Automated Contradiction Detection in Biomedical Literature

Noha S. Tawfik[1,2(✉)] and Marco R. Spruit[2]

[1] Computer Engineering Department, College of Engineering, Arab Academy for Science, Technology, and Maritime Transport (AAST), Abukir, Alexandria 1029, Egypt
noha.abdelsalam@aast.edu

[2] Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands
{n.s.tawfik,m.r.spruit}@uu.nl

Abstract. Medical literature suffers from inconsistencies between reported findings that answer the same research question. This paper introduces an automated two-phase contradiction detection model that integrates semantic properties as input features to a Learning-to-Rank framework, to accurately identify key findings of a research article. It also relies on negation, antonyms and similarity measures to detect contradictions between findings. The proposed technique is implemented and tested on a publicly available contradiction corpus 259 manually annotated abstracts. The performance is compared based on recall, precision and F-measure. Experimental evaluations prove the utility of the model and its contribution to the contradiction classification and extraction task.

Keywords: Biomedical NLP · Answer selection
Contradiction detection · Information extraction · Text mining

1 Introduction

In the last decade, there was a substantial increase in the total number of medical research publications worldwide. Most of the literature publish results on the effectiveness of clinical interventions, and despite the similarity of the scientific experiment designs, not all outcomes are in agreement [12]. Whether published findings are consensual, complimentary, or contradictory facts, many of them get approved, updated or replaced accordingly [23]. Given the varying nature of published findings, it is difficult to fairly assess evidence-based knowledge within articles. More importantly, differences between research outcomes should be highlighted so that further studies do not build assumptions and/or conclusions on prior research that have since been disapproved, and are not valid anymore.

In extreme cases, some published evidence-based facts even get reversed. Prasad et al. [24], reviewed 363 articles published in one high impact factor

P. Perner (Ed.): MLDM 2018, LNAI 10934, pp. 138–148, 2018.
https://doi.org/10.1007/978-3-319-96136-1_12

journal investigating various established medical practices. While 138 (38%) confirmed the practices, 146 (40.2%) found them ineffective and 79 (27.3%) were inconclusive. For example, four studies contradicted the administration of the Aprotinin drug, widely used for treatment in post cardiac surgeries.

Such misinformation, or disinformation, create controversy that is important to researchers and practitioners interested in finding evidence-based answers to clinical queries; whether it is for the benefit of their patients or for the sake of conducting systematic reviews. It is also of great significance to both Comparative Effectiveness Research (CER) and the Precision Medicine (PM) communities. The comparative effectiveness research is interested in the analysis of medical interventions by comparing their benefits and drawbacks, to reach informed evidence-based decisions for a better clinical practice [29]. While Precision Medicine also aims at improving the health care system, PM is different than CRM as it takes into account the genetic, environmental and lifestyle differences between individuals [14]. Highlighting different outcomes to the same medical practice supports the PM claims that there is no "one-size-fits-all" treatment strategy.

However, with the high rate of growth in scientific publications, the task of finding answers, interpreting outcomes and validating them becomes tedious, exhausting and time consuming, even in a specific sub-domain. In result, several text mining tools and frameworks were built and employed to solve the information extraction problem, automatically or semi-automatically, for a variety of research applications. Biomedical text mining faces a number of challenges; the enormous number of existing publications, the unstructured nature of text, and most challengingly, the ambiguity of reporting biomedical or clinical results. Findings can be expressed in long, context-dependent sentences with the usage of a wide variety of terminology.

Contradiction Detection in text is still a relatively new area of research. As in other Natural Language Processing (NLP) sub-domains, it requires a multidisciplinary approach involving text mining, sentiment analysis, opinion mining, knowledge retrieval and information extraction. This paper focuses on the problem of extracting contradicting findings in biomedical texts. In this context, we propose an automated contradiction detection framework that adapts and extends existing NLP tools. The proposed model takes advantage of a recently published corpus, constructed for the same purpose, to validate its accuracy.

2 Related Work

Despite the fact that more research has been conducted on text entailment rather than contradiction detection, the development of two contradiction corporas encouraged more research into the domain [9,20,26]. The corporas were based on direct negation and paraphrasing of sentences from the PASCAL Recognizing Textual Entailment (RTE) dataset [11]. However, contradiction analysis remains a challenging task, mainly due to the different ways in which contradictions can appear (numeric mismatching, negations, contrastive sentences, etc.).

