On the contribution of specific entity detection and comparative construction to automatic spin detection in biomedical scientific publications

Anna Koroleva*, **Patrick Paroubek***

*LIMSI, CNRS, Universit Paris-Saclay
Bt 508, Campus Universitaire F-91405 Orsay
{koroleva, pap}@limsi.fr

Abstract

In this article we address the problem of providing automatized aid for the detection of misrepresentation (spin) of research results in scientific publications from the biomedical domain. For identifying automatically inadequate claims in medical articles, i.e. claims that state the beneficial effect of the experimental treatment to be greater than it is actually proven by the research results, we propose a Natural Language Processing (NLP) approach. We first make a review of related works and an NLP analysis of the problem; then we present our first results obtained on the type of publications most likely amenable to automatic processing: articles which present results of Randomized Controlled Trials (RCTs), i.e. comparisons done by applying the experimental or standard treatment on different registered patient groups. Our results concern the identification of specific entities necessarily present in an RCT description (here outcomes and patient groups), obtained with basic methods (local grammars) on a corpus extracted from the PubMed open archive. Then we describe our findings on the support we could gain by identifying comparative constructions and their relationship to the identified entities as preliminary step for deploying sentiment analysis as one of the constituent functionalities of our automatic spin detection algorithm.

1. Introduction

Results interpretation in research publications is often affected by the presence of spin [1], i.e. beautifying the observed results. This being said without any consideration of the presence or absence of the author's intention to mislead readers, judging only from the text of an article. In the biomedical domain which interests us here, spin consists in claiming that the treatment under study had a positive effect greater than the trial showed. Table1. presents some examples of spin (Boutron et al., 2010) (Boutron et al., 2014) (Yavchitz et al., 2016).

Spin is more frequent in abstracts than in bodies of scientific acticles (Boutron et al., 2010). (Boutron et al., 2014) studied articles on oncology and showed that spin in abstracts influences the way doctors interpret the effects of the treatment studied, making them overestimate its efficacy. Spin can thus influence clinical decisions, causing a serious health problem. Often only the abstract of an article is freely available, so spin in abstracts can impact (bias) more strongly the dissemination of research results.

This considerations lead to the conclusion that it is time to start investigating how to help scientific authors, reviewers and editors to identify probable occurences of spin. We think that Natural Language Processing and Machine Learning can provide helpful solutions for spin identification, both when writing or reading an article. The elaboration of such a solution will require, as a first step, building a corpus with appropriate annotations to model spin.

In this article we present a preliminary feasability study for some basic NLP functionalities required by automatic spin detection: entity extraction for two main spin-related concepts (patient groups and outcomes) and comparative construction identification as an initial step to fine-grained sentiment analysis. The aim in this first phase is to gauge which types of phenomena are amenable to automatic processing with sufficient reliability and to assess their relative frequency in scientific publications. This will support corpus building and annotation in the soon-to-come second phase, before addressing the realization of spin detection algorithm in the final third phase.

In section 2 we present our current model of spin in the biomedical literature and address linguistic characteristics of spin with the corresponding kinds of language processing required. In section 3, we present previous works related to our task. In section 4 we present our entity extraction experiments and analysis of comparatives which precedes our conclusion.

2. A model of Spin

Previous studies proved that several kinds of medical trials are subject to spin (such as *randomized control trials (RCTs)* or *diagnostic accuracy studies*) and identified three main categories of spin whose frequency varies in function of the trial type (Boutron et al., 2010) (Boutron et al., 2014) (Yavchitz et al., 2016):

1. Inappropriate **presentation** of research results, which declines into:
   - Negative effects of a treatment are not presented.
   - Some of the results are not evaluated, e.g. the primary outcome is not presented while the focus is put on significant secondary outcomes.
   - The presentation of the type of trial and its characteristics is incomplete or incorrect.
   - The description of the population studied is fuzzy, the focus is put on particular subgroups for which the results are statistically significant.
   - Linguistic spin (excessive use of positive comparions or superlatives).

[1] The term find its origins in the term "spin doctors", communication agents of public personalities particularly deft at improving the image of their clients.

