="None" Status="Reported">**lymph node metastases**< /Out> than the treatment A group (76.5% v 61.4%; P = .018). < ... > A <Out ID="3" Type="None" Status="Reported"> *survival* < /Out> benefit could not be shown (hazard ratio, 0.84; 95% CI, 0.52 to 1.35; P = .466). < /Res>
<Concl>Conclusion
This trial showed a significantly increased <Out ID="4" Type="None" Status="Reported">**R0 resection rate**< /Out> but failed to demonstrate a <Out ID="5" Type="None" Status="Reported"> *survival* < /Out> benefit. < /Concl>
<BodyText>< ... >
The primary end point of this trial was <Out ID="6" Type="Prim" Status="Declared">**overall survival**< /Out>. < ... >< /BodyText>

Figure 2-6: Example of annotation of spin for an abstract

3. Check and mark if the primary outcome is presented correctly according to its importance: it should be presented in the first place without regard to significance of results; there should be no focus on other outcomes. In this abstract, the insignificant primary outcome is presented after significant secondary ones – we conclude incoherent reporting (focus on secondary outcomes).

Annotation guidelines

We plan to combine automatic annotation as first stage, and manual annotation to correct and complete the annotation. We do not aim at manually annotating all the types of information. Most of the trial data not directly relevant to spin detection will be marked automatically only in trial registry entries, where information is highly structured. We do not thus cover them in the annotation guidelines.

We described our algorithms of automatic pre-annotation in our previous works (Koroleva and Paroubek, 2017b,a). These algorithms aim at extracting/annotating the following:

- text structure: separating results and conclusions sections in abstracts;
- various constructions defining trial outcomes, with special attention to the primary one, for example:

1. The primary outcome is <Out Type="Prim" Status="Declared"> emotional distress (symptoms of depression, anxiety, diabetes-specific stress, and general perceived stress) < /Out>.
2. This project has one primary aim: to measure <Out Type="Prim" Status= "Declared" > the impact of continuity of midwifery care < /Out> compared to routine care on restricting excessive gestational weight gain in obese women.
3. Sample size A power calculation was carried out for the primary outcome (<Out Type="Prim" Status="Declared"> health related quality of life measured on the York version of the SF-12< /Out>).

- comparative constructions that are often used to report the trial results. These constructions usually include some of the following elements: compared patient groups, compared treatments, outcomes that serve as basis for comparison. We mainly focus on extracting outcomes:

  1. <Subj> Patients with TC asthma < /Subj> has significantly higher<Out Type= "None" Status="Reported"> AQLQ scores < /Out> compared to those with NTC asthma.
  2. Muscarinic agonists appear to reduce <Out Type="None" Status= "Reported"> the potency of beta-agonist bronchodilation < /Out>, possibly through an effect on adenylyl cyclase 17.
  3. <Out Type="None" Status="Reported"> Levels of hs-CRP < /Out> increased modestly in the ABC / 3TC arm compared with the TDF / FTC the arm.

- Description of studied population:

  1. We studied <Subj> <Aim Type="None"> 19 consecutive unselected patients who met the ARDS criteria of the American European Consensus Conference 21. < /Aim> < /Subj>
  2. A total of <Subj> 32 patients aged 12 to 17 years with severe, active and refractory JoAS < /Subj> were enrolled in a multicenter, randomized, double-blind, placebo-controlled parallel study of 12 weeks.

These annotations, although not perfectly correct and complete, are hoped to reduce workload for annotators: in case pre-annotation is completely correct or completely erroneous, the annotators will simply need to validate/reject it, reducing the number of cases requiring manual annotation.

The current pre-annotated corpus includes 3938 articles on randomized controlled trials in various medical domains, extracted from PubMed Central. This corpus will serve as basis for manual annotation.

We will split manual annotation into several stages that would differ regarding their complexity and thus the skills required from the annotators.

Some of the tasks are relatively easy and can be done by annotators who do not have special knowledge in medical domain. We consider that the tasks that fall into this group are: explicit descriptions of outcomes, mentions of patient population, statistical measures (p-value), confidence intervals.

Some other types of information require some special knowledge of medical domain as understanding of medical terms is needed to correctly interpret the meaning of sentences and categorize text fragments as representing a certain type of trial data/construction. Following tasks fall into this category: reported outcomes, similarity statements, within-group comparisons, evaluations related to treatment.

The final task of spin annotation (i.e. marking parts of the text that represent coherent and complete reporting for chosen concepts, and marking cases when there is incoherence/incompleteness) is an even more difficult task. The concept of spin in biomedical domain is not completely formally defined yet, experts in the domain often disagree on classifying a certain phenomenon as spin or not. For example, some experts regard absence of explicit definition of the primary outcome in the abstract of an article as definite spin, while others consider it to represent incomplete reporting but less important than spin. Besides, mismatch between information in the abstract and in the article (e.g. change of outcomes studied and reported) is not spin if it has valid scientific justification, which should be provided in the article. Extraction of such justifications and assessment of their validity would be necessary to conclude on absence or presence of spin, but it falls outside scope of our work.

Thus, there are several difficulties that we should take into account when developing annotation guidelines:

1. Some of the tasks require at least some level of special medical knowledge, so it is likely that the annotators will not be linguists and will not have experience in corpus creation/annotation. This fact should be taken into account when choosing terminology (no specific linguistic terms) and when defining the task (e.g., be clear about annotating coordinated elements as separate elements and not one element).

2. Choice of the annotation tool to be used should take into account the complexity of the task but also the involvement of non-linguists in annotation process. From the point of view of functionality, the tool should at the very least be able to capture relations,

potentially embedded. This requirement makes tools not allowing relation annotation, such as WebAnnotator (Tannier, 2012), not appropriate. After testing and comparing several tools, we chose the Glozz platform (Widlöcher and Mathet, 2012) as the one that best corresponds to the needs of the task of full linguistic annotation of spin. Glozz is a flexible and powerful tool that allows to annotate units (text fragments), their relations and schemes (which can be seen as higher-level relations that can include one or more units, relations or other schemes) which covers all possible instances of incompleteness or incoherence in reporting.

However, demonstration of text annotation with Glozz to a medical expert showed that it does not meet the requirements of non-linguist annotators: ease of installation of the tool, amount of time needed for training for a person without previous experience in corpus annotation, complexity of guidelines describing the task. Consequently, we decided to replace the task of linguistic annotation of texts by a set of simpler tasks (in the form of questions) such as the following:

- validation/correction of primary outcomes found at the pre-annotation stage;
- validation/correction of reported outcomes found at the pre-annotation stage;
- establishing if two given (extracted at previous stages) outcomes refer to the same concept;
- identification of similarity statements in the Results and Conclusions of abstracts;
- identification of within-group comparisons in the Results and Conclusions of abstracts;
- identification of other positive evaluation of the studied treatment in the Results and Conclusions of abstracts.

We plan to use a web-based survey tool (such as LimeSurvey[6]) to create questionnaires containing these questions, generated on the basis of pre-annotation. Using survey tools for corpus annotation is not typical. Our decision is motivated by several reasons: survey tools are usually available online and thus do not require any complex installation procedures (survey participants can access the survey simply by following a link received by email); survey tools are widely used in medical community and are familiar to the community. This fact reduces time needed for annotators to learn how to use the tool. Besides, breaking the task into simple questions, independent one from another, allows to include into each question a brief guideline on how to answer, thus in most cases

[6]https://www.limesurvey.org/

annotators will not need to refer to an extensive external annotation guide. Answering simple questions is also likely to cause fewer discrepancies between annotators than full annotation of spin.

3. In case of full linguistic annotation of spin, we should clearly define which pieces of text to annotate. We anticipate some difficulties in cases when elements of trial data get embedded one into another. The guidelines should explain whether to annotate these elements as embedded or as two separate instances linked by a certain type of relation (e.g. outcome and its method of measurement).

4. Given the complexity of the task, we need to clarify the definition of what should be considered to be spin. For this, we need to strictly define the types of spin that we focus on, describe in detail which pieces of information are relevant to these types of spin. Taking into account lack of agreement between experts in detailed definition of spin, for our current annotation project we decided to avoid using the notion "spin" and focus on tasks that are relatively simpler and clearer, such as: annotating outcomes; marking if pairs of extracted outcomes refer to the same concept; annotating specific constructions of interest, such as similarity statements or within-group comparisons. This information would allow to estimate with a certain probability that an article does or does not contain spin, but the final decision is left to the human readers of the article.

5. The task of developing guidelines must be fulfilled in close collaboration with experts in medical domain and in the domain of spin in medical texts, in order to verify that all the definitions regarding medical concepts and spin are correct.

Conclusions and future work

In this paper we described our approach to creation of a corpus of biomedical articles annotated for spin (distorted reporting) and its supporting information. We briefly outlined the proposed algorithm of spin detection and summarized our work on automatic pre-annotation of the corpus. Consequently, we described the annotation scheme that we developed for spin annotation. We discussed the process of creating the annotation guidelines, provided some thoughts as for choice of annotation tool and outlined expected challenges.

Our future tasks include running a pilot survey to validate usability of survey format for our task and evaluate the adequacy and clarity of the questions for annotators. Consequently we will proceed to a full-scale survey project.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

References

I. Boutron, S. Dutton, P. Ravaud, and D. Altman. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*, 2010.

I. Boutron, D. Altman, S. Hopewell, F. Vera-Badillo, I. Tannock, and P. Ravaud. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *Journal of Clinical Oncology*, 2014.

J. T. Hancock, D. I. Beaver, C. K. Chung, J. Frazee, J. W. Pennebaker, A. Graesser, and Z. Cai. Social language processing: A framework for analyzing the communication of terrorists and authoritarian regimes. *Behavioral Sciences of Terrorism and Political Aggression*, 2 (2):108–132, 2010. doi: 10.1080/19434471003597415. URL `https://doi.org/10.1080/19434471003597415`.

R. Haneef, C. Lazarus, P. Ravaud, A. Yavchitz, and I. Boutron. Interpretation of results of studies evaluating an intervention highlighted in google health news: a cross-sectional study of news. *PLoS ONE*, 2015.

A. Koroleva and P. Paroubek. On the contribution of specific entity detection and comparative construction to automatic spin detection in biomedical scientific publications. In *Proceedings of The Second Workshop on Processing Emotions, Decisions and Opinions at The 8th Language and Technology Conference*, 2017a.

A. Koroleva and P. Paroubek. Automatic detection of inadequate claims in biomedical articles: first steps. In *Proceedings of Workshop on Curative Power of MEdical Data*, 2017b.

C. Lazarus, R. Haneef, P. Ravaud, and I. Boutron. Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. *BMC Med Res Methodol.*, 2015.

O. Litvinova, P. Seredin, T. Litvinova, and J. Lyell. Deception detection in Russian texts. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 43–52, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-4005.

R. Mihalcea and C. Strapparava. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312, Suntec, Singapore, Aug. 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P09-2078.

N. Nakashole and T. M. Mitchell. Language-aware truth assessment of fact candidates. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1019, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1095. URL https://www.aclweb.org/anthology/P14-1095.

A. Pak. Automatic, adaptive, and applicative sentiment analysis. 2012.

J. Pustejovsky, J. Castano, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, and G. Katz. Timeml: Robust specification of event and temporal expressions in text. In *Fifth International Workshop on Computational Semantics (IWCS-5*, 2003.

X. Tannier. WebAnnotator, an annotation tool for web pages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 316–319, Istanbul, Turkey, May 2012. European Languages Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/148_Paper.pdf.

A. Widlöcher and Y. Mathet. The glozz platform: A corpus annotation and mining tool. In *Proceedings of the 2012 ACM Symposium on Document Engineering*, DocEng '12, pages 171–180, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1116-8. doi: 10.1145/2361354.2361394. URL http://doi.acm.org/10.1145/2361354.2361394.

J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'05, pages 486–497, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3-540-24523-5, 978-3-540-24523-0. doi: 10.1007/978-3-540-30586-6_53. URL http://dx.doi.org/10.1007/978-3-540-30586-6_53.

A. Yavchitz, I. Boutron, A. Bafeta, I. Marroun, P. Charles, J. Mantz, and P. Ravaud. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS Med*, 2012.

A. Yavchitz, P. Ravaud, D. G. Altman, D. Moher, A. Hróbjartsson, T. J. Lasserson, and I. Boutron. A new classification of spin in systematic reviews and meta-analyses was developed and ranked according to the severity. *Journal of clinical epidemiology*, 75:56–65, 2016.

Part II

Algorithms

Chapter 3

Extracting outcomes from articles reporting randomized controlled trials using pre-trained deep language representations. Anna Koroleva, Sanjay Kamath, Patrick Paroubek. Submitted

Context

Spin is often related to incorrect reporting of trial outcomes, such as outcome switching (unjustified change of pre-defined outcomes of a trial). Outcomes are thus one of the most important information elements for spin detection. Extracting trial outcomes is a key task in a spin detection pipeline. Chapter 1 described our first experiments on this task using simple local grammars. The following chapter reports on the further work on outcome extraction.

In this chapter, we reviewed the state of the art, including works on sentence classification to extract sentences describing outcomes and works on outcome extraction as a sequence labelling task. We defined two types of outcomes (declared and reported), according to the context in which they are mentioned, relevant for spin detection.

The previous chapter reported on our annotation efforts. Initially we planned to run an annotation project with several annotators who should be experts in reporting of clinical trials or in the clinical research in general. However, recruiting several annotators with sufficient level of expertise and training them to perform linguistic annotation proved to be infeasible within the given time frame. As an annotated corpus was nevertheless required for our further experiments, we ran a small-scale annotation project: a single annotator (AK) annotated

corpora for each of the core tasks of the spin detection pipeline. The volume of the annotated data is limited but proved to be sufficient to train machine learning algorithms. Annotation by a single annotator has certain disadvantages, as it is difficult to assess the quality of annotation, but in the context of our project it had the advantage of allowing to obtain training data within a reasonable time frame.

In the following chapter, we report on the collection and annotation of corpora for these two types of outcomes and on our experiments on building rule-based and machine-learning algorithms. The best performing algorithms for declared and reported outcomes detection were included into our spin detection pipeline.

Authors' contributions

AK designed the study described in this chapter and interpreted the data. AK collected and annotated the corpus. AK and SK conducted the experiments, supervised by PP. AK drafted the manuscript. SK and PP revised the draft critically for important intellectual content.

Abstract

Objective: Outcomes are the variables monitored during clinical trials to assess the impact of the intervention studied on the subjects' health. Automatic extraction of trial outcomes is essential for automating systematic review process and for checking the completeness and coherence of reporting to avoid bias and spin. In this work, we provide an overview of the state-of-the art for outcome extraction, introduce a new freely available corpus with annotations for two types of outcomes —declared (primary) and reported —and present a deep learning approach to outcome extraction.

Dataset: We manually annotated a corpus of 2,000 sentences with declared (primary) outcomes and 1,940 sentences with reported outcomes.

Methods: We used deep neural word embeddings derived from the publicly available BERT (Bidirectional Encoder Representations from Transformers) pre-trained language representations to extract trial outcomes from the section defining the primary outcome and the section reporting reporting the results for an outcome. We compared a simple fine-tuning approach and an approach using CRF and Bi-LSTM. We assessed the performance of several pre-trained language models: general domain (BERT), biomedical (BioBERT) and scientific (SciBERT).

Results: Our algorithm achieved the token-level F-measure of 88.52% for primary outcomes and 79.42% for reported outcomes.

Conclusion: Fine-tuning of language models pre-trained on large domain-specific corpora show operational performance for automatic outcome extraction.

Keywords: Natural Language Processing, Randomized Controlled Trials, Outcome extraction, Deep Neural Networks, Pre-trained Language Representations

Introduction

Outcomes of clinical trials are the dependent variables monitored during a trial in order to establish how they are influenced by independent variables such as the intervention taken, dosage, or patient characteristics. Outcomes are a key element of trial design, reflecting its main goal and determining the trial's statistical power and sample size.

Previous works have shown that outcome extraction is a difficult task because of the diversity of outcome mentions and contexts in which they occur. No common textual markers exist (e.g. capitalization, numerical symbols, cue phrase). Recent research proved that the use of deep language representations pre-trained on large corpora, such as BERT (Devlin et al., 2018), outperforms the state-of-the-art results for several natural language processing tasks, including entity extraction. Pre-training on large domain-specific data can further improve the results (Lee et al., 2019; Beltagy et al., 2019).

We propose a deep learning approach, using language representations pre-trained on general domain corpora and on domain-specific datasets to extract trial outcomes. We report on creating a publicly available annotated corpus for outcome extraction.

Definitions

There are substantial discrepancies in the use of the words "outcome", "endpoint", "outcome measure" etc., that we describe in detail elsewhere. In brief, there is no agreement between researchers as for the differences in the meaning and usage of these terms, in practice they are often considered to be synonyms. In our work, we follow the common practice and do not distinguish between these notions. We prefer to use the term "outcome".

Following the accepted usage[1], we define an outcome as *a variable (or measure, or parameter) monitored during a clinical trial*. Outcome in this sense is a type of *entity*, as it is understood by the standard entity recognition task. Our definition differs from that given by Demner-Fushman et al. (2006) who defined an outcome as "*sentence(s) that best summarizes the consequences of an intervention*". In our definition an outcome is usually shorter than a sentence, and it does not refer to trial results ("consequences of an intervention"): the results are the *values* of outcomes. Our definition is in line with other works on outcome extraction (see the Related Works section), as many applications, such as summarization of trial results, require extracting outcomes on the entity level to allow for further analysis of data for each individual outcome, which is not possible if extracting only sentences containing outcomes (cf. Blake and Kehm (2019)).

[1] e.g. https://rethinkingclinicaltrials.org/chapters/design/choosing-specifying-end-points-outcomes/choosing-and-specifying-endpoints-and-outcomes-introduction/

We introduce here the definitions for two types of outcome mentions that are important for our work: declared and reported outcomes.

Declared outcomes are the mentions of outcomes that occur in contexts that explicitly state which variables were measured in a trial, e.g. (outcomes are in bold):

The primary outcome of this study was ***health-related quality of life****.*

Secondary outcomes included ***changes in the 6-minute walk distance (6MWD)*** *and* ***adverse events****.*

In our study, we were most interested in ***changes in PHQ-9 scores after the 12-week trial****.*

Declared outcomes can be further classified according to their importance as stated by the authors (primary, secondary, tertiary, or undefined).

Reported outcomes are the mentions of outcomes that occur in contexts that report the results for the outcomes, e.g. (outcomes are in bold):

The HRQoL *was higher in the experimental group.*

The mean incremental QALY *of intervention was 0.132 (95% CI: 0.104—0.286).*

Applications

Extraction of trial outcomes is an important part of systematic review process (Jonnalagadda et al., 2015), clinical question answering (Demner-Fushman and Lin, 2007), assessment of an article for distorted reporting practices such as bias (Higgins et al., 2011), outcome switching (Goldacre et al., 2016) and spin (Boutron et al., 2010). For us, the main application of interest is spin detection.

In general, spin is defined as presenting research results as being more positive than the experiments proved. In particular, in randomized controlled trials (RCTs) assessing a new intervention, spin consists in exaggerating the beneficial effects (efficacy and/or safety) of the studied intervention. As RCTs are the main source of information for Evidence-Based Medicine, spin in RCTs presents a serious threat to the quality of healthcare. The presence of spin makes clinicians overestimate the effects of the treatment in question (Boutron et al., 2014), and provokes spin in health news and press releases (Haneef et al., 2015; Yavchitz et al., 2012), which can affect public expectations regarding the treatment.

One of the most common forms of spin is *selective reporting of trial outcomes* – reporting only the outcomes that prove the hypothesis of the authors. To automatically detect this form of spin, declared and reported trial outcomes need to be extracted, and declared outcomes must be compared to the reported outcomes to check for mismatches: declared outcomes that are not reported, or reported outcomes that were not declared.

Our current work presents the first step of the algorithm of selective outcome reporting detection and deals with the extraction of declared and reported outcomes. The second step consists in assessing semantic similarity between pairs of declared and reported outcomes and will be presented elsewhere.

Related work

As the volume of published biomedical articles grows exponentially (Khare et al., 2014), manual extraction of clinical trial information becomes infeasible. Several works addressed the extraction of outcome-related information for facilitating systematic reviews or supporting clinical question-answering systems.

A number of works addressed extraction of information on clinical trials using the PICO - Patient/Problem, Intervention, Comparison, Outcome - framework (Richardson et al., 1995) or its extensions. The majority of works using the PICO framework treat the task as sentence classification (Demner-Fushman and Lin, 2005; Boudin et al., 2010; Huang et al., 2011; Kim et al., 2011; gang Cao et al., 2010; Verbeke et al., 2012; Huang et al., 2013; Hassanzadeh et al., 2014; Jin and Szolovits, 2018). F-measure for outcome sentence extraction varies between 54% and 88% for different methods (see the systematic review Jonnalagadda et al. (2015)).

Demner-Fushman and Lin (2007); Demner-Fushman et al. (2006) also treated the task of outcome extraction as text classification. The authors trained several classifiers on a dataset of 633 MEDLINE citations. Naive Bayes classifier outperformed linear SVM and decision-tree classifier. The accuracy of outcome sentence identification ranged from 88% to 93%.

However, for some tasks (including spin detection), identification of relevant sentences is not enough and extracting outcomes at the entity level is required. This task has been addressed by fewer works than the PICO classification. It is important to distinguish the works addressing the extraction of declared (primary and secondary) outcomes from those targeting reported outcomes.

Bruijn et al. (2008) aimed at extracting declared (primary and secondary) outcomes and their time points, along with other elements of trial design. They point out that the outcomes of a trial can be poorly defined by referring to "main outcomes" instead of primary and secondary ones. The authors also note that it is necessary to analyse the whole article, not only the abstract, e.g. to find secondary outcomes. The system uses a two-step approach: first, a classifier is applied to identify sentences containing a given type of information; second, regular expression rules are used to find text fragments corresponding to the target information. The dataset used in this work consists of 88 randomly selected full-text articles from five medical journals. For primary and secondary outcomes, only the first step (sentence classification) was

implemented. Performance for identification of sentences containing outcomes is reported to be lower that for the other elements. For the primary outcomes, the sentence classification reaches precision of 87% and recall of 90%; for secondary outcomes, precision was 57% and recall was 90%.

In their following work (Kiritchenko et al., 2010), the authors further develop their approach and add rules for extracting text fragments for primary and secondary outcomes. This work used a different dataset: the initial corpus consisted of 78 manually annotated articles from five clinical journals that were considered to be representative of general medicine, to which 54 articles from a wider selection of journals were added, resulting in a final training set of 132 articles from 22 clinical journals. The test set contained 50 full-text articles reporting RCTs from 25 journals. The results were assessed at the sentence and fragment levels. The sentence classification performance for outcomes is as follows: precision was 66%, recall was 69% for primary outcomes; precision was 69% and recall 79% for secondary outcomes. For fragment extraction, the authors report for primary outcomes a precision and recall of 97% for both overlapping and exact matches; for secondary outcomes, precision and recall for exact matches are 93% and 88% respectively, and for overlapping matches, both precision and recall are 100%.

Summerscales et al. (2009) addressed the task of identifying treatments, patient groups and reported outcomes in abstracts of medical articles. The authors created a corpus of 100 abstracts of articles published in the BMJ[2], extracted from PubMed[3]. The corpus of 1,344 sentences contained 1,131 outcomes, 494 out of which were unique. Outcomes vary in length from 1 to 14 tokens (mean = 3.6). The examples of outcomes given in the article are noun phrases, but the authors did not specify whether they annotated only noun phrases or included other syntactic constituents (e.g. verb phrases, adjectives). The authors note that the boundaries of entities are often ambiguous and annotating each variant is not optimal; thus, they suggest to evaluate both exact and partial matches. The authors trained a Conditional Random Field (CRF) classifier to label each word, using features such as the word form, its POS tag, corresponding Medical Subject Heading ID, its semantic tag(s) (anatomy, time, disease, symptom, drug, procedure and measurement terms, assigned using lists of terms), the title of the enclosing section, and four words to the left and right of the word with their POS and semantic tags. The token-levels results for outcomes are: precision 75%, recall 62%, and F-measure 68%.

In their following work (Summerscales et al., 2011), the authors enlarge their dataset to 263 abstracts of BMJ articles. A first-order linear-chain CRF classifier on this set yielded a

[2]https://www.bmj.com/
[3]http://www.ncbi.nlm.nih.gov/pubmed/

precision of 56%, recall of 34%, and F-measure of 42% for outcome extraction.

Blake and Lucic (2015) aimed at extracting noun phrases for three items in comparative sentences: two compared entities (the agent and the object) and the ground for comparison (endpoint, or outcome). The dataset for this work included the sentences containing all the three items (agent, object and endpoint), selected from over 2 million sentences from full-text medical articles. 100 sentences with 656 noun phrases constituted the training set. First the algorithm finds comparative sentences with the use of a set of adjectives and lexico-syntactic patterns. Then two classifiers - SVM and Generalized Linear model (GLM) —are used to predict the roles (agent, object, endpoint) of noun phrases. SVM showed better results that GLM on the training set (for endpoint, precision=67%, recall=94% and F-measure=78%). However, on the test set the results were significantly lower: SVM achieved precision of 42%, recall of 64% and F-measure 51% for endpoint detection. The performance was evaluated separately on shorter sentences (up to 30 words), where it was higher than on longer sentences.

The following work (Lucic and Blake, 2016) used the information whether the head noun of the candidate noun phrase denotes an amount or a measure, in order to improve the detection of the first entity and of the endpoint. The annotation of the corpus was enriched by the corresponding information, which resulted in an improvement of endpoint detection: precision was 56% on longer sentences and 58% on shorter ones; recall was 71% on longer sentences and 74% on shorter ones.

A recent work of Nye et al. (2018) describes the development of a crowd-sources corpus of nearly 5000 abstracts with annotations for patients, interventions and outcomes. The authors provide the results of two baseline algorithms for extracting these entities. A linear CRF model, using current, previous and next tokens, pos-tags, and character information as features, achieved the precision of 83%; recall of 17% and F-measure of 29%. A neural model, based on a bi-directional LSTM passing distributed vector representations of input tokens to a CRF, yielded the precision of 69%, recall of 58% and F-measure of 63%.

Dataset

In the course of our work on spin detection, we annotated a corpus of declared and reported outcomes. The reason for creating this new corpus is the absence of any available resource with the annotation for these two types of outcome mentions. The only currently available corpus with outcome annotation, to our knowledge, is that introduced by Nye et al. (2018), which was not available at the time of the beginning of our work and which does not distinguish between declared and reported outcomes, while this distinction is of vital importance for our project.

Our corpus is based on a dataset of 3,938 PMC[4] articles, selected from a larger corpus (119,339) of PMC articles on the basis of being assigned the PubMed publication type "Randomized controlled trial". We annotated a corpus of sentences from full-text articles for two types of entities: declared outcomes and reported outcomes. For declared outcomes, we annotated only primary outcomes, as they are the most important for our final goal of spin detection (omission or change of the primary outcome is one of the most common types of spin). The annotation and extraction of secondary outcomes is one of the directions of future work on this task.

As it proved to be impossible to run a large-scale annotation project with several expert annotators, the annotation was performed by AK with guidance by domain experts. We developed an annotation tool (Koroleva and Paroubek, 2019) for the sake of simplicity, ease of format conversion and customizing. The annotation uses a CoNLL-like representation scheme with B (begin) - I (inside) - O (outside) elements.

Declared outcome annotation

Misreporting of the trial outcomes is most often related to the primary outcome of a trial, thus, the primary outcome presents the highest interest for spin detection. We annotated the declared outcomes only in the contexts that explicitly state that the outcome was the primary one in the given trial, e.g. (the outcome is in bold):

*The primary outcome was **the PHQ-9**.*

Information about the primary outcome can sometimes be expressed implicitly, in statements about objectives or in descriptions of measured variables. For example, in the absence of explicit definition of the primary outcome (while secondary outcomes are clearly defined), the readers can infer that "*the human gastrointestinal microbiota*" and "*metabolic markers of health*" were the primary outcomes in:

*We aimed to assess the impact of walnut consumption on **the human gastrointestinal microbiota** and **metabolic markers of health**. Fecal and blood samples were collected at baseline and at the end of each period to assess secondary outcomes of the study, including effects of walnut consumption on fecal microbiota and bile acids and metabolic markers of health.*

In the following example "*Cognitive functioning*" can be interpreted as the trial's primary outcome as no other outcome is defined in the abstract:

***Cognitive functioning** was measured at baseline and after 12 weeks.*

To assess the need of including these types of statement in our corpus, we conducted a qual-

[4]https://www.ncbi.nlm.nih.gov/pmc/

itative corpus study and consulted our medical advisors (Isabelle Boutron, Patrick Bossuyt and Liz Wager) in the course of the supporting project (2016-2019). We compared the variables mentioned in the statements on measured variables ("***X*** *was measured*", "*We aimed at measuring* ***X***", etc.) to the variables explicitly stated to be the primary outcomes in the same text. Our corpus study showed that the variables described in this type of statements differ from the explicitly declared primary outcomes: in particular, statements of objectives usually contain more general description of what was studied (e.g. "efficacy") compared to outcomes (e.g. "survival" or "quality of life"). Thus, we concluded that these types of statement do not define a primary outcome. Furthermore, absence of an explicit definition of the trial's primary outcome does not conform with good reporting practices (Rennie, 2001; Schulz et al., 2010). Hence we excluded these types of sentences from our corpus.

To create the corpus for declared primary outcome annotation, we searched the full-text articles for sentences where both the word "*primary*" (or its synonyms: "*principal*", "*main*", etc.) and the word "*outcome*" (or its synonyms: "*end-point*", "*measure*", etc.) occur, the former precedes the latter, and the distance between them is no more than 3 tokens. Regular expressions were used to search for the terms and Python NLTK library[5] was used for sentence splitting. Out of the sentences corresponding to our criteria, we randomly selected 2,000 sentences, coming from 1,672 articles.

We created two versions of the corpus which differ in annotation of coordinated outcomes. In the first version of our corpus, we annotated coordinated outcomes as one entity for the sake of simplifying the annotation task. This version contains 1,253 occurrences of declared primary outcomes. Further, we created a more elaborated version of the annotation, re-marking coordinated outcomes as separate entities. This second version contains 1,694 occurrences of declared primary outcomes. This version will be further used in our target application. The two versions serve to assess the capabilities of our algorithms to correctly analyse coordinated entities.

A definition of a primary outcome can include time points, measurement tool, etc. We annotated the longest continuous text span containing all the relevant information about the trial's primary outcome. Declared primary outcomes are most typically represented by noun phrases, but can also be expressed by verb phrases or clauses:

Our primary outcome measures will be (a) ***whether the TUPAC guideline recommendations are implemented****, and (b) if implemented, the estimated time used for the counselling.*

[5]http://www.nltk.org

Reported outcome annotation

We annotated reported outcomes in the abstracts of the articles for which we annotated the primary outcomes, to allow for further annotation of spin related to incomplete outcome reporting. We extracted the Results and Conclusions sections of the corresponding articles (using rules and regular expressions). A number of articles in our corpus are not RCT reports but trial protocols, thus their abstracts did not contain Results and Conclusions. These abstracts were excluded from the reported outcomes corpus. The final corpus contains 1,940 sentences from 402 articles. A total of 2,251 reported outcomes was annotated.

The ways of reporting outcomes differ, and the same outcome can be reported in several ways, e.g. the sentence:

Mean total nutrition knowledge score *increased by 1.1 in intervention (baseline to follow-up : 28.3 to 29.2) and 0.3 in control schools (27.3 to 27.6).*

can be rewritten as:

The increase in ***mean total nutrition knowledge score*** *was 1.1 in intervention (baseline to follow-up : 28.3 to 29.2) and 0.3 in control schools (27.3 to 27.6).*

While both sentences report the same outcome, the structure is different, and for the second sentence both "*The increase in mean total nutrition knowledge score*" and "*mean total nutrition knowledge score*" can be considered to represent the outcome. Besides, there is a choice whether to include the aggregation method ("*mean*") into the outcome, or annotate simply "*total nutrition knowledge score*". In order to preserve uniformity throughout the annotation of reported outcomes, we decided to annotate the smallest possible text fragment referring to an outcome ("*total nutrition knowledge score*" for the given example) as it allows to annotate the same text fragment for all the variants of outcome reporting.

Reported outcomes are characterized by high variability from the syntactic point of view. They can be represented either by a noun phrase:

Overall ***response rate*** *was 39.1% and 33.3% in 3-weekly and weekly arms.*

a verb phrase:

No patients were ***reintubated****.*

or an adjective:

The CSOM and MA appeared less ***responsive*** *following a GLM-diet.*

One of the challenges in annotating reported outcomes is classifying reported variable either as a trial outcome or as a independent variables or covariates[6]. We decided to annotate all the mentions of variables (outcomes or not) unless the context of the sentence or the semantics of the phrase allows to classify it as a non-outcome variable. For example, in the sentence:

[6]https://methods.sagepub.com/Reference//encyc-of-research-design/n85.xml

*Adjustments for **age**, **gender**, and **treatment group** were performed, but did not change the results.*

the context allows to categorize all the variables as covariates.

It should be noted that this annotation decision leads to some counter-intuitive annotations: e.g. it can be expected that in a set of coordinated entities, either all the entities should be annotated as outcomes, or none of them. However, consider the following example:

***Age**, **gender** and **disease status** distribution was similar in both groups.*

Here "*Age*" and "*gender*" are considered to be independent variables due to their semantics, while "*disease status*" can be a dependent variable and will be the only entity annotated as outcome in this sentence.

There are a few differences between our corpus and the datasets used in previous works on outcome extraction. We address both declared (primary) and reported outcomes (annotation and extraction of secondary outcomes has not been covered yet and is a direction for our future work). We do not limit our dataset in terms of specific types of sentences (e.g. comparative) or constituents to be annotated (e.g. noun phrases). Our corpus is not limited to specific journals or topics. Our corpus is publicly available (Koroleva, 2019).

Methods

Baseline

We developed a simple rule-based baseline system, combining syntactic and sequential rules that cover the most typical patterns in which declared and reported outcomes can occur. For declared outcomes, sequential rules search for patterns such as:

*DET ADJ outcome was **DET ADJ* NN***

where DET denotes a determiner, ADJ is an adjective, and NN is a common noun. The sequence matched by "*DET ADJ* NN*" here is considered to be the outcome. Sequential rules use the information on tokens, lemmas and pos-tags of the words in the input text.

Syntactic rules search for similar patterns, but use the syntactic dependency graph (tags and directions of syntactic relations) instead of sequential information, to capture the cases where the target phrase is separated from the cue phrase (e.g. "*DET ADJ outcome was*") by other words.

Our rule-based baseline was designed to detect the declared outcomes in the first version of the corpus only (coordinated outcomes annotated as single entity). A simple rule-based approach can hardly be successful in a complex task such as dividing coordinated entities, hence we did not build a rule-based baseline for the second version of the declared outcomes

corpus.

For reported outcomes, the searched patterns include the expressions with comparative meaning, e.g.:

DET ADJ* NN *increased.*

DET ADJ* NN *was higher in the NN arm.*

DET ADJ* NN *was NUM*,

where NUM denotes a numeral. Fragments matched by the patterns in bold are tagged as reported outcomes.

For pos-tagging and dependency parsing, we used spaCy dependency parser (Honnibal and Johnson, 2015).

Bi-LSTM-CRF-char algorithm

Our second approach is inspired by the work of Ma and Hovy (2016) and uses the implementation of this method proposed by G.Genthial[7]. First, the model gets character-level representations of words from character embeddings using a bi-directional LSTM (bi-LSTM). After that, the model combines the character-level representation with a GloVe (Pennington et al., 2014) word vector representation and passes the combined representations to a bi-LSTM to build contextual representations of words. Finally, a linear chain CRF is applied to decode the labels. Table 3.1 shows the values of the parameters used in the configuration of the model.

Parameter	**Value**
dim_word	300
dim_char	100
train_embeddings	False
nepochs	15
dropout	0.5
batch_size	20
lr_method	"adam"
lr	0.001
lr_decay	0.9
clip	-1
nepoch_no_imprv	3
hidden_size_char	100
hidden_size_lstm	300

Table 3.1: Training parameters

[7]https://github.com/guillaumegenthial/sequence_tagging

BERT-based algorithms: the use of deep pre-trained Language models

Recently, language models pre-trained on large corpora with complex neural network architectures have been shown to be useful for several downstream NLP tasks such as question answering, named entity recognition, natural language inference, etc. by ELMO (Peters et al., 2018), OpenAI's GPT (Radford et al., 2018) and Google's BERT (Devlin et al., 2018). Intuition is to build a model trained on a large corpus for a relatively simple task of language modelling, which can further be modified for complex NLP tasks. There are two approaches to employing these pre-trained models for supervised downstream tasks:

1. feature-based approach (used in ELMO) relies on task-specific architecture, where pre-trained representations are included as additional features to existing neural network models;

2. fine-tuning approach (used in OpenAI GPT and BERT) does not require extensive task-specific parameters, it simply fine-tunes the pre-trained parameters on a downstream task.

We compared a number of recent deep pre-trained language models. First, we employed the BERT (Bidirectional Encoder Representations from Transformers) models which are well documented with openly available pre-trained weights for the models[8]. In brief, BERT uses a masked language model (MLM), randomly masking some input tokens, which allows to pre-train a deep bidirectional Transformer using both left and right contexts. Representation of a token combines the corresponding token, segment and position embeddings. The advantage of BERT compared to ELMO and OpenAI GPT is the deep bi-directionality of the representations and the size of the training corpus. We chose to use BERT because it outperformed ELMO and OpenAI GPT on a number of tasks (Devlin et al., 2018). There are several versions of BERT models: cased and uncased models, differing in the preprocessing of the input data (lower-cased vs unchanged); and base and large models, differing in the model sizes.

BioBERT (Lee et al., 2019), a domain-specific analogue of BERT, was pre-trained on a large (18B words) biomedical corpus: PubMed abstracts and PMC full-text articles, in addition to BERT training data. BioBERT is based on the cased BERT base model. Another domain-specific version of BERT is SciBERT (Beltagy et al., 2019), trained on a corpus of scientific texts (3.1B) added to BERT training data. SciBERT provides both cased and uncased models, with two versions of vocabulary: BaseVocab (the initial BERT general-domain vocabulary) and SciVocab (the vocabulary built on the scientific corpus). Both BioBERT and SciBERT outperformed BERT on some tasks of biomedical natural language processing.

Table 3.2 summarizes the training data of BERT, BioBERT and SciBERT.

[8]https://github.com/google-research/bert

	BERT	**SciBERT**	**BioBERT**
Training data	BooksCorpus English Wikipedia	BooksCorpus English Wikipedia + Semantic Scholar (Biomedical, Computer Science)	BooksCorpus English Wikipedia + PubMed Abstracts PMC Full-text articles
Volume of training data (words)	3.3B	6.4B	21.3B

Table 3.2: Training data for BERT/BioBERT/Scibert

The models we evaluated in our experiments include: BERT-base models, both cased and uncased; BioBERT model; SciBERT models, both cased and uncased, with the SciVocab vocabulary (recommended by the authors). We did not perform experiments with BERT-Large due to the lack of resources.

We explored two approached of employing the BERT-based language models for our sequence labelling task. The first approach, suggested by the developers of BERT and BioBERT (Devlin et al., 2018; Lee et al., 2019), employs a simple fine-tuning of the models on our annotated datasets. The principle behind this approach is that the pre-trained BERT models can be fine-tuned for a supervised task with one additional output layer. The one additional layer parameters along with the whole BERT model parameters are fine-tuned for the intended task. Table 3.3 summarizes the hyperparameters used for BERT-based models training and evaluation.

The second approach, suggested by the SciBERT developers (Beltagy et al., 2019), uses minimal task-specific architecture on top of BERT-based embeddings. A representation of each token in this model is built by concatenating its BERT embedding and a CNN-based character embedding. Similarly to the method of Ma and Hovy (2016), a multilayer bi-LSTM is applied to token embeddings, and a CRF is used on top of the bi-LSTM[9].

We compared performance of all the models both with unaltered input data and with lower-cased input data. It is expected that cased models perform better with unaltered input data, while uncased models perform better with lower-cased data. All the algorithms were evaluated on the token level. Machine-learning algorithms were assessed using 10-fold cross-validation

[9]The configuration and hyperparameters used for training the model can be found at: https://github.com/allenai/scibert/blob/master/allennlp_config/ner.json.