The importance of extracting contradictions has been exploited in other domains, most commonly in news and rumors text processing. Due to the generic type of negation found in normal text, it is difficult to adapt any of the former models to the biomedical domain directly. The language used to express biomedical facts is usually rich in clinical semantics and conceptual overlaps, and involves complex sentence structures. To the best of our knowledge, there has been minimal research conducted on the biomedical contradiction analysis.

In 2011, Sarafraz et al. [27] investigated both rule based and machine learning methods to identify negated molecular events through lexical, syntactic and semantic features, the model was evaluated on the BioNLP09 challenge corpus. Alamri et al. [2,3] explored the use of four features: negation, directionality sentiment and uni+bi-grams combined with an SVM classifier, to extract contrasted findings reported in cardiovascular research literature. More recently, Preum et al. [25] presented *Preclude*, a rule-based system that highlights conflicts in wellness advice, found in on-line health forums. The system constructs a polarity lexicon from verbs , in the training set, and their synonyms using WordNet, for labeling actions found in text as positive or negative. In another attempt to discover the ambiguities in the biomedical literature, de Silva et al. [28] proposed an ontology-based system to extract inconsistencies found in miRNA research articles in the PubMed repository. The system relies on OLLIE "Open Language Learning for Information Extraction" framework to extract all relevant triples (subject, object, and relationship) from abstracts. Triple entries are then compared against each other to find inconsistencies, based on an *oppositeness metric* suggested by the authors.

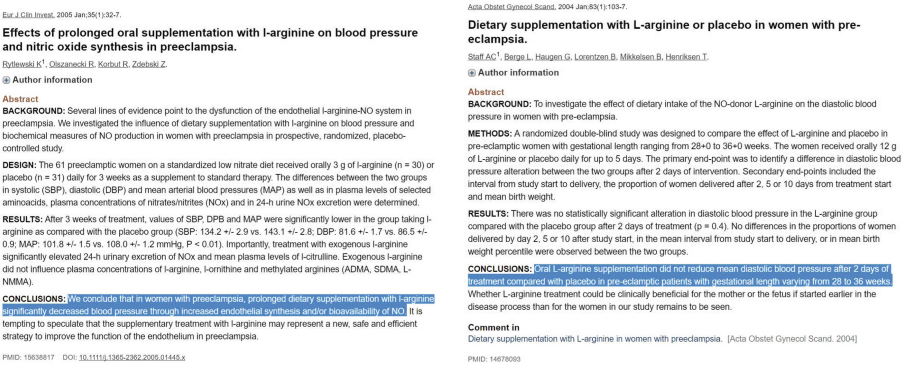
3 Dataset

The lack of annotated data has led to the unavailability of comparison and evaluation of contradiction detection systems in the biomedical literature. However, this may change with the recent availability of Manual Contradiction Corpus (*ManConCorpus*), a corpora of contradictory research claims[1]. The corpus is constructed out of 24 systematic reviews on four important cardiovascular disease topics: Cardiomyopathy, Coronary artery, Hypertensive and Heart failure. Each review article is mapped to a closed PICO (Population, Intervention, Comparison and Outcome) question that could be answered only by Yes or No. The mapping process was conducted manually by a medical expert, after reviewing all research abstracts of studies included in the systematic review. These abstracts include research claims with answers to the questions. *A research claim* is a one-sentence summary of the research findings that the authors find important, either to affirm old information or to introduce new ones. Two annotators were asked separately to find one correct claim per abstract and label it *YES*, if it positively answers the question and *NO* otherwise. It is worth mentioning that despite the fact that multiple sentences in the abstract might hold the answer

[1] Corpus available at http://staffwww.dcs.shef.ac.uk/people/M.Stevenson/resources/bio_contradictions/.

to the query, only the most informative one is chosen as per the annotator's opinion. The corpus has a total of 259 abstracts, out of which 180 introduce positive claims and 79 introduce negative claims. All claims included in the corpus are either evaluative or causal. The former is an assessment of the biomedical concept presented in the research topped by a judgment, while the latter is a statement that describes the relation type between two concepts and whether one affects the other or not. More details on the annotation process and the corpus statistics can be found in [4].

In literature, there is no standard definition of 'contradiction', and it is usually task-dependent according to the nature of the contradiction instances. Therefore, we adopt the authors' definition of contradiction that better matches the corpus and human intuitions: "*Two texts, T_1 and T_2, are said to contradict when, for a given fact F, information inferred about F from T_1 is unlikely to be true at the same time as information about F inferred from T_2*". As per the definition, if both a positive and a negative claim answer the same query, they are considered contradictory as shown in the example in Fig. 1.