Abstract with inappropriate claims	Abstract rewritten without spin
Treatment A **may be useful** in controlling cancer-related fatigue **in patients who present with severe fatigue**.	Treatment A **was not more effective than placebo** in controlling cancer-related fatigue.
This study demonstrated **improved PFS and response** for the treatment A compared with comparator B alone, although this did not result in improved survival.	The treatment A was **not more effective** than comparator B on overall survival in patients with metastatic breast cancer previously treated with anthracycline and taxanes.

Table 1: Examples of abstracts containing inappropriate claims and their version without spin rewritten by domain experts so that conclusions correspond to the actual research results (provided by I. Boutron from (Boutron et al., 2014))

- The limitations of the trial are not presented
- Previous studies are partially cited (important articles are missing)

2. Inappropriate **interpretation** of research results, which may take the form of:
   - The studied treatment is claimed to have a positive effect or an effect equivalent to the standard treatment in spite of non-significant results.
   - The treatment is presented as safe while the results for safety are not significant.
   - The treatment is presented as having positive effects without any comparative trial performed.
   - Only the statistical significance is considered instead of the clinical pertinence.
3. Inappropriate **extrapolation**, which include:
   - Instead of the population, treatment or result evaluated, the author presents a different population, treatment or result.
   - The conclusions are inappropriate for clinical practice, for instance an advice to use the treatement not substanciated by sufficient evidences

We focus on spin in RCTs because they are the main source of information for Evidence-Based Medicine (EBM).
Spin is a complex phenomenon with heterogenous aspects addressing syntax, semantics and pragmatics knowledge as well as inference. We focus here on the types of spin identifiable using only the text of the article, without using extra sources of information such as research protocol. From the listed types of spin we can deduce the following NLP functionalities required for automatic spin identification:

1. classification of biomedical articles according to the type of trial (up to now we have addressed the distinction between RCTs vs. other types);
2. extracting the treatment evaluation (positive/neutral/negative);
3. analysis of document structure (title/abstract/body);
4. entity extraction: studied outcomes (primary and secondary), population (with patients groups), statistical significance of results, trial restrictions, negative effects, treatments compared;
5. extracting the relations between the entities (e.g. between the results and their statistical significance);
6. paraphrase identification for comparing entity mentions from the abstract with those from the body;
7. syntactic analysis: identifying spin specific constructions, for instance concessive propositions often associated with a focus change: "This study demonstrates improved PFS and response for the treatment A compared with comparator B, **although** this did not result in improved survival" (*focus on secondary results*).

3. Previous works

In this section we report on previous works addressing: (1) bias assessment (a task linked to spin detection), (2) entity and relation extraction and (3) comparatives.

3.1. Bias assessment

In biomedical domain, the task closest to spin detection is bias assessment. Bias is defined as a systematic error or deviation with respect to truth in results or conclusions which may lead to under- or overestimation of the effect of the treatment evaluated (Higgins and Green, 2008). Errors can concern study design, research implementation or analysis/presentation of results. The types of bias include: selection bias (generation of the random sequence, masking treatment assignment); performance bias (blind allocation of treatment to patient); detection bias (anonymizing results evaluation); attrition bias (uncomplete data about outcomes); reporting bias (selective presentation of outcomes). We have a special interest in reporting bias because it falls under the definition of spin.
According to (Higgins et al., 2011) the evaluation of bias is more often done by experts who use scales or checklists than with an NLP approach. Examples of the latter are found in (Marchall et al., 2015), the authors describe a corpus of systematic reviews archived by the Cochrane network[2]. The bias evaluation from the systematic reviews is used as gold standard annotation. Bias evaluation is subjective; the authors report that discrepancies are the largest for assessing reporting bias. An SVM based on words is used to classify articles according to their bias level.
The difference between bias assessment and spin detection is that the former matches the research protocol with the

[2]Cochrane is an independent international network of researchers, health professionals and patients whose aim is to improve decision making in health care (http://www.cochrane.org).

article, while for spin detection the abstract is compared to the body text.