Hyperparameter	Value	Definition
init_checkpoint	None	Initial checkpoint (usually from a pre-trained BERT model)
do_lower_case	True/False	Whether to lower case the input text
max_seq_length	128	The maximum total input sequence length after WordPiece tokenization
do_train	True	Whether to run training
use_tpu	False	Whether to use TPU or GPU/CPU
train_batch_size	32	Total batch size for training
eval_batch_size	8	Total batch size for eval
predict_batch_size	8	Total batch size for predict
learning_rate	5e-5	The initial learning rate for Adam
num_train_epochs	10.0	Total number of training epochs to perform
warmup_proportion	0.1	Proportion of training to perform linear learning rate warmup for
save_checkpoints_steps	1000	How often to save the model checkpoint
iterations_per_loop	1000	How many steps to make in each estimator call
master	None	TensorFlow master URL

Table 3.3: BERT/BioBERT/SciBERT hyperparameters

(train-dev-test split was done in proportion 8:1:1). We report the averaged results. We used Tensorflow for our experiments.

Results and discussion

Tables 3.4, 3.5, 3.6 show the performance of the tested algorithms (all evaluations are done at the token level). The True value of the do_lower_case flag indicates lower-cased input data. The suffix "_biLSTM-CRF" for BERT-based model indicates the results of the approach using CRF on top of bi-LSTM.

Comparison of approaches

Our rule-based system showed reasonable performance for extracting primary outcomes (on the first version of the corpus), but not reported outcomes, as the latter are highly diverse and thus rule-based approach is not optimal. The Bi-LSTM-CRF-char algorithm using character and GloVe token embeddings did not show high performance for our tasks. All BERT-based models outperformed the Bi-LSTM-CRF-char algorithm and the rule-based baseline with a large absolute improvement (see Tables 3.4, 3.5, 3.6).

Algorithm	do_lower_case	Precision	Recall	F1
SciBERT-cased	False	89.32	87.87	88.52
BioBERT	True	88.83	88.24	88.45
BioBERT	False	88.44	88.11	88.2
SciBERT-uncased	True	88.74	87.51	88.06
SciBERT-cased	True	88.34	87.82	88
BERT-cased	False	87.73	86.94	87.23
SciBERT-uncased	False	88.05	86.24	87.06
BERT-uncased	True	87.19	86.55	86.71
BERT-cased	True	88.46	85.12	86.68
BERT-uncased	False	86.97	86.08	86.42
SciBERT-uncased_biLSTM-CRF	True	85.01	83.76	84.3
SciBERT-cased_biLSTM-CRF	True	83.93	83.88	83.88
BioBERT_biLSTM-CRF	False	83.43	84.25	83.79
SciBERT-cased_biLSTM-CRF	False	83.37	83.64	83.49
BioBERT_biLSTM-CRF	True	83.12	83.78	83.42
SciBERT-uncased_biLSTM-CRF	False	80.59	81.76	81.15
BERT-uncased_biLSTM-CRF	True	80.26	81.52	80.87
BERT-cased_biLSTM-CRF	False	80.04	80.87	80.38
BERT-cased_biLSTM-CRF	True	78.3	80.97	79.58
BERT-uncased_biLSTM-CRF	False	78.49	79.37	78.87
Rule-based	-	78.6	68.98	73.51
Bi-LSTM-CRF-char	-	59.14	63.41	61.07

Table 3.4: Primary outcome extraction - version 1: results

The fine-tuning approach employing BERT-based models consistently showed better performance than the approach using CRF on top of bi-LSTM with BERT-based embeddings. For all the tasks, even the best results of the model with CRF were inferior to the lowest results achieved by fine-tuning. This fact shows that a simple architecture (fine-tuning) can be superior to more complex (bi-LSTM-CRF) architectures for entity extraction task.

Comparison of BERT-based models

Out of all the tested approaches, fine-tuned BioBERT model showed the best performance for the second version of primary outcome extraction; fine-tuned SciBERT model outperformed

Algorithm	do_lower_case	Precision	Recall	F1
BioBERT	False	86.99	90.07	88.42
SciBERT-cased	False	87.52	89.07	88.21
SciBERT-uncased	True	87.49	88.92	88.1
SciBERT-cased	True	87.39	88.64	87.92
BioBERT	True	87.01	88.96	87.9
SciBERT-uncased	False	86.57	88.3	87.35
BERT-cased	False	86.96	87.41	87.14
BERT-uncased	True	86.6	87.39	86.91
BERT-uncased	False	86.96	86.87	86.84
BERT-cased	True	86.71	87.12	86.81
BioBERT_biLSTM-CRF	False	78.82	82	80.34
SciBERT-uncased_biLSTM-CRF	True	77.52	81.15	79.22
SciBERT-cased_biLSTM-CRF	False	77.23	80.89	78.95
BioBERT_biLSTM-CRF	True	77.86	80.12	78.9
BERT-cased_biLSTM-CRF	False	78.2	78.84	78.47
SciBERT-cased_biLSTM-CRF	True	77.05	79.73	78.29
SciBERT-uncased_biLSTM-CRF	False	77.54	78.67	78.07
BERT-uncased_biLSTM-CRF	True	76.72	78.73	77.63
BERT-cased_biLSTM-CRF	True	75.35	77.61	76.41
BERT-uncased_biLSTM-CRF	False	75.23	76.62	75.79
Bi-LSTM-CRF-char	-	49.16	52.21	50.55

Table 3.5: Primary outcome extraction - version 2: results

other systems for the first version of primary outcome extraction (cased model) and reported outcome extraction (uncased model).

As expected, BERT and SciBERT uncased models performed better with lower-cased input, while cased models performed better with unchanged input data. On the contrary, BioBERT model (cased) performed slightly better with lower-cased input for two out of three tasks. A possible explanation is that BioBERT has the largest amount of training data, where the majority of the input is naturally in lower case, thus the learnt representations show similar performance for lower-cased or unchanged input.

Overall, SciBERT and BioBERT outperformed BERT, supporting the hypothesis that motivated their creation: while pre-training of language representations on large corpora gives

Algorithm	do_lower_case	Precision	Recall	F1
SciBERT-uncased	True	81.17	78.09	79.42
BioBERT	True	80.38	77.85	78.92
BioBERT	False	79.61	77.98	78.6
SciBERT-cased	False	79.6	77.65	78.38
SciBERT-cased	True	79.24	76.61	77.64
SciBERT-uncased	False	79.51	75.5	77.26
BERT-uncased	True	78.98	74.96	76.7
BERT-cased	False	76.63	74.25	75.18
BERT-cased	True	76.7	73.97	75.1
BERT-uncased	False	77.28	72.25	74.46
SciBERT-uncased_biLSTM-CRF	True	68.44	73.47	70.77
BioBERT_biLSTM-CRF	False	70.18	71.43	70.63
BioBERT_biLSTM-CRF	True	69.09	71.57	70.24
SciBERT-cased_biLSTM-CRF	False	67.98	72.52	70.11
SciBERT-cased_biLSTM-CRF	True	66.11	71.16	68.37
SciBERT-uncased_biLSTM-CRF	False	67.25	69.59	68.18
BERT-cased_biLSTM-CRF	False	65.98	65.54	65.64
BERT-uncased_biLSTM-CRF	True	64.6	66.73	65.4
BERT-cased_biLSTM-CRF	True	64.73	66.49	65.37
BERT-uncased_biLSTM-CRF	False	62.07	64.98	63.29
Bi-LSTM-CRF-char	-	51.12	44.6	47.52
Rule-based	-	26.69	55.73	36.09

Table 3.6: Reported outcome extraction: results

good results, adding domain-specific corpora to the pre-training data further improves the performance they yield. SciBERT and BioBERT show comparable performance, demonstrating that addition of domain-specific corpus of 3.1B words to the training data (as done by SciBERT) is sufficient and leads to similar improvements as adding 18B words (as done by BioBERT).

The performance for the second version of the primary outcomes corpus is very close to the performance of corresponding models for the first version of the task. These results show that deep pre-trained language representations successfully handle the extraction of coordinated entities, which makes it a promising approach to extraction of secondary outcomes, most often

represented by coordinated syntactic groups.

It is difficult to compare our results directly with the previous works on outcome extraction: all the works used corpora that vary in volume and that were built on different principles regarding the selection of sentences and text fragments to annotate; besides, evaluation in different approaches was performed on different level (sentence, entity, token). Taking into account this limitations, we can still state that our results are better than the reported results in the previous comparable works (Summerscales et al., 2009, 2011; Blake and Lucic, 2015; Lucic and Blake, 2016; Nye et al., 2018). To allow for transparency and reproducibility of outcome extraction, we released our corpus annotated for declared (primary) and reported outcomes. It has, however, some limitations: e.g., annotating only explicit definitions of declared primary outcomes; annotating reported outcome in abstracts only; annotation done by one annotator.

Conclusions

Automatic extraction of primary and reported outcomes of clinical trials is a vital task for automating systematic review process, clinical question answering, and assessment of biomedical articles for bias and spin.

We proposed a deep learning approach to trial outcome extraction and tested a number of pre-trained language representations. Our results show that language models pre-trained on large general-domain corpora can be successfully employed for extracting complex and varied entities, even with limited amount of domain specific training data. Pre-training language models on domain-specific data further improves the performance. Our approach does not require manual feature engineering or any other task-specific settings.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

References

I. Beltagy, A. Cohan, and K. Lo. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

C. Blake and R. Kehm. Comparing breast cancer treatments using automatically detected surrogate and clinically relevant outcomes entities from text. *Journal of Biomedical Informatics: X*, 1:100005, 2019. ISSN 2590-177X. doi: 10.1016/j.yjbinx.2019.100005. URL `http://www.sciencedirect.com/science/article/pii/\S2590177X19300046`.

C. Blake and A. Lucic. Automatic endpoint detection to support the systematic review process. *J. Biomed. Inform*, 2015.

F. Boudin, J.-y. Nie, J. C Bartlett, R. Grad, P. Pluye, and M. Dawes. Combining classifiers for robust pico element detection. *BMC medical informatics and decision making*, 10:29, 05 2010. doi: 10.1186/1472-6947-10-29.

I. Boutron, S. Dutton, P. Ravaud, and D. Altman. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*, 2010.

I. Boutron, D. Altman, S. Hopewell, F. Vera-Badillo, I. Tannock, and P. Ravaud. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *Journal of Clinical Oncology*, 2014.

B. D. Bruijn, S. Carini, S. Kiritchenko, J. Martin, and I. Sim. Automated information extraction of key trial design elements from clinical trial publications. In *Proceedings of the AMIA Annual Symposium*, 2008.

D. Demner-Fushman and J. Lin. Knowledge extraction for clinical question answering : Preliminary results. In *Proc of the AAAI 2005 Workshop on Question Answering in Restricted Domains*, 2005.

D. Demner-Fushman and J. Lin. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103, 2007. doi: 10.1162/coli.2007.33.1.63. URL `https://doi.org/10.1162/coli.2007.33.1.63`.

D. Demner-Fushman, B. Few, S. Hauser, and G. Thoma. Automatically identifying health outcome information in medline records. *Journal of the American Medical Informatics Association*, 2006.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Y. gang Cao, J. J. Cimino, J. Ely, and H. Yu. Automatically extracting information needs from complex clinical questions. *Journal of Biomedical Informatics*, 43(6):962 – 971, 2010. ISSN 1532-0464. doi: 10.1016/j.jbi.2010.07.007. URL http://www.sciencedirect.com/science/article/pii/\S1532046410001061.

B. Goldacre, H. Drysdale, A. Powell-Smith, A. Dale, I. Milosevic, E. Slade, P. Hartley, C. Marston, K. Mahtani, and C. Heneghan. The compare trials project. 2016. URL www.COMPare-trials.org.

R. Haneef, C. Lazarus, P. Ravaud, A. Yavchitz, and I. Boutron. Interpretation of results of studies evaluating an intervention highlighted in google health news: a cross-sectional study of news. *PLoS ONE*, 2015.

H. Hassanzadeh, T. Groza, and J. Hunter. Identifying scientific artefacts in biomedical literature: The evidence based medicine use case. *Journal of Biomedical Informatics*, 49:159 – 170, 2014. ISSN 1532-0464. doi: 10.1016/j.jbi.2014.02.006. URL http://www.sciencedirect.com/science/article/pii/\S1532046414000422.

J. P. T. Higgins, D. G. Altman, P. C. Gøtzsche, P. Jüni, D. Moher, A. D. Oxman, J. Savović, K. F. Schulz, L. Weeks, J. A. C. Sterne, C. B. M. Group, and C. S. M. Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, 343, 2011. doi: 10.1136/bmj.d5928.

M. Honnibal and M. Johnson. An improved non-monotonic transition system for dependency parsing. In *Proc. of EMNLP 2015*, pages 1373–1378, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL https://aclweb.org/anthology/D/D15/D15-1162.

K.-C. Huang, I.-J. Chiang, F. Xiao, C.-C. Liao, C. C.-H. Liu, and J.-M. Wong. Pico element detection in medical text without metadata: Are first sentences enough? *Journal of Biomedical Informatics*, 46(5):940 – 946, 2013. ISSN 1532-0464. doi: 10.1016/j.jbi.2013.07.009. URL http://www.sciencedirect.com/science/article/pii/\S153204641300110X.

M. Huang, A. Névéol, and Z. Lu. Recommending mesh terms for annotating biomedical articles. *Journal of the American Medical Informatics Association : JAMIA*, 18:660–7, 05 2011. doi: 10.1136/amiajnl-2010-000055.

D. Jin and P. Szolovits. PICO element detection in medical text via long short-term memory neural networks. In *Proceedings of the BioNLP 2018 workshop*, pages 67–75, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-2308.

S. R. Jonnalagadda, P. Goyal, and M. D. Huffman. Automating data extraction in systematic reviews: a systematic review. In *Systematic reviews*, 2015.

R. Khare, R. Leaman, and Z. lu. Accessing biomedical literature in the current information landscape. *Methods in molecular biology (Clifton, N.J.)*, 1159:11–31, 05 2014. doi: 10.1007/978-1-4939-0709-0_2.

S. Kim, D. Martinez, L. Cavedon, and L. Yencken. Automatic classification of sentences to support evidence based medicine. *BMC bioinformatics*, 12 Suppl 2:S5, 03 2011. doi: 10.1186/1471-2105-12-S2-S5.

S. Kiritchenko, B. D. Bruijn, S. Carini, J. Martin, and I. Sim. Exact: automatic extraction of clinical trial characteristics from journal publications. *BMC Med Inform Decis Mak.*, 2010.

A. Koroleva. Annotated corpus for primary and reported outcomes extraction, May 2019. URL https://doi.org/10.5281/zenodo.3234811.

A. Koroleva and P. Paroubek. Demonstrating construkt, a text annotation toolkit for generalized linguistic contructions applied to communication spin. In *The 9th Language and Technology Conference (LTC 2019) Demo Session*, 2019.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*, 2019.

A. Lucic and C. Blake. Improving endpoint detection to support automated systematic reviews. In *AMIA Annu Symp Proc.*, 2016.

X. Ma and E. Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1101. URL https://www.aclweb.org/anthology/P16-1101.

B. Nye, J. J. Li, R. Patel, Y. Yang, I. Marshall, A. Nenkova, and B. Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 197–207, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1019. URL https://www.aclweb.org/anthology/P18-1019.

J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proc. of EMNLP 2014*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://www.aclweb.org/anthology/D14-1162.

M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018. doi: 10.18653/v1/n18-1202. URL http://dx.doi.org/10.18653/v1/N18-1202.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding with unsupervised learning. *Technical report, OpenAI*, 2018.

D. Rennie. Consort revised – improving the reporting of randomized trials. *JAMA*, 285:2006–7, 2001.

W. Richardson, M. Wilson, J. Nishikawa, and R. Hayward. The well-built clinical question: a key to evidence-based decisions. *ACP J Club*, 123(3), 1995.

K. F. Schulz, D. G. Altman, and D. Moher. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340, 2010. ISSN 0959-8138. doi: 10.1136/bmj.c332. URL https://www.bmj.com/content/340/bmj.c332.

R. Summerscales, S. Argamon, J. Hupert, and A. Schwartz. Identifying treatments, groups, and outcomes in medical abstracts. In *Proceedings of the Sixth Midwest Computational Linguistics Colloquium (MCLC)*, 2009.

R. L. Summerscales, S. E. Argamon, S. Bai, J. Hupert, and A. Schwartz. Automatic summarization of results from clinical trials. *2011 IEEE International Conference on Bioinformatics and Biomedicine*, pages 372–377, 2011.

M. Verbeke, V. Van Asch, R. Morante, P. Frasconi, W. Daelemans, and L. De Raedt. A statistical relational learning approach to identifying evidence based medicine categories. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learnin*, pages 579–589, 07 2012.

A. Yavchitz, I. Boutron, A. Bafeta, I. Marroun, P. Charles, J. Mantz, and P. Ravaud. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS Med*, 2012.

Chapter 4

Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations. Anna Koroleva, Sanjay Kamath, Patrick Paroubek. Journal of Biomedical Informatics - X

Context

The identification of outcomes that we addressed in the previous chapter is one of the steps towards a solution for automatic spin detection, but it does not suffice for detection of outcome-related spin. To identify potential spin, we need to be able to compare various mentions of outcomes occurring in an article, as one of the most common types of spin is outcome switching —unjustified change of trial outcomes. The change of the primary outcome is the most alarming case of outcome switching, as it implies that the main objective of a given trial is not reported. Identification of primary outcome switching is thus one of the main elements of our spin detection pipeline.

Outcome switching detection consists of two steps: extraction of declared (primary) and reported outcomes, as described in the previous chapter, and comparing the extracted declared (primary) outcome to the extracted reported outcomes to check if the declared (primary) outcome is present among the reported ones.

The second step of this algorithm consists in assessing the semantic similarity of pairs of outcomes. This chapter reports on building a corpus of pairs of outcomes annotated for semantic similarity (on a binary scale) and on our experiments about assessing the semantic

similarity of trial outcomes. We reviewed the state of the art for semantic similarity assessment in the biomedical domain. We tested a number of single similarity measures, machine learning classifiers using a combination of the single measures, and a deep learning approach consisting in fine-tuning pre-trained language models on our annotated data. The best performing algorithm was included in our spin detection pipeline.

Authors' contributions

AK designed the study described in this chapter and interpreted the data. AK collected and annotated the corpus. AK and SK conducted the experiments, supervised by PP. AK drafted the manuscript. SK and PP revised the draft critically for important intellectual content.

Abstract

Background: Outcomes are variables monitored during a clinical trial to assess the impact of an intervention on humans' health. Automatic assessment of semantic similarity of trial outcomes is required for a number of tasks, such as detection of outcome switching (unjustified changes of pre-defined outcomes of a trial) and implementation of Core Outcome Sets (minimal sets of outcomes that should be reported in a particular medical domain).

Objective: We aimed at building an algorithm for assessing semantic similarity of pairs of primary and reported outcomes. We focused on approaches that do not require manually curated domain-specific resources such as ontologies and thesauri.

Methods: We tested several approaches, including single measures of similarity (based on strings, stems and lemmas, paths and distances in an ontology, and vector representations of phrases), classifiers using a combination of single measures as features, and a deep learning approach that consists in fine-tuning pre-trained deep language representations. We tested language models provided by BERT (trained on general-domain texts), BioBERT and SciBERT (trained on biomedical and scientific texts, respectively). We explored the possibility of improving the results by taking into account the variants for referring to an outcome (e.g. the use of a measurement tool name instead on the outcome name; the use of abbreviations). We release an open corpus with annotation for similarity of pairs of outcomes.

Results: Classifiers using a combination of single measures as features outperformed the single measures, while deep learning algorithms using BioBERT and SciBERT models outperformed the classifiers. BioBERT reached the best F-measure of 89.75%. The addition of variants of outcomes did not improve the results for the best-performing single measures nor for the classifiers, but it improved the performance of deep learning algorithms: BioBERT achieved an F-measure of 93.38%.

Conclusions: Deep learning approaches using pre-trained language representations outperformed other approaches for similarity assessment of trial outcomes, without relying on any manually curated domain-specific resources (ontologies and other lexical resources). Addition of variants of outcomes further improved the performance of deep learning algorithms.

Keywords: Trial outcomes, Semantic similarity, Natural Language Processing, Deep learning, Pre-trained language representations, Spin detection

Introduction

Outcomes in clinical research are the variables monitored during clinical trials to assess how they are affected by the treatment taken or by other parameters. Outcomes are one of the most important elements of trial design: they represent the objectives of the trial; the primary outcome (the main monitored variable) is used to determine the trial's statistical power and to calculate the needed sample size.

There are several data sources that contain information on trial outcomes. First, outcomes of clinical trials are recorded in trial registries - open online databases that store information on planned, ongoing or completed research. Second, outcomes are defined in protocols of clinical trials. Last, outcomes are presented in texts of medical research articles, where they can occur in two main types of contexts: 1) definition of outcomes that were assessed in the trial ("*Primary outcome will be* ***overall survival****.*") – context similar to that in protocols; and 2) reporting of results for an outcome ("*Patients of the treatment condition showed significantly greater reduction of* ***co-morbid depression*** *and* ***anxiety*** *as compared to the waiting list condition.*"). We will refer to the outcomes occurring in the first type of contexts as *pre-defined outcomes*, and to the outcomes occurring in the second type of context as *reported outcomes*.

A number of tasks require comparing two outcomes (from the same or different sources) to establish if they refer to the same concept.

First of all, assessing similarity between pairs of outcomes is vital to detect outcome switching. Outcomes should normally be clearly defined before the start of a trial, usually at the moment of the first registration (Smith et al., 2015; Ghert, 2017), and should not be changed without a justification. Consistency in trial outcome definition and reporting is essential to ensure reliability and replicability of a trial's findings and to avoid false positives based on reporting only the variables that showed statistically significant results confirming the researchers' hypothesis. Despite the widely acknowledged importance of proper reporting of outcomes, outcome switching – omitting pre-defined outcomes of a trial or adding new ones – remains a common problem in reporting clinical trial results. The COMPare Trials project (Goldacre et al., 2016, 2019) showed that, in 67 assessed trials, 354 pre-defined outcomes were not reported, while 357 outcomes that had not been defined in advance were added to the trial's report. Outcome switching can occur at several points: pre-defined outcomes in a medical article may be changed compared to those recorded in trial registry/protocol; reported outcomes in an article may differ compared to those recorded in trial registry/protocol or to those pre-defined in the article.

Outcome switching is directly related to two well-known problems of medical research reporting: bias, i.e. choosing only the outcomes supporting the trial hypothesis (Slade et al.,

2015; Weston et al., 2016; Altman et al., 2017), and spin, i.e. reporting only favourable outcomes and thus making research results seem more positive than the evidence justifies (Boutron et al., 2010; Lockyer et al., 2013; Lazarus et al., 2015; Chiu et al., 2017; Diong et al., 2018; Boutron and Ravaud, 2018). Spin in clinical trials assessing an intervention poses a serious threat to the quality of health care: clinicians reading trial reports with spin tend to overestimate the effects of the intervention studied (Boutron et al., 2014). Besides, spin in research articles causes spin in health news coverage and press releases (Haneef et al., 2015; Yavchitz et al., 2012), that can raise unjustified positive expectations regarding the intervention among the public.

Checking an article for outcome switching is a part of assessment for bias and spin. The checks can be performed at several levels: the outcomes recorded in the corresponding trial protocol/registry entry should be compared to the primary and secondary outcomes defined in the article; the pre-defined primary and secondary outcomes (in the protocol/registry and in the article) should be compared to the outcomes reported in the article. To perform all these comparisons, it is necessary to assess pairs of outcomes for their semantic similarity.

Another task that requires comparing outcomes concerns the core outcome sets (COS) - agreed minimum sets of outcomes to be measured in trials in particular domains[1]. The core outcome set for a domain that a trial belongs to should be compared to the outcomes defined in a trial protocol/registry entry, to identify gaps in the trial planning at an early stage and improve the trial design. Besides, the COS can be compared to the article reporting a trial to check if results for all the core outcomes are reported.

In this paper, we propose an approach to measuring semantic similarity between phrases referring to outcomes of clinical trials. It is important to note that an outcome is a complex notion that is characterized by several aspects:

- outcome name: "***depression severity***";
- measurement tool used if the outcome cannot be measured directly: "*depression severity measured by* ***the Beck Depression Inventory-II (BDI-II)***",
- time points at which the outcome is measured: "*differences in the Symptom Index of Dyspepsia* ***before randomization***, ***2 weeks*** *and* ***4 weeks after randomization***, *and* ***1 month*** *and* ***3 months after completing treatment***";
- patient-level analysis metric, e.g., change from baseline, final value, time to event: "***change from baseline*** *in body mass index (BMI)*"

[1] http://www.comet-initiative.org/glossary/cos/

- population-level aggregation method, e.g. mean, median, proportion: "*the **mean** number of detected polyps*", "***the proportion of patients** suffering from postoperative major morbidity and mortality*";

- type of analysis of results based on the population included, i.e. intention-to-treat analysis (all the enrolled patients are analyzed, even those who dropped out) or per-protocol analysis (only the patients who followed the protocol are analyzed): "*the change in IOP from baseline to week 4 at 8 a.m. and 4 p.m. **for the per protocol (PP) population** using a "worse eye" analysis*";

- covariates that the analysis of the outcome is adjusted for: "*whole body bone mineral content of the neonate, adjusted for **gestational age** and **age at neonatal DXA scan***";

- reasons for using a particular outcome (explanation of relevance, references to previous works using the outcome): "*the physical and mental component scores (PCS and MCS) of the Short Form 36 (SF-36), **a widely used general health status measure***".

Outcome mentions necessarily contain the outcome name or the measurement tool name, which are used to refer to the outcome. However, all the other items are not mandatory. The level of detail in an outcome mention can differ between different data sources: e.g. registry outcomes tend to be longer and described in more detail than those defined in the articles. Thus, an inherent problem for establishing the similarity between two outcomes is comparing detailed outcome descriptions to under-specified ones. Besides, it is questionable whether two outcomes differing in e.g. type of analysis (intention-to-treat vs per-protocol) are different outcomes or different aspects of the same outcome. In this work, we consider two outcomes to refer to the same concept if the outcome/measurement tool names of the two are the same, disregarding the other aspects.

To the best of our knowledge, automatic outcome similarity assessment has not been addressed yet. We present the first corpus of sentences from biomedical articles from PubMed Central (PMC)[2] annotated for outcomes and their semantic similarity. This corpus has been created in the context of a project aimed at automating spin detection in clinical articles, which is a part of the Methods in Research on Research (MiRoR) programme[3], an international multi-disciplinary research project aiming at reducing the waste in biomedical research.

We propose deep learning methods using pre-trained language representations to evaluate similarity between pairs of outcomes. We compare a number of representations, pre-trained on general-domain and on domain-specific datasets. We compare the deep learning approach to some simple baseline similarity measures.

[2]https://www.ncbi.nlm.nih.gov/pmc/
[3]http://miror-ejd.eu/

Related work

The previous work distinguished between the notions of semantic similarity and semantic relatedness. Pedersen et al. (2005) define relatedness as "the human judgments of the degree to which a given pair of concepts is related", and state that it is a more general concept of semantics of two concepts, while similarity is a type of relatedness, usually defined via the "is-a" relation between the concepts in a taxonomy or ontology. Measuring semantic similarity of clinical trial outcomes has not been addressed as a separate task before, but semantic similarity and relatedness assessment and paraphrase recognition attracts substantial attention as it is required in a wide range of domains and applications. Similarity is measured between long or short texts or concepts. Measures used are often based on specialized lexical resources (thesauri, taxonomies). In this section, we provide on overview of several works on similarity and relatedness in the biomedical domain.

The measures of semantic similarity and relatedness can be divided into the following groups: string similarity measures, path-based measures, information content-based measures, and vector-based measures. Similarity and relatedness can be measured on different levels: word, term, concept, or sentence.

String similarity measures

String-based similarity measures are the simplest similarity measures based only on the surface form of the compared phrases, without taking into account the semantics. Still, they find their use in measuring the semantic similarity in the biomedical domain, e.g. the work of Soganciogl u et al. (2017) used, among other measures of similarity, a number of string-based measures: q-gram similarity (the number of q-grams from the first string over the q-grams obtained from the other string), block distance (the sum of the differences of corresponding components of two compared items), Jaccard similarity (the number of common terms in two sets over the number of unique terms in them), overlap coefficient (the number of common terms in two sets divided by the size of the smaller set), and Levenshtein distance (the minimum number of changes required to transform one string into another).

Ontology-based measures

Path-based measures

Ontologies contain a formal, structured representation of knowledge. A number of similarity measures based on paths between the concepts in ontologies exist, such as the path similarity (the shortest path connecting the concepts in the hypernym-hyponym taxonomy); the Leacock-

Chodorow similarity score (Leacock and Chodorow, 1998) (the shortest path connecting the concepts and the maximum depth of the taxonomy used); the Wu-Palmer similarity score (Wu and Palmer, 1994) (the depth of the senses of the concepts in the taxonomy and that of their most specific ancestor node); a metric of distance in a semantic net, introduced by Rada et al. (1989), calculated as the average minimum path length between all combinations of pairs of nodes corresponding to concepts; the minimum number of parent links between the concepts (Caviedes and Cimino, 2004). The most commonly used ontology in the general domain is WordNet (Fellbaum, 1998), however, similarity measures based on general-domain resources are stated to be ineffective for domain-specific tasks (Pedersen et al., 2005). A number of works proposed to adapt the existing measures of semantic similarity, which are based on WordNet, to the biomedical domain using the available medical ontologies, in particular SNOMED CT[4], MeSH[5] (Medical Subject Headings), or the Gene Ontology (Rada et al., 1989; Pedersen et al., 2005; McInnes et al., 2009; Lord et al., 2003; Caviedes and Cimino, 2004; Soganciogl u et al., 2017). Importantly, when similarity is assessed on the sentence level, tools such as Metamap (Aronson, 2001) are needed to map the sentence text to concepts from the Unified Medical Language System (UMLS) (Soganciogl u et al., 2017). Metamap finds both words and phrases corresponding to medical concepts, which makes this approach more reliable that assuming that each word is a concept.

Information content-based measures

Information content (IC) reflects the amount of information carried by a term in a discourse. The notion of IC was introduced by Resnik (1995) who proposed to measure the IC of a concept as $IC(c) = -logp(c)$, where c denotes a concept and p(c) denotes the probability of the concept c occurring in a corpus. IC can be used to measure the similarity of two concepts by calculating the amount of information shared by them. Resnik (1995) proposed to measure the similarity of concepts as the IC of their least common subsumer (the most specific taxonomical ancestor of the two terms).

IC-based similarity measures have been used in the biomedical domain. Pedersen et al. (2005) assessed IC-based measures introduced by Resnik (1995) and Lin (1998) on a set of pairs of medical terms. Sánchez and Batet (2011) proposed an overview of IC-based similarity measures (e.g. Resnik (1995); Lin (1998)) and developed a method of computing IC from the taxonomical knowledge in biomedical ontologies, in order to propose new IC-based semantic similarity measures. Aouicha and Taieb (2016) proposed to measure semantic similarity based on IC, using topological parameters of the MeSH taxonomy.

[4]http://www.snomed.org/

[5]https://www.nlm.nih.gov/mesh/meshhome.html

A notable work of Harispe et al. (2014) provides a more systematic view at ontology-based similarity measures. The authors analyzed a number of ontology-based semantic similarity measures to assess whether some of the existing measures are equivalent and which measures should be chosen for a particular application. The authors classify the similarity measures into a few categories: edge-based measures (similarity of two concepts is calculated according to the strength of their interlinking in an ontology); node-based measures, divided into feature-based approaches (evaluating a concept by a set of features made of its ancestors) and approaches based on information theory (similarity of concepts is calculated according to the amount of information they provide, as a function of their usage in a corpus); and hybrid approaches, combining edge-based and node-based approaches.

Apart from representing the compared concepts, ontologies can be used to exploit contextual features to assess the similarity of new terms. Spasić and Ananiadou (2004) proposed to represent the context of a term by syntactic elements annotated with information retrieved from a medical ontology. The sequences of contextual elements are compared using the edit distance (number of changes needed to transform one sequence into another).

Vector-based measures

Distributional models of semantics, representing term information as high-dimensional vectors, are successfully used in a number of tasks, including semantic similarity assessment (e.g. Blacoe and Lapata (2012)). In the biomedical domain, Soganciogl u et al. (2017) used distributed vector representations of sentences built with the word2vec (Mikolov et al., 2013) model to compute sentence-level semantic similarity. Henry et al. (2018) compared a number of multi-word term aggregation methods of distributional context vectors for measuring semantic similarity and relatedness. The methods assessed include summation or mean of component word vectors, construction of compound vectors using the compoundify tool (a part of the Perl word2vec interface package[6]), and construction of concept vectors using MetaMap. None of the evaluated multi-word term aggregation methods was significantly better than the others. Park et al. (2019) developed a concept-embedding model of a semantic relatedness measure, combining the UMLS and Wikipedia as an external resource to obtain contexts texts for words not presented in the UMLS. Concept vector representations were built upon the context texts of the concepts. The degree of relatedness of concepts was calculated by the cosine similarity between corresponding vectors. This approach is stated to overcome the issue of limited word coverage, which the authors state to pose problems for earlier approaches.

[6]https://sourceforge.net/projects/word2vec-interface/

Methods combining several measures

Some approaches combine several of the above-listed measures of similarity and/or relatedness. Soganciogl et al. (2017) developed a supervised regression-based model combining the string similarity measures, ontology-based measures, and distributed vector representations as features. Henry et al. (2019) developed an approach combining statistical information on co-occurrences of UMLS concepts with structured knowledge from a taxonomy, based on concept expansion using hierarchical information from the UMLS.

The common feature of the majority of the listed approaches to semantic similarity assessment is the use of domain-specific resources such as ontologies, that require laborious curation. Recently, Blagec et al. (2019) suggested an alternative approach to evaluating semantic similarity of sentences from biomedical literature. The authors employed neural embedding models that are trained in an unsupervised manner on large text corpora without any manual curation effort needed. The models used in this work were trained on 1.7 million PubMed articles. The models were evaluated on the BIOSSES dataset of 100 sentence pairs (Soganciogl et al., 2017). The unsupervised model based on the Paragraph Vector Distributed Memory algorithm showed the best results, outperforming the state-of-the-art results for the BIOSSES dataset. The authors also proposed a supervised model including string-based similarity metrics and a neural embedding model. It was shown to outperform the existing ontology-dependent supervised state-of-the-art approaches.

Existing datasets

A few datasets annotated for semantic similarity of biomedical concepts or texts exist. Pedersen et al. (2005) were the first to introduce a set of 30 pairs of medical terms annotated for semantic relatedness by 12 annotators on a 10-point scale.

Pakhomov et al. (2010b) created a set of 101 medical term pairs that were rated for semantic relatedness on a 10-point scale by 13 medical coding experts. The set was initially compiled by a practicing Mayo Clinic physician.

Pakhomov et al. (2010a) compiled a set of 724 pairs of medical terms from the UMLS, belonging to the categories of disorders, symptoms and drugs. The dataset included only concepts with at least one single-word term, to control for impact of term complexity on the judgements on similarity and relatedness. Further, a practicing physician selected pairs of terms for four categories: completely unrelated, somewhat unrelated, somewhat related, and closely related. Each category comprised approximately 30 term pairs. The pairs were rated for semantic similarity and relatedness by 8 medical residents.

The BIOSSES dataset (Soganciogl et al., 2017) contains 100 pairs of sentences selected

Paper	Similarity /relatedness	Items	Number of pairs	Scale	Selection process	Number of annotators	Competence of annotators
Pedersen et al. (2005)	relatedness	medical terms	30	1-10	manual selection	12	physicians and medical coders
Pakhomov et al. (2010b)	relatedness	medical terms	101	1-10	manual selection	13	medical coding experts
Pakhomov et al. (2010a)	similarity and relatedness	medical terms	724	0-1600 (pixel offsets)	two-step (automated + manual)	8	medical residents
Soganciogl et al. (2017)	similarity	sentences	100	0-4	manual selection	5	unspecified
Wang et al. (2018)	similarity	sentences	1068	0-5	automatic selection	2	medical experts

Table 4.1: Existing datasets annotated for semantic similarity/relatedness in the biomedical domain

from the Text Analysis Conference Biomedical Summarization Track Training Dataset. The sentence pairs were rated for similarity on a 5-point scale by five human experts.

Wang et al. (2018) aimed at creating a resource for semantic textual similarity assessment in the clinical domain. The authors assembled MedSTS, a set of 174,629 sentence pairs from a clinical corpus at Mayo Clinic. Two medical experts annotated a subset of 1,068 sentence pairs with similarity scores in the range from 0 to 5.

Table 4.1 summarizes the characteristics of the existing datasets.

Annotation of outcome pairs

For us the application of interest is detection of spin related to incorrect reporting of the primary outcome in abstracts of articles reporting randomized controlled trials (RCTs), in particular, omission of the primary outcome. This task is very specific and requires a corpus with annotations for semantic similarity of pairs of primary and reported outcomes. The task of semantic similarity assessment of outcomes can be regarded as a subtask of semantic similarity assessment of medical term pairs, which has been explored in previous works and for which a few datasets exist. However, there is an inherent difference between a corpus of outcome pairs and the existing corpora of medical term pairs: while the existing corpora of medical term

pairs contain terms belonging to different categories (e.g. drugs, symptoms and disorders), all the terms in a corpus of outcome pairs belong to the same class (outcomes, i.e. measures or variables). In a corpus containing several categories, it can be expected that the items of the same category are judged to be more similar to each other than to the items of other categories (e.g. all the drug names are more similar to each other than to the names of disorders), while in a corpus with a single category this criterion does not apply. The relation of semantic similarity is simpler for outcomes: two outcome mentions are either same (refer to the same measure/variable), or different, hence the relation is binary and can be annotated on a 0-1 scale. On the contrary, in the existing corpora multi-item scales were necessary to annotate similarity/relatedness (drugs names are more similar to each other than disorder names, but the level of similarity within the category of drug names vary).

As no corpus with annotation for semantic similarity of outcomes exists, we created and annotated our own, that we release as a freely available dataset (Koroleva, 2019). It is based on a set of 3,938 articles from PMC[7] with the publication type "Randomized controlled trial". The corpus annotation proceeded in two steps: annotation of primary and reported outcomes, and annotation of semantic similarity between them. As it proved to be impossible to recruit within a reasonable time frame several annotators with sufficient level of expertise in the domain of medical research reporting, the annotation work was performed by one single annotator with expertise in NLP, trained and consulted by three experts in clinical research reporting.

Annotation of outcomes

The annotation and extraction of primary and reported outcomes is the subject of a separate paper, here we only present in brief the annotation principles that are important for the topic of this paper.

For primary outcome annotation, we aimed at annotating contexts that explicitly define the primary outcome of a trial, e.g.:

"We selected ***the shortened version of the Chedoke Arm & Hand Activity Inventory (CAHAI-7)*** *as the primary outcome measure."*

To find these contexts, we randomly selected 2,000 sentences that contain the word *"primary"* or its synonyms, followed by the word *"outcome"* or its synonyms, with the distance no more than 3 token between them. The synonyms of the words "*primary*" and "*outcome*" used in sentence selection are shown in Table 4.2. The sentences were selected from full-text articles. We annotated the longest continuous text span that includes all the relevant information about the trial's outcome, such as measurement tool used, time points, etc.

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

Word	Synonyms
primary	main, first, principal, final, key
outcome	endpoint/end-point/end point, measure, variable, assessment, parameter, criterion

Table 4.2: Synonyms of the words "*primary*" and "*outcome*" used in sentence selection

For reported outcomes annotation, we selected the Results and Conclusions sections of the abstracts of the articles for which we previously annotated the primary outcomes. 1,940 sentences constituted the corpus for reported outcomes annotation.

Reporting outcomes are characterized by high diversity: they can be expressed by a noun phrase, a verb phrase or an adjective. The same outcomes can be reported in different ways, e.g. the following sentences report the same outcome:

1. *"At 12-month follow-up, the intervention group showed a significant positive change (OR = 0.48) in* ***receiving information on healthy computer use*** *compared to the usual care group."*

2. *"The intervention group showed a significant positive change (OR = 0.48) in* ***receiving information on healthy computer use*** *at 12-month follow-up, compared to the usual care group."*

3. *"****Receiving information on healthy computer use*** *in the intervention group showed a significant positive change (OR = 0.48) at 12-month follow-up, compared to the usual care group."*

In different variants of the sentence, it is possible to annotate as the outcome either:

1. *"change (OR = 0.48) in receiving information on healthy computer use"*,

2. *"receiving information on healthy computer use at 12-month follow-up"*, or

3. *"Receiving information on healthy computer use"*.

However, it appears reasonable to have the same outcome annotated in all of the variants. Thus, we annotated the shortest possible text span for reported outcomes.