(a) PMID: 15638817 with assertion value YES

(b) PMID: 14678093 with assertion value NO

Fig. 1. An example of two contradictory claims found in literature that answer the query: *In women with pre-eclampsia, does treatment with L-Arginine, compared to a placebo, reduce blood pressure?*

4 Methods

Identifying inconsistencies in text is a two-phase problem, claim retrieval and claim assertion. During the first phase, we need to identify potential sentences relevant to the query. In the claim assertion phase, we have to evaluate whether sentences infer text entailment or contradiction.

4.1 Identification of Abstract Claims

Finding relevant sentences that answer the query is a key component in the biomedical contradiction detection system, as the performance of the system

is dependent on the accuracy of the extracted key phrase. Several methods and techniques have been employed for passage retrieval in general, and answer identification in specific. Nevertheless, it still remains a challenge in the biomedical language processing field, mainly due to the complex nature of the text. In this research, we address the claim extraction process as a ranking problem, where each sentence in the input text is scored according to its relevance to the query.

Input Preprocessing. We split all abstract text included in the corpus into sentences using The Natural Language Toolkit (NLTK). All sentences with less than three words are considered an error of the splitting process, and thus eliminated. Afterwards, a set of potential claims that answer the query correctly is compiled for each abstract. As in any text mining application, the input text might be totally unstructured or semi-structured, and the same applies for literature abstracts. For slightly structured abstracts, i.e abstracts where text is divided into subsections such as Title, Introduction/Background, Methods/Aims, Results, and Conclusion, we take advantage of this information and include all sentences within the headings, Results and Conclusion, as candidate sentences. If the text is unstructured, all sentences in the second half of the abstract are included the candidate set, following the assumption that important findings are most probably reported by the end of the abstract. The candidate set is filtered out from any stop words, symbols and punctuations. All 24 PICO questions went through the same filtering process as the candidate sentence collection.

Feature Extraction. A fixed length feature vector representing sentences included in the candidate set is derived. These features combine both semantic and syntactic properties of the sentence. They capture relevance to the query, as well as relatedness to domain-specific concepts. In our model, we rely on easy-to-compute features, which have proven successful in other retrieval tasks. All of the following features are extracted for each of the candidate sentences.

1. *Sentence Length.* The count of terms per sentence after removal of stop words.
2. *Sentence Location.* The relative position of the sentence within the abstract as it highlights the importance of a sentence. The feature is calculated as the location divided by the total number of sentences. However, instead of using the original location of the sentence, we use its position in the candidate set.
3. *Term Overlap.* This measures the number of terms that are found in both the query and the sentence, after removing of stopping words, and also stemming all terms using Porter stemmer [22].
4. *Synonyms Overlap.* The fraction of overlap between the query terms and sentence terms or their synonyms fetched from WordNet.
5. *BM25 score.* The Okapi BM25 framework is a Bag-of-Words model with a collection of scoring functions combined. For a query Q containing n terms $\{q_1, q_2, q_3, ..., q_n\}$ and a sentence length S, the similarity score between Q and S is calculated as

$$\text{BM25score}(S,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,S)\cdot(k_1+1)}{f(q_i,S)+k_1\cdot(1-b+b\cdot\frac{|S|}{\text{avgSl}})}$$

where IDF is the *inverse document frequency*, *avgSl* is the average sentence length and *k*-*b* are two tuning parameters set to 0.75 and 1.5 respectively in our implementation.

6. *Word Embeddings*. Cosine similarity between query and candidate sentences' word vectors pre-trained using a set of over 10 million PubMed abstracts.

The first four features are a subset of the features suggested and used by Metzler and Kanungos work on text summarization [18]. The last two features give more insight into the context of the sentences. In general, word embeddings are a vector representation of words co-occurrences that regard words as contexts, and hence gives a better apprehension of the word meanings. The vectors are generated from a large collection of data through Neural Networks. We take advantage of the word vectors, provided by the BioASQ challenge team [13], trained on a corpus of 10,876,004 English abstracts of biomedical articles from PubMed with 1,701,632 distinct words (types). The implementation of the above features was accomplished using *Summaryrank*[2], a reference package released for a similar task [7,30].

