Automatic detection of inadequate claims in biomedical articles: first steps

Anna Koroleva[1] and Patrick Paroubek[1]

[1] LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
koroleva@limsi.fr, pap@limsi.fr

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.

1 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 [2; 3; 10; 27]. Two examples of conclusions with spin and the same conclusions rewritten by experts to remove spin are given in Table 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 [2]. 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 [9, 25].

Table 1. Examples of conclusions with spin and the same conclusions rewritten without spin

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.

Spin occurs in articles reporting various types of trials (non-randomized controlled trials, randomized controlled trials, diagnostic accuracy studies) [2; 3; 16; 27]. 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 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.

2 Types of spin

Spin in medical articles can be classified into the following types [3; 15; 26]:

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.

3 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?

4 Text classification according to study design

4.1 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 [8; 11; 19]. 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 [8].

Cohen et al. [5] 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

4.2 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

[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/

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 [9]. 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.

4.3 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.

5 Abstracts sentence classification

5.1 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 [12; 13; 19; 25]. Simple bag-of-words approach was explored and showed good performance [18]. Other features used to enhance the classification performance include: structural information (position of a sentence within an abstract) [19], semantic information (semantic categories of words and phrases, obtained through MetaMap [1]), sequential information (features of preceding/following sentences) [12; 13]. Classifiers used for this task include SVM and CRF. Classifiers are trained on manually annotated corpora [13; 19; 25] or use structured abstracts as gold standard [12; 19; 25].

5.2 Experiments

We seek to classify sentences in the abstracts into three categories: Results, Conclusions and Other. Following the approach adapted in [12; 18; 24], 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

5.3 Future work

Our current results are relatively good and comparable to some of the previously reported approaches [19, 25], but they are still lower than the best results obtained for this task, e.g. [12]. 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.

6 Information extraction: outcomes and population

6.1 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 [6; 14] 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 [22; 23]; descriptions and sizes of patient groups, outcomes examined, and numerical data for outcomes [22]; intervention arms [4]; patient population including general description, sample sizes, medical condition [21; 24].

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 [6; 14]. 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 [4; 6; 14; 21; 22; 24]; second, the sentences identified at the first stage are searched for entity mentions, which is done by means of rule based approaches [6; 14; 21; 24] or CRF-based automatic classifiers [4; 21]. 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 [4]. 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 [4; 22; 23]. Semantic information is reported to be more useful than information about word shape [22; 23].

6.2 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 [20], following the successful reports of previous experience along this approach in [7; 17].

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 [23]:

(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.

6.3 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 [15], 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 [22]. 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

1. Aronson A. Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. In: Proc. AMIA Symposium (2001).
2. Boutron I., Altman D.G., Hopewell S., Vera-Badillo F., Tannock I., Ravaud P. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of Cancer: the SPIIN randomized controlled trial. J Clin Oncol, 32, 4120–4126 (2014).
3. Boutron I., Dutton S., Ravaud P., Altman D.G. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. JAMA, 303, 2058–2064 (2010).
4. Chung G.Y. Towards identifying intervention arms in randomized controlled trials: extracting coordinating constructions. J Biomed Inform, 42(5), 790-800 (2009).
5. Cohen A.M., Smalheiser N.R., McDonagh M.S., Yu C., Adams C.E., Davis J.M., Yu P.S. Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine. J Am Med Inform Assoc., 22(3):707–17 (2015).
6. De Bruijn B., Carini S., Kiritchenko S., Martin J., Sim I. Automated information extraction of key trial design elements from clinical trial publications. In: Proceedings of the AMIA Annual Symposium, 141-145 (2008).
7. Friburger N., Maurel D. Finite-state transducer cascade to extract named entities in texts. Theoretical Computer Science, 313, 94-104 (2004).
8. Glanville J. M., Lefebvre C., Miles J. N., Camosso-Stefinociv J. How to identify randomized controlled trials in MEDLINE: 10 years on. Journal of the Medical Library Association, 94, 130–136 (2006).
9. Hall M., Frank E., Holmes G., Pfahringer B., Peter R., Witten I. H. The weka data mining software: An update. SIGKDD Explorations, 11(1) (2009).
10. Haneef R., Lazarus C., Ravaud P., Yavchitz A., Boutron I. Interpretation of results of studies evaluating an intervention highlighted in Google Health News: a cross-sectional study of news. PLoS ONE, 10(10) (2015).
11. Higgins J.P., Green S., eds. Cochrane handbook for systematic reviews of interventions. Wiley & Sons Ltd., West Sussex (2008).
12. Hirohata K., Okazaki N., Ananiadou S., Ishizuka M. Identifying sections in scientific abstracts using conditional random fields. In: Proceedings of the Third International Joint Conference on Natural Language Processing. Hyderabad, 381–388 (2008).
13. Kim S.N., Martinez D., Cavedon L., Yencken L.. Automatic classification of sentences to support evidence based medicine. BMC bioinformatics, 12(Suppl 2):S5 (2011).

14. Kiritchenko S., De Bruijn B., Carini S., Martin J., Sim I. ExaCT: automatic extraction of clinical trial characteristics from journal publications. BMC Med Inform Decis Mak., 10: 56-10.1186/1472-6947-10-56 (2010).
15. Kouylekov M., Magnini B. Tree Edit Distance for Textual Entailment. RANLP (2005).
16. Lazarus C., Haneef R., Ravaud P., Boutron I. Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. BMC Med Res Methodol., 15:85 (2015).
17. Maurel D., Friburger N., Antoine J.-Y., Eshkol-Taravella I., Nouvel D. Cascades autour de la reconnaissance des entités nommées. TAL 52-1 (2011).
18. McKibbon K.A., Wilczynski N.L., Haynes R.B. Retrieving randomized controlled trials from medline: a comparison of 38 published search filters. Health Information and Libraries Journal, 26(3), 187-202 (2009).
19. McKnight L., Srinivasan P. Categorization of sentence types in medical abstracts. In: AMIA Annu. Symp. Proc., 440–444 (2003).
20. Paumier S. (2016). Unitex 3.1 User Manual. http://unitexgramlab.org/releases/3.1/man/Unitex-GramLab-3.1-usermanual-en.pdf, last accessed 2017/07/12.
21. Raja K., Dasot N., Tech B., Goyal P., Jonnalagadda S.R. Towards evidence-based precision medicine: extracting population information from biomedical text using binary classifiers and syntactic patterns. In: AMIA Jt Summits Transl Sci Proc, 203-212 (2016).
22. Summerscales R.L., Argamon S., Bai S., Hupert J., Schwartz A. Automatic summarization of results from clinical trials. In: The 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 372–377 (2011).
23. Summerscales R.L., Argamon S., Hupert J., Schwartz A. Identifying treatments, groups, and outcomes in medical abstracts. In: The Sixth Midwest Computational Linguistics Colloquium (MCLC 2009) (2009).
24. Xu R., Garten Y., Supekar K.S., Das A.K., Altman R.B., Garber A.M. Extracting subject demographic information from abstracts of randomized clinical trial reports. In: Proceedings of the 12th World Congress on Health (Medical) Informatics, 550-554 (2007).
25. Yamamoto Y., Takagi T. A sentence classification system for multi biomedical literature summarization. In: Proceedings of the 21st International Conference on Data Engineering Workshops (2005).
26. Yavchitz A., Boutron I., Bafeta A., Marroun I., Charles P., Mantz J., et al. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. PLoS Med, 9:e1001308 (2012).
27. Yavchitz A., Ravaud P., Altman D.G., Moher D., Hrobjartsson A., Lasserson T., Boutron I. 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).