Annotation of semantic similarity of pairs of outcomes

To annotate the similarity between primary and reported outcomes, we took pairs of sentences from the corpora annotated for outcomes: the first sentence in each pair comes from the corpus of primary outcomes, the second sentence comes from the corpus of reported outcomes,

and both sentences are from the same article (to ensure that primary and reported outcomes exist in the same document, in order to avoid a too high percentage of dissimilar pairs in the final corpus). We used a binary flag to annotate the pairs of outcomes: if both outcomes in a pair are considered to refer to the same outcome, the pair is assigned the 'similar' label; otherwise the 'dissimilar' label. Interestingly, outcomes can refer to the same concept by using antonyms: e.g. "*ICP (Intracranial Pressure) control*" vs. "*uncontrollable intracranial pressure*".

It is important to note that the annotated primary outcomes included all the possible information items present in the sentence (time points, measurement methods, etc.), while the annotated reported outcomes contain the minimal information (usually, the outcome or measurement tool name). Thus, primary outcomes typically contain more information than reported outcomes. When annotating semantic similarity, we disregarded possible differences in additional information such as time points: outcomes were annotated as similar if the outcome/measurement tool used is the same. Table 4.3 shows some examples of the outcome pairs that were judged to refer to the same (similarity = 1) or different (similarity = 0) concept.

Differences in additional information items (time points, analysis metrics, etc.) are important for a more fine-grained assessment of outcome similarity. However, annotating this information would make the annotation much more complex. We regard comparing additional information on outcomes as a separate task and thus do not include it in the current approach.

Absence of medical knowledge can cause difficulties in annotating outcome similarity. In cases of doubt, the annotator referred to the whole article text or conducted additional research to make the final decision. The total of 3,043 pairs of outcomes were annotated: 701 (612 after deduplication) "similar" and 2,342 (2,187 after deduplication) "dissimilar" pairs.

Expanded dataset

The ways of referring to an outcome may differ: e.g., the outcome defined as "*the quality of life of people with dementia, as assessed by QoL-AD*" may be referred to by the outcome name ("*the quality of life of people with dementia*") or by the measurement tool name ("*QoL-AD*"), which can in turn be used in the abbreviated or full ("*Quality of Life-Alzheimer's Disease*") form. We expect the variability in choosing one of these options to negatively affect the performance of the similarity assessment. Thus, we tried to account for this variability in two ways.

First, we searched for abbreviations and their expansions in the full text of the article where a given outcome occurs, using regular expressions. We chose this approach instead of using medical thesauri and automated tools such as Metamap (Aronson, 2001) based on the thesauri, because abbreviations can have several possible expansions depending on the

Primary outcome	Reported outcome	Similarity
the change relative to baseline in the multiple sclerosis functional composite score (MSFC)	MSFC score	1
the recruitment rate	the overall recruitment yield	1
the maximum % fall in FEV1 7 hours after the first AMP challenge	FEV1	1
ICP control	uncontrollable intracranial pressure	1
body weight	body composition	0
the volume of blood loss between T1 and T4	bleeding duration	0
tube dependency at one-year	hospital admission days	0
HbA1c	Attendance at yoga classes	0

Table 4.3: Examples of outcomes that are judged as similar(similarity = 1)/different (similarity = 0)

particular medical domain. Thus, selecting the correct expansion from a thesaurus would require some additional steps such as detecting the topic of the article. On the contrary, in the text of an article abbreviation expansions are unambiguous. After extracting abbreviations and their expansions, we replace the abbreviations in the outcome mentions by their expansions. For example, for the outcome "*EBM knowledge*" we obtain the expanded variant "*evidence-based medicine knowledge*".

Second, we looked for measurement tool names within outcome mentions, using linguistic markers such as *"measured by"*. We keep the text fragment preceding such markers as the outcome name, and the text following them as the measurement tool name, e.g. for the outcome *"cognitive functioning, as measured by the ADAS-Cog, a 0—70 point scale with a higher score indicating worse cognition"*, we add two variants: *"cognitive functioning"* and *"the ADAS-Cog, a 0—70 point scale with a higher score indicating worse cognition"*.

By applying these algorithms, we obtain an expanded version of the corpus which contains 5,050 pairs of outcomes (1,222 similar and 3,828 dissimilar pairs).

Methods

Many existing approaches to semantic similarity assessment rely on manually curated domain-specific resources, such as ontologies or other lexical resources. Although this kind of approach can show good results, its disadvantage consists in the limited word coverage of existing resources and in the need to use tools such as Metamap to map a text to biomedical concepts, resulting in a complex multi-step system with many dependencies.

Deep learning approach

In the general domain, it was recently shown that unsupervised pre-training of language models on a large corpus, followed by fine-tuning of the models for a particular task, improves the performance of many NLP algorithms, including semantic similarity assessment (Radford et al., 2018; Devlin et al., 2018). In the biomedical domain, Blagec et al. (2019) showed that neural embedding models trained on large domain-specific data outperform the state-of-the-art approaches for similarity assessment.

We explored these novel methods in order to propose an algorithm for assessment of semantic similarity that does not rely on domain-specific resources such as ontologies and taxonomies. We adopt the approach that was recently introduced by Devlin et al. (2018) and has already been shown to be highly performant. It consists in fine-tuning language representations that were pre-trained on large datasets, on a limited amount of task-specific annotated data.

Devlin et al. (2018) proposed a new method of pre-training language representations, called BERT (Bidirectional Encoder Representations from Transformers). The principle consists in pre-training language representations with the use of a masked language model (MLM) that randomly masks some of the input tokens, allowing pre-training of a deep bidirectional Transformer on both the left and right context. BERT-based pre-trained models can be easily fine-tuned for a supervised task by adding an additional output layer. For our semantic similarity assessment task, we employ the similar architecture as that used for sentence pair classification by Devlin et al. (2018) in BERT: a self-attention mechanism is used to encode a concatenated text pair. The task-specific input is fed to the output layer of BERT model, and the end-to-end fine-tuning of all the model parameters is performed. The details on the implementation can be found in Devlin et al. (2018).

BERT models were pre-trained on the joint general-domain corpus of English Wikipedia and BooksCorpus, with the total of 3.3B tokens. Two domain-specific version of BERT are of interest for our task: BioBERT (Lee et al., 2019), pre-trained on a large biomedical corpus of PubMed abstracts and PMC full-text articles comprising 18B tokens, added to the initial BERT training data; and SciBERT (Beltagy et al., 2019), pre-trained on a corpus of scientific texts with the total of 3.1B tokens, in addition to the initial BERT training corpus.

BERT provides several models: cased and uncased (differing with regard to the input data preprocessing); base and large (differing with regard to the model size). We fine-tuned and tested both cased and uncased base models. We did not perform experiments with BERT-Large due to limited computational resources. BioBERT has only cased model, with a few versions with different pre-training data (PubMed abstracts only, PMC full-text articles only, or both). We used the model pre-trained on both datasets. SciBERT provides both cased and uncased models and has two versions of vocabulary: BaseVocab (the initial BERT general-domain

vocabulary) and SciVocab (the vocabulary built on the scientific corpus). The uncased model with SciVocab is recommended by the authors, as this models showed the best performance in their experiments. We tested both cased and uncased models with SciVocab.

The hyperparameters used for fine-tuning of BERT-based models are shown in the Table 4.4.

Hyperparameter	**Value**	**Definition**
do_lower_case	True (uncased models)/False (cased models)	Whether to lower case the input text
max_seq_length	128	The maximum total input sequence length after WordPiece tokenization
train_batch_size	32	Total batch size for training
eval_batch_size	8	Total batch size for eval
predict_batch_size	8	Total batch size for predict
learning_rate	5e-5	The initial learning rate for Adam
num_train_epochs	3.0	Total number of training epochs to perform
warmup _proportion	0.1	Proportion of training to perform linear learning rate warmup for
save_checkpoints_steps	1000	How often to save the model checkpoint
iterations_per_loop	1000	How many steps to make in each estimator call
use_tpu	False	Whether to use TPU or GPU/CPU
master	None	TensorFlow master URL

Table 4.4: BERT/BioBERT/SciBERT hyperparameters

Baseline approach

We compare the BERT-based approaches to a few simple domain-independent baseline measures that fall into the following categories:

1. string measures:

   - normalized Levenshtein distance (Miller et al., 2009) (in Tables referred to as *levenshtein_norm*) - the minimal edit distance between two strings (number of edits needed to change one string into the other). We calculate the Levenshtein distance using the Python Levenshtein package and normalize it by dividing it by the length of the longer string.
   - a measure based on the Ratcliff and Obershelp algorithm (Ratcliff and Metzener, 1998) (in Tables referred to as *difflib*) which calculates the number of matching

characters in two strings divided by the total number of characters. We use the implementation proposed by the Python difflib library (SequenceMatcher function).

2. lexical measures reflecting the number of lexical items shared by the compared phrases:

   - the proportion of lemmas occurring in both compared outcomes (in Tables referred to as *lemmas*), calculated as the proportion of the lemmas shared by the compared phrases divided by the length (in lemmas) of the shorter outcome. Lemmatization was performed with the help of WordNetLemmatizer function of Python NLTK library.
   - the proportion of stems occurring in both compared outcomes (in Tables referred to as *stems*), calculated as the proportion of the stems shared by the compared phrases divided by the length (in stems) of the shorter outcome. Stemming was performed using the PorterStemmer function of Python NLTK library.

   In both lexical measures, stop-words and digits were excluded, as well as some words with general semantics typical for outcome mentions (e.g. "change", "increase", "difference").

3. vector-based measures:

   - a cosine similarity between the compared outcomes (in Tables referred to as *gensim*), using vector representation obtained with Latent Semantic Analysis using singular value decomposition. We use the implementation proposed by the Python gensim (Rehurek and Sojka, 2010) library[8].
   - a cosine similarity between the compared outcomes (in Tables referred to as *spacy*), using an average of word vectors. We use the implementation proposed by the Python spaCy (Honnibal and Johnson, 2015) library.

4. ontology-based measures:

   - path similarity score (in Tables referred to as *path*) is a WordNet-based measure of similarity of two word senses calculated as the shortest path connecting them in the hypernym-hyponym taxonomy.
   - Leacock-Chodorow similarity score (Leacock and Chodorow, 1998) (in Tables referred to as *lch*) is a WordNet-based measure of similarity of two word senses based on the shortest path connecting them and the maximum depth of the taxonomy in which they are found.

[8]https://radimrehurek.com/gensim/tut3.html

- Wu-Palmer similarity score (Wu and Palmer, 1994) (in Tables referred to as *wup*) is a WordNet-based measure of similarity of two word senses based on the depth of the senses in the taxonomy and that of their most specific ancestor node.

For all three measures, we use the functions implemented in the Python NLTK library. The final scores are calculated as proposed by Mihalcea et al. (2006).

Each of these measures returns a similarity score on a certain scale (most typically, between 0 and 1). After testing several cut-off values, we manually set a threshold for each measure to maximize the F-measure: pairs of outcomes with the similarity measure above the threshold are considered similar. The thresholds chosen for each measure are shown in Table 4.5.

Measure	**Threshold**
difflib	0.4
levenshtein_norm	0.3
lemmas	0.6
spacy	0.6
gensim	0.9
stems	0.6
path	0.4
wup	0.5
lch	2.5

Table 4.5: Thresholds set for the similarity measures

Feature-based machine-learning approach

Following the approach proposed by Sogancioglu et al. (2017), we trained and tested a number of classifiers, taking the above-listed single similarity measures as the input features. We evaluated several classifiers: Support Vector Machine (SVM) (Cortes and Vapnik, 1995); Decision Tree Classifier (Rokach and Maimon, 2008); MLP Classifier (von der Malsburg, 1986); K-neighbors Classifier (Altman, 1992); Gaussian Process Classifier (Rasmussen and Williams, 2005); Random Forest Classifier (Breiman, 2001); Ada Boost Classifier (Freund and Schapire, 1997); Extra Trees Classifier (Geurts et al., 2006); Gradient Boosting Classifier (Friedman, 2002). We used the implementation provided by Python scikit-learn library (Pedregosa et al., 2011). We performed hyperparameters tuning via exhaustive grid search (with the help of the scikit-learn GridSearchCV function). The chosen hyperparameters are shown in Table 4.6 (for the experiments on the original corpus) and Table 4.7 (for the experiments on the expanded corpus).

Classifier	Hyperparameters
RandomForest	max_depth = 25, min_samples_split = 5, n_estimators = 300
MLP	activation = 'tanh', alpha = 0.0001, hidden_layer_sizes = (50, 100, 50), learning_rate = 'constant', solver = 'adam'
GaussianProcess	1.0 * RBF(1.0)
GradientBoosting	default
KNeighbors	n_neighbors = 13, p = 1
ExtraTrees	default
AdaBoost	default
DecisionTree	default
SVC	C = 1000, gamma = 0.001, kernel = 'rbf'

Table 4.6: Hyperparameters for classifiers on the original corpus

Experiments on the expanded dataset

The expanded dataset (with expanded abbreviations and added variants of referring to an outcome by the measurement tool name or by the outcome name) is used in the experiments in the following way. For individual similarity measures, we compare all the combinations of variants for both outcomes. Out of the similarity scores obtained for all the variants, we take the maximum value as the final evaluation score. For machine learning approaches, we expanded the original annotated corpus by the extracted variants of the outcomes. We trained and tested the machine learning and deep learning approaches on both the original corpus an on the expanded corpus.

Results and discussion

For the deep learning approach, we performed the evaluation using 10-fold cross-validation, with the dataset split into train and development sets in the proportion 9:1. The performance is reported for the development set. For scikit-learn classifiers, we performed 10-fold cross-validation using the scikit-learn built-in cross_validate function.

Table 4.8 below presents the results of our experiments on the original and expanded corpus, respectively. We use the following notations in the results tables: *BioBERT*, *SciBERT uncased*, *SciBERT cased*, *BERT uncased* and *BERT cased* refer to the results of fine-tuning of the corresponding language model. *RandomForest*, *MLP*, *GaussianProcess*, *GradientBoosting*, *KNeighbors*, *ExtraTrees*, *AdaBoost*, *DecisionTree*, and *SVC* refer to the results of the corresponding scikit-learn classifier. *stems* and *lemmas* refer to the lexical similarity measures (the proportion of stems/lemmas occurring in both compared outcomes). *gensim* and *spacy* refer to vector-based measures (cosine similarity as implemented by gensim and spacy packages,

Classifier	Hyperparameters
RandomForest	max_depth = 25, min_samples_split = 5, n_estimators = 300
KNeighbors	n_neighbors = 9, p = 5
GradientBoosting	learning_rate = 0.25, max_depth = 23.0, max_features = 7, min_samples_leaf = 0.1, min_samples_split = 0.2, n_estimators = 200
MLP	activation = 'relu', alpha = 0.0001, hidden_layer_sizes = (50, 100, 50), learning_rate = 'adaptive', solver = 'adam'
GaussianProcess	1.0 * RBF(1.0)
ExtraTrees	default
AdaBoost	learning_rate = 0.1, n_estimators = 500
SVC	kernel='linear', C=1, random_state=0
DecisionTree	max_depth = 1.0, max_features = 2, min_samples_leaf = 0.1, min_samples_split = 1.0

Table 4.7: Hyperparameters for classifiers on the expanded corpus

respectively). *levenshtein_norm* refers to the normalized Levenshtein distance, *difflib* refers to the the Ratcliff and Obershelp algorithm-based measure. *path* refers to the path similarity score; *lch* refers to the Leacock-Chodorow similarity score; *wup* refers to the Wu-Palmer similarity score.

Among the single similarity measures tested on our original (non-expanded) corpus, the best performance was shown by the stem-based measure (F-measure=71.35%). Among the classifiers using the combination of measures as features, the best results were achieved by the Random Forest Classifier (F-measure=84.73%). Among the deep learning models, the fine-tuned BioBERT model showed the highest performance (F-measure=89.75%).

These results clearly show that, out of the three tested approaches (baseline single similarity measures, machine learning classifiers using the single measures as features, and deep learning), the best results on the original corpus were shown by the deep learning approaches. All the single measures were inferior to the classifiers based on the combination of the single measures. Thus, we can state the measures complement each other. Further, all the deep learning BERT-based approaches showed better performance than each of the classifiers, which indicates that the pre-trained representations are more powerful in reflecting semantic similarity than the measures used.

On the expanded corpus, the performance of single measures changed slightly compared to that on the original corpus (cf. Table 4.8). The best result, achieved by the stem-based measure, was not improved. The performance of machine learning classifiers on the expanded

	On the original corpus			On the expanded corpus		
Algorithm	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
BioBERT	88.93	90.76	89.75	92.98	93.85	93.38
SciBERT uncased	87.99	90.78	89.3	91.3	91.79	91.51
SciBERT cased	87.31	91.53	89.3	89	92.54	90.69
BERT uncased	85.76	88.15	86.8	89.31	89.12	89.16
RandomForest	86.76	82.92	84.73	74.09	60.12	66.13
BERT cased	83.36	85.2	84.21	88.25	90.1	89.12
MLP	87.79	80.61	83.95	72.21	58.05	63.87
GaussianProcess	86.69	81.11	83.74	72.08	57.13	63.58
GradientBoosting	87.84	79.96	83.63	72.94	58.4	64.72
KNeighbors	87.35	78.81	82.75	75.24	58.13	65.31
ExtraTrees	85.26	79.29	82.08	71.83	57.14	63.47
AdaBoost	86.08	77.99	81.79	72.66	55.87	62.97
DecisionTree	81.66	79.62	80.53	62.73	63.09	60.61
SVC	82.3	78.32	80.19	73.2	54.42	62.26
stems	64.03	80.56	71.35	64.03	80.56	71.35
lemmas	64.75	77.45	70.54	63.18	78.23	69.91
gensim	55.71	83.66	66.88	54.98	79.14	64.89
path	60.06	65.36	62.6	58.04	69.47	63.24
wup	53.26	68.14	59.78	52.15	73.35	60.96
levenshtein_norm	65.87	49.84	56.74	64.64	56.14	60.09
difflib	47.08	71.08	56.64	63.84	61.73	62.77
lch	59.42	53.59	56.36	62.95	25.02	35.81
spacy	35.86	75.65	48.66	35.86	75.65	48.66

Table 4.8: Results

corpus dropped significantly (the highest F-measure was 66.13% vs. 84.73% on the original corpus). On the contrary, the performance of all the fine-tuned deep learning models was better on the expanded corpus than on the original corpus. The best result, similarly to the original corpus, was shown by the fine-tuned BioBERT model: F-measure was 93.38%.

Error analysis

We provide here the error analysis of the best-performing model (fine-tuned BioBERT) on the original corpus. The most common cases of errors are as follows:

1. Use of abbreviations which leads to false negatives, e.g.:

   - *Uncontrollable intracranial pressure – **ICP** control*
   - *sickness absence – **SA** days*
   - *pain catastrophising – global **PC***
   - *controlling intracranial pressure – **ICP** control*

- *the Yale-Brown Obsessive-Compulsive Scale – the change in **YBOCS** score from baseline to endpoint*

2. Terms that are semantically close but refer to different measured variables result in false positives, e.g.:

   - *coma **recovery** time – total coma **duration***
   - *patient **satisfaction** – patient **comfort***
   - *time to **azoospermia** time to severe **oligozoospermia***

   In particular, this type of error can be observed when the terms are hyponyms of the same term, e.g.:

   - ***child** body mass index (BMI) z-score – **parent** BMI*
   - ***foot** pain – **'first-step'** pain*
   - *the proportion of delivered compressions **within** target depth compared over a 2-minute period within the groups and between the groups – the proportion of delivered compressions **below** target depth*

   Besides, this type of error occurs when the outcomes refer to different aspects of one parameter, e.g.(words indicating the differences in semantics of the phrases are in bold):

   - *the GSRS **subscores** for abdominal pain – the GSRS **total score***
   - *The **frequency** of acute exacerbation – **duration** of acute exacerbation*
   - ***costs** per quality adjusted life years (cost/QALY) – Quality adjusted life years*
   - ***time** needed to perform the motor task – **degree of help** needed to perform the task*
   - *the mean time to onset of the **first 24-h** heartburn-free period after initial dosing – The mean number of heartburn-free days **by D7***
   - *the proportion of patients with plasma HIV-1 RNA levels <**200** copies/mL at week 24 – HIV-1 RNA <**50** copies/mL*

3. Use of terms for which the similarity can only be established based on domain knowledge but not by their textual features leads to false negatives, e.g.:

   - *HSCL-25 – the severity of symptoms of depression and anxiety* (HSCL-25 is a check-list measuring the symptoms of anxiety and depression[9])

[9] http://hprt-cambridge.org/screening/hopkins-symptom-checklist/

- *response rate – took part in the Link-Up Study*
- *return of final follow-up questionnaire or reminder by the participant – the response rates*

4. Significantly different level of detail in two mentions of the same measure can lead to false negatives, e.g.:

   - *the incidence of oxygen desaturations defined as a decrease in oxygen saturation ≥ 5%, assessed by continuous pulse oxymetry, at any time between the start of the induction sequence and two minutes after the completion of the intubation – oxygen desaturations*

The best method for assessing semantic similarity

On the original outcome pairs corpus, the best-performing single similarity measure is the stem-based one (F1 = 71.35%), followed by the lemmas-based and gensim measures (Table 4.8). The gensim measure shows the best recall (83.66%).

All the scikit-learn classifiers trained on the original corpus using the combination of the single measures as features outperformed single measures (Table 4.8). The best results were achieved by the Random Forest Classifier (F-measure of 84.73%).

When trained on the original corpus, all the BERT-based models, except for the one using the BERT cased model, outperformed the feature-based classifiers and single similarity measures (Table 4.8). The best results were shown by the fine-tuned BioBERT model, reaching the F-measure of 89.75%. Results of fine-tuned SciBERT models (both cased and uncased) reached the F-measure of 89.3%, closely following BioBERT; the SciBERT cased model demonstrated the best recall (91.53%).

These results show that fine-tuned models using deep pre-trained language representations can outperform all the other tested similarity measures, with an additional advantage of not requiring any specialized resources or specific text preprocessing such as mapping to the UMLS concepts. Pre-training of language models on biomedical texts proves to be an advantageous approach as it allows to learn representations for domain-specific words, including abbreviations, from the available large unstructured data.

Does the addition of variants of referring to an outcome help?

For the single measures of similarity, expansion of the corpus by the variants of outcomes improved the performance of Wordnet-based and string-based measures, but did not improve

the results of the three best-performing measures - stem- and lemma-based ones and the gensim measure (cf. Table 4.8).

A possible explanation for the absence of improvement in the stem-and lemma-based measures is that the primary outcomes are usually rather lengthy and detailed, and tend to include all the variants: abbreviations and their expansions, measurement method. Thus, additional variants are not in fact needed. For example, the primary outcome "*depression severity measured by the Beck Depression Inventory-II (BDI-II)*" may be reported as "*depression severity*", "*the Beck Depression Inventory-II*" or "*BDI-II*", but all these variants are already present within the primary outcome mention, thus, measuring the intersection in terms of stems or lemmas will return a high similarity score. At the same time, for string-based and WordNet-based measures, addition of variants is useful: for the example above, if the outcome is reported as "*BDI-II*", it will be expanded as "*the Beck Depression Inventory-II*", which will have high similarity scores with the variant "*the Beck Depression Inventory-II (BDI-II)*" of the primary outcome.

For the classifiers using single similarity measures as features, adding outcome variants to the training corpus did not prove useful: the results of the classifiers trained on the corpus expanded by the outcome variants dropped significantly (cf. Table 4.8).

It should be highlighted that single measures and classifiers in our approach account for outcome variants in different ways: single measures compare all the pairs of variants and take the highest score as the final result, thus, low similarity between some of the variants does not affect the results. On the contrary, the classifiers use the expanded corpus to train, and thus, pairs of variants with low similarity scores but with the 'similar' label can negatively impact the results.

Interestingly, the addition of the variants to the training corpus can be useful: performance of all the BERT-based systems improved on the corpus expanded by outcome variants (cf. Table 4.8). The best result was achieved by the fine-tuned BioBERT model, with the F-measure of 93.38%.

The difference between the results of classifiers using single measures as features and the fine-tuned BERT-based models on the expanded corpus demonstrates differences between these approaches. BERT-based models successfully train on the expanded corpus as they use deep pre-trained language representations and fine-tune to learn the features required for a given task, while the training of classifiers is likely to be undermined by the pairs of outcome variants with low scores on the single similarity measures.

The results of these experiments should, however, be taken with caution, as the expansion of the corpus by outcome variants was performed automatically. We manually checked the quality of the algorithm, but it does not exclude presence of some noise. Still, we believe that

this approach is promising for our task.

What metrics are best able to identify similar or dissimilar outcomes?

Out of single similarity measures, the best ability to distinguish between similar and dissimilar outcomes, in both the original and the expanded corpora, was shown by the stem-based measure, followed by the lemma-based measure (Table 4.8).

What classifiers are best able to distinguish between similar and dissimilar outcome pairs?

In our experiments, the Random Forest Classifier showed the best results in the task of distinguishing between similar and dissimilar outcome pairs, compared to a range of other classifiers (MLP, Gaussian Process Classifier, Gradient Boosting Classifier, K-neighbors Classifier, Extra Trees Classifier, Ada Boost Classifier, Decision Tree Classifier, and SVM) (Table 4.8).

What language representation is best able to represent outcomes?

Our experiments showed that the best performance for semantic similarity assessment of outcomes is shown by the fine-tuned BioBERT model, i.e. a language model pre-trained on a large (18B tokens) biomedical corpus in addition to a 3.3.B tokens general domain-corpus. This model outperformed the models trained on the general-domain corpus only (BERT) and the models trained on a smaller (3.1) corpus of scientific paper in addition to the general domain corpus (SciBERT) (Tables 4.8).

Conclusion

Evaluation of similarity assessment of trial outcomes is a vital part of tasks such as assessment of an article for outcome switching, reporting bias and spin; besides, it can be used to improve the adherence to Core Outcomes Sets use. In this work, we introduced a first open-access corpus of pairs of primary and reported outcomes, annotated on a binary scale as similar or different. We presented our experiments on developing an algorithm of semantic similarity assessment not using domain-specific resources such as ontologies and taxonomies. We tested a number of single similarity measures, classifiers using the combination of single measures as features, and a number of deep learning models. We explored the possibility of using variants of referring to outcomes (abbreviations, measurement tool names) to improve the performance of similarity assessment.

The best results were shown by the deep learning approach using the BioBERT fine-tuned model, both on the original corpus and on the corpus expanded by the outcome variants.

Acknowledgements

We thank prof. Isabelle Boutron from the University Paris Descartes, prof. Patrick Bossuyt from the University of Amsterdam, and Dr. Liz Wager from SideView who provided valuable insight and expertise as our consultants in the domain of reporting of clinical trials.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

References

D. Altman, D. Moher, and K. Schulz. Harms of outcome switching in reports of randomised trials: Consort perspective. *BMJ: British Medical Journal (Online)*, 2017.

N. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician - AMER STATIST*, 46:175–185, 08 1992. doi: 10.1080/00031305.1992.10475879.

M. B. Aouicha and M. A. H. Taieb. Computing semantic similarity between biomedical concepts using new information content approach. *Journal of Biomedical Informatics*, 59:258 – 275, 2016. ISSN 1532-0464. doi: 10.1016/j.jbi.2015.12.007. URL `http://www.sciencedirect.com/science/article/pii/S1532046415002877`.

A. Aronson. Effective mapping of biomedical text to the umls metathesaurus: The metamap program. *AMIA Annual Symposium*, 2001:17–21, 02 2001.

I. Beltagy, A. Cohan, and K. Lo. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

W. Blacoe and M. Lapata. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 546–556,

Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D12-1050.

K. Blagec, H. Xu, A. Agibetov, and M. Samwald. Neural sentence embedding models for semantic similarity estimation in the biomedical domain. In *BMC Bioinformatics*, page 178, 2019.

I. Boutron and P. Ravaud. Misrepresentation and distortion of research in biomedical literature. *Proc Natl Acad Sci U S A.*, 2018. doi: 10.1073/pnas.1710755115.

I. Boutron, S. Dutton, P. Ravaud, and D. Altman. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*, 2010.

I. Boutron, D. Altman, S. Hopewell, F. Vera-Badillo, I. Tannock, and P. Ravaud. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *Journal of Clinical Oncology*, 2014.

L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, Oct. 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

J. E. Caviedes and J. J. Cimino. Towards the development of a conceptual distance metric for the umls. *Journal of Biomedical Informatics*, 37(2):77 – 85, 2004. ISSN 1532-0464. doi: 10.1016/j.jbi.2004.02.001. URL http://www.sciencedirect.com/science/\article/pii/S1532046404000218.

K. Chiu, Q. Grundy, and L. Bero. 'Spin' in published biomedical literature: A methodological systematic review. *PLoS Biol.*, 2017. doi: 10.1371/journal.pbio.2002173.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995. ISSN 1573-0565. doi: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

J. Diong, A. Butler, S. Gandevia, and M. Héroux. Poor statistical reporting, inadequate data presentation and spin persist despite editorial advice. *PLoS One*, 2018. doi: 10.1371/journal.pone.0202121.

C. Fellbaum. *WordNet: An electronic lexical database (Language, Speech, and Communication)*. Cambridge, MA: The MIT Press, 05 1998.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139, 1997. ISSN 0022-0000. doi: 10.1006/jcss.1997.1504. URL http://www.sciencedirect.com/science/article/pii/S002200009791504X.

J. H. Friedman. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378, Feb. 2002. ISSN 0167-9473. doi: 10.1016/S0167-9473(01)00065-2. URL http://dx.doi.org/10.1016/S0167-9473(01)00065-2.

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Mach. Learn.*, 63(1):3–42, Apr. 2006. ISSN 0885-6125. doi: 10.1007/s10994-006-6226-1.

M. Ghert. The reporting of outcomes in randomised controlled trials: The switch and the spin. *Bone and Joint Research*, 6:600–601, 10 2017. doi: 10.1302/2046-3758.610.BJR-2017-0296.

B. Goldacre, H. Drysdale, A. Powell-Smith, A. Dale, I. Milosevic, E. Slade, P. Hartley, C. Marston, K. Mahtani, and C. Heneghan. The compare trials project. 2016. URL www.COMPare-trials.org.

B. Goldacre, H. Drysdale, A. Dale, I. Milosevic, E. Slade, P. Hartley, C. Marston, A. Powell-Smith, C. Heneghan, and K. R. Mahtani. Compare: a prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*, 20(1):118, Feb 2019. ISSN 1745-6215. doi: 10.1186/s13063-019-3173-2. URL https://doi.org/10.1186/s13063-019-3173-2.

R. Haneef, C. Lazarus, P. Ravaud, A. Yavchitz, and I. Boutron. Interpretation of results of studies evaluating an intervention highlighted in google health news: a cross-sectional study of news. *PLoS ONE*, 2015.

S. Harispe, D. Sánchez, S. Ranwez, S. Janaqi, and J. Montmain. A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. *Journal of Biomedical Informatics*, 48:38 – 53, 2014. ISSN 1532-0464. doi: 10.1016/j.jbi.2013.11.006. URL http://www.sciencedirect.com/science/article/pii/S1532046413001834.

S. Henry, C. Cuffy, and B. T. McInnes. Vector representations of multi-word terms for semantic relatedness. *Journal of Biomedical Informatics*, 77:111 – 119, 2018. ISSN 1532-0464. doi: 10.1016/j.jbi.2017.12.006. URL http://www.sciencedirect.com/science/\article/pii/S1532046417302769.

S. Henry, A. McQuilkin, and B. T. McInnes. Association measures for estimating semantic similarity and relatedness between biomedical concepts. *Artificial Intelligence in*

Medicine, 93:1 – 10, 2019. ISSN 0933-3657. doi: 10.1016/j.artmed.2018.08.006. URL http://www.sciencedirect.com/science/\article/pii/S0933365717304475. Extracting and Processing of Rich Semantics from Medical Texts.

M. Honnibal and M. Johnson. An improved non-monotonic transition system for dependency parsing. In *Proc. of EMNLP 2015*, pages 1373–1378, Lisbon, Portugal, September 2015. ACL. URL https://aclweb.org/anthology/D/D15/D15-1162.

A. Koroleva. Annotated corpus for the relation between reported outcomes and their significance levels, May 2019.

C. Lazarus, R. Haneef, P. Ravaud, and I. Boutron. Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. *BMC Med Res Methodol.*, 2015.

C. Leacock and M. Chodorow. *Combining Local Context and WordNet Similarity for Word Sense Identification*, volume 49, pages 265–. MITP, 01 1998.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*, 2019.

D. Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-556-8. URL http://dl.acm.org/citation.cfm?id=645527.657297.

S. Lockyer, R. Hodgson, J. Dumville, and N. Cullum. "Spin" in wound care research: the reporting and interpretation of randomized controlled trials with statistically non-significant primary outcome results or unspecified primary outcomes. *Trials*, 2013. doi: 10.1186/1745-6215-14-371.

P. Lord, R. Stevens, A. Brass, and C. Goble. Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation. *Bioinformatics*, 19: 1275–1283, 01 2003.

B. T. McInnes, T. Pedersen, and S. V. S. Pakhomov. Umls-interface and umls-similarity : Open source software for measuring paths and semantic similarity. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2009:431–5, 2009.

R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pages 775–780. AAAI Press, 2006. ISBN 978-1-57735-281-5. URL http://dl.acm.org/citation.cfm?id=1597538.1597662.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999792.2999959.

F. P. Miller, A. F. Vandome, and J. McBrewster. *Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau?Levenshtein Distance, Spell Checker, Hamming Distance*. Alpha Press, 2009. ISBN 6130216904, 9786130216900.

S. Pakhomov, B. Mcinnes, T. Adam, Y. Liu, T. Pedersen, and G. B Melton. Semantic similarity and relatedness between clinical terms: An experimental study. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2010:572–6, 11 2010a.

S. Pakhomov, T. Pedersen, B. Mcinnes, G. B Melton, A. Ruggieri, and C. Chute. Towards a framework for developing semantic relatedness reference standards. *Journal of biomedical informatics*, 44:251–65, 10 2010b. doi: 10.1016/j.jbi.2010.10.004.

J. Park, K. Kim, W. Hwang, and D. Lee. Concept embedding to measure semantic relatedness for biomedical information ontologies. *Journal of Biomedical Informatics*, 94:103182, 2019. ISSN 1532-0464. doi: 10.1016/j.jbi.2019.103182. URL http://www.sciencedirect.com/science/\article/pii/S1532046419301005.

T. Pedersen, S. Pakhomov, and S. Patwardhan. Measures of semantic similarity and relatedness in the medical domain. *Journal of Biomedical Informatics - JBI*, 01 2005.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Trans. Systems, Man, and Cybernetics*, 19:17–30, 1989.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding with unsupervised learning. *Technical report, OpenAI*, 2018.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

J. Ratcliff and D. Metzener. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, Jul 1998.

R. Rehurek and P. Sojka. Software framework for topic modelling with large corpora. In *Proc. LREC Workshop on New Challenges for NLP Frameworks*, pages 2216–2219, 2010.

P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1-55860-363-8, 978-1-558-60363-9. URL `http://dl.acm.org/citation.cfm?id=1625855.1625914`.

L. Rokach and O. Maimon. *Data Mining with Decision Trees: Theroy and Applications*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2008. ISBN 9789812771711, 9812771719.

D. Sánchez and M. Batet. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *Journal of Biomedical Informatics*, 44 (5):749 – 759, 2011. ISSN 1532-0464. doi: 10.1016/j.jbi.2011.03.013. URL `http://www.sciencedirect.com/science/article/pii/\S1532046411000645`.

E. Slade, H. Drysdale, and B. Goldacre. Discrepancies between prespecified and reported outcomes. *BMJ*, 2015. URL `http://www.bmj.com/content/351/bmj.h5627/rr-12`.

P. Smith, R. Morrow, and D. Ross. Outcome measures and case definition. In *Field Trials of Health Interventions: A Toolbox. 3rd edition*. OUP Oxford, 2015.

G. Soganciogl u, H. Öztürk, and A. Özgür. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33 14, 2017. doi: 10.1093/bioinformatics/btx238.

I. Spasić and S. Ananiadou. A flexible measure of contextual similarity for biomedical terms. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 197–208, 2004.

C. von der Malsburg. Frank rosenblatt: Principles of neurodynamics: Perceptrons and the theory of brain mechanisms. *Brain Theory*, pages 245–248, 01 1986. doi: 10.1007/978-3-642-70911-1_20.

Y. Wang, N. Afzal, S. Fu, L. Wang, F. Shen, M. Rastegar-Mojarad, and H. Liu. Medsts: A resource for clinical semantic textual similarity. *CoRR*, abs/1808.09397, 2018. URL `http://arxiv.org/abs/1808.09397`.

J. Weston, K. Dwan, D. Altman, M. Clarke, C. Gamble, S. Schroter, P. Williamson, and J. Kirkham. Feasibility study to examine discrepancy rates in prespecified and reported outcomes in articles submitted to the bmj. *BMJ Open*, 2016. doi: 10.1136/bmjopen-2015-010075. URL `http://bmjopen.bmj.com/content/6/4/e010075`.

Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics. doi: 10.3115/981732.981751.

A. Yavchitz, I. Boutron, A. Bafeta, I. Marroun, P. Charles, J. Mantz, and P. Ravaud. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS Med*, 2012.

Chapter 5

Towards Automatic Detection of Primary Outcome Switching in Articles Reporting Clinical Trials. Anna Koroleva, Patrick Paroubek. Submitted

Context

As we stated in the previous chapter, outcome switching (adding new outcomes or excluding pre-defined outcomes without a justification) is one of the most common and alarming types of spin. Outcome switching is the main type of spin addressed by our spin detection pipeline. This chapter is aimed at describing how we identify outcome switching using the algorithms presented in the previous chapters (3 and 4).

Chapter 3 reported on annotating a corpus with primary and reported outcomes and developing rule-based and machine learning algorithms for detecting trial outcomes. The experiments showed that a deep learning approach consisting in fine-tuning pre-trained language models on our annotated corpora obtained the best performance. Based on the results of our experiments, we selected the best fine-tuned model for primary outcome detection (BioBERT) and for reported outcome detection (SciBERT uncased) to be included as a part of the pipeline for spin and outcome switching detection.

Chapter 4 reported on annotating a corpus of pairs of outcomes for semantic similarity and developing algorithms for assessing the similarity of outcomes. The best performance was shown by BioBERT fine-tuned model, which was included into the spin/outcome switching detection pipeline.

The following chapter describes the integration of the two functionalities presented above

into a single algorithm, that is to be used as a decision-supporting system for spin detection. The presented algorithm is the main contribution of this thesis, since it provides an operational solution to help authors and peer reviewers in performing outcome switching detection —a task that, before this PhD project, needed to be done entirely by hand. This thesis is aimed at establishing a proof that automatic spin detection is possible, even though we only address here a subset of the possible types of spin, that is frequent and whose potential negative impact on public health is high.

Authors' contributions

The work reported in this chapter was conducted by AK under supervision of PP. AK was responsible for data collection and analysis. SK took part in the conduct of experiments, as reflected in the Acknowledgements section. AK drafted the manuscript. PP revised the draft critically for important intellectual content.

Abstract

Background: Outcome switching – changing the pre-defined outcomes of a trial – is one of the distorted reporting practices in articles presenting results of clinical trials.

Objective: We present the first approach towards automatic detection of primary outcome switching.

Method: We created a first corpus with outcome switching annotations. We propose to use a combination of information extraction (deep learning), structured data parsing, and phrase similarity estimation techniques to detect outcome switching. We assessed the semantic similarity assessment algorithms on the original corpus and a corpus expanded with variants of referring to an outcome (using abbreviations/their expansions, outcome name or measurement tools name).

Results: The annotated corpus contains 2,000 sentences with 1,694 primary outcomes; 1,940 sentences with 2,251 reported outcomes; 3,043 pairs of outcomes annotated for semantic similarity. Our models achieved the F-measure of 88.42% for primary outcome extraction, 79.42% for reported outcome extraction, and 89.75% and 93.38% for original an expanded versions of the corpus for semantic similarity evaluation.

Conclusions: We proposed the first algorithm for detecting primary outcome switching. The algorithm can be used as an aid tool for authors and peer-reviewers.

Keywords: Outcome Switching, Randomized Controlled Trials, Information Extraction, Natural Language Processing, Semantic Similarity

Background

The variables measured in clinical trials, usually referred to as "outcomes", are one of the most essential elements of a trial. Outcomes are dependent variables, and trials assess the

impact on the outcomes of some other (independent) variables such as the treatment used, dosage, patient characteristics, etc. For scientific soundness, outcomes must be pre-defined before a trial starts and should not be changed without a valid reason, particularly "the primary outcome" – the main goal of the experiment. The outcomes of a trial are directly related to its main objective, e.g. the trial's statistical power and sample size depend on the primary outcome chosen. However, a common problem identified when reviewing clinical trial reports is outcome switching, i.e. the omission of some pre-defined outcomes or the addition of new outcomes (Goldacre et al., 2016).