3.2. Entities and relations extraction

In the biomedical domain the research on entity extraction deals mainly with gene, protein and medicine names extraction, less attention was given to the entities which interest us (Summerscales et al., 2011). The entities we address encompass specific Named Entities (e.g. medicine brands etc.) and all nominal phrases associated to particular semantic roles in the description of an RCT, such as outcomes, patients groups, statistical indicators (p-value), cf. (Kiritchenko et al., 2010), (Nguyen et al., 2013).
Entities representing characteristics of a trial or clinical situation have been addressed because they concern many NLP tasks: summarization (Summerscales et al., 2009), systematic review elaboration [3], decision making aids, question answering systems, databases creation and querying. Different approaches have been used depending on the task addressed. For systematic reviews, at least four basic elements of a clinical study (known as "PICO framework") need to be identified: (1) Population/Problem, (2) "Intervention" (treatment), (3) Comparator treatment and (4) Outcome. This type of analysis does not always rely on entity extraction per se, since it is enough to identify the sentences which contain the wanted information (Wallace et al., 2016). With a focus on automatic summarization and with the limitation of using a small corpus of 20 abstracts, (Dawes et al., 2007) looked at a larger set of elements: patient-population-problem, exposure-intervention, comparison, outcome, duration and results and their associated co-texts. (Bruijn et al., 2008) further extended the set of addressed elements to all the elements from the CONSORT statement (http://www.consort-statement.org) which contains among others: patients eligibility criteria, the treatment studied and the comparator treatment, the intervention parameters (dosing, frequency, etc.), financing information, publication metadata, etc. This system differs from the majority of other similar systems because it works on the entirety of an article, not only on its abstract. The approach has been further developed in (Kiritchenko et al., 2010) with the ExaCT system. (Summerscales et al., 2011) try to compute automatically summary statistics (reduction of absolute risk, number of patients to treat) from the articles on RCTs using Conditionnal Random Fields. They show that outcomes are among the most difficult elements to extract. (Chung, 2009) addresses the extraction of the "branches" of an intervention (i.e. the application of the studied treatment and of the comparator treatment respectively) from coordinated constructions (using a maximum entropy classifier, a parser and the UMLS[4] accessed through the MetaMap application[5]) present in the methods section of the abstracts and the objectives, results and conclusion sections where the information is often explicitely mentioned. Other research looked at information about patient population: (Xu et al., 2007) extract the desciption of the population, the number of patients examined and the description of symtoms/diseases from RCTs using a Hidden Markov approach and parsing.
From this state of the art, we conclude that most of the research is focused on RCTs and work mainly with abstracts. A two-step approach is commonly used: first classifying the sentences then extracting the entities from the selected sentences, using a combination of Machine Learning and symbolic rule based algorithms. Systems like MetaMap to match the article contents with the UMLS are often used. The definition of the exact limits of an entity remains a difficult task; among various entities, the outcome is the most difficult to identify.

3.3. Comparatives

In NLP, comparative constructions have been studied early and mainly for English, e.g. (Ryan, 1981), (Ballard, 1988), (Friedman, 1989) or (Olawsky, 1989). (Li et al., 2010) extracted comparable entities from a corpus of questions using a minimal bootstrap approach. (Hatzivassiloglou and Wiebe, 2000) focuses on gradable adjectives as subjectivity markers (defining a measure of gradability). (Ganapathibhotla and Liu, 2008) mines opinions expressed about entities from comparative sentences, trying to determine which entity in a comparison is preferred by its author. More recently (Yang and Ko, 2011) addressed comparatives in Korean and extracted comparatives and comparisons between entities from a corpus of questions. (Gupta et al., 2017) addressed extracting compared entities and compared features (the feature with respect to which the entities are compared) from biomedical texts.
Comparative constructions can be divided into three types: morpho-lexical (e.g. with adjectives in comparative or superlative form), syntactic (with patterns like "as ADJ1 as ADJ2") or semantic (e.g. verbs or nouns indicating a change of state like "improved" or "improvement" comparing implicitly the current state of affairs with what it was in the past).

4. Experiments

4.1. Entity extraction

In this first set of experiments we used an approach combining manual exploration of the corpus and finite state automata (Unitex environment (Paumier, 2016)) filtering in a bootstrapping approach alternatively relying on the phenomena targeted and on their cotext. The most important information for spin detection is the outcome. We provide here two examples of the outcome identification results performed with 9 Unitex graphs and using the following markup: PROL for the outcome marker and OUT for the mention of the OUTCOME:
The <PROL>primary outcome was< /PROL> <OUT type=PRIM>the remission of depressive symptoms at the 2-month follow up visit< /OUT>, defined as a HDRS score of 7 or less.
<PROL>Secondary outcome parameters are< /PROL>

[3] A systematic review is a type of scientific articles aimed at an exhaustive summary of the literature about a particular problem with statistical evaluation of the results.