Outcome switching presents a serious problem because it leads to bias, i.e. reporting only the outcomes that prove the hypothesis of the authors (Slade et al., 2015; Weston et al., 2016; Altman et al., 2017), along with spin, i.e. beautifying of research results by presenting only favourable outcomes (Boutron et al., 2010; Lockyer et al., 2013; Lazarus et al., 2015; Chiu et al., 2017; Diong et al., 2018; Boutron and Ravaud, 2018). It has been proved that the presence of spin can pose a threat to the quality of healthcare: spin makes clinicians overestimate the effects of the treatment in question, and provokes spin in health news and press releases (Boutron et al., 2014; Haneef et al., 2015; Yavchitz et al., 2012), which can influence public perception and expectations regarding the treatment.

Although the problem is known to the medical community (Kay et al., 2012; Delgado and Delgado, 2017; Jones et al., 2018; Goldacre et al., 2019), to the best of our knowledge, no attempts have been made yet to automate the process of checking an article for outcome switching. We report here on the building of a first annotated corpus of sentences from scientific articles from PubMed Central (PMC)[1] with annotations relevant for outcome switching detection and on the development and evaluation of outcome switching detection algorithms in the context of the Methods in Research on Research (MiRoR) project[2], an international multi-disciplinary research project aiming at improving the quality of biomedical research.

Identifying outcomes and detecting outcome switching is a vital task for a range of applications, such as systematic reviews, peer reviewing or scientometrics studies. In our case, the application of interest is detection of spin in texts of articles reporting randomized controlled trials (RCTs) – a clinical trial comparing two interventions (a standard treatment and an experimental one).

Outcome: definition

Outcome are difficult to define precisely as they can be of many different types and their description may comprise several items (Chan et al., 2013) involving lexical, syntactic, semantic

[1] https://www.ncbi.nlm.nih.gov/pmc/
[2] http://miror-ejd.eu/

and pragmatic information:

- outcome name – general term referring to a measure/variable, which can be numerical (*blood pressure*), binary (*death*), or qualitative (*quality of life*).
- name of the measurement tool (*questionnaire*, *score*, etc.) used to measure the outcome, if it cannot be measured directly;
- time points at which the outcome is recorded (*baseline*, *after treatment*, etc.);
- patient-level analysis metrics (*change from baseline*, *final value*, *time to event*);
- population-level method of aggregation (*mean*, *median*, *proportion of patients*).

It is important to distinguish two contexts in which outcome mentions occur (outcomes in the examples are in bold):

- definition of trial outcomes, often specifying the type (primary or secondary):

  Primary outcome will be ***rate of expulsion at 1 year****.*

- reporting of results for an outcome, usually in numerical form or in the form of comparison between two treatments:

  The mean ***serum 25(OH)D*** *increase in the intervention group was 25 ng/ml (range 1-47 ng/ml).*

  SFC is superior to FP in reducing ***airway resistance*** *in mild asthmatics with near normal FEV 1 values.*

State-of-the-art

To our best knowledge, no works has yet tackled the task of outcome switching detection as a whole. However, its parts (outcome extraction and phrase similarity assessment) have been addressed.

Outcome extraction

A number of works approached the task of automatic extraction of outcome information from medical articles. Some of them treated it as a sentence classification task, with Outcome being one of the target classes, along with Population, Intervention, etc. (Kim et al., 2011; Amini et al., 2012; Lui, 2012; Mollá, 2012), or with outcome being the only target class (Demner-Fushman et al., 2006).

Bruijn et al. (2008); Kiritchenko et al. (2010) addressed the extraction of a number of trial design elements, including primary and secondary outcomes. The authors used a two-step approach, combining a classifier to detect relevant sentences for each type of information, and, for most of the information elements: rules combining regular expressions to extract text fragments of interest. However, for outcome there were no extraction rules created, thus, the algorithm is limited to sentence extraction. Precision and recall for sentence extraction are 87% and 90%, respectively.

Summerscales et al. (2009) addressed the extraction of treatments, patient groups and reported outcomes from abstracts of medical articles. Their corpus includes 100 abstracts of articles from BMJ[3]. The authors used a named-entity recognizer based on a Conditional Random Field (CRF) classifier to classify each token as belonging to a certain entity type. While testing different entity sets, the authors noted that entity boundaries can be ambiguous, posing problems for both manual annotation and automatic identification of their precise boundaries. For exact matches, the best result achieved is a recall of 46%, a precision of 59% and an F-measure of 51%. For partial matches, the best result is a recall of 64%, a precision of 82% and an F-measure of 72%.

In a more recent work of 2015, Blake and Lucic (2015) addressed the task of analyzing comparative sentences to extract the entities being compared (the agent and the object) and the ground for comparison, i.e. the characteristic with respect to which the entities are compared, which represents the outcome, or endpoint. Unlike previous work, in this one the authors aim at extracting noun phrases for outcomes. The training set included 100 sentences with 656 noun phrases. Comparative sentences were identified using a set of adjectives and lexico-syntactic patterns. Further, two classifiers, one using an SVM model and the other a generalized linear model (GLM), were built for noun phrase classification, for each element (agent, object, endpoint). For the endpoint detection, SVM achieved an F-measure 0.78 on the training set, but only 0.51 on the test set. The performance was better on shorter sentences (up to 30 words). In a following work, Lucic and Blake (2016) proposed to improve endpoint identification by determining if the head noun of the candidate noun phrase denotes an amount or a measure. The authors used for the training set the same 100 sentences as in their previous work, with enriched annotation. The test set consisted of 132 sentences with up to 40 words, containing 939 noun phrases. The best results achieved were a precision of 56%, a recall of 71% and an F-measure of 62%.

[3]https://www.bmj.com/

Semantic similarity assessment

To our knowledge, assessing the similarity of outcomes has not been addressed yet. For phrase and short text similarity evaluation, there is a substantial volume of work, proposing different approaches, from historical symbolic approaches using semantic networks like WordNet (Miura and Takagi, 2015) to more recent ones using word embeddings (Martinez-Gil, 2012; Kenter and de Rijke, 2015; Torabi Asr et al., 2018). Distributional language models representing terms as high-dimensional vectors proved to be useful for a number of NLP tasks and have recently been successfully applied for semantic similarity assessment in the biomedical domain. Soganciogl et al. (2017) proposed to calculate semantic similarity of sentences using distributed vector representations based on the word2vec (Mikolov et al., 2013) model. Henry et al. (2018) assessed several methods for aggregation of distributional context vectors for multi-word expressions for the task of estimating semantic similarity and relatedness. The authors evaluated methods such as sum or mean of component word vectors, construction of compound vectors, and construction of concept vectors using the MetaMap (Aronson, 2001) tool. Park et al. (2019) proposed a concept-embedding model of a semantic relatedness measure. They used the Unified Medical Language System (UMLS)[4] and Wikipedia as an external data sources for context texts, that were used to build concept vector representations. Cosine similarity between vectors was used to evaluate semantic relatedness of concepts.

Dataset

To our knowledge, there is no open corpus available for the task of outcome switching detection. Still, a very useful data source for this task is provided by trial registries – open online databases where the information about clinical trials, including primary and secondary outcomes, is recorded. Preventing outcome switching is one of the purposes that trial registries serve. Trial registration is mandatory in some countries and encouraged in others. Trials are assigned a unique identification number upon registration, which should be reported when the trial results are published. This way, it is possible to associate the description of the primary and secondary outcomes found in the registry to the outcomes presented in the article. We use the information from trial registries in our experiments as an external additional data source.

Our main experiments are based on analysing the texts of medical articles. An annotated corpus for the task of outcome switching needs to include the following: annotation for text structure (abstract, sections within the abstract), trial registration number, primary and secondary outcomes, reported outcomes; "entity linking" for outcomes (marking chains of

[4]https://www.nlm.nih.gov/research/umls/index.html

outcome mentions that refer to the same outcome). The annotation of such a corpus requires substantial time and effort and annotators with medical expertise familiar with the concept of outcome switching. We initially planned to recruit annotators with the needed expertise. However, running a large-scale annotation project was not feasible because of the unavailability of annotators with the appropriate skills and the cost of the annotation task inherent in its complexity. The annotation presented here was done by one researcher with NLP expertise, consulted by medical advisors. It constitutes a minimal corpus for bootstrapping the first experiments for outcome switching detection that we present here.

Our initial corpus comprises 3,938 articles from PMC, that have the PubMed publication type RCT. In our current work, we focused on detection switching of the primary outcome, as it is the most important outcome in clinical trials. We annotated subsets of sentences from our corpus for each of the following tasks:

1. Primary outcome annotation: we randomly selected 2,000 sentences containing the word *outcome* or its contextual synonyms (*end-point*, *measure*, etc.) and the word *primary* or its contextual synonyms (*principal*, *main*, etc.), where the latter precedes the former with the distance no more than 3 tokens. Sentence splitting was performed with the help of Python NLTK library[5]. Search for the terms was done with regular expressions.

   We decided to annotate only the sentences where the context explicitly defines the primary outcome. In some types of statements (statements of objectives, description of measured variables), information about the primary outcome can be expressed implicitly and can potentially be inferred from the context, e.g.:

   This study investigated ***the efficacy (trough forced expiratory volume in 1 second [FEV1] response) and safety*** *of additional treatment with once-daily tiotropium 18g via the HandiHaler in a primary care COPD population. Secondary endpoints included: trough forced vital capacity (FVC) response, weekly use of rescue short-acting beta-agonist, and exacerbation of COPD (complex of respiratory symptoms/events of >3 days in duration requiring a change in treatment).*

   Response *was assessed quantitatively through health status questionnaires, measures of breathing control, exercise capacity and physical activity and qualitatively, through structured interviews with a clinical psychologist.*

   Our objective was to evaluate ***the feasibility*** *of comparing these two modes in a randomized trial.*

   However, after consulting our medical advisors, we decided not to include these types of sentences in our corpus as they do not necessarily define a (primary) outcome; besides,

[5]http://www.nltk.org

absence of explicit primary outcome definition does not conform with proper trial reporting practices, reflected in reporting guidelines (Schulz et al., 2010), and may itself be a marker of outcome switching.

A definition of a primary outcome can be complex and include timepoints, measurement tool, etc. For each instance, we decided to annotate the longest continuous text span possible containing all the relevant information about an outcome in a given trial. In our corpus, we identified 1,694 primary outcome occurrences.

2. Reported outcomes annotation: we selected the Results and Conclusions sections of the abstracts of the articles for which we annotated the primary outcomes. Search for the abstract in the article and the subdivision of the abstract into sections was done using a set of rules and regular expressions. Some of the articles in our corpus turned out to be not RCT reports but protocols and were missing Results and Conclusion sections.

   Ways of reporting an outcome are varied; most typically, it is by means of a noun phrase, but it can also be done using a verb phrase (*"10% of patients **died**"*) or an adjective (*"Treatment was **cost-effective**"*). It can be difficult to choose a text span for annotation, e.g. in a phrase *"The difference between group in increase of X at 1 year was observed"*, all the phrases *"the difference between group in increase of X at 1 year"*, *"the difference between group in increase of X"*, *"increase of X"* and *"X"* can be considered to be reported outcomes. As it seems reasonable to consider that in the phrase *"X increased more in group A than in group B at 1 year"* the reported outcome is the same as in the first example, we decided to annotate the shortest possible text span for reported outcomes, i.e. only the outcome or measurement tool name (*"X"* for the discussed example).

   2,251 reported outcomes were annotated.

3. Annotation of relations between a primary outcome and a corresponding reported outcome: we created a set of pairs of sentences, where the first sentence comes from the corpus with annotated primary outcomes, the second sentence comes from the corpus with annotated reported outcomes, and both sentences belong to same text. We considered outcomes to be similar if the outcome or measurement tool are the same, disregarding timepoints, analysis metrics, etc. This approach is arguable, since the difference in timepoints between two otherwise similar outcomes can lead consider them as different outcomes. However, detecting a change in timepoints is a separate task, thus we did not mark this information, considering it to be a part of the future work.

   In total, 3,043 pairs of outcomes were annotated: 701 (612 after deduplication) pairs were

considered to be "similar" and 2,342 (2,187 after deduplication) pairs were considered not to be "similar".

There can be several ways of referring to the same outcome. First, abbreviations can be used instead of full outcome names: e.g. the outcome "patient-perceived recovery (PPR)" can be referred to as "patient-perceived recovery" or "PPR". Besides, measurement tool name can be used to refer to the outcome it was used to measure: e.g. the outcome "perceived stress, as measured by the Perceived Stress Scale" can be referred to as "perceived stress" or "the Perceived Stress Scale". This variability is expected to pose difficulties for semantic similarity assessment. In an attempt to account for it, we created an expanded version of the outcome pairs corpus. First, using regular expression rules, we extracted abbreviation expansions and replaced the abbreviations by the expansions. For an outcome outcome mention "PPR" we obtained a variant "patient-perceived recovery". Second, using regular expressions for phrases such as "measured/assessed as/by", we extracted outcome names and measurement tool names. For the outcome "perceived stress, as measured by the Perceived Stress Scale", we obtained variants "perceived stress" or "the Perceived Stress Scale". These new variants were used to add new pairs of similar/dissimilar to the corpus, thus forming the expanded corpus. We ran our experiments on both the initial and the expended versions of the corpus.

Our annotated corpora are freely available (Koroleva, 2019a,b).

Methods

Our proposed algorithm for outcome switching detection comprises three steps:

1. Extraction of entities (primary and reported outcomes) from the text of a given article;
2. Extraction the information about outcomes from trial registries;
3. Semantic similarity assessment of pairs of outcomes, taking the outcomes from the first step as input.

Deep learning methods

Our experiments on outcome extraction and semantic similarity assessment of outcomes are described in detail elsewhere. Here we briefly present the best performing method, selected for the implementation in our spin detection pipeline, released as open source code (available at: https://github.com/aakorolyova/DeSpin). Our selected approach is based on the use of

language models pre-trained on large corpora, that are further fine-tuned on a limited amount of annotated data for a downstream task. Approaches using pre-trained language models, such as ELMO (Peters et al., 2018), OpenAI's GPT (Radford et al., 2018) and Google's BERT (Devlin et al., 2018), achieve the state-of-the-art (SOTA) performance, or even outperform the current SOTA, for a number of tasks, such as sentence or sentence pair classification, entity extraction, natural language inference.

We chose to use the BERT-based approach as it was shown to achieve better results on relevant tasks than ELMO and OpenAI GPT (Devlin et al., 2018). BERT (Bidirectional Encoder Representations from Transformers) language models are based on a masked language model (MLM), that randomly masks some input tokens, thus pre-training a deep bidirectional transformer on both left and right contexts. Token representation used in BERT consists of the token itself, and segment and position embeddings. BERT models were pre-trained on a general-domain corpus of 3.3B tokens, combining BooksCorpus and English Wikipedia. BERT models can be easily fine-tuned on a downstream task by adding one additional output layer.

Two domain-specific versions of BERT were recently released: BioBERT (Lee et al., 2019), pre-trained on the BERT training data and a biomedical corpus of 18B tokens from PubMed abstracts and PMC full-text articles; and SciBERT (Beltagy et al., 2019), pre-trained on the BERT training data and a corpus of scientific (including biomedical) articles of 3.1B tokens.

BERT and SciBERT provide several versions of models: cased (input data unaltered) and uncased (input data lower-cased); BioBERT has only a cased model. SciBERT provides two versions of vocabulary: general-domain and scientific (recommended by the authors). In our experiments, we compared the performance of BERT cased and uncased models, BioBERT model, and SciBERT cased and uncased models with the scientific vocabulary.

Primary and reported outcome extraction was modelled as sequence labelling task. We compared two approaches: a simple fine-tuning of pre-trained models, proposed by (Devlin et al., 2018; Lee et al., 2019); and an approach employing a CRF on top of a Bi-LSTM applied to token embeddings, as proposed by Beltagy et al. (2019). We used BioBERT[6] and SciBERT[7] code for named entity recognition.

Outcome similarity assessment was modelled similarly to sentence pair classification task[8]. We used the fine-tuning approach for this task. Our similarity assessment algorithm assumes that the pairs of outcomes have been previously extracted and are given as input.

We assessed the models using 10-fold cross-validation (splitting the data into 10 train-dev-test sets), we report the averaged results.

[6]https://github.com/dmis-lab/biobert

[7]https://github.com/allenai/scibert

[8]We used BERT code for sentence air classification: run_classifier.py from https://github.com/google-research/bert

Trial registry data extraction

An additional step in our algorithm is extraction outcome data from trial registries. We extract trial registration numbers with the use of regular expressions. Registration numbers were found in 2796 (71%) text of our initial corpus. We further download and parse registry entries for a given trial, obtaining data about primary and secondary outcomes. These data is used in an additional level of checking outcome switching: besides checking coherence of outcome reporting within the article text (primary outcomes defined in the article vs. reported outcome in the article), we also check the coherence between the registry and text (primary outcomes defined in the registry vs. primary outcomes defined in the article; primary outcomes defined in the registry vs. reported outcomes in the article).

Results

Entity extraction

The results of extraction outcomes using BERT-based models are shown in Tables 5.1 and 5.2. The suffix "_biLSTM-CRF" in model names denoted the approach using CRF and Bi-LSTM, model names without the suffix refer to the fine-tuning approach.

The fine-tuning approach consistently outperformed the CRF+BiLSTM approach. The BioBERT model showed the best performance for primary outcome extraction (F-measure = 88.42%), while the SciBERT uncased model showed the best performance for reported outcome extraction (F-measure = 79.42%). Overall, models pre-trained on domain-specific data (BioBERT an SciBERT) outperform BERT, trained on a general-domain corpus.

Algorithm	**Precision**	**Recall**	**F1**
BioBERT	86.99	90.07	88.42
SciBERT-cased	87.52	89.07	88.21
SciBERT-uncased	87.49	88.92	88.1
BERT-cased	86.96	87.41	87.14
BERT-uncased	86.6	87.39	86.91
BioBERT_biLSTM-CRF	78.82	82	80.34
SciBERT-uncased_biLSTM-CRF	77.52	81.15	79.22
SciBERT-cased_biLSTM-CRF	77.23	80.89	78.95
BERT-cased_biLSTM-CRF	78.2	78.84	78.47
BERT-uncased_biLSTM-CRF	76.72	78.73	77.63

Table 5.1: Primary outcome extraction: results

It is difficult to directly compare our results with the results of previous research on outcome extraction: some of the previous works addressed extraction of outcome information as

Algorithm	Precision	Recall	F1
SciBERT-uncased	81.17	78.09	79.42
BioBERT	79.61	77.98	78.6
SciBERT-cased	79.6	77.65	78.38
BERT-uncased	78.98	74.96	76.7
BERT-cased	76.63	74.25	75.18
SciBERT-uncased_biLSTM-CRF	68.44	73.47	70.77
BioBERT_biLSTM-CRF	70.18	71.43	70.63
SciBERT-cased_biLSTM-CRF	67.98	72.52	70.11
BERT-cased_biLSTM-CRF	65.98	65.54	65.64
BERT-uncased_biLSTM-CRF	64.6	66.73	65.4

Table 5.2: Reported outcome extraction: results

sentence classification task only, without actual extraction of outcomes; other works chose a particular type of sentences (such as comparative sentences) to annotate and extract outcomes. In other cases, annotation conventions are not reported in sufficient detail to compare them with our annotation (e.g. whether verb phrases and adjectives were annotated as outcomes). However, our system achieves operational performance for outcome extraction, outperforming the comparable previous works.

Semantic similarity assessment

Tables 5.3 and 5.4 show the results of semantic similarity assessment of pairs of outcomes on the original and expanded versions of the corpus. The BioBERT fine-tuned model showed the best performance on both versions of the corpus; both BioBERT and SciBERT outperform BERT.

The results of all the models on the expanded corpus are better than the results of the corresponding models on the original corpus, which proves that addition of outcomes variants to the corpus improves the fine-tuning performance.

Algorithm	Precision	Recall	F1
BioBERT	88.93	90.76	89.75
SciBERT-uncased	87.99	90.78	89.3
SciBERT-cased	87.31	91.53	89.3
BERT-uncased	85.76	88.15	86.8
BERT-cased	83.36	85.2	84.21

Table 5.3: Results of BERT-based systems on the original corpus

Algorithm	Precision	Recall	F1
BioBERT	92.98	93.85	93.38
SciBERT-uncased	91.3	91.79	91.51
SciBERT-cased	89	92.54	90.69
BERT-uncased	89.31	89.12	89.16
BERT-cased	88.25	90.1	89.12

Table 5.4: Results of BERT-based systems on the expanded corpus

Proposed algorithm

Table 5.5 summarizes the results of the best-performing models for each task. Fine-tuned BioBERT model achieved the best results for primary outcome extraction (F-measure = 88.42%) and semantic similarity assessment (F-measure = 89.75% for the initial corpus and 93.38% for the expanded corpus). For reported outcome extraction, SciBERT uncased model showed the best performance (F-measure = 79.4%). These models are used in our implementation of an outcome switching detection algorithm, which forms a part of spin detection pipeline. The pipeline also incorporates the algorithm for extraction the trial registration number, and accessing, downloading and parsing the outcome data from the trial registry.

Our implementation of an outcome switching detection algorithm includes the following steps: extracting primary and reported outcomes from the article, extracting primary outcomes from the registry, assessing semantic similarity between pairs of outcomes (primary outcomes in the registry vs. primary outcomes in the article; primary outcomes in the registry vs. reported outcome in the abstract; primary outcomes in the article vs. reported outcome in the abstract). The system reports the unmatched primary outcomes. The source code, implemented in Python, is available at https://github.com/aakorolyova/DeSpin.

Algorithm	Model	Precision	Recall	F1
Primary outcomes extraction	BioBERT	86.99	90.07	88.42
Reported outcomes extraction	SciBERT-uncased	81.17	78.09	79.42
Outcome similarity assessment (initial corpus)	BioBERT	88.93	90.76	89.75
Outcome similarity assessment (expanded corpus)	BioBERT	92.98	93.85	93.38

Table 5.5: The best achieved results

Discussion

Possible use of the algorithm

We envisage the usage of our system as a semi-automated aid tool, combining the described algorithms with the functionality for manual annotation of text spans and relations between them. This way, the tool would be able to use annotated data from users for cases when our algorithms make errors, which has two benefits: first, the performance of the system based on the data provided by users will be higher than the default performance based on the outcome extraction algorithms; second, if users agree to provide results of their annotation to us, this would allow to obtain more manually annotated data and to further use machine-learning algorithms with more data to make the results more accurate.

Limitations

Our work has some limitations. First of all, the annotation was performed by a single annotator due to infeasibility of running a large annotation project involving several annotators. We hope that this issue can be resolved in the future by collecting annotations from the users of the system.

Another limitation is the step-by-step approach that we used in our current implementation, with the semantic similarity assessment algorithm taking the results of the outcome extraction as input. We chose this approach for number of reasons: it is relatively simple; it is a standard approach for similar tasks; models developed for each step (primary outcome extraction, reported outcomes extraction, outcome similarity assessment) can be used separately for other tasks and applications. However, the final performance of this approach can be hindered by the errors of included algorithms. A possible direction of the future work, addressing this problem, could be an implementation of an end-to-end system for outcome switching detection.

Future work

The most challenging task in outcome switching detection pipeline appears to be the extraction of reported outcomes, currently showing the lowest performance among all the algorithms. Thus, our future work will aim at improving reported outcome extraction.

Another major objective for the future work is adding algorithms for extraction secondary outcomes and checking article for secondary outcomes switching.

Conclusion

Outcome switching is a common and important problem in clinical trials reporting that can lead to overestimation of the treatment studied and thus can have a detrimental effect on the clinical practice. Switching (non-reporting or change) of the primary outcome of a clinical trial is an issue of particular importance as it can cause erroneous interpretation by the trial results by clinicians and health media.

Assessment for outcome switching is now performed manually. Automated or semi-automated systems were proved to be useful in various tasks of text analysis. Outcome switching can be addressed as a task combining entity extraction and semantic similarity assessment.

In this paper, we proposed a first implementation for an outcome switching detection algorithm. Our proposed pipeline includes algorithms for extraction of primary and reported outcomes, extracting trial registry data, and evaluation of similarity between the extracted outcomes. The algorithms achieve operational results and can be further improved over time with more data collected. The developed algorithms and models are a part of our spin detection system and are freely available at: https://github.com/aakorolyova/DeSpin.

Acknowledgements

We thank prof. Isabelle Boutron from the University Paris Descartes, prof. Patrick Bossuyt from the University of Amsterdam, and Dr. Liz Wager from SideView who provided valuable insight and expertise as our advisors in the domain of reporting of clinical trials.

We thank Sanjay Kamath from LRI, Université Paris-Sud for his help in conducting experiments with BERT.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

References

D. Altman, D. Moher, and K. Schulz. Harms of outcome switching in reports of randomised trials: Consort perspective. *BMJ: British Medical Journal (Online)*, 2017.

I. Amini, D. Martinez, and D. Molla. Overview of the alta 2012 shared task. In *Australasian L. T. Association Workshop*, page 124, 2012.

A. Aronson. Effective mapping of biomedical text to the umls metathesaurus: The metamap program. *AMIA Annual Symposium*, 2001:17–21, 02 2001.

I. Beltagy, A. Cohan, and K. Lo. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

C. Blake and A. Lucic. Automatic endpoint detection to support the systematic review process. *J. Biomed. Inform*, 2015.

I. Boutron and P. Ravaud. Misrepresentation and distortion of research in biomedical literature. *Proc Natl Acad Sci U S A.*, 2018. doi: 10.1073/pnas.1710755115.

I. Boutron, S. Dutton, P. Ravaud, and D. Altman. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*, 2010.

I. Boutron, D. Altman, S. Hopewell, F. Vera-Badillo, I. Tannock, and P. Ravaud. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *Journal of Clinical Oncology*, 2014.

B. D. Bruijn, S. Carini, S. Kiritchenko, J. Martin, and I. Sim. Automated information extraction of key trial design elements from clinical trial publications. In *Proceedings of the AMIA Annual Symposium*, 2008.

A.-W. Chan, J. Tetzlaff, D. Altman, A. Laupacis, P. Gøtzsche, K. Krleža-Jerić, A. Hróbjartsson, H. Mann, K. Dickersin, J. Berlin, C. Doré, W. Parulekar, W. Summerskill, T. Groves, K. Schulz, H. Sox, F. Rockhold, D. Rennie, and D. Moher. Spirit 2013 statement: Defining standard protocol items for clinical trials. *Ann Intern Med*, 2013.

K. Chiu, Q. Grundy, and L. Bero. 'Spin' in published biomedical literature: A methodological systematic review. *PLoS Biol.*, 2017. doi: 10.1371/journal.pbio.2002173.

A. F. Delgado and A. F. Delgado. Outcome switching in randomized controlled oncology trials reporting on surrogate endpoints: a cross-sectional analysis. In *Scientific Reports*, 2017.

D. Demner-Fushman, B. Few, S. Hauser, and G. Thoma. Automatically identifying health outcome information in medline records. *Journal of the American Medical Informatics Association*, 2006.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

J. Diong, A. Butler, S. Gandevia, and M. Héroux. Poor statistical reporting, inadequate data presentation and spin persist despite editorial advice. *PLoS One*, 2018. doi: 10.1371/journal.pone.0202121.

B. Goldacre, H. Drysdale, A. Powell-Smith, A. Dale, I. Milosevic, E. Slade, P. Hartley, C. Marston, K. Mahtani, and C. Heneghan. The compare trials project. 2016. URL `www.COMPare-trials.org`.

B. Goldacre, H. Drysdale, A. T. Dale, I. Milosevic, E. S. Slade, P. D. Hartley, C. A. Marston, A. Powell-Smith, C. J. Heneghan, and K. R. Mahtani. Compare: a prospective cohort study correcting and monitoring 58 misreported trials in real time. In *Trials*, 2019.

R. Haneef, C. Lazarus, P. Ravaud, A. Yavchitz, and I. Boutron. Interpretation of results of studies evaluating an intervention highlighted in google health news: a cross-sectional study of news. *PLoS ONE*, 2015.

S. Henry, C. Cuffy, and B. T. McInnes. Vector representations of multi-word terms for semantic relatedness. *Journal of Biomedical Informatics*, 77:111 – 119, 2018. ISSN 1532-0464. doi: 10.1016/j.jbi.2017.12.006. URL `http://www.sciencedirect.com/science/\article/pii/S1532046417302769`.

C. W. Jones, B. S. Misemer, T. F. Platts-Mills, R. Ahn, A. Woodbridge, A. Abraham, S. Saba, D. Korenstein, E. Madden, and S. Keyhani. Primary outcome switching among drug trials with and without principal investigator financial ties to industry: a cross-sectional study. *BMJ Open*, 8(2), 2018. ISSN 2044-6055. doi: 10.1136/bmjopen-2017-019831. URL `https://bmjopen.bmj.com/content/8/2/e019831`.

A. Kay, J. D. Higgins, A. G. Day, R. M. Meyer, and C. Booth. Randomized controlled trials in the era of molecular oncology: methodology, biomarkers, and end points. *Annals of oncology : official journal of the European Society for Medical Oncology*, 23 6:1646–51, 2012.

T. Kenter and M. de Rijke. Short text similarity with word embeddings. In *CIKM*, 2015.

S. Kim, D. Martinez, L. Cavedon, and L. Yencken. Automatic classification of sentences to support evidence based medicine. *BMC bioinformatics*, 12 Suppl 2:S5, 03 2011. doi: 10.1186/1471-2105-12-S2-S5.

S. Kiritchenko, B. D. Bruijn, S. Carini, J. Martin, and I. Sim. Exact: automatic extraction of clinical trial characteristics from journal publications. *BMC Med Inform Decis Mak.*, 2010.

A. Koroleva. Annotated corpus for primary and reported outcomes extraction, May 2019a. URL `https://doi.org/10.5281/zenodo.3234811`.

A. Koroleva. Annotated corpus for semantic similarity of clinical trial outcomes, May 2019b.

C. Lazarus, R. Haneef, P. Ravaud, and I. Boutron. Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. *BMC Med Res Methodol.*, 2015.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*, 2019.

S. Lockyer, R. Hodgson, J. Dumville, and N. Cullum. "Spin" in wound care research: the reporting and interpretation of randomized controlled trials with statistically non-significant primary outcome results or unspecified primary outcomes. *Trials*, 2013. doi: 10.1186/1745-6215-14-371.

A. Lucic and C. Blake. Improving endpoint detection to support automated systematic reviews. In *AMIA Annu Symp Proc.*, 2016.

M. Lui. Feature stacking for sentence classification in evidence-based medicine. In *Proc. of Australasian L. T. Association Workshop*, 2012.

J. Martinez-Gil. An overview of textual semantic similarity measures based on web intelligence. *Artificial Intelligence Review*, 42, 12 2012. doi: 10.1007/s10462-012-9349-8.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc. URL `http://dl.acm.org/citation.cfm?id=2999792.2999959`.

N. Miura and T. Takagi. WSL: Sentence similarity using semantic distance between words. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*,

pages 128–131, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-2023. URL https://www.aclweb.org/anthology/S15-2023.

D. Mollá. Experiments with clustering-based features for sentence classification in medical publications: Macquarie test's participation in the ALTA 2012 shared task. In *Proc. of Australasian L. T. Association Workshop*, pages 139–142, Dunedin, New Zealand, Dec. 2012. URL https://www.aclweb.org/anthology/U12-1020.

J. Park, K. Kim, W. Hwang, and D. Lee. Concept embedding to measure semantic relatedness for biomedical information ontologies. *Journal of Biomedical Informatics*, 94:103182, 2019. ISSN 1532-0464. doi: 10.1016/j.jbi.2019.103182. URL http://www.sciencedirect.com/science/\article/pii/S1532046419301005.

M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018. doi: 10.18653/v1/n18-1202. URL http://dx.doi.org/10.18653/v1/N18-1202.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding with unsupervised learning. *Technical report, OpenAI*, 2018.

K. F. Schulz, D. G. Altman, and D. Moher. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340, 2010. ISSN 0959-8138. doi: 10.1136/bmj.c332. URL https://www.bmj.com/content/340/bmj.c332.

E. Slade, H. Drysdale, and B. Goldacre. Discrepancies between prespecified and reported outcomes. *BMJ*, 2015. URL http://www.bmj.com/content/351/bmj.h5627/rr-12.

G. Soganciogl u, H. Öztürk, and A. Özgür. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33 14, 2017. doi: 10.1093/bioinformatics/btx238.

R. Summerscales, S. Argamon, J. Hupert, and A. Schwartz. Identifying treatments, groups, and outcomes in medical abstracts. In *Proceedings of the Sixth Midwest Computational Linguistics Colloquium (MCLC)*, 2009.

F. Torabi Asr, R. Zinkov, and M. Jones. Querying word embeddings for similarity and relatedness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*,

pages 675–684, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1062. URL https://www.aclweb.org/anthology/N18-1062.

J. Weston, K. Dwan, D. Altman, M. Clarke, C. Gamble, S. Schroter, P. Williamson, and J. Kirkham. Feasibility study to examine discrepancy rates in prespecified and reported outcomes in articles submitted to the bmj. *BMJ Open*, 2016. doi: 10.1136/bmjopen-2015-010075. URL http://bmjopen.bmj.com/content/6/4/e010075.

A. Yavchitz, I. Boutron, A. Bafeta, I. Marroun, P. Charles, J. Mantz, and P. Ravaud. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS Med*, 2012.

Chapter 6

Extracting relations between outcomes and significance levels in Randomized Controlled Trials (RCTs) publications.

Anna Koroleva, Patrick Paroubek. Proceedings of ACL BioNLP Workshop 2019, Florence, Italy, August 2019

Context

As shown by the chapter 5, we established that automatic detection of spin can be feasible for the problem of outcome switching. After that, we can start looking at some complementary issues concerning spin detection, one of them being the significance level of an outcome.

The presence of spin is often related to significance levels of trial outcomes. In particular, spin is often found in clinical trials with non-significant results for the primary outcome; besides, spin can consist in selecting only significant outcomes for presentation in the abstract of an article. Hence, identifying significance levels for trial outcomes could be helpful as a part of a spin detection pipeline. We regard this task as a complementary one as it does not directly contribute to detection of any type of spin for the moment, but it extracts the information related to an important risk factor for spin (non-significant outcomes) that can help human experts in spin detection.

This chapter describes our work on extracting the relation between reported trial outcomes and significance levels. As a part of the relation extraction algorithm, we used the algorithms

of reported outcome extraction presented in the chapter 3. This chapter reports on annotating a corpus and creating machine learning algorithms for relation extraction. The best performing algorithm was included into our spin detection prototype.

Authors' contributions

The work reported in this chapter was conducted by AK under supervision of PP. AK was responsible for data collection and analysis. SK took part in the conduct of experiments, as reflected in the Acknowledgements section. AK drafted the manuscript. PP revised the draft critically for important intellectual content.

Abstract

Randomized controlled trials assess the effects of an experimental intervention by comparing it to a control intervention with regard to some variables - trial outcomes. Statistical hypothesis testing is used to test if the experimental intervention is superior to the control. Statistical significance is typically reported for the measured outcomes and is an important characteristic of the results. We propose a machine learning approach to automatically extract reported outcomes, significance levels and the relation between them. We annotated a corpus of 663 sentences with 2,552 outcome - significance level relations (1,372 positive and 1,180 negative relations). We compared several classifiers, using a manually crafted feature set, and a number of deep learning models. The best performance (F-measure of 94%) was shown by the BioBERT fine-tuned model.

Introduction

In clinical trials, outcomes are the dependent variables that are monitored to assess how they are influenced by other, independent, variables (treatment used, dosage, patient characteristics). Outcomes are a central notion for clinical trials.

To assess the impact of different variables on the outcomes, statistical hypothesis testing is commonly used, giving an estimation of statistical significance – the likelihood that a relationship between two or more variables is caused by something other than a chance (Schindler, 2015). Statistical significance levels are typically reported along with the trial outcomes as p-values, with a certain set threshold, where a p-value below the threshold means that the results are statistically significant, while a p-value above the threshold presents non-significant results. Hypothesis testing in clinical trials is used in two main cases:

1. In a trial comparing several treatments given to different groups of patients, a difference in value of an outcome observed between the groups at the end of the trial is evaluated by hypothesis testing to determine if the difference is due to the difference in medication.

If the difference is statistically significant, the null hypothesis (the difference between treatments is due to a chance) is rejected, i.e. the superiority of one treatment over the other is considered to be proved.

2. When an improvement of an outcome is observed within a group of patients taking a treatment, hypothesis testing is used to determine if the difference in the outcome at different time points within the group is due to the treatment. If the results are statistically significant, it is considered to be proven that the treatment has a positive effect on the outcome in the given group of patients.

Although p-values are often misused and misinterpreted (Head et al., 2015), extracting significance levels for trial outcomes is still vital for a number of tasks, such as systematic reviews, detection of bias and spin. In particular, our application of interest is automatic detection of spin, or distorted reporting of research results, that consists in presenting an intervention studied in a trial as having higher beneficial effects than the research has proved. Spin is an alarming problem in health care as it causes overestimation of the intervention by clinicians (Boutron et al., 2014) and unjustified positive claims regarding the intervention is health news and press releases (Haneef et al., 2015; Yavchitz et al., 2012).

Spin is often related to a focus on significant outcomes, and occurs when the primary outcome (the main variable monitored during a trial) is not significant. Thus, to detect spin, it is important to identify the significance of outcomes, and especially of the primary outcome. To our best knowledge, no previous work addressed the extraction of the relation between outcomes and significance levels. In this paper, we present our approach towards extracting outcomes, significance levels and relations between them, that can be incorporated into a spin detection pipeline.

State of the art

Extraction of outcome - significance level relations consists of two parts: entity extraction (reported outcomes and significance levels) and extraction of the relationship between the entities. In this section, we present the previous works on these or similar tasks.

Entity extraction

The number of works addressing automatic extraction of significance levels is limited.

Hsu et al. (2012) used regular expressions to extract statistical interpretation, p-values, confidence intervals, and comparison groups from sentences categorized as "outcomes and

estimation". The authors report precision of 93%, recall of 88% and F-measure of 90% for this type of information.

Chavalarias et al. (2016) applied text mining to evaluate the p-values reported in the abstracts and full texts of biomedical articles published in 1990 – 2015. The authors also assessed how frequently statistical information is presented in ways other than p-values. P-values were extracted using a regular expression; the system was evaluated on a manually annotated dataset. The reported sensitivity (true positive rate) is 96.3% and specificity (true negative rate) is 99.8%. P-values and qualitative statements about significance were more common ways of reporting significance than confidence intervals, Bayes factors, or effect sizes.

A few works focused on extracting outcome-related information, addressing it either as a sentence classification, or as entity extraction task.

Demner-Fushman et al. (2006) defined an outcome as *"The sentence(s) that best summarizes the consequences of an intervention"* and thus adopted a sentence classification approach to extract outcome-related information from medical articles, using a corpus of 633 MEDLINE citations. The authors tested Naive Bayes, linear SVM and decision-tree classifiers. Naive Bayes showed the best performance. The reported classification accuracy ranged from 88% to 93%.

One of the notable recent works addressing outcome identification as an entity extraction task, rather than sentence classification, is (Blake and Lucic, 2015). The authors addressed a particular type of syntactic constructions – comparative sentences – to extract three items: the compared entities, referred to as the agent and the object, and the ground for comparison, referred to as the endpoint (synonymous to outcome). The aim of this work was to extract corresponding noun phrases. The dataset was based on full-text medical articles and included only the sentences that contain all the three entities (agent, object and endpoint). The training set comprised 100 sentences that contain 656 noun phrases. The algorithm proceeds in two steps: first, comparative sentences are detected with the help of a set of adjectives and lexico-syntactic patterns. Second, the noun phrases are classified according to their role (agent, object, endpoint) using SVM and generalized linear model (GLM). On the training set, SVM showed better performance than GLM, with an F-measure of 78% for the endpoint. However, on the test set the performance was significantly lower: SVM showed an F-measure of only 51% for the endpoint. The performance was higher on shorter sentences (up to 30 words) than on the longer ones.

A following work (Lucic and Blake, 2016) aimed at improving the recognition of the first entity and of the endpoint. The authors propose to use in the classification the information on whether the head noun of the candidate noun phrase denotes an amount or a measure. The annotation of the corpus was enriched by the corresponding information. As a result, precision

of the endpoint detection improved to 56% on longer sentences and 58% on shorter ones; recall improved to 71% on longer sentences and 74% on shorter ones.

Relation extraction

To our knowledge, extraction of the relation between outcomes and significance levels has not been addressed yet. In this section, we overview some frameworks for relation extraction and outline some common features of different approaches in the biomedical relation extraction.

A substantial number of works addressed extracting binary relations, such as protein-protein interactions or gene-phenotype relation, or complex relations, such as biomolecular events. A common feature of the works in this domain, noted by Zhou et al. (2014); Lever and Jones (2017) and still relevant for recent works e.g. Peng and Lu (2017); Asada et al. (2017), consists in assuming that entities of interest are already extracted and provided to the relation extraction system as input. Thus, the relation extraction is assessed separately, without taking into account the performance of entity extraction. We adopt this approach for relation extraction evaluation in our work, but we provide separate assessment for our algorithms of entity extraction.

One of the general frameworks for relation extraction in the biomedical domain is proposed by Zhou et al. (2014). The authors suggest using trigger words to determine the type of a relation, noting that for some relation types trigger words can be extracted simply with a dictionary, while for other types, rule-based or machine-learning approaches may be required. For relation extraction, rule-based methods can be applied, often employing regular expressions using words or POS tags. Rules can be crafted manually or learned automatically. The machine learning approaches to binary relation extraction, as the authors note, usually treat the task as a classification problem. Features for classification often use output of textual analysis algorithms such as POS-tagging and syntactic parsing. Machine learning approaches can be divided into feature-based approaches (using syntactic and semantic features) and kernel approaches (calculating similarity between input sequences based on string or syntactic representation of the input). Supervised machine learning is a highly successful approach for binary relation extraction, but its main drawback consists in the need of large amount of annotated data.

A framework for pattern-based relation extraction is introduced by Peng et al. (2014). The approach aims at reducing the need for manual annotation. The approach is based on a user-provided list of trigger words and specifications (the definition of arguments for each trigger). Variations of lexico-syntactic patterns are derived using this information and are matched with the input text, detecting the target relations. Some interesting features of the framework include the following: the use of text simplification to avoid writing rules for all

existing constructions; the use of referential relations to find the best phrase referring to an entity. The authors state that their system is characterized by good generalizability due to the use of language properties and not of task-specific knowledge.

A recent work (Björne and Salakoski, 2018) reports on the development of convolutional neural networks (CNNs) for event and relation extraction, using Keras (Chollet et al., 2015) with Tensorflow backend (Abadi et al., 2016). Parallel convolutional layers process the input, using sequence windows centered around the candidate entity, relation or event. Vector space embeddings are built for input tokens, including features such as word vectors, POS, entity features, relative position, etc. The system was tested on several tasks and showed improved performance and good generalizability.

Our dataset

Corpus creation and annotation

In our previous work on outcome extraction, we manually annotated a corpus for reported outcomes comprising 1,940 sentences from the Results and Conclusions sections of PMC article abstracts. We used this corpus as a basis for a corpus with annotations for outcome —significance level relations.

Our corpus contains 2,551 annotated outcomes. Out of the sentences with outcomes, we selected those where statistical significance levels are supposedly reported (using regular expressions) and manually annotated relations between outcomes and significance levels. The annotation was done by one annotator (AK), in consultation with a number of domain experts, due to infeasibility of recruiting several annotators with sufficient level of expertise within a reasonable time frame.

The final corpus contains 663 sentences with 2,552 annotated relations, out of which 1,372 relations are "positive" (the significance level is related to the outcome) and 1,180 relations are "negative" (the significance level is not related to the outcome). The corpus is publicly available (Koroleva, 2019).

Data description

There are three types of data relevant for this work: outcomes, significance levels, and relationship between them. In this section, we describe these types of data and the observed variability in the ways of presenting them.

1. Outcomes

A trial outcome is, in broad sense, a measure or variable monitored during a trial. It can be binary (presence of a symptom or state), numerical ("*temperature*") or qualitative ("*burden of disease*"). Apart from the general term denoting the outcome, there are several aspects that define it: a measurement tool (questionnaire, score, etc.) used to measure the outcome; time points at which the outcome is measured; patient-level analysis metrics (change from baseline, time to event); population-level aggregation method (mean, median, proportion of patients with some characteristic).

Generally, there are two main contexts in which outcomes of a clinical trial can be mentioned: a definition of what the outcomes of a trial were ("***Quality of life*** *was selected as the primary outcome.*"), and reporting results for an outcome ("***Quality of life*** *was higher in the experimental group than in the control group.*"). In both cases, a mention of an outcome may contain the aspects listed above, but does not necessarily include all of them. In this work, we are interested in the second type of context.

The ways of reporting outcomes are highly diverse. Results for an outcome may be reported as a value of the outcome measure: for binary outcomes, it refers to presence/absence of an event or state; for numerical outcome, it is a numerical value; for qualitative outcome, it is often a value obtained on the associated measurement tool. As the primary goal of RCTs is to compare two or more interventions, results for an outcome can be reported as a comparison between the interventions/patient groups, with or without actual values of the outcome measure. Syntactically, an outcome may be represented by a noun phrase, a verb phrase, an adjective or a clause. We provide here some examples of outcome reporting, to give an idea of variability of expressions.

The outcome is reported as a numerical value:

a) *The median* ***progression-free survival*** *was 32 days.*

The outcome is reported as a comparison between groups, without the values for groups:

b) *MMS resulted in more* ***stunting*** *than standard Fe60F ($p = 0.02$).*

The outcome is reported as a numerical value with comparison between groups:

c) *The average* ***birth weight*** *was 2694 g and* ***birth length*** *was 47.7 cm, with no difference among intervention groups.*

d) *The crude incidence of* ***late rectal toxicity*** $\geq$ ***G2*** *was 14.0% and 12.3% for the arm A and B, respectively.*

e) *More than 96% of patients who received DPT were* ***apyrexial*** *48 hours after treatment compared to 83.5% in the AL group ($p < 0.001$).*

f) *The proportion of patients who **remained relapse-free at Week 26** did not differ significantly between the placebo group (5/16, 31%) and the IFN beta-1a 44 mcg biw (6/17, 35%; p = 0.497), 44 mcg tw (7/16, 44%; p = 0.280) or 66 mcg tw (2/18, 11%; p = 0.333) groups.*

In the latter case, the variation is especially high, and the same outcome may be reported in several different ways (cf. the examples **d**, **e** and **f** that all talk about a percentage of patients in which a certain event occurred, but the structure of the phrases differs).

Identifying the textual boundaries of an outcome presents a challenge: for the example **d**, it can be "*the crude incidence of late rectal toxicity ≥ G2*" or "*late rectal toxicity ≥ G2*"; for the example **f**, it can be "*the proportion of patents who remained relapse-free at Week 26*", or "*remained relapse-free at Week 26*", or simply "*relapse-free*". This variability poses difficulties for both annotation and extraction of reported outcomes. In our annotation, we aimed at annotating the minimal possible text span describing an outcome, not including time points, aggregation and analysis metrics.

2. Significance levels

The ways of presenting significance levels are less diverse than the ways of reporting outcomes. Typically, significance levels are reported via p-values. Another way of determining significance of the results is the confidence interval (CI), where a CI comprising zero denotes non-significant results. In this work, we do not address CIs as they are less frequently reported (Chavalarias et al., 2016).

Statistical significance can be reported as an exact value of P ("*p=0.02*"), as P-value relative to a pre-set threshold ("*p<0.05*"), or in qualitative form ("*significant*"/"*non-significant*"). We address all these forms of reporting significance.

Although in general the ways of presenting statistical significance are rather uniform, there are a few cases to be noted:

- Coordinated p-values:

  *For the non-HPD stratum, the intent-to-treat relative risks of spontaneous premature birth at < 34 and < 37 weeks' gestation were 0.33 (0.03, 3.16) and 0.49 (0.17, 1.44), respectively, and they were non-significant (ns) with **p = 0.31 and 0.14**.*

- Significance level in score of a negation:

  *The respiratory rate, chest indrawing, cyanosis, stridor, nasal flaring, wheeze and fever in both groups recorded at enrollment and parameters **did not differ significantly** between the two groups.*

A particular difficulty is presented by the cases in which a negation marker occurs in the main clause and a significance level in the dependent clause, thus the significance level is within the scope of the negation, but there is a big linear distance between them:

Results There was ***no evidence*** *that an incentive (52% versus 43%, Risk Difference (RD) -8.8 (95%CI -22.5, 4.8); or abridged questionnaire (46% versus 43%, RD -2.9 (95%CI -16.5, 10.7);* ***statistically significantly*** *improved dentist response rates compared to a full length questionnaire in RCT A.*

3. Relationship between outcomes and significance levels

The correspondence between outcomes and significance levels in a sentence is often not one-to-one: multiple outcomes can be linked to the same significance level, and vice versa. Several outcomes are linked to one significance level when outcomes are coordinated:

No significant *improvements in* ***lung function***, ***symptoms***, *or* ***quality of life*** *were seen.*

Several significance levels can be associated to one outcome in a number of cases:

- one outcome is linked to two significance levels when a significance level is presented in both qualitative and numerical form:

  Results ***The response rates*** *were* ***not significantly*** *different Odds Ratio 0.88 (95% confidence intervals 0.48 to 1.63)* $\boldsymbol{p = 0.69}$.

- in the case of comparison between patient groups taking different medications, when there are more than 2 groups, significance can be reported for all pairs of groups;

- significance level for difference observed within groups of patients receiving a particular medication:

  [***Na***] *increased* ***significantly*** *in the 0.9% group (+0.20 mmol/L/h [IQR +0.03, +0.4];* $\boldsymbol{P = 0.02}$*) and increased, but* ***not significantly****, in the 0.45% group (+0.08 mmol/L/h [IQR -0.15, +0.16];* $\boldsymbol{P = 0.07}$*).*

- significance reported for both between- and within-group comparison:

  PTEF *increased* ***significantly*** *both after albuterol and saline treatments but the difference between the two treatments was* ***not significant*** ($\boldsymbol{P = 0.6}$*).*

- significance for differences within subgroups of patients (e.g. gender or age subgroups) receiving a medication;

- significance for different types of analysis: intention-to-treat / per protocol:

*Results For **BMD**, **no** intent-to-treat analyses were **statistically significant**; however, per protocol analyses (ie, only including TC participants who completed ≥ 75% training requirements) of **femoral neck BMD** changes were **significantly** different between TC and UC (+0.04 vs -0.98%; P = 0.05).*

- significance for several time points:

  *Results A **significant** main effect of time (p < 0.001) was found for **step-counts** attributable to significant increases in steps/day between: pre-intervention (M = 6941, SD = 3047) and 12 weeks (M = 9327, SD = 4136), t (78) = - 6.52, **p < 0.001**, d = 0.66; pre-intervention and 24 weeks (M = 8804, SD = 4145), t (78) = - 4.82, **p < 0.001**, d = 0.52; and pre-intervention and 48 weeks (M = 8450, SD = 3855), t (78) = - 4.15, **p < 0.001**, d = 0.44.*

- significance level for comparison of various analysis metrics (mean, AUC, etc.)

Methods

To extract the relation between an outcome and its significance level, we propose a 3-step algorithm: 1) extracting reported outcomes; 2) extracting significance levels; 3) classification of pairs of outcomes and significance levels to detect those related to each other.

As significance levels are not characterized by high variability, we follow the previous research in using rules (regular expressions and sequential rules using information from pos-tagging) to extract significance levels.

We present our methods and results for outcome extraction in detail elsewhere, here we provide a brief summary. We tested several approaches: a baseline approach using sequential rules using information from pos-tagging; an approach using rules based on syntactic structure provided by spaCy dependency parser (Honnibal and Johnson, 2015); a combination of bi-LSTM, CNN and CRF using GloVe (Pennington et al., 2014) word embeddings and character-level representations (Ma and Hovy, 2016); and a fine-tuned bi-LSTM using BERT (Devlin et al., 2018) vector word representations.

BERT (Bidirectional Encoder Representations from Transformers) is a recently introduced approach to pre-training language representations, using a masked language model (MLM) which randomly masks some input tokens, allowing to pre-train a deep bidirectional Transformer using both left and right context. The pre-trained BERT models can be fine-tuned for supervised downstream tasks by adding one output layer.

BERT was trained on a dataset of 3.3B words combining English Wikipedia and BooksCorpus. Two domain-specific versions of BERT are available, pre-trained on a combination of the

initial BERT corpus and additional domain-specific datasets: BioBERT (Lee et al., 2019), adding a large biomedical corpus of PubMed abstracts and PMC full-text articles comprising 18B tokens; and SciBERT (Beltagy et al., 2019), adding a corpus of 1.14M full-text papers from Semantic Scholar with the total of 3.1B tokens. Both BioBERT and SciBERT outperform BERT on biomedical tasks.

BERT provides several models: uncased (trained on lower-cased data) and cased (trained on unchanged data); base and large (differing in model sizes). BioBERT is based on the BERT-base cased model and provides three versions of models: pre-trained on PubMed abstracts, on PMC full-text articles, or on combination of both. SciBERT has both cased and uncased models and provides two versions of vocabulary: BaseVocab (the initial BERT vocabulary) and SciVocab (the vocabulary from the SciBERT corpus). We fine-tuned and tested the BioBERT model trained on the whole corpus, and both cased and uncased base models for BERT and SciBERT (using SciVocab). We did not perform experiments with BERT-Large as we do not have enough resources. We used the code provided by BioBERT for the entity extraction task[1].

The relation extraction assumes that the entities have already been extracted and are given as an input to the algorithm, with the sentence in which they occur. To predict the tag for outcome - significance level pair, we use machine learning.

As the first approach, we compared several classifiers available in the Python scikit-learn library (Pedregosa et al., 2011): Support Vector Machine (SVM) (Cortes and Vapnik, 1995); DecisionTreeClassifier (Rokach and Maimon, 2008); MLPClassifier (von der Malsburg, 1986); KneighborsClassifier (Altman, 1992); GaussianProcessClassifier (Rasmussen and Williams, 2005); RandomForestClassifier (Breiman, 2001); AdaBoostClassifier (Freund and Schapire, 1997); ExtraTreesClassifier (Geurts et al., 2006); GradientBoostingClassifier (Friedman, 2002). Feature engineering was performed manually and was based on our observations on the corpus.

Evaluation was performed using 10-fold cross-validation. To account for different random states, the experiments were run 10 times, we report the average results of the 10 runs. We performed hyperparameters tuning via exhaustive grid search (with the help of the scikit-learn GridSearchCV function).

As the second approach, we employed a deep learning approach to relation extraction, fine-tuning BERT-based models on this task. We tested the same models as for the outcome extraction. We used the code provided by BioBERT for relation extraction task[2]. The algorithm takes as input sentences with the two target entities replaced by masks ("@outcome$" and "@significance$") and positive/negative relation labels assigned to the sentence.

Hyperparameters for entity and relation extraction with BERT-based algorithms are shown

[1]https://github.com/dmis-lab/biobert/blob/master/run_ner.py
[2]https://github.com/dmis-lab/biobert/blob/master/run_re.py

in the Table 6.1. We tested both possible values (True/False) of the hyperparameter "do_lower _case" (lower-casing the input) for all the models.

Hyperparameter	**Entity extraction**	**Relation extraction**
max_seq_length	128	
train_batch_size	32	
eval_batch_size	8	
predict_batch_size	8	
use_tpu	False	
learning_rate	5e-5	2e-5
num_train_epochs	10.0	3.0
warmup_proportion	0.1	
save_checkpoints_steps	1000	
iterations_per_loop	1000	
tf.master	None	

Table 6.1: BERT/BioBERT/SciBERT hyperparameters

Algorithm	**do_lower_case**	**Precision**	**Recall**	**F1**
SciBERT uncased	True	81.17	78.09	79.42
BioBERT	True	80.38	77.85	78.92
BioBERT	False	79.61	77.98	78.6
SciBERT cased	False	79.6	77.65	78.38
SciBERT cased	True	79.24	76.61	77.64
SciBERT uncased	False	79.51	75.5	77.26
BERT uncased	True	78.98	74.96	76.7
BERT cased	False	76.63	74.25	75.18
BERT cased	True	76.7	73.97	75.1
BERT uncased	False	77.28	72.25	74.46
Bi-LSTM-CNN-CRF	–	51.12	44.6	47.52
Rule-based	–	26.69	55.73	36.09

Table 6.2: Reported outcome extraction results

Features

Features are calculated for each pair of outcome and significance level. They are based both on the information about these entities (their position, text, etc.) and on the contextual information (presence of other entities in the sentence, etc.). We used the following binary (True/False) features:

1. only_out: whether the outcome is the only outcome present in the sentence. If yes, it is the only candidate that can be related to the present statistical significance values.

2. only_signif: whether the significance level is the only significance level in the sentence. If yes, it is the only candidate that can be related to the present outcomes.

3. signif_type_num: whether the significance level is expressed in the numerical form;

4. signif_type_word: whether the significance level is expressed in the qualitative form;

5. signif_exact: whether the exact value of significance level is given ("$P = 0.049$"), or it is presented only as comparison to a threshold ("$P < 0.05$"). Significance levels expressed in the word form always have "False" value for this feature. We assumed that significance levels with exact numerical value are less likely to be related to several outcomes that significance levels with inexact value: obtaining exactly same significance level for several outcomes seems unlikely.

6. signif_precedes: whether the significance level precedes the outcome. It is especially pertinent for numerical significance values as they most often follow the related outcome.

7. out_between: whether there is another outcome between the outcome and significance level in the given pair. The outcome that is closer to a significance level is a more likely candidate to be related to it.

8. signif_between: whether there is another significance level between the outcome and the significance level in a given pair. The significance level that is closer to an outcome is a more likely candidate to be related to it.

9. concessive_between: whether there are words (conjunctions) with consessive semantics ("*but*", "*however*", "*although*", etc.) between the outcome and the significance level in the pair.

We used the following numerical features:

1. dist: the distance in characters between the outcome and the significance level in the pair;

2. dist_min_graph: the minimal syntactic distance between the words in the outcome and the words in the significance level;

3. dist_min_out_preceding: the distance from the outcome of the pair to the nearest preceding outcome.

4. dist_min_out_following: the distance from the outcome of the pair to the nearest following outcome. The two last features are designed to reflect the information about

coordination of outcomes (the distances between coordinated entities is typically small), as coordinated outcomes are likely to be related to the same significance level.

We assessed the importance of the features with the attribute "feature_importances_" of the RandomForestClassifier. The results are presented in the Table 6.4.

Evaluation

Entity extraction

The rule-based extraction of significance levels shows the following per-token performance: precision of 99.18%, recall of 96.58% and F-measure of 97.86%.

The results of all the tested approaches to the extraction of reported outcomes are reported in the Table 6.2. The best performance was achieved by the fine-tuned SciBERT uncased model: precision was 81.17%, recall was 78.09% and F-measure was 79.42%.

Relation extraction

The baseline value is based on assigning the majority (positive) class to all the entity pairs. Baseline precision is 53.76%, recall is 100% and F-measure is 69.95%.

The results of the classifiers are presented in the Table 6.3. We present the performance of the default classifiers and of the classifiers with tuned hyperparameters. All the classifiers outperformed the baseline. Random Forest Classifier with tuned hyperparameters (max_depth = 15, min_samples_split = 10, n_estimators = 300) showed the best results, with F-measure of 91.33%, which is by 21.41% higher than the baseline.

It is interesting to compare the deep learning approach using BERT-based fine-tuned models (Table 6.5) to the feature-based classifiers: none of the Google BERT models outperformed the Random Forest Classifier, neither did BioBERT with unchanged input data. However, all the SciBERT fine-tuned models and the BioBERT model with lower-cased input outperformed the Random Forest Classifier. Interestingly, BioBERT, which only has a cased model pre-trained on unchanged data and is thus meant to work with unchanged input, showed the best performance on lower-cased input for the relation extraction task, achieving the F-measure of 94%.

Conclusion and future work

In this paper, we presented a first approach towards the extraction of the relation between outcomes of clinical trials and their reported significance levels. We presented our annotated

corpus for this task and described the ways of reporting outcomes, significance levels and their relation in a text. We pointed out the difficulties posed by the high diversity of the data.

We crafted a feature set for relation extraction and trained and tested a number of classifiers for this task. The best performance was shown by the Random Forest classifier, with the F-measure of 91.33%. Further, we fine-tuned and evaluated a few deep learning models (BERT, SciBERT, BioBERT). The best performance was achieved by the BioBERT model fine-tuned on lower-cased data, with F-measure of 94%.

Our relation extraction algorithm assumes that the entities have been previously extracted and provided as input. An interesting direction for future experiments is building an end-to-end system extracting both entities and relations, as proposed by Miwa and Bansal (2016) or Pawar et al. (2017).

As in our algorithm the extraction of the relevant entities (reported outcomes and significance levels) is essential for extracting the relations, we reported the results of our experiments for extracting this task. Extraction of significance levels reaches the F-measure of 97.86%, while the extraction of reported outcomes shows the F-measure of only 79.42%. Thus, improving the outcome extraction is the main direction of the future work.

Besides, a very important task for clinical trial data analysis consists in determining the significance level for the primary outcome. This task requires two additional steps: 1) identifying the primary outcome, and 2) establishing the correspondence between the primary outcome and a reported outcome. We will present our algorithms for these tasks in a separate paper.

Acknowledgements

We thank Sanjay Kamath for his help in conducting experiments with BERT.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

Classifier	**Hyperparameters**	**Precision**	**Recall**	**F1**
RandomForestClassifier	max_depth = 15, min_samples_split = 10, n_estimators = 300	90.16	92.6	91.33
ExtraTreesClassifier	default	89.74	88.53	89.08
GradientBoostingClassifier	learning_rate = 0.25, max_depth = 23.0, max_features = 7, min_samples_leaf = 0.1, min_samples_split = 0.2, n_estimators = 200	88.44	89.8	89.07
RandomForestClassifier	default	89.54	88.64	89.03
GaussianProcessClassifier	1.0 * RBF(1.0)	86.99	90.38	88.64
GradientBoostingClassifier	default	87.75	89.14	88.4
SVC	C = 1000, gamma = 0.0001, kernel = 'rbf'	86.14	89.65	87.79
DecisionTreeClassifier	default	87.85	86.83	87.27
MLPClassifier	activation = 'tanh', alpha = 0.0001, hidden_layer_sizes = (50, 100, 50), learning_rate = 'constant', solver = 'adam'	84.06	85.15	84.44
MLPClassifier	default	84.4	83.34	83.47
KNeighborsClassifier	n_neighbors = 7, p = 1	83.37	81.27	82.21
AdaBoostClassifier	learning_rate = 0.1, n_estimators = 500	81.34	83.09	82.16
AdaBoostClassifier	default	80.85	82.36	81.53
KNeighborsClassifier	default	81.39	79.88	80.55
GaussianProcessClassifier	default	79.41	78.86	79.1
SVC	default	87.24	64.06	73.77
baseline (majority class)		53.76	100	69.92

Table 6.3: Results of classifiers

Feature	Weight
only_signif	0.21663222
signif_type_num	0.21341347
signif_exact	0.15207938
signif_type_word	0.10103105
dist_min_out_preceding	0.0919397
out_between	0.05683003
dist_min_out_following	0.04683059
concessive_between	0.04260114
only_out	0.02336161
dist	0.02043495
dist_min_graph	0.01794923
signif_precedes	0.01631646
signif_between	0.00058017

Table 6.4: Feature ranking

Algorithm	do_lower_case	Precision	Recall	F1
BioBERT	True	94.3	94	94
SciBERT cased	True	93.9	93.6	93.8
SciBERT cased	False	93.5	93.1	93.3
SciBERT uncased	False	94.2	92.3	93.3
SciBERT uncased	True	94	92.8	93.2
BioBERT	False	92.8	89.7	91.1
BERT cased	False	91.6	90.2	90.9
BERT uncased	True	90.9	90.9	90.8
BERT uncased	False	90.4	89.8	90
BERT cased	True	89.6	90.5	89.8

Table 6.5: Results of relation extraction with BERT/BioBERT/SciBERT

References

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A system for large-scale machine learning, 2016.

N. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician - AMER STATIST*, 46:175–185, 08 1992. doi: 10.1080/00031305.1992.10475879.

M. Asada, M. Miwa, and Y. Sasaki. Extracting drug-drug interactions with attention CNNs. In *BioNLP 2017*, pages 9–18, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2302. URL `https://www.aclweb.org/anthology/W17-2302`.

I. Beltagy, A. Cohan, and K. Lo. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

J. Björne and T. Salakoski. Biomedical event extraction using convolutional neural networks and dependency parsing. In *BioNLP 2018 workshop*, pages 98–108, Melbourne, Australia, July 2018. ACL. URL `https://www.aclweb.org/anthology/W18-2311`.

C. Blake and A. Lucic. Automatic endpoint detection to support the systematic review process. *J. Biomed. Inform*, 2015.

I. Boutron, D. Altman, S. Hopewell, F. Vera-Badillo, I. Tannock, and P. Ravaud. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *Journal of Clinical Oncology*, 2014.

L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, Oct. 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL `https://doi.org/10.1023/A:1010933404324`.

D. Chavalarias, J. D. Wallach, A. H. T. Li, and J. P. A. Ioannidis. Evolution of reporting p values in the biomedical literature, 1990-2015. *JAMA*, 315 11:1141–8, 2016.

F. Chollet et al. Keras, 2015. URL `https://github.com/fchollet/keras`.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995. ISSN 1573-0565. doi: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

D. Demner-Fushman, B. Few, S. Hauser, and G. Thoma. Automatically identifying health outcome information in medline records. *Journal of the American Medical Informatics Association*, 2006.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139, 1997. ISSN 0022-0000. doi: 10.1006/jcss.1997.1504. URL http://www.sciencedirect.com/science/article/pii/S002200009791504X.

J. H. Friedman. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378, Feb. 2002. ISSN 0167-9473. doi: 10.1016/S0167-9473(01)00065-2. URL http://dx.doi.org/10.1016/S0167-9473(01)00065-2.

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Mach. Learn.*, 63(1):3–42, Apr. 2006. ISSN 0885-6125. doi: 10.1007/s10994-006-6226-1.

R. Haneef, C. Lazarus, P. Ravaud, A. Yavchitz, and I. Boutron. Interpretation of results of studies evaluating an intervention highlighted in google health news: a cross-sectional study of news. *PLoS ONE*, 2015.

M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions. The extent and consequences of p-hacking in science. In *PLoS biology*, 2015.

M. Honnibal and M. Johnson. An improved non-monotonic transition system for dependency parsing. In *Proc. of EMNLP 2015*, pages 1373–1378, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL https://aclweb.org/anthology/D/D15/D15-1162.

W. Hsu, W. Speier, and R. K. Taira. Automated extraction of reported statistical analyses: Towards a logical representation of clinical trial literature. *AMIA Annual Symposium*, 2012: 350–359, 2012.

A. Koroleva. Annotated corpus for the relation between reported outcomes and their significance levels, May 2019.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*, 2019.

J. Lever and S. Jones. Painless relation extraction with kindred. In *BioNLP 2017*, pages 176–183, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2322. URL https://www.aclweb.org/anthology/W17-2322.

A. Lucic and C. Blake. Improving endpoint detection to support automated systematic reviews. In *AMIA Annu Symp Proc.*, 2016.

X. Ma and E. Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1101. URL https://www.aclweb.org/anthology/P16-1101.

M. Miwa and M. Bansal. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1105. URL https://www.aclweb.org/anthology/P16-1105.

S. Pawar, P. Bhattacharyya, and G. Palshikar. End-to-end relation extraction using neural networks and Markov logic networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 818–827, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-1077.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Y. Peng and Z. Lu. Deep learning for extracting protein-protein interactions from biomedical literature. In *BioNLP 2017*, pages 29–38, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2304. URL https://www.aclweb.org/anthology/W17-2304.

Y. Peng, M. Torii, C. Wu, and K. Vijay-Shanker. A generalizable nlp framework for fast development of pattern-based biomedical relation extraction systems. *BMC bioinformatics*, 15:285, 08 2014. doi: 10.1186/1471-2105-15-285.

J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proc. of EMNLP 2014*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL `https://www.aclweb.org/anthology/D14-1162`.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

L. Rokach and O. Maimon. *Data Mining with Decision Trees: Theroy and Applications*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2008. ISBN 9789812771711, 9812771719.

T. M. Schindler. Hypothesis testing in clinical trials. *AMWA Journal*, 30(2), 2015.

C. von der Malsburg. Frank rosenblatt: Principles of neurodynamics: Perceptrons and the theory of brain mechanisms. *Brain Theory*, pages 245–248, 01 1986. doi: 10.1007/978-3-642-70911-1_20.

A. Yavchitz, I. Boutron, A. Bafeta, I. Marroun, P. Charles, J. Mantz, and P. Ravaud. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS Med*, 2012.

D. Zhou, D. Zhong, and Y. He. Biomedical relation extraction: From binary to complex. In *Comp. Math. Methods in Medicine*, 2014.

Part III

Implementation and interface

Chapter 7

DeSpin: a prototype system for detecting spin in biomedical publications. Anna Koroleva, Sanjay Kamath, Patrick MM Bossuyt, Patrick Paroubek. Submitted

Context

This chapter is aimed at presenting the spin detection prototype that resulted from our work. We describe the types of spin that we addressed and their linguistic characteristics. We explain how the algorithms introduced in the previous chapters (3, 4 and 6) are combined in our spin detection pipeline. The following chapter addresses the question of integrating all the information elements addressed before into a single functionality that could be provided as an additional functionality to the editing/viewing software used by authors and peer reviewers of scientific articles.

Authors' contributions

AK designed the study described in this chapter and interpreted the data. AK collected and annotated the corpus. AK and SK conducted the experiments, supervised by PP. AK and PP developed the system core and the interface. AK drafted the manuscript. SK, PMMB and PP revised the draft critically for important intellectual content.

Abstract

Background: Improving the quality of medical research reporting is crucial in all efforts to reduce avoidable waste in research and to improve the quality of health care and health information. Despite various initiatives aiming at improving research reporting – guidelines, checklists, authoring aids, peer review procedures, etc. – overinterpretation of research results, also known as spin, is still a serious issue in research reporting.

Methods: We propose a Natural Language Processing (NLP) system for semi-automatic detection of spin in scientific articles, here applied to randomized controlled trial (RCTs) reports. We use a combination of rule-based and machine learning approaches to extract important information on trial design and to detect potential spin. Along with our entity extraction algorithms, our system incorporates a simple but powerful interface for manual annotation of spin, which can also be used as an authoring or peer-reviewing aid.

Results: Our algorithms achieved operational performance for detecting relevant phenomena, F-measure ranging from 79.42 to 97.86% for different tasks. The most difficult task is extracting reported outcomes.

Conclusion: The proposed tool is the first semi-automated tool for spin detection. It can be used by both authors and reviewers to detect potential spin, helping to improve the quality of research results reporting. The tool and the annotated dataset are freely available.

Keywords: Spin, Randomized Controlled Trials, Information Extraction, Prototype, Automated Aid Tool

Background

Several authors have observed that the quality of reporting research results in the clinical domain is suboptimal. As a consequence, research findings can often not be replicated, and billions of euros may be wasted yearly (Ioannidis, 2005).

Numerous initiatives aim at improving the quality of research reporting. Guidelines and checklists have been developed for every type of clinical research. Still, the quality of reporting remains low: authors fail to choose and follow a correct guideline/checklist (Samaan et al., 2013). Automated tools, such as Penelope[1], are introduced to facilitate the use of guidelines/checklists. It was proved that authoring aids improve the completeness of reporting (Barnes et al., 2015).

Enhancing the quality of peer reviewing is another step to improve research reporting. Peer reviewing requires assessing a large number of information items. Nowadays, Natural Language Processing (NLP) is applied to facilitate laborious manual tasks such as indexing of medical literature (Huang et al., 2011) and systematic review process (Ananiadou et al., 2009). Similarly, the peer reviewing process can be partially automated with the help of NLP.

Our project tackles a specific issue of research reporting:spin, also referred to as overinter-

[1]https://www.penelope.ai/

pretation of research results. In the context of clinical trials assessing a new (experimental) intervention, spin consists in exaggerating the beneficial effects of the studied intervention (Boutron et al., 2010).

Spin was shown to be common in abstracts of articles reporting randomized controlled trials (RCTs) - clinical trials comparing health interventions, to which participants are allocated randomly to avoid biases - with non-significant primary outcome in surgical research (40%) (Fleming, 2016), cardiovascular diseases (57%) (Khan et al., 2019), cancer (47%) (Vera-Badillo et al., 2016), obesity (46.7%) (Austin et al., 2018), otolaryngology (70%) (Cooper et al., 2018), anaesthesiology (32.2%) (Kinder et al., 2018), and wound care (71%) (Lockyer et al., 2013). Although the problem of spin has started to attract attention in the medical community in the recent years, the shown prevalence of spin proves that it often remains unnoticed by editors and peer reviewers.

Abstracts are often the only part of the article available to readers, and spin in abstracts of RCTs poses a serious threat to the quality of health care by causing overestimation of the intervention by clinicians (Boutron et al., 2014), which may lead to the use of an ineffective of unsafe intervention in clinical practice. Besides, spin in research articles is linked to spin in press releases and health news (Yavchitz et al., 2012), which has the negative impact of raising false expectations regarding the intervention among the public.

The importance of the problem of spin motivated our project, which aims at developing NLP algorithms to aid authors and readers in detecting spin. We focus on randomized controlled trials (RCTs) as they are the most important source of evidence for Evidence-based medicine, and spin in RCTs has high negative impact.

In this paper, we introduce the first prototype of a system, called DeSpin (Detector of Spin), that automatically detects potential spin in abstracts of RCTs and relevant supporting information. This prototype comprises a number of algorithms that show operational performance, and a simple interface. We have paid particular attention at developing portable algorithms that can be built in other clinical text analysis systems.

Our work lies within the scope of the Methods in Research on Research (MiRoR) project[2], a large international collaborative project devoted to improving the planning, conduct, reporting and peer reviewing of health care research. For the design and development of our toolkit, we benefited from advice from the MiRoR consortium members experts in medical reporting.

This paper is organized as follows: first, we overview some existing semi-automated aid systems for authors and readers/reviewers of biomedical articles. Second, we introduce in more detail the notion of spin, the types of spin that we address, and the related information that is needed to assess an article for spin. After that, we describe our current algorithms, methods

[2]http://miror-ejd.eu/

employed and provide their evaluation. Finally, we discuss the potential future development of the prototype.

Related work

Although there has been no attempt to automate spin detection in biomedical articles, a number of works addressed developing automated aid tools to assist authors and readers of scientific articles in performing various other tasks. Some of these tools were tested and were shown to reduce the workload and improve the performance of human experts on the corresponding task.

Authoring aid tools

Barnes et al. (2015) assessed the impact of a writing aid tool based on the CONSORT statement (Schulz et al., 2010) on the completeness of reporting of RCTs. The tools was developed for six domains of the Methods section (trial design, randomization, blinding, participants, interventions, and outcomes) and consisted of reminders of the corresponding CONSORT item(s), bullet points enumerating the key elements to report, and good reporting examples. The tool was assessed in an RCT in which the participants were asked to write a Methods section of an article based on a trial protocol, either using the aid tool ('intervention' group) or without using the tool ('control' group). The results of 41 participants showed that the mean global score for reporting completeness was higher with the use of the tool than without it.

Aid tools for readers and reviewers

Kiritchenko et al. (2010) developed a system called ExaCT to automatically extract 21 key characteristics of clinical trial design, such as treatment names, eligibility criteria, outcomes, etc. ExaCT consists of an information extraction algorithm that looks for text fragments corresponding to the target information elements, a web-based user interface through which human experts can view and correct the suggested fragments.

The National Library of Medicine's Medical Text Indexer (MTI) is a system providing automatic recommendations based on the Medical Subject Headings (MeSH) terms for indexing medical articles (Mork et al., 2013). MTI is used to assist human indexers, catalogers, and NLM's History of Medicine Division in their work. Its use by indexers was shown to grow over years (used to index 15.75% of the articles 2002 vs 62.44% in 2014) and to improve the performance (precision, recall and F-measure) of indexers (Mork et al., 2017).

Marshall et al. (2015) addressed the task of automating assessment of risk of bias in clinical trials. Bias is phenomenon related to spin: it is a systematic error or a deviation from the

truth in the results or conclusions that can cause an under- or overestimation of the effect of the examined treatment (Higgins and Green, 2008). The authors developed a system called RobotReviewer that used machine learning to assess an article for the risk of different types of bias and to extract text fragments that support these judgements. These works showed that automated risk of bias assessment can be achieve reasonable performance, and the extraction of supporting text fragments reached similar quality to that of human experts. Marshall et al. (2017) further developed RobotReviewer, adding functionality for extracting the PICO (Population, Interventions/Comparators, Outcomes) elements from articles and detecting study design (RCT), for the purpose of automated evidence synthesis. Soboczenski et al. (2019) assessed RobotReviewer in a user study involving 41 participants, evaluating time spent for bias assessment, text fragment suggestions by machine learning, and usability of the tool. Semi-automation in this study was shown to be quicker than manual assessment; 91% of the automated risk of bias judgments and 62% of supporting text suggestions were accepted by the human reviewers.

The cited works demonstrate that semi-automated aid tools can prove useful for both authors and readers/reviewers of medical articles and has a potential to improve the quality of the articles and facilitate the analysis of the texts.

Spin: definition and types

We adopt the definition and classification of spin introduced by Boutron et al. (2010) and Lazarus et al. (2015), who divided instances of spin into several types and subtypes.In our project, we address the following types of spin:

1. Outcome switching – unjustified change of the pre-defined trial outcomes, leading to reporting only the favourable outcomes that support the hypothesis of the researchers (Goldacre et al., 2019). Outcome switching is one of the most common types of spin. It can consist in omitting the primary outcome in the results / conclusions of the abstract, or in the focus on significant secondary outcomes, e.g.:

   *The primary end point of this trial was **overall survival**. <...> This trial showed a significantly increased **R0 resection rate** although it failed to demonstrate a **survival** benefit.*

   In this example, the primary outcome ("overall survival"), the results for which were not favourable, is mentioned in the conclusion, but it is not reported in the first place and occurs within a concessive clause (starting by "although"). This way of reporting puts the focus on the other, favourable, outcome ("R0 resection rate").

2. Interpreting non-significant outcome as a proof of equivalence of the treatments, e.g.:

   *The median PFS was 10.3 months in the XELIRI and 9.3 months in the FOLFIRI arm (**p = 0.78**). Conclusion: The XELIRI regimen showed **similar PFS** compared to the FOLFIRI regimen.*

   The results for the outcome "median PFS" are not significant, which is often erroneously interpreted as a proof of similarity of the treatments. However, a non-significant result means that the null hypothesis of a difference could not be rejected, which is not equivalent to a demonstration of similarity of the treatments. This would require the rejection of the null hypothesis of a difference, or a substantial difference, in outcomes between treatments.

3. Focus on within-group comparisons, e.g.:

   ***Both groups showed robust improvement** in both symptoms and functioning.*

   The goal of randomized controlled trials is to compare two treatments with regard to some outcomes. If the superiority of the experimental treatment over the control treatment was not shown, within-group comparisons (reporting the changes within a group of patients receiving a treatment, instead of comparing patients receiving different treatments) can be used to persuade the reader of beneficial effects of the experimental treatment.

Two concepts are vital for spin detection and play a key role in our algorithms:

1. The primary outcome of a trial – the most important variable monitored during the trial to assess how the studied treatment impacts it. Primary outcomes are recorded in trial registries (open online databases storing the information about registered clinical trials), and should be defined in the text of clinical articles, e.g.:

   *The primary end point was **a difference of > 20% in the microvascular flow index of small vessels among groups**.*

2. Statistical significance of the primary outcome. Statistical hypothesis testing is used to check for a significant difference in outcomes between two patient groups, one receiving the experimental treatment and the other receiving the control treatment. Statistical significance is often reported as a P-value compared to predefined threshold, usually set to 0.05. Spin most often occurs when the results for the primary outcome are not significant (Boutron et al., 2010; Fleming, 2016; Khan et al., 2019; Vera-Badillo et al., 2016; Austin et al., 2018; Cooper et al., 2018; Kinder et al., 2018; Lockyer et al., 2013), although trials with significant effect on the primary outcome may also be prone to spin (Beijers et al., 2017).

Trial results are commonly reported as an effect on the (primary) outcome[3], along with the p-value.

Microcirculatory flow indices of small and medium vessels *were* ***significantly*** *higher in the levosimendan group as compared to the control group (**$p < 0.05$**).*

Statistical significance levels of trial outcomes are vital for spin detection, as spin is commonly related to non-significant results for the primary outcome, or to selective reporting of significant outcomes only.

Implementation

Our prototype allows the user to load a text (with or without annotations), run algorithms, visualize their output, correct, add or remove annotations. The expected input is an article reporting an RCT in the text format, including the abstract.

Figure 7-1 shows the interface with an example of a processed text.

The main items of the drop-down menu on the top of the page are **Annotations**, which allows to visualize and manage the annotations, and **Algorithms**, which allows to run the algorithms listed and explained below to detect potential spin and the related information. The text fragments identified by the algorithms can be highlighted in the text. A report is saved into the Metadata section of Annotations menu and can be saved into a file via the **Generate report** item of the **Algorithms** menu.