[4] UMLS (Unified Medical Language System) is a compendium of several medical controlled vocabularies, https://www.nlm.nih.gov/research/um

[5] https://metamap.nlm.nih.gov/

<OUT type=SEC>overall mortality, severity of BPD, number of days on the ventilator, number of treatment failures, ventilation-induced lung injury and pulmonary hypertension< /OUT> according to clinical parameters.
From a subcorpus of 3,938 articles on RCTs from PubMed Central[6], we have identified 6,292 outcome occurences.

4.2. Comparatives

For spin detection in RCTs, comparative constructions are of high interest because the main goal of RCTs is comparing two or more treatments with respect to a number of outcomes, and thus the results are most often presented in the form of a comparative sentence.
Our first goal is to identify comparative sentences that state superiority of the experimental treatment over the control treatment, similarity between the two treatments, or some positive changes occurred under the experimental treatment. These sentences are considered as containing positive evaluation of the experimental treatment.
Our second goal is to extract the components of a comparison including a comparative word (such as "better"), compared entities and compared features (Ganapathibhotla and Liu, 2008). In RCT reports, compared entities belong most often to one of three types: compared treatments (example 1); patient groups that received the treatments (example 2); value of an outcome before and after a treatment (example 3). In first and second cases, the compared feature is typically an outcome, as "efficacy" in example 1 and "response rate" in example 2.

1. Treatment A was better that treatment B in terms of efficacy.

2. The group receiving treatment A showed better response rate than the group receiving treatment B.

3. PANSS score improved with treatment A.

The novelty of our work is that our aim is not only to extract the compared entities and features, but also to detect their type (treatment, patient group, outcome).
We performed our experiments on a corpus of 3934 abstracts of articles from Pubmed and 5005 abstracts of articles from Cochrane Schizophrenia group database.
We proceeded in two steps: 1) we collected concordances for a set of words (verbs and nouns with the semantics of change of a parameter such as "improve"/"improvement") and constructions (such as comparative adjective/adverb + "than"/"versus", etc.) that are likely to convey a comparative meaning; 2) we studied the concordances to identify typical ways of expressing different types of components of a comparison. Our analysis shows that each type of components is associated with a set of morphological, lexical, morphosyntactic and syntactic features of a phrase. Each feature may be associated with several types of components, but a whole set of features can determine the type with sufficient accuracy. For example, a group of preposition "in" may represent patient groups ("... improved **in aged subjects**") or outcomes ("improvement **in PANSS total score**"). Treatments may be subjects of the active transitive verbs (with an outcome as the direct object) or occur within groups of prepositions "by"/"with"/"on"/"over". Outcomes may be active or passive subjects, direct objects. Patient groups mentions have additional lexical feature: words "group", "population" or words denoting humans such as "patients" etc. Words "medication", "agent", etc. and suffixes "-one", "-ine" etc. are associated with treatments. Outcomes in general have fewer lexical and morphological features compared to patients and treatments.
We created a set of finite-state automata to extract the comparison components (noun and prepositional phrases) and detect their type using the above-mentioned features.
Our data show that verbs and nouns denoting a change occur more often that constructions with comparative adjectives/adverbs (21721 vs. 3660 occurrences), so the first set of experiments concerns these verbs and nouns. Extraction of components from constructions with adjectives/adverbs, as well as from statements of similarity, is our future work.
In the Table4.2. we provide a preliminary statistics for constructions with outcome as a component of a comparison (where Out is outcome, Int is intervention). More precise evaluation is still work-in-progress.

5. Conclusion

The current system provides a baseline approach against which methods based on word embeddings will be tested, once the first versions of the reference annotated corpus currently under development will be available.

6. Acknowledgement

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

7. References

Ballard, Bruce W., 1988. A general computational treatment of comparatives for natural language question answering. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*. Buffalo, New York, USA: Association for Computational Linguistics.

Boutron, I., D. G. Altman, S. Hopewell, F. Vera-Badillo, I. Tannock, and P. Ravaud, 2014. 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*, 32.

Boutron, I., S. Dutton, P. Ravaud, and D.G. Altman, 2010. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*, 303:20582064.