Algorithms

Detection of spin and related information is a complex task which cannot by fully automated. Our system is designed as a semi-automated tool that finds potential instances of some types of spin and extracts the supporting information that can help the user to make the final decision on the presence of spin. In this section, we present the current functions of the system according to the types of spin that they are used to detect.

As we aim at detecting spin in the Results and Conclusions sections of articles' abstracts, the detection of spin requires an algorithm analyzing the given article to detect its abstract and the results and conclusions sections within the abstract. We will not mention this algorithm in the list of algorithms for each spin type to avoid repetition. If we talk about extracting some information from the abstract, it implies that the text structure analysis algorithm was applied.

[3]It is important to distinguish between the notions of outcome, effect and result in this context: an outcome is a measure/variable monitored during a clinical trial; effect refers to the change in an outcome observed during a trial; trial results refer to the set of effects for all measured outcomes.

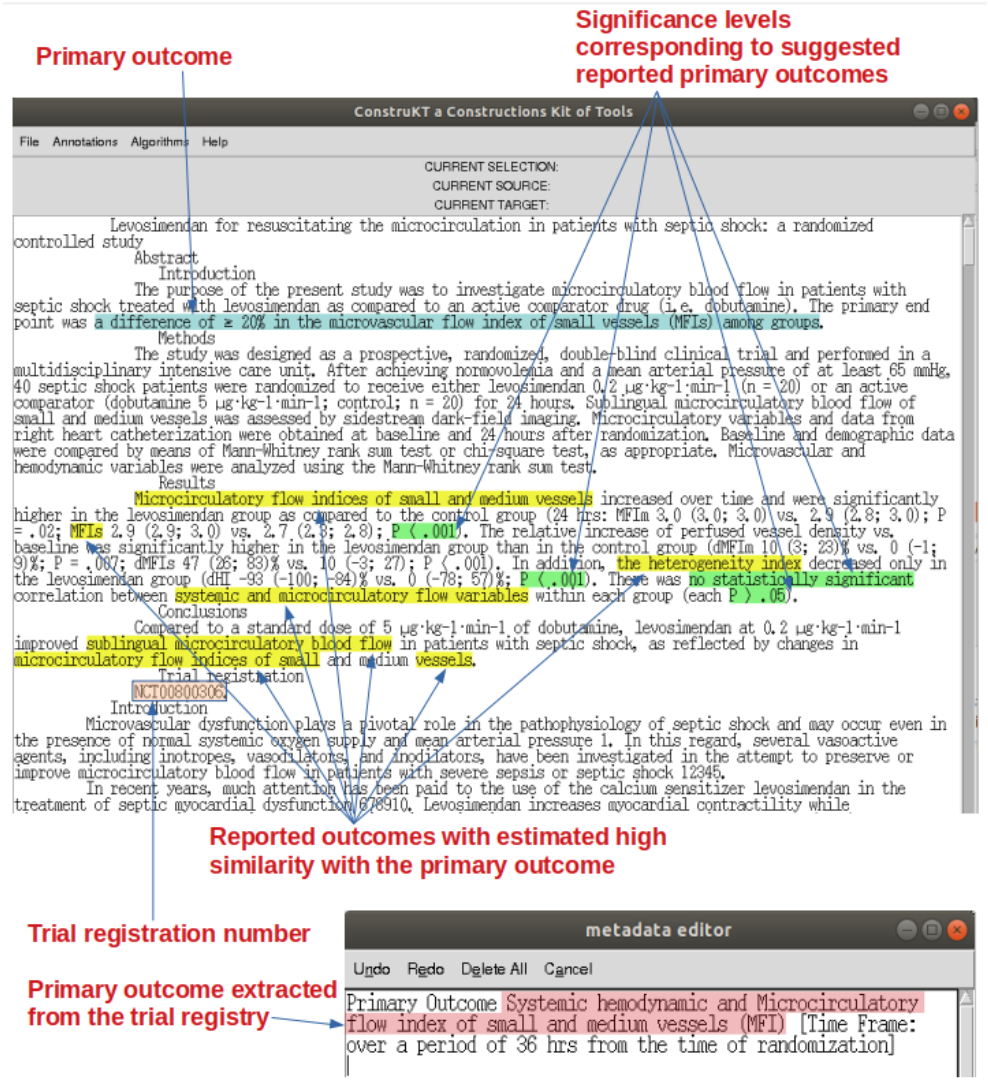


Figure 7-1: Example of a processed text

Outcome switching

We focus on the switching (change/omission) of the primary outcome. Primary outcome switching can occur at several points:

- the primary outcome(s) recorded in the trial registry can differ from the primary outcome(s) declared in the article;
- the primary outcome(s) declared in the abstract can differ from the primary outcome(s) declared in the body of the article;
- the primary outcome(s) recorded in the trial registry can be omitted when reporting the results for the outcomes in the abstract;
- the primary outcome(s) recorded in the article can be omitted when reporting the results

for the outcomes in the abstract.

Primary outcome switching detection involves the following algorithms:

1. Identification of primary outcomes in trial registries and in the article's text.

2. Identification of reported outcomes from sentences reporting the results, e.g. (reported outcomes are in bold):

   *The results of this study showed that **symptom Scores** in massage group were improved significantly compared with control group, and the rate of **dyspnea**, **cough** and **wheeze** in the experimental group than the control group were reduced by approximately 45%, 56% and 52%.*

3. Assessment of semantic similarity of pairs of outcomes extracted by the above algorithms to check for missing outcomes. We perform the assessment for the following sets of outcomes:

   - The primary outcome extracted from the registry are compared to the primary outcome(s) declared in the article;
   - The primary outcome extracted from the abstract are compared to the primary outcome(s) declared in the body of the article;
   - The primary outcome extracted from the article are compared to the outcomes reported in the abstract;
   - The primary outcome extracted from the registry are compared to the outcomes reported in the abstract.

   These assessments allow to detect switching of the primary outcome at all the possible stages. If the primary outcome in the registry and in the article, or in the abstract and body of the article differ, we conclude that there is potential outcome switching, which is reported to the user. Similarly, if the primary outcome (from the article or from the registry) is missing from the list of the reported outcomes, we suspect selective reporting of outcomes, and the system reports it to the user.

   In the example on the page 149, the system should extract "overall survival" as the primary outcome, and "R0 resection rate" and "survival" as reported outcomes. The similarity between "overall survival" and "R0 resection rate" is low, while the similarity between "overall survival" and "survival" is high, thus, we conclude that the primary outcome "overall survival" is reported as "survival".

As semantic similarity often depends on the context, the conclusions of the system are presented to the user, who can check them to make the conclusions on correctness of the analysis.

4. Assessing the discourse prominence of the reported primary outcome (detected by the previous algorithms) by checking if it is reported the first place among all the outcomes; if it is reported in a concessive clause.

   In the example above, the system will detect that the primary outcome "survival" is reported within a concessive clause (starting by "although") and will flag the sentence as potentially focusing on secondary outcomes.

Interpreting non-significant outcome as a proof of equivalence of the treatments

As we stated above, conclusions on the similarity/equivalence of the studies treatments are justified only if the trial was of non-inferiority or equivalence type. Thus, we employ two algorithms to detect this type of spin:

1. Identification of statements of similarity between treatments, e.g.:

   *Both products caused **similar** leukocyte counts diminution and had **similar** safety profiles.*

2. Identifying the markers of non-inferiority or equivalence trial design, e.g.:

   *ONCEMRK is a phase 3, multicenter, double-blind, **noninferiority** trial comparing raltegravir 1200mg QD with raltegravir 400mg BID in treatment-naive HIV-1–infected adults.*

If there is a statement of similarity of treatments while no markers of non-inferiority / equivalence design are found, we conclude the presence of spin and report it to the user.

Focus on within-group comparisons

Any statement in the results and conclusions of the abstract that presents a comparison of two states of a patient group without comparing it to another group is a within-group comparison. This type of spin is detected by a single algorithm that identifies within-group comparisons that are further reported to the user:

*Young Mania Rating Scale total scores **improved with ritanserin**.*

Other algorithms

We support extraction of some information that is not directly involved in detection of spin, but that can currently help user in spin assessment and that can be used in the future when new spin types are added. The algorithms include:

1. Extraction of measures of statistical significance, both numerical and verbal (in bold):

   Study group patients had a ***significant*** *lower reintubation rate than did controls; six patients (17%) versus 19 patients (48%),* ***P<0.05****; respectively.*

2. Extraction of the relation between the reported outcomes and their statistical significance, extracted at the previous stages. For the example above, we extract pairs ("reintubation rate", "significant") and ("reintubation rate", "P<0.05").

   These algorithms, in combination with the assessment of semantic similarity of extracted outcomes, allows to identify the significance level for the primary outcome.

3. Identifying "hedge": presenting the findings with a certain level of uncertainty ("hedging") may result from the absence of sufficient evidence, which is also one cause of spin. Even if the relation between hedging and spin has not been studied yet, we hypothesize its existence and thus we decided to extract also markers of hedging (expressions reducing the certainty of a statement, such as modal verbs, verbs like "suggest", "appear", etc.).

Methods and Results

In this section, we briefly outline the methods used in our algorithms, the datasets used for evaluation, and the current performance of the algorithms. The details on development of the algorithms, annotating the data and testing different approaches are described in detail elsewhere; we provide here only a brief description of the best-performing method for each task.

The methods we employ can be divided into two groups: machine learning, including deep learning, used for the core tasks for which we have sufficient training data, and rule-based methods, used for the simpler tasks or for tasks where we do not have enough data for machine learning.

Rule-based methods

We developed rules for the following tasks:

Algorithm	Method	Annotated dataset	Precision	Recall	F1
Primary outcomes extraction	Deep learning	2,000 sentences / 1,694 outcomes	86.99	90.07	88.42
Reported outcomes extraction	Deep learning	1,940 sentences / 2,251 outcomes	81.17	78.09	79.42
Outcome similarity assessment	Deep learning	3,043 pairs of outcomes	88.93	90.76	89.75
Similarity statements extraction	Rules	180 abstracts / 2402 sentences whole abstract results and conclusions	 77.8 85.1	 87.5 87.5	 82.4 86.3
Within-group comparisons	Rules	180 abstracts / 2402 sentences whole abstract results and conclusions	 53.2 71.9	 90.6 90.6	 67.1 80.1
Abstract extraction	Rules	3938 abstracts	94.7	94	94.3
Text structure analysis: sections of abstract	Deep learning	PubMed200k	97.82	95.81	96.8
Extraction of significance levels	Rules	664 sentences / 1,188 significance level markers	99.18	96.58	97.86
Outcome - significance level relation extraction	Deep learning	2,678 pairs of outcomes and significance level markers	94.3	94	94

Table 7.1: Overview of algorithms, methods, results and annotated datasets

- To find the abstract, we use regular expression rules that are evaluated on the set of 3938 PubMed Central (PMC)[4] articles in XML format with a specific tag for the abstract, used as the gold standard.

- To extract outcomes from trial registries, we use regular expressions to extract the trial registration number from the article; using it, we find on the web, download and parse the registry entry corresponding to the trial.

- To extract significance levels, we use rules based on regular expressions and token, lemma and pos-tag information.

- To assess the discourse prominence of an outcome, to detect "hedge", statements of similarity between treatments, within-group comparisons and markers of non-inferiority design, we employ rules based on token, lemma and pos-tag information.

[4]https://www.ncbi.nlm.nih.gov/pmc/

We annotated abstracts of 180 articles (2402 sentences) for similarity statements and within-group comparisons. The proportion of these types of statements in our corpus is low: we identified only 72 similarity statements and 127 within-group comparisons. The evaluation of statements of similarity between treatments and within-group comparisons was performed with two settings: 1) using the whole text of abstracts; 2) using only the Results and Conclusions sections of the abstract, which expectedly raised the precision (Table 7.1).

Machine learning methods

For the core tasks of our system, we use a deep learning approach that was recently proved to be highly successful in many NLP applications. It employs language representations pre-trained on large unannotated data and fine-tuned on a relatively small amount of annotated data for a specific downstream task. The language representations that we tested include: BERT (Bidirectional Encoder Representations from Transformers) models (Devlin et al., 2018), trained on a general-domain corpus of 3.3B words; BioBERT model (Lee et al., 2019), trained on the BERT corpus and a biomedical corpus of 18B words; and SciBERT models (Beltagy et al., 2019), trained on the BERT corpus and a scientific corpus of 3.1B words. For each task, we chose the best-performing model:

- to identify sections within the abstract, we use BioBERT model fine-tuned on the annotated dataset introduced in Dernoncourt and Lee (2017).

- to extract primary outcomes, we use the BioBERT model fine-tuned on our manually annotated corpus of 2000 sentences with 1694 primary outcomes.

- to extract reported outcomes, we use the SciBERT model fine-tuned on our manually annotated corpus of 1940 sentences with 2251 reported outcomes.

- to assess the similarity between outcomes, we use the BioBERT model fine-tuned on 3,043 pairs of outcomes annotated for semantic similarity.

- to extract the relation between reported outcome and statistical significance levels, we use the BioBERT model fine-tuned on 2,552 annotated relations.

The current functionality, methods in use, annotated datasets and the best achieved results are outlined in Table 7.1. Performance is assessed per-token for outcome and significance level extraction and per-unit for other tasks.

Conclusions

We presented a first prototype tool for assisting authors and reviewers to detect spin and related information in abstracts of articles reporting RCTs. The employed algorithms show operational performance in complex semantic tasks, even with low volume of available annotated data. We envisage two possible applications of our system: as an authoring aid or as peer-reviewing tool. The authoring aid version can be further developed into an educational tool, explaining the notion of spin and its types to the user.

Possible directions for future work are improving the implementation (adding prompts for interaction with the user; facilitating installation process), algorithms (improving current performance, adding detection of new spin types), application (promoting the tool among the target audience; encouraging users to submit their manually annotated data, to be used to retrain and improve the algorithms), and optimization (parallel processing of multiple input text files). Our system cal be easily incorporated into other text processing tools.

Availability and requirements

Project name: DeSpin

Project home page: The source code and models for the system, together with a short info file describing how to set it ut, are available at:

`https://github.com/aakorolyova/DeSpin`

(to be released on GitHub).

Operating system(s): Linux

Programming language: Python3

Other requirements:

Python packages required:

- nltk3.4
- numpy1.16.4
- pandas0.24.1
- pickle
- sklearn0.20.3
- spacy2.0.18

- tensorflow1.13.1
- tkinter
- unicodedata
- urllib

Language models required:

- BioBERT – v1.0-pubmed-pmc – Pre-trained weight of BioBERT v1.0 (+PubMed 200K +PMC 270K):

  `https://github.com/naver/biobert-pretrained/releases`

- SciBERT – scibert-scivocab-uncased model for tensorflow:

  `https://github.com/allenai/scibert`

License: TBA

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

References

S. Ananiadou, B. Rea, N. Okazaki, R. Procter, and J. Thomas. Supporting systematic reviews using text mining. *Social Science Computer Review - SOC SCI COMPUT REV*, 27:509–523, 10 2009. doi: 10.1177/0894439309332293.

J. Austin, C. Smith, K. Natarajan, M. Som, C. Wayant, and M. Vassar. Evaluation of spin within abstracts in obesity randomized clinical trials: A cross-sectional review: Spin in obesity clinical trials. *Clinical Obesity*, 9:e12292, 12 2018. doi: 10.1111/cob.12292.

C. Barnes, I. Boutron, B. Giraudeau, R. Porcher, D. Altman, and P. Ravaud. Impact of an online writing aid tool for writing a randomized trial report: The cobweb (consort-based web tool) randomized controlled trial. *BMC medicine*, 13:221, 09 2015. doi: 10.1186/s12916-015-0460-y.

L. Beijers, B. F. Jeronimus, E. H. Turner, P. de Jonge, and A. M. Roest. Spin in rcts of anxiety medication with a positive primary outcome: a comparison of concerns expressed by the us fda and in the published literature. *BMJ Open*, 7(3), 2017. ISSN 2044-6055. doi: 10.1136/bmjopen-2016-012886. URL `https://bmjopen.bmj.com/content/7/3/e012886`.

I. Beltagy, A. Cohan, and K. Lo. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

I. Boutron, S. Dutton, P. Ravaud, and D. Altman. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*, 2010.

I. Boutron, D. Altman, S. Hopewell, F. Vera-Badillo, I. Tannock, and P. Ravaud. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *Journal of Clinical Oncology*, 2014.

C. M. Cooper, H. M. Gray, A. E. Ross, T. A. Hamilton, J. B. Downs, C. Wayant, and M. Vassar. Evaluation of spin in the abstracts of otolaryngology randomized controlled trials: Spin found in majority of clinical trials. *The Laryngoscope*, 12 2018. doi: 10.1002/lary.27750.

F. Dernoncourt and J. Y. Lee. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *8th IJCNLP (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan, Nov. 2017. Asian Federation of NLP. URL `https://www.aclweb.org/anthology/I17-2052`.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

P. S. Fleming. Evidence of spin in clinical trials in the surgical literature. *Ann Transl Med.*, 4,19(385), Oct 2016. doi: 10.21037/atm.2016.08.23.

B. Goldacre, H. Drysdale, A. Dale, I. Milosevic, E. Slade, P. Hartley, C. Marston, A. Powell-Smith, C. Heneghan, and K. R. Mahtani. Compare: a prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*, 20(1):118, Feb 2019. ISSN 1745-6215. doi: 10.1186/s13063-019-3173-2. URL `https://doi.org/10.1186/s13063-019-3173-2`.

J. P. Higgins and S. Green, editors. *Cochrane handbook for systematic reviews of interventions*. Wiley & Sons Ltd., West Sussex, 2008.

M. Huang, A. Névéol, and Z. Lu. Recommending mesh terms for annotating biomedical articles. *Journal of the American Medical Informatics Association : JAMIA*, 18:660–7, 05 2011. doi: 10.1136/amiajnl-2010-000055.

J. Ioannidis. Why most published research findings are false. *PLoS medicine*, 2:e124, 09 2005. doi: 10.1371/journal.pmed.0020124.

M. Khan, N. Lateef, T. Siddiqi, K. Abdur Rehman, S. Alnaimat, S. Khan, H. Riaz, M. Hassan Murad, J. Mandrola, R. Doukky, and R. Krasuski. Level and prevalence of spin in published cardiovascular randomized clinical trial reports with statistically nonsignificant primary outcomes: A systematic review. *JAMA Network Open*, 2:e192622, 05 2019. doi: 10.1001/jamanetworkopen.2019.2622.

N. Kinder, M. Weaver, C. Wayant, and M. Vassar. Presence of 'spin' in the abstracts and titles of anaesthesiology randomised controlled trials. *British Journal of Anaesthesia*, 122, 11 2018. doi: 10.1016/j.bja.2018.10.023.

S. Kiritchenko, B. D. Bruijn, S. Carini, J. Martin, and I. Sim. Exact: automatic extraction of clinical trial characteristics from journal publications. *BMC Med Inform Decis Mak.*, 2010.

C. Lazarus, R. Haneef, P. Ravaud, and I. Boutron. Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. *BMC Med Res Methodol.*, 2015.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*, 2019.

S. Lockyer, R. W. Hodgson, J. C. Dumville, and N. Cullum. "Spin" in wound care research: the reporting and interpretation of randomized controlled trials with statistically non-significant primary outcome results or unspecified primary outcomes. In *Trials*, 2013.

I. Marshall, J. Kuiper, E. Banner, and B. C. Wallace. Automating biomedical evidence synthesis: RobotReviewer. In *Proceedings of ACL 2017, System Demonstrations*, pages 7–12, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P17-4002`.

I. J. Marshall, J. Kuiper, and B. C. Wallace. Robotreviewer: Evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association : JAMIA*, 23, 06 2015. doi: 10.1093/jamia/ocv044.

J. Mork, A. Jimeno-Yepes, and A. Aronson. The nlm medical text indexer system for indexing biomedical literature. *CEUR Workshop Proceedings*, 1094, 01 2013.

J. Mork, A. Aronson, and D. Demner-Fushman. 12 years on – is the NLM medical text indexer still useful and relevant? *Journal of Biomedical Semantics*, 8(1), feb 2017. doi: 10.1186/s13326-017-0113-5.

Z. Samaan, L. Mbuagbaw, D. Kosa, V. Borg Debono, R. Dillenburg, S. Zhang, V. Fruci, B. Dennis, M. Bawor, and L. Thabane. A systematic scoping review of adherence to reporting guidelines in health care literature. *Journal of multidisciplinary healthcare*, 6:169–88, 05 2013. doi: 10.2147/JMDH.S43952.

K. F. Schulz, D. G. Altman, and D. Moher. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340, 2010. ISSN 0959-8138. doi: 10.1136/bmj.c332. URL `https://www.bmj.com/content/340/bmj.c332`.

F. Soboczenski, T. Trikalinos, J. Kuiper, R. G. Bias, B. Wallace, and I. J. Marshall. Machine learning to help researchers evaluate biases in clinical trials: A prospective, randomized user study. *BMC Medical Informatics and Decision Making*, 19, 12 2019. doi: 10.1186/s12911-019-0814-z.

F. E. Vera-Badillo, M. Napoleone, M. K. Krzyzanowska, S. M. Alibhai, A.-W. Chan, A. Ocana, B. Seruga, A. J. Templeton, E. Amir, and I. F. Tannock. Bias in reporting of randomised clinical trials in oncology. *European Journal of Cancer*, 61:29 – 35, 2016. ISSN 0959-8049. doi: 10.1016/j.ejca.2016.03.066. URL `http://www.sciencedirect.com/science/article/pii/S0959804916320287`.

A. Yavchitz, I. Boutron, A. Bafeta, I. Marroun, P. Charles, J. Mantz, and P. Ravaud. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS Med*, 2012.

Part IV

Difficulties

Chapter 8

What is a primary outcome? A corpus study. Anna Koroleva, Elizabeth Wager, Patrick MM Bossuyt. Submitted

Context

We have shown that spin, at least some types of it such as outcome switching, can be detected automatically. We found that it can be useful to return to the starting point of our work with a more informed point of view in order to perform a refinement approach, looking at improving the definition of what is spin and related notions.

In the course of our work on outcome extraction, we realised that the use of the notion "outcome" in both articles and trial registries is far from uniform. What is meant by an "outcome" differs substantially from text to text and from author to author. Although there are medical dictionaries that define the notion "outcome", there does not seem to have been any attempt to address the variability that is observed in the real world data. We believe that describing and understanding the variability in use of the term "outcome" is important in order to ensure its correct understanding and more uniform use in the future.

Hence, in this chapter we report on a study we did to collect and systematise the data on the use of the term "outcome" and its possible synonyms ("end point", "measure"), that we observed in the articles of our corpus and trial registries.

Authors' contributions

All authors (AK, EW, and PMMB) made substantial contributions to the design and implementation of the study and to the interpretation of data. AK collected and analyzed the

data and drafted the manuscript. EW and PMMB revised the draft critically for important intellectual content.

Abstract

Background: Outcomes are a key element for clinical trials. Despite the existing guidance on how to report trial outcomes in an informative and complete manner, the reporting quality remains suboptimal, which poses challenges to both manual and automatic analysis of reports of clinical trials.

Methods: We conducted a corpus study of 1694 outcomes extracted from research articles and 6221 outcome entries extracted from trial registries. We used qualitative and quantitative methods to study the ways primary outcomes are described in the texts of articles reporting randomized controlled trials and in trial registry entries. We assessed the structure of the fields dedicated to outcomes in different registries, length of outcomes, presence of related items (such as measurement tools, time points, analysis metrics used), the consistency in filling in trial registration fields describing an outcome, the ways of introducing outcomes in the text and the observed ambiguities.

Results: We found a substantial diversity in how primary outcomes are defined in articles and registries, in terms of length and the included information. We observed an ambiguity in the use of the words "outcomes", "end-point", "measure", etc., used to introduce an outcome in the text. We summarised the differences in the structure of trials registries in terms of the presence of separate fields for time points and measurement tools and for each of the outcomes of a trial. We describe inconsistencies in the description of primary outcomes in trial registries, in particular, in introducing the information about the time points: the structured form for describing an outcome provided by many registries is often ignored, with time points being presented both in a separate time point field and in the outcome field, or in the outcome field instead of the time point field.

Conclusions: There is a great inconsistency in the presentation of clinical trial outcomes in trial registries and research articles, which creates challenges to analysis of trial reports. Standardizing the terminology of outcome reporting requires substantial time and effort. Increased consultation of trial registries by editors and peer reviewers handling the report of a trial could help reach more consistency between outcomes in registries and articles and increase the rigour and transparency in reporting clinical trials.

Keywords: Outcomes, Research reporting, Corpus study, Trial registries

Introduction

Outcomes are a key element in clinical trials. Trial outcomes should be pre-defined before the trial starts and should not be changed without a reasonable explanation.

Pre-defining trial outcomes is vital to avoid selective reporting of significant or positive outcomes (Andrade, 2015; Ferreira and Maria Patino, 2017). Unjustified changes in the pre-defined trial outcomes (outcome switching) includes reporting only a number of pre-defined outcomes (i.e. selective reporting) or adding new outcomes (Goldacre et al., 2016, 2019).

Outcome switching is a type of reporting bias and spin and leads to unjustified conclusions and recommendations via reporting only the favourable outcomes that support the hypothesis of the authors (or the message of the sponsors) (Boutron et al., 2010; Lockyer et al., 2013; Lazarus et al., 2015; Slade et al., 2015; Weston et al., 2016; Altman et al., 2017; Chiu et al., 2017; Diong et al., 2018; Boutron and Ravaud, 2018). The presence of spin makes clinicians overestimate the effects of the experimental treatment (Boutron et al., 2014), which poses a serious healthcare issue. Besides, spin in research articles provokes spin in health news and press releases (Haneef et al., 2015; Yavchitz et al., 2012), which can raise false expectations regarding the treatment among the readers.

Therefore, consistent and complete reporting of results for all pre-defined outcomes is a matter of high importance. Clear and unambiguous definition of trial outcomes is essential in any assessment of manuscripts by peer reviewers or systematic reviewers, but it is even more important for automated data analysis techniques (e.g. Natural Language Processing - NLP) which are being developed to automate some parts of the systematic review process (cf. Ananiadou et al. (2009); O'Mara-Eves et al. (2015)) and peer review (cf. Gehringer et al. (2018); Kang et al. (2018)).

Automatic extraction of trial outcomes from research articles and registries is an essential part of these applications (Blake and Lucic, 2015). High diversity and ambiguity of natural language texts have always created difficulties for NLP algorithms; on the contrary, concise and well-structured statements are rather easy to analyse automatically. Thus, in order to allow for automated outcome extraction, it is necessary to understand the complexity of this notion and the related textual expressions, and to try and unify the ways outcomes are presented. An attempt to standardize the reporting of trial outcomes was made by the CONSORT Statement for randomized clinical trials (Begg et al., 1996; Schulz et al., 2010), and SPIRIT guideline (Chan et al., 2013) for protocols of interventional trials. However, a number of studies showed that the quality of reporting of clinical trials remains suboptimal, despite the widespread acceptance of the use of guidelines and checklists (Turner et al., 2012; Samaan et al., 2013).

Our primary work lies in the domain of automatic extraction of information, in particular trial outcomes, from medical articles and trial registries. In the course of our research, we noticed that outcomes of clinical trials are highly diverse with regard to the information included in the definition, the ways of introducing an outcome in a text and even the ways of presenting an outcome in registries.

Outcomes of clinical trials can be represented by a binary, numerical or qualitative measure. Measurement tools (questionnaires, scores) can be used to measure qualitative outcomes. Other aspects important for trial outcome definition, as stated in the SPIRIT guideline (Chan et al., 2013) include: analysis metric (e.g., change from baseline, final value, time to event), method of

aggregation (e.g., median, proportion), and time points. Other items that can be included in an outcome description are: the chosen method of analysing results (intent-to-treat, per protocol); covariates that the analysis is adjusted for; reasons for using a particular outcome (such as explanation of relevance, references to previous works using the outcome). An outcome may be a composite of several measures or may serve as a surrogate for another outcome (Ferreira and Maria Patino, 2017).

We conducted this linguistic study based on a corpus of PMC articles and related trial registry entries, aiming at describing the diversity of the explicit definitions of the primary outcomes in medical articles and in trial registries to answer the following questions:

- what is understood by a primary outcome?
- how are primary outcomes presented in an article?
- how many primary outcomes can a trial have?
- which items are commonly included in a definition of a primary outcome?
- what information structure do trial registries use for primary outcomes?
- how similar/different are different trial registries with regard to the information structure?
- how similar/different are definitions of the primary outcome in journal articles compared to those in registries?

Our aim for describing the diversity in presentation of trial outcomes is two-fold. First, the understanding of the characteristics of outcomes is important for the NLP community working on automating the identification of trial outcomes. Second, consistency of outcome presentation in trial registries and research articles is vital to ensure the rigour and transparency in clinical research reporting, which makes our topic interesting for a broad clinical community.

Methods

We conducted a corpus study of research articles reporting randomized controlled trials (RCTs) and related trial registry entries.

Trial registration entries were obtained automatically: trial registration numbers were found in the articles with the help of regular expressions (patterns, consisting of one or more character literals, operators, or constructs, that are used by a search engine to match the input

text[1]); corresponding registry entries were accessed, downloaded in HTML format, which is a structured format for storing data, and parsed with a Python script to extract the target data elements (outcomes, time points, measurement methods) from the downloaded structured registry entry.

The qualitative methods we used consisted in observing and describing the linguistic phenomena in the corpus. The quantitative methods we used include calculation of different numeric parameters of individual phrases (length measured in words and symbols), of the corpus in general (numbers of occurrences of certain words and patterns). In particular, we calculated the number of outcomes that contain words or phrases referring to the measurement tool used (such as "measured as", "rated using"), to time points (such as "day", "baseline", "follow up"), to analysis metric (such as words denoting a change: "increase", "improvement"; words denoting time to event), to aggregation method (such as "mean", "median", "proportion"), to type of analysis (such as "intention-to-treat", "per-protocol"). We also calculated the number of outcomes that state explicitly the between-group comparison performed. Regular expressions used for each search are presented in the Table 8.1. These expressions were manually crafted and tested on the corpus of outcomes.

For trial registry data, we performed the search for time points in outcomes extracted from the registries that do not have separate fields for these types of information and from all the registries, to check if the difference in structure of registries impacts the way outcomes are defined.

Apart from that, we report our observations on the usage of the notion of outcome in literature. Besides, we developed and pilot-tested a survey aimed at primary outcome extraction. We share some experienced from this pilot.

Data

Articles

Our initial text corpus consisted of 113,339 articles automatically downloaded from PMC. In order to select the articles reporting RCTs, we ran a script to automatically check the Publication type field in the metadata of the articles. 47943 articles had no publication type, and 61458 articles had publication type different from "Randomized controlled trial". This check resulted in a corpus of 3938 articles from PubMed Central with publication type "Randomized controlled trial". The articles were published between 2000 and 2012. From these

[1]https://docs.microsoft.com/en-us/dotnet/standard/base-types/regular-expression-language-quick-reference

articles, we randomly selected 2,000 sentences (using Python NLTK library[2] for sentence splitting) containing the words "outcome", "end-point" or other synonyms ("measure, "variable", etc.) and the word "primary" or its equivalent ("core", "main", etc.), where the latter precedes the former and the distance between them is not more than 3 words. We manually annotated primary outcomes with the help of our developed annotation tool. For example, for a phrase "The primary outcome of our trial was progression-free survival" we annotated the fragment "progression-free survival". We annotated the longest text fragment containing information about the primary outcome in a given trial, including time points, measurement tools, etc. Coordinated outcomes were annotated as separate items, while composite outcomes were annotated as one item. 1,694 primary outcomes were annotated.

Figure 8-1 shows the number of articles downloaded, checked for publication type, and included on the final corpus.

We annotated only explicit definitions of the primary outcome. Statements where the information about the primary outcome can potentially be inferred from the context (statements of objectives, description of measured variables) were excluded, e.g.:

This study investigated ***the efficacy (trough forced expiratory volume in 1 second [FEV1] response)*** *and* ***safety*** *of additional treatment with once-daily tiotropium 18g via the HandiHaler in a primary care COPD population. Secondary endpoints included: trough forced vital capacity (FVC) response, weekly use of rescue short-acting beta-agonist, and exacerbation of COPD (complex of respiratory symptoms/events of >3 days in duration requiring a change in treatment).*

Trial registries

We extracted associated entries for the articles of our corpus. We note that this gives us a limited dataset of registry entries; a larger dataset could be obtained by a large-scale web crawling of registries. However, our primary goal remains the analysis of the 3938 articles of our corpus and a comparison of the ways in which outcomes are presented in the journal articles with those in the registries. Thus, we did not perform extraction of outcomes from registry entries unrelated to the articles of our corpus.

If a trial has corresponding entries in several registries, we extracted all entries. We parsed the following registries: Australian New Zealand Clinical Trials Registry (ANZCTR), Brazilian Clinical Trials Registry (ReBec), Chinese Clinical Trial Registry (ChiCTR), Clinical Research Information Service (CriS) of the Republic of Korea, Clinical Trials Registry of India (CTRI), ClinicalTrials.gov, EU Clinical Trials Register (EU-CTR), German Clinical Trials Register

[2]http://www.nltk.org

(Deutsches Register Klinischer Studien – DRKS), International Standard Randomised Controlled Trial Number Register (ISRCTN), Netherlands Trial Registry (NTR), Pan African Clinical Trials Registry (PACTR), Sri Lanka Clinical Trials Registry (SLCTR), University hospital Medical Information Network Clinical Trials Registry (UMIN), and WHO International Clinical Trials Registry Platform.

The structure of registries differ: registries may have a separate field for each of listed outcomes, for time points for each outcome, or for general time points for all outcomes, for measurement method (tool). The structure of registries with regard to presence of separate fields is shown in the Table 8.2. ChiCTR has a separate field for the outcome type (primary/secondary/additional/adverse event).

The WHO portal does not have a visually separated field for time points, but its HTML structure has a tag for outcome time frames. ClinicalTrials.gov does not have either a visually separated field of HTML tag, but the outcome time frames are commonly put in square brackets after the outcome. We considered that these registries have a time points field. Both these registries have HTML tags for each of listed outcomes.

Usage of the separate fields is not mandatory: the fields for time points and measurement methods can be left empty; at the same time, information about time points can be duplicated in the dedicated field and within a description of the corresponding outcome. Several outcomes can be listed in one field, instead of using several fields provided by a registry.

One registry (EU-CTR) uses the phrase "End point" instead of "outcome" in the name of the corresponding field.

Registration of several primary outcomes is allowed by registries: several registries have a possibility of adding multiple fields for primary outcomes, while others use the plural form of the words "outcome"/"end point" in the field name (ANZCTR, ReBec, CriS, ClinicalTrials.gov, EU-CTR, SLCTR, UMIN).

We extracted the field describing the primary outcome. We obtained 6,221 unique outcome registry entries (Table 8.3 shows the numbers of entries from each trial registry in our dataset), with 3,353 outcomes after normalization and deduplication. Time frames, when they are presented in square brackets in the text, were removed from the outcome text before deduplication.

Analysis

Qualitative analysis

How are primary outcomes introduced in a text?

The most typical way of introducing a primary outcome is to use the words "primary outcome"; however, both these words can be replaced by their partial synonyms. For the word "primary", we identified the words "main", "principal", "key" as potential synonyms. For the word "outcome", we identified the words "end-point" (with spelling variants "end point", "endpoint"), "measure", "variable", "parameter", "criterion" as potential synonyms. The most commonly used combinations in our corpus (defined in our search as when the two words occurred together or with no more than 3 words separating the adjective and noun): "primary outcome" - 1222 occurrences, "primary measure" - 458 occurrences, "primary end point" (with spelling variants) – 387 occurrences, "main outcome" = 151 occurrences, "primary variable" = 75 occurrences. The diversity in term usage may create ambiguity as it is not always clear whether the term introduces the primary outcome or another type of variable.

The usage of the words "end point" vs. "outcome" presents an interesting issue. Some sources state that these terms are not exact synonyms. As stated in Curtis et al. (2019), the word "outcome" usually refers to the measured variable (the authors' example: "PROMIS Fatigue score"), while the word "endpoint" is used to refer to the analyzed parameter (the authors' example: "change-from-baseline at 6 weeks in mean PROMIS Fatigue score"). On the other hand, the NCI Dictionary of Cancer Terms interprets an end point as a particular type of outcome1: "an event or outcome that can be measured objectively to determine whether the intervention being studied is beneficial". However, we did not observe any semantic difference between the terms "outcome", "end-point" or "measure" (e.g. they are all used to introduce the quality of life as the main variable measured in the trial). One out of 14 analyzed trial registries used the word "end point" instead of "outcome", supporting the hypothesis that they have synonymous meaning.

However, the word "end-point" is also used to refer to the final time point in a trial, e.g.:

Our primary outcome measures for OCD symptoms were (1) the change in YBOCS score from baseline ***to endpoint*** *and (2) the clinical global impression of improvement (CGI-I)* ***at endpoint****.*

The primary response variable is change in total MADRS-score ***at endpoint*** *versus baseline.*

The ambiguity in use of the word "end-point", the disagreements in definitions of end-point vs. outcome, and the inconsistency between the definitions and the usage of these terms prove

that the usage of terms in this area is still not completely stable, even for such an important and widely used notion as trial outcomes.

What is considered to be an outcome?

As observed from our data, there are substantial differences in what authors mean by "an outcome". There are several dimensions of the diversity in defining an outcome. It concerns first of all the inclusion/exclusion of the relevant items (measurement tools, time points, etc.), describing what, how, when was measured and how it was analysed. Different combinations of items are common:

1. outcome measure name only: "disease burden", "BMI". Note that even for qualitative outcomes "disease burden") measurement tool may not be defined.
2. outcome measure + measurement tool used: "depression measured by the BDI-II", "the QALY based on the EQ-5D".
3. outcome measure + time points: "claim duration (in days) during 12 months follow-up"
4. outcome measure + analysis metric: "change in SBP", "the change in prevalence of atypical cells", "time to progression", "the time between study inclusion and first competitive employment that lasted three months or longer"
5. outcome measure + analysis metric + time points: "change in HOMA index, from week 0 (pre-treatment) to week 6"
6. outcome measure + aggregation metric: "the proportion of women reporting a live birth defined as the delivery of one or more living infants, >20 weeks gestation or 400 g or more birth weight"
7. outcome measure + analysis metric + aggregation metric: "mean IMT-CCA change"
8. outcome measure + analysis metric + time points + aggregation metric: "The mean decrease in HAM-D score from baseline"

Rarely, the type of analysis can be included: "the change in IOP from baseline to week 4 at 8 a.m. and 4 p.m. for the per protocol (PP) population using a "worse eye" analysis".

An outcome description may state that the outcome is a surrogate measure: "a surrogate marker, Ang-2", - and may also refer to the substituted measure: "the active local radiation dose leading to metastasis infiltrating T cells as a surrogate parameter for antitumor activity".

Apart from aspects describing what was measured and how, an outcome may explicitly state the comparison between groups that was performed: "reversal of metabolic syndrome

in the intervention group subjects compared to controls at 12 months follow-up", "6 MWT in the treatment group as compared to control group at 180 days post randomization", "a comparison of the incidence of clinical episodes of malaria in children in the 2 intervention groups, measured by PCD", "A comparison of intra- and post-dialytic complications among study groups", "the difference in NSCL / P recurrence rates between the two groups", "the differences in birth weight between the 2 groups".

Further, the primary outcome can be understood as some target value of a certain variable to be achieved: "a difference of $\geq$20% in the microvascular flow index of small vessels among groups". In this case, "the microvascular flow index of small vessels" is the actual outcome measure, while "a difference of $\geq$20%" refers to the target that the researchers expect to achieve.

Mention of target values can be combined with other items describing an outcome, in particular with aggregation metric (proportion of patients): "the proportion of subjects who achieved targets for compression depth", "the proportion of OHCA patients that achieve the target temperature within six hours of ED arrival". Alternatively, outcome as a target can be expressed as a definition of an event expected in the intervention group, with no mention of the controls: "a decrease in carer burden in the intervention group three months after receiving the DA".

Besides, there is a tendency observed in registries, among our pilot survey participants, in the literature regarding outcomes, and in the systematic reviews: using the whole sentence describing an outcome ("The primary outcome of our trial was X measured as Y at time points 1, 2") as an outcome. In registries, this tendency is observed through the number of entries in the primary outcome field that contain not only the outcome itself, but a free text describing it. In the pilot survey, some participants, when asked to extract the primary outcome from a given text fragment, extract the whole sentence instead of the expected noun phrase, despite the examples provided in instructions. In literature, some works addressing extraction of the trial outcomes, address in fact extraction of the sentences. In systematic reviews, the researchers conducting the information extraction, often copy a whole sentence describing an outcome to the outcome field of the extraction form.

Of note, one of the works on automatic outcome extraction (Demner-Fushman et al., 2006) defines an outcome as "The sentence(s) that best summarizes the consequences of an intervention". While in this case the authors refer to reported outcomes and not to the primary outcome, it is additional evidence of the fact that the term "outcome" may denote not only a concept, but also a whole sentence where the concept is introduced.

Quantitative analysis: how great the diversity in primary outcome definitions is?

Length

The length of outcomes both in the articles and in the registries varied substantially and differed between the outcomes in the articles and those in the registries (Table 8.4).

Syntactic characteristics

Syntactically, the primary outcome definitions were most often represented by a noun phrase. But they can also be represented by a verb phrase (13 / 1694 outcomes): "to determine the time from beginning the scenario to correct insertion of the laryngeal airway after the students' opinion", - or clause (6 / 1694 outcomes): "whether the GP had provided patients with a written asthma action plan (WAAP yes / no)".

Multiple primary outcomes

There are 793 multiple (coordinated) primary outcomes in the sentences of our corpus. At least 800 registry outcomes contain numbered lists.

The way multiple/single outcomes are introduced in a text was not coherent: plural form of the noun "outcome" (or synonyms) can be used to introduce a single outcome:

*Primary **outcomes** are **death** from causes at study end (follow-up until at least 46 weeks after randomization).*

*Primary endpoints The primary **endpoints** of the study are **the primary patency** at 1-year follow-up.*

At the same time, singular form of the noun "outcome (or synonyms) can be used to introduce multiple outcomes:

*The primary **outcome** was **HbA1c**, **lipid levels**, **blood pressure**, **BMI** after 24 months of follow-up.*

Composite outcomes Composite outcomes were uncommon in our corpus: we identified 15 composite outcomes in texts and 54 in registries.

Registry entries

At least 384 registry outcomes contain the word "primary" or synonyms and "outcome" or synonyms at a distance no more than 3 words (search for with regular expression

```
"(primary|main|principal|key)\\s+(\\w+\\s+){0,3}(outcome
|end[ -]*point|measure|variable|parameter|criterion)")
```

, i.e. they contain free-text sentences of paragraphs describing the primary outcomes.

Information included in an outcome definition

140 out of 6,221 registry entries do not have registered primary outcomes ("Not provided at time of registration").

The table 8.5 provides statistics of items that may be included into a definition of a primary outcome, for outcomes from articles and registries. The percentages for registry outcomes are calculated on the basis of 3430 deduplicated outcomes.

Out of 3312 deduplicated outcomes extracted from registries with separate field for time points, 1214 (36.65%) outcomes still contain indications of time points within the text in the outcome field. This fact means that the researchers either do not use the structure provided by registries (indication time points in the outcome field), or duplicate this information (in both outcome field and time points field).

Out of 6029 non-deduplicated outcomes extracted from registries with separate field for time points,

2731 (45.3%) have non-empty timepoint field. 584 (9.69%) outcomes have a non-empty timepoint field and an indication of time point within the definition of outcome, e.g.:

Change from Baseline in Dietary KAB Score at 12 months [Time Frame: Baseline & 12 months]

1963 (32.56%) outcomes have an empty time point field and do not have any indication of time in the outcome definition.

Out of 963 deduplicated outcomes extracted from registries without a separate field for time points, 629 (65.32%) contain indications of time.

Discussion

When results for outcomes of a trial are reported, they often include several of outcome-related information items, e.g.:

- Time points, analysis metrics:

  *There were similar, significant **improvements** in functional capacity for the RT and NMES groups at **week 8** compared **to week 1** ($p \leq 0.001$) and compared to the control group ($p < 0.005$).*

- Time points, aggregation metrics:

***At 5 min post-dose on Day 1**, the **mean** FEV 1 for both indacaterol doses was significantly higher than placebo (by 120 and 130 mL for indacaterol 150 and 300 μ g, respectively; $p < 0.001$) and tiotropium (by 80 mL for both doses; $p < 0.001$).*

- Covariates:

  *Total nutrition knowledge score at follow-up, adjusted for **baseline score**, **deprivation**, and **school size**, was higher in intervention than in control schools (mean difference = 1.1 ; 95% CI: 0.05 to 2.16; $p = 0.042$).*

- Type of analysis:

  *Results For BMD, no **intent-to-treat analyses** were statistically significant; however, **per protocol analyses** (ie, only including TC participants who completed ≥75% training requirements) of femoral neck BMD changes were significantly different between TC and UC (+0.04 vs -0.98%; $P = 0.05$).*

However, as we have shown, these items are rarely present in the description of outcomes. This fact may pose difficulties when assessing an article for outcome switching and reporting bias, as it is impossible to define whether all the aspects of a pre-defined outcome correspond to a reported outcome.

Limitations

This study has a number of limitations.

First, our data covers only the entries from trial registries that correspond to the articles in our initial corpus. In particular, our sample is imbalanced with regard to the number of entries from different registries. Obtaining more data from registries could be interesting to get a more complete view of the topic but it is outside the scope of our current work.

One difficulty in working with registry data is handling duplicate entries. We collected data from primary registries as well as through the WHO portal. As a result, for some trials we obtained duplicate entries. To account for this, we removed duplicate outcomes when analysing outcomes from registries for occurrence of information elements. However, to analyse the proportion of empty/non-empty time point fields, the analysis was conducted without the deduplication step because it requires analysis of the registry structure, thus entry from each registry is considered separately, even if they refer to the same trial.

Second, we did not annotate all the sentences potentially introducing primary outcomes in the articles. The total number of sentences that contain the words "outcome" or its synonyms and the word "primary" or its equivalent, where the latter precedes the former within a distance

of 3 words, is over 10,000, out of which we annotated 2,000 sentences. Annotation of all the sentences would require significantly more time and effort, which the current annotated set already provides substantial data to observe the variability in outcome definitions.

Third, some observations that we report from our pilot survey aimed at outcome extraction are based on a small sample of participants. To estimate the generalizability of these observations, a large-scale survey would be needed. However, these observations are in line with tendencies in other data sources, which makes them interesting as supporting findings.

Conclusions

In this corpus study, we showed that primary outcome definitions in medical articles and in trial registries vary significantly in terms of both length and the included information. Registry entries tend to contain outcome-related information items, such as time points, measurement methods, analysis metrics, type of analysis, more often than outcomes from journal articles. Information on aggregation metric, type of analysis, and covariates used is rarely included in outcome definitions, although these are vital for understanding the method of analysis of the outcome in question.

Multiple primary outcomes are common: only 53.2% of annotated outcomes in the articles were single. Trial registries do not discourage the practice of having multiple outcomes.

There is some redundancy in defining outcomes in articles and registries. First, outcome definitions may include definition of comparison between treatment groups, which is already implied by study design. Second, some registry outcomes include a full-text definition of outcome ("The primary outcome was ..."), instead of the outcome name only.

The ways of introducing an outcome in the text vary and have some ambiguity regarding the use of the synonyms of the word "outcome" ("end-point", "measure") and of the word "primary". This can cause incorrect classification of an outcome as the primary one.

The understanding of the term "outcome" can also differ, as some authors interpret outcome not as a measured variable, but as a target value for a variable. Besides, the term "outcome" may denote not only a concept, but also a whole sentence where the concept is introduced.

Primary outcomes can be introduced implicitly, as shown by examples, although we have not studied these cases and their prevalence in detail.

We provided an overview of differences in structure of a number of trial registries: some registries have separate fields for certain types of information (time points, measurement methods) or for each of the listed outcomes, thus encouraging a structured description of outcomes, while other registries have only one field for all outcome-related information.

We showed that researchers registering trials are not consistent in filling registry fields:

36.65% of outcomes in registries with separate field for time points have the indication of time points in the outcome field. Some researchers fill the timepoints field but still mention time points in the outcome field (9.69%). At the same time, more than a half of outcomes from these registries have an empty timepoint field and no indication of time points in the outcome field.

All the listed discrepancies in defining an outcome, observed even in a rather small data sample, prove that the notion of a primary outcome is not as clear and well-defined as it may seem to be, given the scale and importance of its use. This may lead to difficulties in manual extraction of trial outcomes by researchers conducting systematic reviews and in assessments of articles for outcome reporting bias and outcome switching. Further, highly diverse ways of defining an outcome create obstacles for automatic extraction of outcomes from medical articles and registries: in particular, to check for outcome switching, the computer needs to extract all the outcome-related types of information separately, while currently they are mixed both in registries (despite the expectation that registries provide a structured presentation of information) and in article texts.

This work revealed a great inconsistency in the ways of how outcomes of clinical trials are defined and represented. We envisage that the lack of standardization in presenting trial outcomes is unlikely to change in the near future, as standardizing the terminology of a domain is always a challenge, it requires substantial time and effort. This fact implies that, in the foreseeable future, any automated tools aiming to detect outcomes of clinical trials and, consequently, recognise outcome-related spin will be facing the challenges resulting from variability and lack of consistency in presentation of outcomes.

As a possible way of tackling the problem of inconsistency in outcome presentation, we can suggest journal editors and peer reviewers to increase the consultation of trial registries when handling the report of a trial, to aim for more consistency between outcomes in registries and articles. This could help to increase the rigour and transparency in reporting clinical trials and to reduce potential spin in reporting trial outcomes, contributing to fair representation of trial results.

Availability of data and material

The data collected and analyzed in this article is freely available in the Zenodo repository:

`http://doi.org/10.5281/zenodo.3343029`

`http://doi.org/10.5281/zenodo.3234811`

Author's contributions

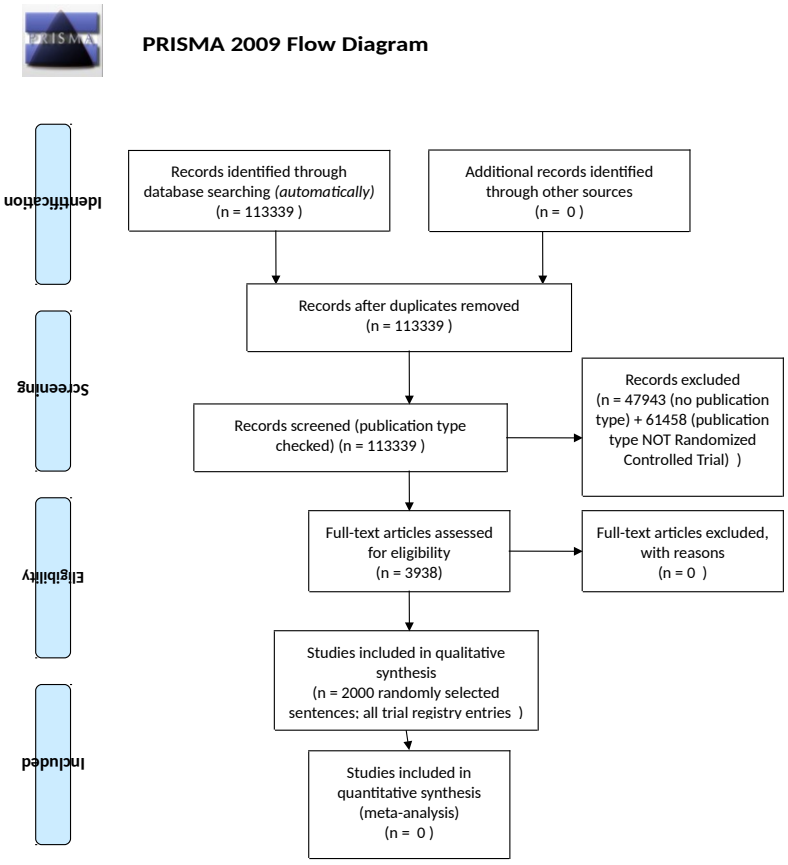
All authors (AK, EW, and PMMB) made substantial contributions to the design and implementation of the study and to the interpretation of data. AK collected and analyzed the data and drafted the manuscript. EW and PMMB revised the draft critically for important intellectual content. All authors read and approved the final manuscript.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

Figures

Figure 8-1: PRISMA flow diagram



From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit www.prisma-statement.org.

Tables

Item	Regular expression
Measure-ment tool	`^.*((assessed|measured|investigated|analyzed|studied|examined|evaluated|estimated|calculated|explored|tested|recorded|based|defined|rated|quantified|marked|determined|considered)\s+(with|by|as|using|on|through)|using).*$`
Time points	`^.*\b(time[ -]*(point|frame)|year|month|week|day|hour|period|baseline|decade|follow[- ]?up|t\d+)s?\b.*$`
Analysis metric	`^.*\b(alteration|amelioration|change|decline|decrease|drop|elevation|enhancement|fall|gain|improvement|increase|loss|raise|reduction|rise|worsening|time\s+(to|between))\b.*$`
Aggregation metric	`^.*\b((average|mean|median|AUC)\b|(proportion|percent|percentage) of).*$`
Type of analysis	`^.*\b(per[- ]?protocol|intent\S*[- ]to[- ]treat|PP|ITT)\b.*$`
Covariates	`^.*(adjust|co[ -]*varia).*$`
Comparison	`^.*(comparison|\b(between|among|in|with|to|compared|difference|relative)\s+(\w+\s+){0,3}(group|arm|treatments|interventions|control)).*$`

Table 8.1: Structure of outcome-related fields in trial registries

Registry	**Separate fields for:**				**No separate fields**
	time points – one for each outcome	time points – one for all outcomes	measurement method	each of listed outcomes	
Australian New Zealand Clinical Trials Registry (ANZCTR)	+	-	-	+	-
Brazilian Clinical Trials Registry (ReBec)	-	-	-	-	+
Chinese Clinical Trial Registry (ChiCTR)	+	-	+	+	-
Clinical Research Information Service (CriS) of the Republic of Korea	+	-	-	+	-
Clinical Trials Registry of India (CTRI)	+	-	-	+	-
ClinicalTrials.gov	+	-	-	+	-
EU Clinical Trials Register (EUCTR)	-	+	-	-	-
German Clinical Trials Register (Deutsches Register Klinischer Studien – DRKS)	-	-	-	-	+
International Standard Randomised Controlled Trial Number Register (ISRCTN)	-	-	-	-	+
Netherlands Trial Registry (NTR)	-	+	-	-	-
Pan African Clinical Trials Registry (PACTR)	-	+	-	-	-
Sri Lanka Clinical Trials Registry (SLCTR)	-	+	-	-	-
University hospital Medical Information Network Clinical Trials Registry (UMIN)	-	-	-	-	+
WHO International Clinical Trials Registry Platform	+	-	-	+	-

Table 8.2: Structure of outcome-related fields in trial registries

Registry	Number of outcomes
WHO	3386
ISRCTN	1146
NCT	646
ACTRN	564
NTR	332
EUDRACT	49
JPRN-UMIN	32
DRKS	27
CHICTR	22
KCT	10
CTRI	4
SLCTR	3
Total	6221

Table 8.3: Number of entries per registry

Registry outcomes				
	min	max	mean	st.dev
characters	3	5009	162.01	286.85
words	1	900	28.36	51.39
sentences	1	35	1.82	2.31
Articles outcomes				
	min	max	mean	st.dev
characters	2	505	53.3	48.91
words	1	116	9.07	8.85
sentences	1	1	1.0	0.0

Table 8.4: Length of outcomes in registries and articles

Item	**Outcomes in articles (number/proportion)**	**Outcomes in registries (number/proportion)**
Measurement tool	196 (11.57%)	781 (23.29%)
Time points	422 (24.91%)	1241 (37.01%)
Analysis metric	212 (12.51%)	541 (16.13%)
Aggregation metric	121 (7.14%)	268 (7.99%)
Type of analysis	2 (0.12%)	17 (0.51%)
Covariates	8 (0.47%)	33 (0.98%)
Comparison	52 (3.07%)	152 (4.53%)

Table 8.5: Number/proportion of information items in outcomes in registries and articles

References

D. Altman, D. Moher, and K. Schulz. Harms of outcome switching in reports of randomised trials: Consort perspective. *BMJ: British Medical Journal (Online)*, 2017.

S. Ananiadou, B. Rea, N. Okazaki, R. Procter, and J. Thomas. Supporting systematic reviews using text mining. *Social Science Computer Review - SOC SCI COMPUT REV*, 27:509–523, 10 2009. doi: 10.1177/0894439309332293.

C. Andrade. The primary outcome measure and its importance in clinical trials. *The Journal of clinical psychiatry*, 76:e1320–e1323, 11 2015. doi: 10.4088/JCP.15f10377.

C. Begg, M. Cho, S. Eastwood, R. Horton, D. Moher, I. Olkin, R. Pitkin, D. Rennie, K. F. Schulz, D. Simel, and D. F. Stroup. Improving the Quality of Reporting of Randomized Controlled Trials: The CONSORT Statement. *JAMA*, 276(8):637–639, 08 1996. ISSN 0098-7484. doi: 10.1001/jama.1996.03540080059030. URL `https://doi.org/10.1001/jama.1996.03540080059030`.

C. Blake and A. Lucic. Automatic endpoint detection to support the systematic review process. *J. Biomed. Inform*, 2015.

I. Boutron and P. Ravaud. Misrepresentation and distortion of research in biomedical literature. *Proc Natl Acad Sci U S A.*, 2018. doi: 10.1073/pnas.1710755115.

I. Boutron, S. Dutton, P. Ravaud, and D. Altman. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*, 2010.

I. Boutron, D. Altman, S. Hopewell, F. Vera-Badillo, I. Tannock, and P. Ravaud. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *Journal of Clinical Oncology*, 2014.

A.-W. Chan, J. Tetzlaff, D. Altman, K. Dickersin, and D. Moher. Spirit: New guidance for content of clinical trial protocols. *Lancet*, 381, 2013.

K. Chiu, Q. Grundy, and L. Bero. 'Spin' in published biomedical literature: A methodological systematic review. *PLoS Biol.*, 2017. doi: 10.1371/journal.pbio.2002173.

L. Curtis, A. Hernandez, and K. Weinfurt. Choosing and specifying endpoints and outcomes: Introduction. In *Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials*. NIH Health Care Systems Research Collaboratory, Bethesda, MD, 2019.

D. Demner-Fushman, B. Few, S. Hauser, and G. Thoma. Automatically identifying health outcome information in medline records. *Journal of the American Medical Informatics Association*, 2006.

J. Diong, A. Butler, S. Gandevia, and M. Héroux. Poor statistical reporting, inadequate data presentation and spin persist despite editorial advice. *PLoS One*, 2018. doi: 10.1371/journal.pone.0202121.

J. Ferreira and C. Maria Patino. Types of outcomes in clinical research. *Jornal Brasileiro de Pneumologia*, 43:5–5, 02 2017. doi: 10.1590/s1806-37562017000000021.

E. F. Gehringer, F. Pramudianto, A. Medhekar, C. Rajasekar, , and Z. Xiao. Board 62: Applications of artificial intelligence in peer assessment. In *2018 ASEE Annual Conference & Exposition*, Salt Lake City, Utah, June 2018. ASEE Conferences. https://peer.asee.org/30073.

B. Goldacre, H. Drysdale, A. Powell-Smith, A. Dale, I. Milosevic, E. Slade, P. Hartley, C. Marston, K. Mahtani, and C. Heneghan. The compare trials project. 2016. URL `www.COMPare-trials.org`.

B. Goldacre, H. Drysdale, A. Dale, I. Milosevic, E. Slade, P. Hartley, C. Marston, A. Powell-Smith, C. Heneghan, and K. R. Mahtani. Compare: a prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*, 20(1):118, Feb 2019. ISSN 1745-6215. doi: 10.1186/s13063-019-3173-2. URL `https://doi.org/10.1186/s13063-019-3173-2`.

R. Haneef, C. Lazarus, P. Ravaud, A. Yavchitz, and I. Boutron. Interpretation of results of studies evaluating an intervention highlighted in google health news: a cross-sectional study of news. *PLoS ONE*, 2015.

D. Kang, W. Ammar, B. Dalvi, M. van Zuylen, S. Kohlmeier, E. Hovy, and R. Schwartz. A dataset of peer reviews (peerread): Collection, insights and nlp applications. 04 2018.

C. Lazarus, R. Haneef, P. Ravaud, and I. Boutron. Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. *BMC Med Res Methodol.*, 2015.

S. Lockyer, R. Hodgson, J. Dumville, and N. Cullum. "Spin" in wound care research: the reporting and interpretation of randomized controlled trials with statistically non-significant primary outcome results or unspecified primary outcomes. *Trials*, 2013. doi: 10.1186/1745-6215-14-371.

A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews*, 4(1):5, Jan 2015. ISSN 2046-4053. doi: 10.1186/2046-4053-4-5. URL https://doi.org/10.1186/2046-4053-4-5.

Z. Samaan, L. Mbuagbaw, D. Kosa, V. Borg Debono, R. Dillenburg, S. Zhang, V. Fruci, B. Dennis, M. Bawor, and L. Thabane. A systematic scoping review of adherence to reporting guidelines in health care literature. *Journal of multidisciplinary healthcare*, 6:169–88, 05 2013. doi: 10.2147/JMDH.S43952.

K. F. Schulz, D. G. Altman, and D. Moher. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340, 2010. ISSN 0959-8138. doi: 10.1136/bmj.c332. URL https://www.bmj.com/content/340/bmj.c332.

E. Slade, H. Drysdale, and B. Goldacre. Discrepancies between prespecified and reported outcomes. *BMJ*, 2015. URL http://www.bmj.com/content/351/bmj.h5627/rr-12.

L. Turner, L. Shamseer, D. Altman, K. F Schulz, and D. Moher. Does use of the consort statement impact the completeness of reporting of randomised controlled trials published in medical journals? a cochrane review. *Systematic reviews*, 1:60, 11 2012. doi: 10.1186/2046-4053-1-60.

J. Weston, K. Dwan, D. Altman, M. Clarke, C. Gamble, S. Schroter, P. Williamson, and J. Kirkham. Feasibility study to examine discrepancy rates in prespecified and reported outcomes in articles submitted to the bmj. *BMJ Open*, 2016. doi: 10.1136/bmjopen-2015-010075. URL http://bmjopen.bmj.com/content/6/4/e010075.

A. Yavchitz, I. Boutron, A. Bafeta, I. Marroun, P. Charles, J. Mantz, and P. Ravaud. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS Med*, 2012.

Summary

In this thesis, we report on our work on developing Natural Language Processing (NLP) algorithms to aid readers and authors of scientific (biomedical) articles in detecting spin (distorted presentation of research results). Our algorithm focuses on spin in abstracts of articles reporting Randomized Controlled Trials (RCTs).

We studied the phenomenon of spin from the linguistic point of view to create a description of its textual features. We annotated a set of corpora for the key tasks of our spin detection pipeline: extraction of declared (primary) and reported outcomes, assessment of semantic similarity of pairs of trial outcomes, and extraction of relations between reported outcomes and their statistical significance levels. Besides, we annotated two smaller corpora for identification of statements of similarity of treatments and of within-group comparisons.

We developed and tested a number of rule-based and machine learning algorithms for the key tasks of spin detection (outcome extraction, outcome similarity assessment, and outcome-significance relation extraction). The best performance was shown by a deep learning approach that consists in fine-tuning deep pre-trained domain-specific language representations (BioBERT and SciBERT models) for our downstream tasks. This approach was implemented in our spin detection prototype system, called DeSpin, released as open source code.

Our prototype includes some other important algorithms, such as text structure analysis (identification of the abstract of an article, identification of sections within the abstract), detection of statements of similarity of treatments and of within-group comparisons, extraction of data from trial registries. Identification of abstract sections is performed with a deep learning approach using the fine-tuned BioBERT model, while other tasks are performed using a rule-based approach.

Our prototype system includes a simple annotation and visualization interface.

Chapter 1 presented our first steps in developing NLP algorithms for automatic detection of spin in biomedical articles. We proposed a scheme for an algorithm for automatic extraction of important statements in the abstracts of biomedical articles and possible supporting information. We addressed three tasks related to spin detection: classification of articles according to the type of clinical trial (to detect RCT reports), classification of sentences in the abstracts

aimed at identifying the Results and Conclusions sections, and entity extraction (for trial outcomes and population studied). We reviewed the state of the art and our first experiments for these tasks. For each task, we suggested some possible directions of future work.

To evaluate our rule-based algorithms described in the articles above and to train machine learning algorithms, a corpus annotated with the relevant information is necessary. Chapter 2 describes our efforts in collecting a corpus of biomedical articles and annotating it for spin and related information. The paper presented an annotation scheme for spin and related information elements, our annotation guidelines, and difficulties that we faced, such as the level of expertise required from annotators, choice of an annotation tool, and the complexity of the task.

Chapter 3 described our experiments on using deep learning for the task of extracting trial outcomes – variables monitored during clinical trials. Extraction of declared (primary) and reported outcomes is a key task for spin detection. In this paper, we reviewed the state of the art for outcome extraction. We introduced our manually annotated corpus of 2,000 sentences with declared (primary) outcomes and 1,940 sentences with reported outcomes, which is freely available. We compared two deep learning approaches: a simple fine-tuning approach and an approach using CRF and Bi-LSTM with character embeddings and embeddings derived from pre-trained language models. We employed and compared several pre-trained language representation models, including BERT (Bidirectional Encoder Representations from Transformers), BioBERT and SciBERT. We compared these approaches to the previously implemented rule-based baseline (described in previous chapters). The best achieved results were the token-level F-measure of 88.52% for primary outcomes (BioBERT fine-tuned model) and 79.42% for reported outcomes (SciBERT fine-tuned model).

Chapter 4 described the development of an algorithm for semantic similarity assessment of pairs of trial outcomes, as a part of spin detecting pipeline. We aimed at building an algorithm that does not require manually curated domain-specific resources such as ontologies and thesauri. Based on the corpus annotated for primary and reported outcomes (described in the previous chapter), we annotated pairs of primary and reported outcomes for semantic similarity (on binary scale). The corpus is freely available. We created an expanded corpus by adding the variants for referring to an outcome (e.g. the use of a measurement tool name instead on the outcome name; the use of abbreviations).

As a baseline, we used a number of single semantic similarity measures, based on strings, tokens and lemmas, distances between phrases in the WordNet semantic network, and vector representations of phrases. We trained and tested a number of machine learning classifiers using a combination of the single similarity measures as features. Finally, we employed a deep learning approach that consists in fine-tuning pre-trained deep language representations on the

corpus of outcome pairs. We tested several language models: BERT (trained on general-domain texts), BioBERT and SciBERT (trained on biomedical and scientific texts, respectively). The deep learning approach proved to be superior to other tested approaches. The best result on the original corpus was shown by the fine-tuned BioBERT model, with the F-measure of 89.75%. On the expanded corpus, the performance of deep learning algorithms improved compared to that on the original corpus: BioBERT achieved an F-measure of 93.38%.

Chapter 5 reported on combining the outcome extraction and semantic similarity assessment algorithms (described in the two previous chapters) for developing an algorithm for detection of outcome switching - a type of spin consisting in unjustified change (omitting or adding) of pre-defined outcomes of a trial. We focused on the primary outcome switching. We annotated a corpus with information needed to detect outcome switching: 2,000 sentences with 1,694 primary outcomes; 1,940 sentences with 2,251 reported outcomes; and 3,043 pairs of outcomes annotated for semantic similarity. We employed a combination of information extraction, structured data parsing, and semantic similarity assessment methods to identify primary outcome switching. The semantic similarity assessment algorithms were evaluated on the original corpus and on a corpus expanded with variants of referring to an outcome. The best performance achieved was the F-measure of 88.42% for primary outcome extraction, 79.42% for reported outcome extraction, and 89.75% and 93.38% on the original and expanded versions of the corpus for semantic similarity evaluation.

Statistical hypothesis testing is commonly used in RCTs to test if the experimental intervention is superior to the control one. Statistical significance levels are often reported for the trial outcomes. Chapter 6 reported on the development of an algorithm for extracting the relation between trial outcomes and statistical significance levels. We annotated a corpus of 663 sentences with 2,552 relations between outcomes and significance levels (1,372 positive and 1,180 negative relations). We briefly described our algorithms for entity extraction (reported outcomes and significance levels) and provided their evaluation. In our relation extraction experiments, we assumed that the entities (reported outcomes and significance levels) were extracted at a previous step and could be given to the relation extraction algorithm as input. We evaluated two approaches for relation extraction: machine learning classifiers, using manually crafted feature set, and a deep learning approach consisting of fine-tuning pre-trained language models (BERT, BioBERT and SciBERT). The deep learning approach proved to be superior to the machine classifiers using manually crafted feature set. The BioBERT fine-tuned model showed the best performance for the relation extraction (F-measure of 94%).

Chapter 7 outlined the textual features of the types of spin we addressed and introduces our spin detection prototype system, called DeSpin (Detector of Spin). DeSpin is a Natural Language Processing (NLP) system for semi-automatic detection of spin in scientific articles,

in particular in reports of randomized controlled trial (RCTs). DeSpin combines rule-based and machine learning approaches to detect potential spin and related information. Our algorithms achieved operational performance with the F-measure ranging from 79.42 to 97.86% for different tasks. The most difficult task is extracting reported outcomes. The proposed tool can be used by both authors and reviewers to detect potential spin. The tool and the annotated dataset are freely available.

Chapter 8 reported on a corpus study of 1,694 primary outcomes extracted from research articles and 6,221 primary outcome entries extracted from trial registries. We studied how primary outcomes are described in the articles reporting RCTs and in trial registries. We assessed the structure of the outcome field in registries, length of outcome descriptions, presence of related items (measurement tools, time points, analysis metrics used, etc.), the consistency of filling in trial registration fields describing an outcome, the ways of introducing outcomes in the text and the observed ambiguities. Our work showed that there is a high diversity in the way outcomes are defined. We observed an ambiguity in the use of the terms "outcomes", "end-point", "measure", and others that are used to introduce an outcome in the text. The structure of trials registries differs in terms of the presence of separate fields for time points and measurement tools and for each of the outcomes of a trial. We observed inconsistencies in introducing the information about the time points in registries: the structured form for describing an outcome provided by registries is often ignored, time points being presented both in a separate time point field and in the outcome field, or in the outcome field instead of the time point field. The observed inconsistencies pose challenges to both manual and automatic analysis of clinical trial reports. There is a need to Standardise the terminology of outcome reporting.

There are a number of important directions for future research on automatic spin detection.

First of all, our prototype system DeSpin needs to be tested with users to estimate its usability and potential gains in spin detection task performed by humans. The prototype can be employed in several ways: it can serve as an aid for editors and peer reviewers in assessing a submitted article for spin; it can serve as a writing aid tool to help authors check and improve their manuscripts before submission; and it can be used as an educational tool, defining various types of spin, explaining why they represent inappropriate reporting practices, and how to avoid them. The algorithms implemented in our prototype can be built in other systems of medical text analysis or writing aid.

Other directions for future work include the following.

1. Implementing new algorithms for detection of types of spin not covered by our work.

   These include e.g. detection of focus on subgroups of patients, or, vice versa, inadequate extrapolation for a wider population than that studied. Detection of the spin of this type

would require extracting the population studied (from the article's text or, alternatively, from the associated trial registry record) and reported in the abstract, and assessment of their semantic similarity. Another algorithm vital for spin detection is assessing the polarity of results and conclusions (positive - neutral - negative) in a given article with regard to the treatment studied. Positive statements about the treatment in the results and conclusions can represent spin when the primary outcome is not significant.

2. Detecting the types of spin we addressed in a broader setting.

   In particular, we addressed the detection of switching of primary outcomes, but we did not address switching of secondary outcomes. Detection of this type of spin would require an additional algorithm for the extraction of secondary outcomes. Our semantic similarity assessment algorithm can consequently be applied to compare secondary and reported outcomes to find mismatches.

3. Identifying spin in other types of medical research articles apart from RCTs.

   Spin can occur in non-randomized trials, systematic reviews and meta-analyses, diagnostic accuracy studies. Certain types of spin can be common for different types of research, while other types are specific for a particular study type and thus require separate algorithms for their detection.

4. Identifying spin in other types of texts related to medical research, such as news items and press releases, that were shown to contain spin, often stemming from spin in the related research article.

5. Exploring and defining the phenomenon of spin in other research domains; e.g. NLP or computer science. Until now, spin in research has mainly been addressed in the medical domain, where it can pose a direct threat to public healthcare. Still, research in other domains is also prone to spin, as the reasons for "spinning" research results are likely to be common for various domains. However, research domains other than medicine often lack precise reporting guidelines, leaving more freedom in what and how to report. This fact can complexify the definition of spin.

Spin is a complex phenomenon that even humans can find difficult to identify. Still, we believe that our project demonstrated that detection of spin and related information an be automated using NLP techniques with sufficient performance to serve as assistance to authors and readers of scientific articles. As our project was the first one to tackle automatic detection of spin, we hope that future research will address this topic, bring better results and provide new tools.

Samenvatting

In dit proefschrift brengen we verslag uit van ons onderzoek rond de ontwikkeling van NLP-algoritmen (Natural Language Processing) om lezers en auteurs van wetenschappelijke (biomedische) artikelen te helpen bij het detecteren van "spin": de vervormde presentatie van onderzoek, waardoor een geflatteerde, te rooskleuriger of zelfs ronduit misleidend beeld ontstaat van de resultaten van dat onderzoek. Ons onderzoek richtte zich primair op spin in artikelen die Randomized Controlled Trials (RCT's) beschrijven.

We bestudeerden het fenomeen spin eerst vanuit taalkundig oogpunt, om op basis hiervan een beschrijving van de tekstuele kenmerken te kunnen maken. We hebben een aantal corpora geannoteerd voor de kerntaken in onze pijplijn voor de detectie van spin: dat zijn de extractie van (primaire) en gerapporteerde uitkomsten, de beoordeling van de semantische gelijkenis binnen paren van uitkomsten, en de extractie van relaties tussen gerapporteerde uitkomsten en bijpassende statistische p-waarden. Daarnaast hebben we twee kleinere corpora geannoteerd, voor de identificatie van verklaringen van gelijkwaardigheid van behandelingen en van binnengroepsvergelijkingen (in tegenstelling tot de gebruikelijke vergelijkingen tussen groepen in RCT).

We hebben een aantal op regels en op machine learning gebaseerde algoritmen ontwikkeld en getest voor de sleuteltaken bij de detectie van spin. De beste prestaties werden waargenomen voor benadering op basis van deep learning, die bestaat uit het finetunen van vooraf getrainde, domeinspecifieke taalrepresentaties (BioBERT- en SciBERT modellen). Met deze aanpak bouwden we ons prototype voor de detectie van spin, genaamd DeSpin.

Ons prototype bevat andere algoritmen, zoals voor tekststructuuranalyse (identificatie van de samenvatting van een artikel, identificatie van secties binnen de samenvatting), voor de detectie van verklaringen van gelijkwaardigheid van behandelingen en van vergelijkingen binnen de groep, en voor de extractie van gegevens uit trialregisters. Identificatie van de onderdelen in een samenvatting bij een onderzoeksverslag wordt uitgevoerd met een benadering op basis van deep learning, met behulp van een aangepast BioBERT-model, terwijl andere taken worden uitgevoerd met behulp van een op regels gebaseerde benadering. Ons prototypes kent een eenvoudige interface voor annotatie en visualisatie.

Hoofdstuk 1 presenteert onze eerste stappen in het ontwikkelen van NLP-algoritmen voor de automatische detectie van spin in biomedische artikelen. Het beschrijft een algoritme voor de automatische extractie van uitspraken in samenvattingen bij biomedische artikelen en eventuele ondersteunende informatie. We hebben drie taken met betrekking tot spin-detectie aangepakt: de classificatie van artikelen over clinical trials (om verslagen van RCT's als dusdanig te herkennen), de classificatie van zinnen in de samenvattingen die resultaten en conclusies betreffen, de extractie van beschrijvingen van de uitkomsten en van de onderzoeksgroep. We hebben de stand van zaken samengevat alsook onze eerste experimenten voor deze taken beschreven. Voor elke taak hebben we ook enkele richtingen voor eventueel vervolgonderzoek voorgesteld.

Om onze op regels gebaseerde algoritmen te evalueren en om algoritmen voor machine learning te trainen hadden we een corpus met de relevante informatie nodig. Hoofdstuk 2 beschrijft onze inspanningen bij het bouwen van een corpus van biomedische artikelen en het annoteren voor spin. We presenteren een annotatieschema voor spin, onze annotatierichtlijnen, en schetsten de moeilijkheden waarmee we te kampen hadden, zoals het kennisniveau waarover de annotatoren moeten kunnen beschikken, de keuze van een annotatietool en de complexiteit van de taak zelf.

Hoofdstuk 3 bevat een beschrijving van onze experimenten rond het gebruik van deep learning voor de extractie van de uitkomsten in een RCT. Een sleuteltaak in dit proces is het detecteren van de beoogde primaire uitkomstmaat en van de feitelijke gerapporteerde uitkomstmaten. We gingen na hoever we staan in de techniek voor de extractie van dit soort uitkomsten. We beschrijven ons handmatig geannoteerd corpus van 2,000 zinnen met de primaire uitkomstmaten en 1,940 zinnen met gerapporteerde uitkomsten. Dit corpus is vrij beschikbaar. We vergeleken twee benaderingen voor deep learning: een eenvoudige fine-tuning en een benadering met behulp van CRF en Bi-LSTM met character embeddings en embeddings van vooraf getrainde taalmodellen. We gebruikten en vergeleken verschillende taalrepresentatiemodellen, waaronder BERT (Bidirectional Encoder Representations from Transformers), BioBERT en SciBERT. We hebben deze benaderingen vergeleken met de vorige, op regels gebaseerde aanpak. De beste resultaten waren de token-level F-measure van 88.52% voor primaire uitkomsten (met het BioBERT fine-tuned model) en 79.42% voor gerapporteerde uitkomsten (met het SciBERT fine-tuned model).

Hoofdstuk 4 beschrijft de ontwikkeling van een algoritme voor beoordeling van semantische overeenkomsten tussen paren uitkomsten in een RCT, een onderdeel van onze pijplijn voor spin-detectie. We wilden een algoritme bouwen dat geen handmatig beheerde, domeinspecifieke bronnen nodig heeft, zoals ontologieën of thesauri. Met het eerder beschreven corpus hebben we paren van primaire en gerapporteerde uitkomsten geannoteerd op basis van hun semantische gelijkenis (op een binaire schaal). We hebben een uitgebreid corpus gecreëerd

door ook varianten in de beschrijving van een uitkomst toe te voegen. Dat kan bijvoorbeeld de naam van een meetinstrument zijn, in plaats van de uitkomst met naam te noemen, of door het gebruik van afkortingen.

We startten met een aantal enkele maten voor semantische overeenkomst, op basis van tekenreeksen, tokens en lemma's, afstanden tussen woordgroepen in een semantisch WordNet netwerk en vectorrepresentaties van zinnen. We hebben een aantal classifiers op basis van machine learning getraind en getest, met behulp van een combinatie van de eerder genoemde maten voor semantische overeenkomst. Tot slot hebben we een methode op basis van deep learning toegepast, die bestaat uit het verfijnen van deep language representaties op het corpus van uitkomstparen. We hebben verschillende taalmodellen getest: BERT (getraind in algemene domeinteksten), BioBERT en SciBERT (getraind op respectievelijk biomedische en wetenschappelijke teksten). De aanpak op basis van deep learning bleek superieur aan de andere geteste benaderingen. De beste resultaten werden bereikt met het verfijnde BioBERT-model, met een F-maat van 89.75%. Op het uitgebreide corpus zagen we betere prestaties van de deep learning algoritmen, vergeleken met die op het oorspronkelijke corpus: BioBERT behaalde daar een F-maat van 93.38%.

Hoofdstuk 5 beschrijft onze inspanningen in het combineren van de uitkomstextractie en algoritmen voor het beoordelen van semantische overeenkomst. Hiermee wilden we een algoritme ontwikkelen voor de detectie van uitkomstruil: een vorm van "spin" die gebaseerd is op het ongerechtvaardigd wijzigen van vooraf gedefinieerde uitkomsten in een RCT (weglaten of toevoegen). We hebben ons primair gericht op het veranderen van de primaire uitkomst. Hiervoor hebben een corpus geannoteerd met de informatie die nodig is om het smokkelen in uitkomsten te detecteren: 2,000 zinnen met 1,694 primaire uitkomsten, 1,940 zinnen met 2,251 gerapporteerde resultaten en 3,043 paren van uitkomsten, waarbij we de semantische gelijkenis hebben beoordeeld. We maakten gebruik van een combinatie van informatie-extractie, gestructureerd data parsing, en methoden om semantische gelijkenis te beoordelen. De algoritmen voor semantische gelijkenis werden geëvalueerd op het oorspronkelijke corpus en op een corpus dat was uitgebreid met varianten voor het vermelden van een uitkomst in een RCT. De beste prestaties waren een F-maat van 88.42% voor de extractie van de primaire uitkomstmaat, 79.42% voor de extractie van de gerapporteerde uitkomst en 89.75% resp. 93.38% de evaluatie van semantische overeenkomsten in op de originele en in de uitgebreide versie van het corpus.