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

Chung, G. Y., 2009. Towards identifying intervention arms in randomized controlled trials: extracting coordinating constructions. *J Biomed Inform*, 42(5):790–800.

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

Type of components	Pattern	N of occurrences / percentage	Example
Outcome	noun.change + PP(in/on/of + OUT)	1991 / 29,9%	Reductions<Out> in plasma estrogen levels< /Out> and increases<Out> in bone-resorption< /Out> markers were comparable in both groups.
Intervention + Outcome	INT.subj + verb.change.active + OUT.obj	711 / 10,7%	<Int>Adjunctive treatment with galantamine< /Int> improves<Out> memory and attention< /Out> in patients with schizophrenia.
Outcome	OUT.subj + verb.change.active/passive	1336 / 20%	<Out>HRQL< /Out> improves after successful treatment.
Outcome	Verb.change.prtcp + OUT	2306 / 34,6%	Three sessions of education led to significantly increased<Out> insight< /Out>.
Outcome	OUT = NP(noun + gain/elevation)	315 / 4,7%	Clozapine treatment is associated with <Out>weight gain< /Out>.

Table 2: Comparative constructions with outcomes

Dawes, M., P. Pluye, L. Shea, R. Grad, A. Greenberg, and J.-N. Nie, 2007. The identification of clinically important elements within medical journal abstracts: Patient-population-problem, exposure-intervention, comparison, outcome, duration and results (pecodr). *Informatics in Primary Care*, 15(1):916.

Friedman, Carol, 1989. A general computational treatment of the comparative. In *27th Annual Meeting of the Association for Computational Linguistics*.

Ganapathibhotla, Murthy and Bing Liu, 2008. Mining opinions in comparative sentences. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Coling 2008 Organizing Committee.

Gupta, Samir, A. S. M. Ashique Mahmood, Karen E. Ross, Cathy H. Wu, and K. Vijay-Shanker, 2017. Identifying comparative structures in biomedical text. In *Proceedings of the BioNLP 2017 workshop*.

Hatzivassiloglou, Vasileios and Janyce M. Wiebe, 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Higgins, J.P., D.G. Altman, and P.C. Gotzsche, 2011. The cochrane collaboration's tool for assessing risk oof bias in randomised trials. *BMJ*, 343:d5928.

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

Kiritchenko, S., B. De Bruijn, S. Carini, J. Martin, and I. Sim, 2010. Exact: automatic extraction of clinical trial characteristics from journal publications. *BMC Med Inform Decis Mak.*, 10:56-10.1186/1472-6947-10-56.

Li, Shasha, Chin-Yew Lin, Young-In Song, and Zhoujun Li, 2010. Comparable entity mining from comparative questions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics.

Marchall, I.J., J. Kuiper, and B.C. Wallace, 2015. Robotreviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, ocv044.

Nguyen, N., M. Miwa, Y. Tsuruoka, and S. Tojo, 2013. Open information extraction from biomedical literature using predicate-argument structure patterns. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine*.

Olawsky, Duane E., 1989. The lexical semantics of comparative expressions in a multi-level semantic processor. In *27th Annual Meeting of the Association for Computational Linguistics*.

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

Ryan, Karen, 1981. Corepresentational grammar and parsing english comparatives. In *Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics*. Stanford, California, USA: Association for Computational Linguistics.

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

Summerscales, R.L., S. Argamon, S. Bai, J. Hupert, and A. Schwartz, 2011. Automatic summarization of results from clinical trials. In *The 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*.

Wallace, B. C., J. Kuiper, A. Sharma, M. Zhu, and I.J. Marchall, 2016. Extracting pico sentences from clinical trial reports using supervised distant supervision. *Journal of Machine Learning Research*, 17(132):125.

Xu, R., Y. Garten, K.S. Supekar, A.K. Das, R.B. Altman, and A.M. Garber, 2007. Extracting subject demographic information from abstracts of randomized clinical trial reports. Amsterdam: IOS Press.

Yang, Seon and Youngjoong Ko, 2011. Extracting comparative entities and predicates from texts using comparative type classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Yavchitz, A., P. Ravaud, D.G. Altman, D. Moher, A. Hrobjartsson, T. Lasserson, and I. Boutron, 2016. 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.