Statistische hypothesetoetsen worden vaak gebruikt in RCT's, om na te gaan of een experimentele behandeling daadwerkelijk superieur aan de controlebehandeling. Gebruikelijk is dat p-waarden worden gerapporteerd bij de uitkomsten. In Hoofdstuk 6 brengen we verslag uit van onze inspanningen om een algoritme te ontwikkelen voor het extraheren van de relatie tussen onderzoeksresultaten en p-waarden. We hebben hiervoor een corpus geannoteerd dat

bestaat uit 663 zinnen met 2,552 relaties tussen uitkomsten en p-waarden (1,372 positieve en 1,180 negatieve relaties). We beschrijven beknopt onze algoritmen voor de extractie en de evaluatie. We gingen er daarbij van uit dat de gerapporteerde uitkomsten en de p-waarden al eerder waren geëxtraheerd, zodat deze vervolgens als input voor een algoritme voor relatie-extractie kunnen dienen. Wij evalueerden twee benaderingen voor relatie-extractie: machine learning classifiers, en een benadering op basis van deep learning, met het tunen van getrainde taalmodellen (BERT, BioBERT en SciBERT). De deep learning aanpak bleek superieur aan de machine learning classifiers. Het aangepaste BioBERT-model presteerde het best, met een F-maat van 94%.

Hoofdstuk 7 schetst de tekstuele kenmerken van de soorten spin die we hebben behandeld en beschijft ook ons prototype voor spin-detectie, genaamd DeSpin (Detector of Spin). DeSpin is een systeem op basis van Natural Language Processing voor de semi-automatische detectie van spin in wetenschappelijke artikelen, en dan in het bijzonder verslagen van gerandomiseerd vergelijkend onderzoek (randomized controlled trials - RCT's). DeSpin combineert regels en machine learning om potentiële spin te detecteren. Onze algoritmen bereikten een F-maat variërend van 79.42 tot 97.86% in de verschillende taken. De moeilijkste taak bleek het identificeren van de gerapporteerde uitkomstem.

Hoofdstuk 8 beschrijft een corpusstudie van 1,694 primaire uitkomsten die uit verslagen van onderzoek zijn gehaald en 6,221 primaire uitkomsten uit trialregisters zijn geplukt. We hebben onderzocht hoe primaire resultaten worden beschreven in verslagen van RCT's en hoe dat gebeurt in trialregisters. We hebben de structuur van het uitkomstveld in registers beoordeeld, de lengte van uitkomstbeschrijvingen, de aanwezigheid van verwante items (zoals meetinstrumenten, tijdstippen, gebruikte maten), de consistentie in het invullen van velden voor uitkomst in trialregisters, de verschillende manieren om uitkomsten in het verslag te introduceren en de daarbij waargenomen onduidelijkheden. Er bleek een grote diversiteit te bestaan in de manier waarop de uitkomsten worden gedefinieerd. We zagen dubbelzinnigheid bij het gebruik van termen als 'uitkomsten', 'eindpunt', 'uitkomstmaat' en alle andere termen die kunnen worden gebruikt om een uitkomst in de tekst te introduceren. We zagen ook dat de structuur van trialregisters verschilt, omdat er afzonderlijke veldenbestaan voor meetpunten en meetinstrumenten, en voor elk van de uitkomsten in een trial. We zagen inconsistenties bij het invoeren van informatie over de meetpunten in de registers: de structuur die er is wordt vaak genegeerd, tijdspunten worden twee keer genoemd, of velden worden verward. Dit soort inconsistenties vormen een behoorlijke uitdaging, zowel voor handmatige als voor een geautomatiseerde analyse van verslagen van klinisch onderzoek. Al bij al is er een duidelijke behoefte om de terminologie voor het rapporteren van uitkomsten te standaardiseren.

Na ons onderzoek kunnen we een aantal aanwijzingen geven voor vervolgonderzoek naar de

automatische detectie van "spin". Om te beginnen moet ons prototype DeSpin verder worden getest door eindgebruikers, om de bruikbaarheid en potentiële voordelen bij de detectie van spin te kunnen evalueren. Ons prototype kan op verschillende manieren worden gebruikt: het kan dienen als hulpmiddel voor de redactie van tijdschriften en voor peer reviewers, als deze een ingediend artikel op spin moeten beoordelen. Het kan ook dienen als een hulpmiddel voor auteurs, om hun manuscript te controleren en eventueel te verbeteren, voordat het ter publicatie bij een tijdschrift wordt aangeboden. Het kan ook educatief worden gebruikt, voor het illustreren van verschillende soorten spin, om uit te leggen waarom het te vermijden vormen van rapportage zijn, en hoe deze ook daadwerkelijk kunnen worden vermeden. De algoritmen die we in ons prototype kan hebben verwerkt kunnen ook worden ingebouwd in andere systemen voor medische tekstanalyse, of bij het bouwen van een schrijfhulp.

We zien ook meer algemene aanknopingspunten voor verder onderzoek.

1. Nieuwe algoritmen voor detectie van spin

   Deze nieuwe algoritmen omvatten bijvoorbeeld het opsporen van een onterechte focus op subgroepen in trials of van een niet-gefundeerde extrapolatie van de resultaten naar een bredere populatie dan die was bestudeerd. Om dit te bereiken moet de bestudeerde populatie worden geïdentificeerd (uit de tekst van het artikel of, anders, uit het trialregister) en de semantische overeenkomst met de conclusie worden beoordeeld. Een andere algoritme dat van vitaal belang is voor spindetectie omvat het beoordelen van de richting van de resultaten en van de conclusies (positief - neutraal - negatief) over de bestudeerde behandeling. Een positief oordeel over een behandeling in de conclusies kan bijvoorbeeld niet te rijmen zijn met een effect dat niet statistisch significant is.

2. Het detecteren van de soorten spin in een bredere setting.

   We hebben ons in het hier gerapporteerde onderzoek vooral gericht op de detectie van een wijziging in de primaire uitkomst in een RCT. Ook met secundaire uitkomsten kan echter worden gesmokkeld. Detectie van deze vorm van spin zou een extra algoritme vereisen, voor de extractie van secundaire uitkomsten. Ons algoritme voor de beoordeling van semantische verwantschap kan dan vervolgens worden toegepast om mismatches tussen vooraf gedefinieerde en feitelijke gerapporteerde secundaire uitkomsten vast te stellen.

3. Identificatie van spin in andere verslagen van klinisch-wetenschappelijk onderzoek.

   Spin komt niet alleen bij RCT voor. Ook in verslagen van niet-gerandomiseerde studies, van systematisch literatuuronderzoek, of bij onderzoek naar diagnostische accuratesse kan worden gespind. Bepaalde vormen van spin komen bij veel vormen van onderzoek

voor, terwijl andere vormen meer specifiek zijn voor een bepaald type onderzoek. In dat geval zijn nieuwe algoritmen nodig voor de signalering.

4. Identificatie van spin in andere teksten over medisch onderzoek, zoals nieuwsitems en persberichten.

   Ook deze kunnen spin bevatten, vaak afkomstig van spin in het onderzoeksverslag zelf.

5. Onderzoek naar spin in andere onderzoeksdomeinen.

   Tot nu toe werd het onderzoek naar Spin voornamelijk uitgevoerd in het medische domein, waar het een directe bedreiging voor de gezondheidszorg kan vormen. Maar ook in andere vormen van wetenschappelijk onderzoek, buiten het medische, kan spin voorkomen.

Spin is een complex fenomeen. Zelfs mensen kunnen het niet altijd makkelijk identificeren. Toch zijn we ervan overtuigd dat ons onderzoek rond de detectie van spin heeft aangetoond dat het mogelijk is een geautomatiseerd systeem te bouwen dat goed genoeg is om auteurs en lezers van wetenschappelijke artikelen te ondersteunen bij het herkennen van spin. We hopen dat toekomstige onderzoekers de draad oppakken, na deze eerste proeve, en nieuwe systemen gaan bouwen, met nog betere resultaten en nieuwe instrumenten.

Résumé

Dans cette thèse, nous présentons notre travail sur le développement d'algorithmes de traitement automatique des langues (TAL) pour aider les lecteurs et les auteurs d'articles scientifiques (biomédicaux) à détecter le spin (présentation inadéquate des résultats de recherche). Notre algorithme se concentre sur le spin dans les résumés d'articles rapportant des essais contrôlés randomisés.

Nous avons étudié le phénomène de " spin " du point de vue linguistique pour créer une description de ses caractéristiques textuelles. Nous avons annoté des corpus pour les tâches principales de notre chaîne de traitement pour la détection de spin: extraction des résultats —en anglais " outcomes " —déclarés (primaires) et rapportés, évaluation de la similarité sémantique des paires de résultats d'essais et extraction des relations entre les résultats rapportés et leurs niveaux de signification statistique. En outre, nous avons annoté deux corpus plus petits pour identifier les déclarations de similarité des traitements et les comparaisons intra-groupe.

Nous avons développé et testé un nombre d'algorithmes d'apprentissage automatique et d'algorithmes basés sur des règles pour les tâches principales de la détection de spin (extraction des résultats, évaluation de la similarité des résultats et extraction de la relation résultat-signification statistique). La meilleure performance a été obtenues par une approche d'apprentissage profond qui consist à adapter les représentations linguistiques pré-apprises spécifiques à un domaine (modèles de BioBERT et SciBERT) à nos tâches. Cette approche a été mise en œuvre dans notre système prototype de détection de spin, appelé DeSpin, dont le code source est librement accessible sur un serveur public.

Notre prototype inclut d'autres algorithmes importants, tels que l'analyse de structure de texte (identification du résumé d'un article, identification de sections dans le résumé), la détection de déclarations de similarité de traitements et de comparaisons intra-groupe, l'extraction de données de registres d'essais. L'identification des sections des résumés est effectuée avec une approche d'apprentissage profond utilisant le modèle BioBERT, tandis que les autres tâches sont effectuées à l'aide d'une approche basée sur des règles.

Notre système prototype a une interface simple d'annotation et de visualisation.

Le chapitre 1 a présenté nos premières pas dans le développement d'algorithmes de TAL pour la détection automatique du spin dans des articles biomédicaux. Nous avons proposé un schéma pour un algorithme d'extraction automatique d'affirmations importants dans les résumés d'articles biomédicaux et des informations d'appoint possibles. Nous avons abordé trois tâches liées à la détection du spin: la classification des articles en fonction du type d'essai clinique (pour détecter les rapports d'essais randomisés), la classification des phrases dans les résumés pour identifier les sections Résultats et Conclusions, et l'extraction d'entités (pour les résultats des essais et la population étudié). Nous avons présenté l'état de l'art et nos premières expériences pour ces tâches. Pour chaque tâche, nous avons suggéré quelques directions possibles pour les travaux futurs.

Pour évaluer nos algorithmes à base de règles décrits dans les articles ci-dessus et pour entrainer des algorithmes d'apprentissage automatique, un corpus annoté avec les informations pertinentes est nécessaire. Le chapitre 2 décrit nos efforts pour collecter un corpus d'articles biomédicaux et les annoter pour le spin et les informations d'appoint. Le papier a présenté un schéma d'annotation pour le spin et l'information liée, nos guidelines d'annotation et les difficultés que nous avons rencontré, telles que le niveau d'expertise requis des annotateurs, le choix d'un outil d'annotation et la complexité de la tâche.

Le chapitre 3 a décrit nos expériences d'utilisation de l'apprentissage profond pour extraire les résultats d'un essai - variables surveillées au cours d'essais cliniques. L'extraction des résultats déclarés (primaires) et rapportés est une tâche principale pour la détection de spin. Dans cet article, nous avons examiné l'état de l'art pour l'extraction des résultats. Nous avons présenté notre corpus annoté manuellement de 2 000 phrases avec résultats déclarés (primaires) et 1 940 phrases avec résultats rapportés, qui est disponible gratuitement. Nous avons comparé deux approches d'apprentissage profond: une approche d'adaptation (" fine-tuning ") simple et une approche utilisant des champs aléatoires conditionnels (CRF) et des réseaux récurrents bi-directionnels avec mémoire à long terme (Bi-LSTM), en conjonction avec des plongements lexicaux ("embeddings") de caractères et des plongements lexicaux dérivées de modèles linguistiques pré-appris. Nous avons utilisé et comparé plusieurs modèles de représentation linguistique pré-appris, notamment le modèle général BERT (Bidirectional Encoder Representations from Transformers) ansi que le modèles pour le domaine biologie-médecine BioBERT et le modèle pour le domaine scientifique général SciBERT. Nous avons comparé ces approches à l'approche de base implementée précédemment (décrite dans les chapitres précédents). Les meilleurs résultats obtenus ont été la F- mesure au niveau des tokens de 88.52% pour les résultats primaires (fine-tuned BioBERT) et de 79.42% pour les résultats rapportés (fine-tuned SciBERT).

Le chapitre 4 a décrit le développement d'un algorithme d'évaluation de la similarité sé-

mantique de paires de résultats d'essais, qui fait partie du pipeline de détection de spin. Notre objectif a été de construire un algorithme qui ne nécessit pas de ressources spécifiques à un domaine créés manuellement, telles que des ontologies et des thésaurus. Sur la base du corpus annoté pour les résultats primaires et rapportés (décrit dans le chapitre précédent), nous avons annoté des paires de résultats primaires et rapportés pour la similarité sémantique (sur une échelle binaire). Le corpus est disponible gratuitement. Nous avons créé un corpus étendu en ajoutant les variantes permettant de faire référence à un résultat (par exemple, l'utilisation d'un nom d'outil de mesure à la place de l'expression dénommant le résultat; l'utilisation d'abréviations).

Pour l'approche de base, nous avons utilisé un nombre de mesures de similarité sémantique, basées sur des charactères, des tokens et des lemmes, des distances entre des expressions dans le réseau sémantique WordNet et des représentations vectorielles des expressions. Nous avons entraîné et testé différents classificateurs d'apprentissage automatique en utilisant une combinaison de mesures de similarité en tant que traits. Enfin, nous avons utilisé une approche d'apprentissage profond consistant à adpater ("fine tuning") les représentations linguistiques profondes pré-apprises sur le corpus des paires de résultats. Nous avons testé plusieurs modèles de langue: BERT (entraîné sur des textes de domaine général), BioBERT et SciBERT (entraînés respectivement sur le domaines biomédical et le domaine scientifique général). L'approche d'apprentissage profond a été supérieure aux autres approches testées. Le meilleur résultat sur le corpus original a été montré par le modèle BioBERT, avec une F- mesure de 89.75%. Sur le corpus étendu, la performance des algorithmes d'apprentissage profond s'est améliorée par rapport à celle du corpus initial: BioBERT a obtenu une F-mesure de 93.38%.

Le chapitre 5 a décrit les expériences d'utilisation des algorithmes d'extraction de résultats et d'évaluation de la similarité sémantique (décrits dans les deux chapitres précédents) pour développer un algorithme de détection de la substitution de résultat —en anglais "outcome switching" —un type de spin consistant en un changement injustifié (en omettant ou en ajoutant) des résultats définis d'un essai. Nous nous sommes concentrés sur la substitution de résultat primaire. Nous avons annoté un corpus avec les informations nécessaires à la détection de substitution de résultat : 2 000 phrases avec 1 694 résultats primaires; 1 940 phrases avec 2 251 résultats rapportés; et 3 043 paires de résultats annotés pour la similarité sémantique. Nous avons utilisé une combinaison d'extraction d'informations, d'analyse de données structurées et de méthodes d'évaluation de la similarité sémantique pour identifier la substitution du résultat primaire. Les algorithmes d'évaluation de la similarité sémantique ont été évalués sur le corpus d'origine et sur un corpus étendu avec des variantes de référence à un résultat. La meilleure performance obtenue était la F-mesure de 88.42% pour l'extraction des résultats

primaires, de 79.42 % pour l'extraction des résultats rapportée, et de 89.75% et 93.38% pour les versions originale et étendue du corpus d'évaluation de similarité sémantique.

Le test d'hypothèse statistique est souvent utilisé dans les essais randomisés pour déterminer si l'intervention expérimentale est supérieure à celle du groupe de controle. Des niveaux de signification statistique sont souvent rapportés pour les résultats de l'essai. Le chapitre 6 a décrit le développement d'un algorithme pour extraire la relation entre les résultats des essais et les niveaux de signification statistique. Nous avons annoté un corpus de 663 phrases avec 2 552 relations entre les résultats et les niveaux de signification (1 372 relations positives et 1 180 relations négatives). Nous avons brièvement décrit nos algorithmes d'extraction d'entités (résultats rapportés et niveaux de signification) et fourni leur évaluation. Dans nos expériences d'extraction de relations, nous avons supposé que les entités (résultats rapportés et niveaux de signification) avaient été extraites à une étape précédente et pouvaient être passer comme input à l'algorithme d'extraction de relations. Nous avons évalué deux approches pour l'extraction de relations: les classificateurs d'apprentissage automatique, en utilisant un ensemble de traits créé manuellement, et une approche d'apprentissage profond consistant à adapter ("fine-tuning") des modèles de langue pré-appris (BERT, BioBERT et SciBERT). L'approche d'apprentissage profond a été supérieure aux classificateurs utilisant des traits créé manuellement. Le modèle de BioBERT a montré les meilleures performances pour l'extraction de relations (F-mesure de 94%).

Le chapitre 7 a décrit les caractéristiques textuelles des types de spin que nous avons abordés et a présenté notre système prototype de détection de spin, appelé DeSpin (Detector of Spin). DeSpin est un système de traitement automatiique de langue (TAL) pour la détection semi-automatique du spin dans les articles scientifiques, en particulier dans les rapports d'essais contrôlés randomisés. DeSpin combine des approches d'apprentissage automatique et approches basées sur des règles pour détecter le spin potentiel et les informations associées. Nos algorithmes ont atteint des performances opérationnelles avec la F-mesure de 79.42 à 97.86% pour tâches différentes. La tâche la plus difficile consiste à extraire les résultats rapportés. L'outil proposé peut être utilisé à la fois par les auteurs et les relecteurs pour détecter le spin potentiel. L'outil et les corpus annotés sont disponibles gratuitement.

Le chapitre 8 a présenté une étude de corpus comprenant 1 694 résultats primaires extraits d'articles de recherche et 6 221 résultats primaires extraites des registres d'essais. Nous avons étudié comment les résultats primaires sont décrits dans les articles sur les essais randomisés et dans les registres d'essais. Nous avons évalué la structure du champ résultat dans les registres, la longueur des descriptions de résultat, la présence d'éléments liés (outils de mesure, points de temps, métriques d'analyse utilisées, etc.), la cohérence du remplissage des champs dans les registres décrivant un résultat, les manières d'introduire des résultats dans le texte et les

ambiguïtés observées. Nos travaux ont montré qu'il y a une grande diversité dans la définition des résultats. Nous avons observé une ambiguïté dans l'utilisation des termes "outcomes", "end point", "measure" et d'autres utilisés pour introduire un résultat dans le texte. La structure des registres d'essais diffère en ce qui concerne la présence de champs distincts pour les points de temps et les outils de mesure et pour chacun des résultats d'un essai. Nous avons observé des incohérences dans l'introduction des informations sur les points de temps dans les registres: la forme structurée pour décrire un résultat fourni par les registres est souvent ignorée, les points de temps étant présentés à la fois dans un champ de point de temps séparé et dans le champ résultat, ou dans le champ résultat au lieu du champ de point de temps. Les incohérences observées constituent un défi pour l'analyse manuelle et automatique des rapports d'essais cliniques. Il est nécessaire de normaliser la terminologie utilisée pour rapporter des résultats.

Il y a un nombre de directions importantes pour les recherches futures sur la détection automatique du spin.

Tout d'abord, notre système prototype DeSpin doit être testé auprès des utilisateurs pour estimer son utilité et les gains potentiels pour la tâche de détection du spin effectuée par des humains. Le prototype peut être utilisé de plusieurs manières: il peut aider les rédacteurs et les relecteurs à évaluer un article soumis pour la présence du spin; il peut servir comme un outil d'aide à la rédaction pour aider les auteurs à vérifier et à améliorer leurs manuscrits avant de les soumettre; et il peut être utilisé comme un outil pédagogique, définissant différents types de spin, expliquant pourquoi ils représentent des pratiques inappropriées et comment les éviter. Les algorithmes mis en œuvre dans notre prototype peuvent être intégrés à d'autres systèmes d'analyse de texte médical ou d'aide à l'écriture.

Les autres directions pour les travaux futurs sont les suivantes.

1. Implémentation de nouveaux algorithmes de détection des types de spin non adressés par nos travaux.

   Ceux-ci incluent par exemple détection de la focalisation sur des sous-groupes de patients ou, à l'inverse, extrapolation inadéquate pour une population plus large que celle étudiée. La détection du spin de ce type nécessiterait d'extraire la population étudiée (du texte de l'article ou du registre de l'essai correspondant) et preséntée dans le résumé, ainsi que d'évaluer leur similarité sémantique. Un autre algorithme essentiel pour la détection du spin consiste à évaluer la polarité des résultats et des conclusions (positif - neutre - négatif) dans un article donné en ce qui concerne le traitement étudié. Les déclarations positives sur le traitement dans les résultats et les conclusions peuvent représenter du spin si le résultat principal n'est pas significatif.

2. Détecter les types de spin que nous avons adressés dans un contexte plus large.

En particulier, nous nous sommes concentrés sur la détection de substitution de résultat primaire, mais pas sur la substitution des résultats secondaires. La détection de ce type de spin nécessiterait un algorithme supplémentaire pour l'extraction des résultats secondaires. Notre algorithme d'évaluation de la similarité sémantique peut donc être appliqué pour comparer les résultats secondaires et les résultats rapportés afin de détecter les discordances.

3. Identifier le spin dans d'autres types d'articles de recherche médicale en dehors des essais randomisés.

   Le spin peut apparaître dans des essais non randomisés, des revues systématiques et des méta-analyses, des études de précision du diagnostic. Certains types de spin peuvent être communs à différents types de recherche, alors que d'autres types sont spécifiques à un type d'étude particulier et nécessitent donc des algorithmes distincts pour leur détection.

4. Identifier le spin dans d'autres types de textes liés à la recherche médicale, tels que des articles de presse et des communiqués de presse, qui peuvent contenir du spin, souvent lié à spin dans l'article de recherche correspondant

5. Explorer et définir le phénomène de spin dans d'autres domaines de recherche; par exemple TAL ou informatique. Jusqu'à présent, spin dans les recherches ont été principalement abordées dans le domaine médical, où il peut constituer une menace directe pour la santé publique. Néanmoins, la recherche dans d'autres domaines peut aussi avoir du spin, car les raisons de "spinning" les résultats de la recherche sont communes à différents domaines. Cependant, les domaines de recherche autres que la médecine manquent souvent de directives précises sur la présentations des résultats de la recherche, ce qui laisse plus de liberté quant à ce qu'il faut présented et comment. Ce fait peut complexifier la définition du spin.

Le spin est un phénomène complexe, difficile à identifier même pour les humains. Néanmoins, nous croyons que notre projet a montré que la détection de spin et d'informations d'appoint pouvait être automatisée à l'aide de techniques du TAL avec de le performance suffisante pour aider les auteurs et les lecteurs d'articles scientifiques. Comme notre projet a été le premier à adresser la détection automatique du spin, nous espérons que les recherches futures aborderont ce sujet, apporteront de meilleurs résultats et fourniront de nouveaux outils.

PhD Portfolio

Name PhD student: Anna Koroleva PhD period: Nov 2016 – Feb 2020 Name PhD supervisor: Patrick Paroubek, Patrick Bossuyt		
PhD training		
	Year	**Work-load (Hours)**
General courses		
Reproducibility - experiences from natural language processing and bioinformatics	2016	0,5
Protocol writing	2016	1
Data Management Plan	2016	1
General introduction to communication	2016	1,75
Social media	2016	1.75
Tools for reproducibility – Github	2016	0,5
Tools for reproducibility – Markdown	2016	0,5
Training through action and collaboration	2016	1
Letters to editors	2016	1
Scientific journals in the 21st century. The editor-in-chief's perspective	2017	3,5
Science in transition	2017	3,5
Informative bibliometrics	2017	3
From Government to Bench: How a funding agency spends government's science budget	2017	3
Integrity in Science	2017	3,5
Patient involvement in research	2017	1,5
Communication of projects	2017	0,5
Writing a plain language summary for one of our recently completed trials	2017	0,5
Evidence-based clinical guideline development	2017	1
Transparency in research	2017	0,5
Exercise following on from publication ethics webinar	2017	0,75
Statistical Analysis Plans	2017	0,5
EQUATOR	2017	0,5
Data sharing	2017	0,5
Implementation science	2017	2
Ethics & STICs	2018	12
Writing grant proposals	2018	4
Facilitating Reflection on Responsible Research and Innovation	2018	3,5

Parcours de l'après-thèse / what about post-thesis?	2019	5
Peer-review - answering to reviewers and editors	2019	2
Devising your career plan: an alliance between your mind, your heart and your guts	2019	1
Preparing job applications outside academia: optimizing your written and oral communication	2019	4
Perfecting your elevator pitch	2019	3,5
The relationships of young scientists with newspapers	2019	2,5
GRADE framework from Evidence to Decision	2019	2,5
Specific courses		
Introduction to research on research - Waste in research	2016	1
A review of basic statistical concepts: variability, uncertainty, confidence intervals	2016	1,5
A review of basic statistical concepts: p values, replicability	2016	1,5
Using causal diagrams to understand problems of confounding and selection bias	2016	1
Effect measures, Effect modification and non-collapsibility. Adjustment for confounding	2016	1
Identify causal effect parameters which our research is targeting / What assumptions are reasonable, how might we approach it?	2016	2
13th EUROLAN School on Natural Language Processing	2017	31
Introduction to Python	2019	3,25
Advanced methods in Research on Research: Use of specific experimental study design in Research on Research	2017	1,5
Value of Qualitative Research; Introduction to Qualitative Research; Reflexivity; Planning and Designing a Qualitative Study	2017	1,8
Qualitative Research Methods, Collecting Data; Writing an Interview Guide	2017	1,33
Conducting an Interview; Qualitative Analysis; Analysing Transcribed interview Data	2017	1,8
Quality in Qualitative Research	2017	1,5
Introduction to core outcome sets	2017	0,5
From quantitative to qualitative: ORBIT - a case study	2017	1,5
Critique a COS paper	2017	0,75
Fundamentals of Natural Language Processing	2018	2
Natural Language Processing - Medical terminology	2018	2
Biomedical text classification (assigning labels to texts)	2018	2
Information extraction (detecting biomedical entities in text)	2018	2
Content analysis	2018	2
Seminars, workshops and master classes		
Webinar on research integrity	2017	1
Webinar on meta-analyses and meta-epidemiology	2017	1
Open access and data management	2017	1
Séminaire (post)-doctorant at LIMSI	2017	
Webinar on Entrepreneurship	2018	1
Webinar 'The current research climate: changing culture and the incentive systems'	2018	1,5
P-value workshop	2019	1
Organisational change and identity	2019	1
Webinar on mediation analysis	2019	1

Presentations	
Vers la détection automatique des affirmations inappropriées dans les articles scientifiques, 18e REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL)	2017
Automatic detection of inadequate claims in biomedical articles: first steps. Workshop on Curative Power of MEdical Data	2017
On the contribution of specific entity detection and comparative construction to automatic spin detection in biomedical scientific publications. The Second Workshop on Processing Emotions, Decisions and Opinions (EDO 2017), The 8th Language and Technology Conference (LTC)	2017
Annotating Spin in Biomedical Scientific Publications: the case of Randomized Controlled Trials (RCTs). LREC	2018
Scientific rigour versus power to convince: an NLP approach to detecting distorted conclusions in biomedical literature	2018
Demonstrating ConstruKT, a text annotation toolkit for generalized linguistic constructions applied to communication spin. LTC 2019: Demo Session	2019
Extracting relations between outcome and significance level in Randomized Controlled Trials (RCTs) publications. ACL BioNLP workshop	2019
Analysing clinical trial outcomes in trial registries. TTM-TOTh	2019
A machine learning algorithm and tools for automatic detection of spin (distorted presentation of results) in articles reporting randomized controlled trials. ICTMC 2019	2019
(Inter)national conferences	
18e REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017), Orléans, France, 26 - 30 June 2017	2017
Workshop on Curative Power of MEdical Data, Constanta, Romania, 12 - 13 September 2017	2017
The Second Workshop on Processing Emotions, Decisions and Opinions (EDO 2017), The 8th Language and Technology Conference (LTC), Poznan, Poland, 2017	2017
LREC 2018, Miazaki, Japan, May 2018	2018
LTC 2019: Demo Session, Poznan, Poland, 2019	2019
ACL BioNLP workshop, Florence, Italy, 2019	2019
TTM-TOTh, Bourget-le-Lac, France, 2019	2019
ICTMC 2019, Brighton, UK, 2019	2019

Other		
Secondment at the Cochrane Schizophrenia Group, University of Nottingham, Nottingham, UK	2017	3 months
Secondment at the AMC, University of Amsterdam, Amsterdam, Netherlands	2018	2 months
Secondment at the UK EQUATOR Network, University of Oxford, Oxford, UK	2018-2019	2 months
Secondment at the AMC, University of Amsterdam, Amsterdam, Netherlands	2019	3 months
MiRoR Journal Club	2019	1

Publications

Peer reviewed

Anna Koroleva. Vers la détection automatique des affirmations inappropriées dans les articles scientifiques ("Towards automatic detection of inadequate claims in scientific articles"), 18e REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017), 2017, pp. 135 – 148

Anna Koroleva, Patrick Paroubek. Automatic detection of inadequate claims in biomedical articles: first steps. Proceedings of Workshop on Curative Power of MEdical Data, Constanta, Romania, September 12-13, 2017

Anna Koroleva, Patrick Paroubek. On the contribution of specific entity detection and comparative construction to automatic spin detection in biomedical scientific publications. Proceedings of The Second Workshop on Processing Emotions, Decisions and Opinions (EDO 2017), The 8th Language and Technology Conference (LTC), 2017

Anna Koroleva, Patrick Paroubek. Annotating Spin in Biomedical Scientific Publications: the case of Random Controlled Trials (RCTs). LREC 2018

Anna Koroleva, Patrick Paroubek. Demonstrating ConstruKT, a text annotation toolkit for generalized linguistic constructions applied to communication spin. LTC 2019: Demo Session

Anna Koroleva, Patrick Paroubek. Extracting relations between outcome and significance level in Randomized Controlled Trials (RCTs) publications. Proceedings of ACL BioNLP workshop, 2019

Anna Koroleva, Sanjay Kamath, Patrick Paroubek. Extracting primary and reported outcomes from articles reporting randomized controlled trials using pre-trained deep language representations. Under revision

Anna Koroleva, Patrick Paroubek. Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations. Journal of Biomedical Informatics – X, 2019

Anna Koroleva, Patrick Paroubek. Towards Automatic Detection of Primary Outcome Switching in Articles Reporting Clinical Trials. Under revision

Anna Koroleva, Sanjay Kamath, Patrick MM Bossuyt, Patrick Paroubek. DeSpin: a prototype system for detecting spin in biomedical publications. Under revision

Anna Koroleva, Elizabeth Wager, Patrick Bossuyt. Can computers be taught to peer review and detect spin? Issues and challenges of a novel application of Natural Language Processing: a case study. Under revision

Anna Koroleva, Elizabeth Wager, Patrick Bossuyt. What is an outcome? A corpus study. Under revision

Anna Koroleva, Camila Olarte Parra, Patrick Paroubek, On improving the implementation of automatic updating of systematic reviews, JAMIA Open, 2019, ooz044

Other

Anna Koroleva, Corentin Masson, Patrick Paroubek. Analysing clinical trial outcomes in trial registries. TTM-TOTh 2019

Anna Koroleva, Patrick Paroubek. A machine learning algorithm and tools for automatic detection of spin (distorted presentation of results) in articles reporting randomized controlled trials. ICTMC 2019, Brighton, UK

Anna Koroleva, Patrick Paroubek. Automating the detection of communication spin in scientific articles reporting Randomized Controlled Trials. In preparation

Contributing authors

- Patrick Paroubek (PP), supervisor
  LIMSI, CNRS, Université Paris-Saclay, Orsay, France
- Patrick MM Bossuyt (PMMB), co-supervisor
  Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands
- Elizabeth Wager (EW), mentor
  Sideview, Princes Risborough, UK
  School of Medicine, University of Split, Split, Croatia
- Sanjay Kamath (SK), PhD student
  LIMSI, CNRS, Université Paris-Saclay, Orsay, France
  LRI Univ. Paris-Sud, CNRS, Université Paris-Saclay, Orsay, France

Acknowledgements

First of all, I would like to thank my supervisors Patrick Paroubek and Patrick Bossuyt for introducing me to the world of scientific research and for leading me through these three years of work on my difficult but amazingly interesting topic. Their expertise in their respective research fields was invaluable for me. Their guidance and constructive criticism made this thesis possible.

I am very grateful to my mentor Liz Wager for her openness and support, for the warm welcome I had during our meetings at her house, for suggesting great ideas for the research, and for our talks on non-research topics, the most important of which are cats, music and poetry.

I feel indebted to Isabelle Boutron and the MiRoR project for providing the opportunity to conduct this research project. I am grateful to Laura De Nale for fulfilling the challenging task of being the MiRoR project coordinator, providing help and support to us the students.

I would like to thank Sanjay Kamath, a (former) PhD student at LIMSI, for his patience and sense of humour when we were working together on the experiments with BERT.

My secondment at the Cochrane Schizophrenia Group was extremely valuable for my research, and an immense pleasure. I am grateful to Clive Adams for hosting my stay and for always being curious and enthusiastic about my topic (even when I had doubts myself). I thank Jackie Patrick for her help with every question I had, from getting the campus card to finding cycling routes from my home to the office. I also thank all the Cochrane Schizophrenia Group members (Tracey, Farhad and all) and interns (Lena, Johannes) for providing a great, friendly and inspiring environment for my stay.

I would like to acknowledge the UK EQUATOR Centre and its head Gary Collins for hosting my other secondment, with Jen de Beyer providing the invaluable administrative support. I am thankful to Caroline Struthers and Michael Schlussel for letting me get involved into their research projects, it was a great chance to familiarize myself with topics that were new for me and to get a broader view of the domain of research reporting. Last but not least, it was a wonderful Christmas, with quizzes, dinners, and other extracurricular activities that made this time such a good fun!

I am thankful for AMC-UvA for hosting two my secondments. It was a great pleasure working side by side with the PhD students and researchers that I met there (Maria, Mona, Bada, Yasaman and others).

I would like to thank my ex-colleagues at Megaputer Intelligence Moscow for letting me start my career in Natural Language Processing, obtain my first experience with medical text mining, and inspiring me to continue my professional development and enroll in a PhD programme.

Most importantly, I want to thank my family and friends who supported my decision to do a PhD abroad and who were by my side all this time (even despite being in different countries). I thank my parents and my grandfather for their help and understanding. I also thank my brother and his wife Ksenia who showed by their example that moving abroad is not as scary as it seems and that all difficulties can be overcome.

I thank Ira for always believing in me, for interest in my work, good sense of humour and emotional support at any time, wherever I am.

I thank James, who has been with me through good and bad times and without whom I would hardly be able to finish this PhD.

Finally, I dedicate this thesis to my grandmother who passed away two years ago. Without her, I would not have been who I am and where I am.

About the author

Anna Koroleva obtained her Master's degree in Theoretical and Applied Linguistics from the Moscow State University in 2013. During her studies, she developed an interest for computational linguistics and took part in two research projects devoted to evaluation of morphological and syntactical parsers for Russian language.

After graduating, she worked for 2.5 years as an analytical linguist at Megaputer Intelligence in Moscow, being a part of the team developing PolyAnalyst data and text mining software. During this time, she gained the first experience of developing algorithms for automatic analysis of medical text, addressing tasks such as classification, entity extraction and sentiment analysis.

In 2016, she started a PhD project at LIMSI-CNRS (France) and the AMC-UvA (Netherlands) under supervision of dr. Patrick Paroubek and prod. dr. Patrick Bossuyt. Her research project addressed the task of automatically identifying spin (distorted presentation of research results) in articles reporting randomized controlled trials. During this period, she had a few secondments (at the Academic Medical Center, University of Amsterdam; at the Cochrane Schizophrenia Group, University of Nottingham; and at the UK EQUATOR Centre, University of Oxford). She took part in the GoodReports pilot study and the EQUATOR library audit.

Anna plans to continue her research in the biomedical Natural Language Processing at the Zurich University of Applied Sciences (ZHAW) in the coming years.

université PARIS-SACLAY

ECOLE DOCTORALE
Sciences et Technologies de l'Information et de la Communication (STIC)

Titre : Rédaction assistée pour éviter les affirmations inappropriées dans la rédaction scientifique

Mots clés : Traitement Automatique du Langage Naturel, Informatique Biomédicale, Spin

Résumé : Dans cette thèse, nous présentons notre travail sur le développement d'algorithmes de traitement automatique des langues (TAL) pour aider les lecteurs et les auteurs d'articles scientifiques (biomédicaux) à détecter le spin (présentation inadéquate des résultats de recherche). Notre algorithme se concentre sur le spin dans les résumés d'articles rapportant des essais contrôlés randomisés. Nous avons étudié le phénomène de " spin " du point de vue linguistique pour créer une description de ses caractéristiques textuelles. Nous avons annoté des corpus pour les tâches principales de notre chaîne de traitement pour la détection de spin: extraction des résultats — en anglais " outcomes " — déclarés (primaires) et rapportés, évaluation de la similarité sémantique des paires de résultats d'essais et extraction des relations entre les résultats rapportés et leurs niveaux de signification statistique. En outre, nous avons annoté deux corpus plus petits pour identifier les déclarations de similarité des traitements et les comparaisons intra-groupe. Nous avons développé et testé un nombre d'algorithmes d'apprentissage automatique et d'algorithmes basés sur des règles pour les tâches principales de la détection de spin (extraction des résultats, évaluation de la similarité des résultats et extraction de la relation résultat-signification statistique). La meilleure performance a été obtenues par une approche d'apprentissage profond qui consist à adapter les représentations linguistiques pré-apprises spécifiques à un domaine (modèles de BioBERT et SciBERT) à nos tâches. Cette approche a été mise en oeuvre dans notre système prototype de détection de spin, appelé DeSpin, dont le code source est librement accessible sur un serveur public. Notre prototype inclut d'autres algorithmes importants, tels que l'analyse de structure de texte (identification du résumé d'un article, identification de sections dans le résumé), la détection de déclarations de similarité de traitements et de comparaisons intra-groupe, l'extraction de données de registres d'essais. L'identification des sections des résumés est effectuée avec une approche d'apprentissage profond utilisant le modèle BioBERT, tandis que les autres tâches sont effectuées à l'aide d'une approche basée sur des règles. Notre système prototype a une interface simple d'annotation et de visualisation.

Title : Assisted authoring for avoiding inadequate claims in scientific reporting

Keywords : Natural Language Processing, Biomedical Informatics, Spin

Abstract : In this thesis, we report on our work on developing Natural Language Processing (NLP) algorithms to aid readers and authors of scientific (biomedical) articles in detecting spin (distorted presentation of research results). Our algorithm focuses on spin in abstracts of articles reporting Randomized Controlled Trials (RCTs). We studied the phenomenon of spin from the linguistic point of view to create a description of its textual features. We annotated a set of corpora for the key tasks of our spin detection pipeline: extraction of declared (primary) and reported outcomes, assessment of semantic similarity of pairs of trial outcomes, and extraction of relations between reported outcomes and their statistical significance levels. Besides, we anno- tated two smaller corpora for identification of statements of similarity of treatments and of within-group comparisons. We developed and tested a number of rule-based and machine learning algorithms for the key tasks of spin detection (outcome extraction, outcome similarity assessment, and outcome-significance relation extraction). The best performance was shown by a deep learning approach that consists in fine-tuning deep pre-trained domain-specific language representations (BioBERT and SciBERT models) for our downstream tasks. This approach was implemented in our spin detection prototype system, called De- Spin, released as open source code. Our prototype includes some other important algorithms, such as text structure analysis (identification of the abstract of an article, identification of sections within the abstract), detection of statements of similarity of treatments and of within-group comparisons, extraction of data from trial registries. Identification of abstract sections is performed with a deep learning approach using the fine-tuned BioBERT model, while other tasks are performed using a rule-based approach. Our prototype system includes a simple annotation and visualization interface.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France