Sentence Ranking. As mentioned above, there might be multiple sentences that answer the query, but only the most suitable should be extracted. For example, while the next two phrases positively answer the question, only the second one is chosen as a claim. *In patients with hypertension, does treatment with ACE inhibitors, compared to placebo, reduce risk of cardiovascular event or improve blood pressure?*

- The vascular pathophysiologic alterations of ISH-a decreased aortic distensibility-can be improved with long-term lisinopril treatment, whereas values deteriorate further in placebo-treated subjects. [PMID: 11336102][assertion= YES]
- These results, in one of the first studies including subjects with previously untreated ISH only, indicate that lisinopril treatment might favorably influence the cardiovascular risk of ISH. [PMID: 11336102][assertion= YES]

Learning to Rank (LTR) is better suited for such a task since it differs from traditional machine learning techniques; the latter solves it as a classification problem while the aim is an optimal order of the instances in the list [17]. In our research, we evaluated two of the popular state-of-art learning to rank algorithms,*LambdaRank* and *LambdaMART*. LambdaRank is a succesor of RankNet that only uses the gradient of the costs instead of the model score. LambdaMART benefits from the strengths of MART, Multiple Additive Regression Trees, and LambdaRank by combining regression trees boosting, used in MART with a cost function derived from LambdaRank [5]. Our proposed model implements

[2] https://github.com/rmit-ir/SummaryRank.

a LambdaMART function, because it outperformed lambdaRank, with training metric NDCG@10. The Normalized Discounted Cumulative Gain (NDCG) is a cumulative measure of the ranking quality truncated at a particular rank level [15]. The model is trained on the generated feature vectors using the RankLib library[3] and the top ranked answer sentence is regarded as the output.

4.2 Contradiction Detection

The contradiction detection component is not regarded as a yes/no question answering system, but more as a semantic relation analyzer between two sentences. The system determines whether the input text has an entailment or contradictory relation.

Query Reformulation. This step aims at modifying the PICO-format question into a reduced list of keywords in a declarative form. In our approach, we consider each word in the question as a keyword unless it is a stop word, question word, or the substring "compared to placebo" is removed as it adds no value when identifying entailment or contradiction. Following that, we apply ClausIE [10], an open information extractor, to identify relations and corresponding arguments found in input question.

Features. Three features are used to identify the assertion values of claims.

Negation. The presence of negation is still the most effective feature of identifying oppositeness. Instead of relying only on the odd count of negation words in the sentence, our proposed model uses NegEx [6]. NegEx takes as input a keyword/concept and a sentence, and uses regular expressions and a predefined trigger word list to decide whether the concept is negated or affirmed. This module iterates three times over each question triple (*left argument, relation, right argument*).

Antonyms. The model includes direct and indirect antonyms; for two words w_1 and w_2, it checks if w_1 is an antonym of w_2 or an antonym of any of its synonyms and vice versa. Instead of comparing raw words, we use lemmas of words for better detection. However, even though the occurrence of antonym pairs in text is a direct and reliable indication of contradiction, it is limited by the low number of antonym pairs in current lexicons. Trying to overcome this limitation, we expand the antonym coverage by using two lexical resources, WordNet [19] and VerbOcean [8]. Below is an example that contains an antonym in *ManConCorpus*:

[3] https://sourceforge.net/p/lemur/wiki/RankLib/.

In women with pre-eclampsia, does treatment with L Arginine, compared to placebo, reduce blood pressure or pre-eclampsia

- L-Arginine load in pregnant women is associated with increased nitric oxide (NO) production and hypotension. [PMID: 10486782 - Assertion value: NO]

In this example, 'reduce' and 'increased' are not direct antonyms like 'good' and 'bad' but are still detected in model. This feature is computed as the count of antonyms per sentence.

Alignment. It is also important to include features that model text entailment. Alignment between sentences relies on mapping dependency graphs of two sentences with each other. The algorithm uses SpaCy[4] to generate dependencies, and a built-in similarity score is calculated for each word node in the query related to a similar one in the claim. Finally, the total alignment score is the sum of all output scores.

Classification. A linear support vector machine classifier is used to determine the relation of each input sentence, based on the output feature values. The model implements the classifier using the Scikit library [21].

5 Results

5.1 Claim Extraction Results

To evaluate the performance of the Learning to Rank framework and the efficiency of the features employed, we conduct two experiments. We first test the model using the first 5 features mentioned in Sect. 4.1 and then we repeat the test after adding the domain-based features covered by the word embedding trained on biomedical articles. For that purpose, we split the *ManConCorpus* into two sets for training and testing purposes. The training set consists of all abstracts with structured format, while the test set includes all unstructured abstracts. After the preprocessing phase, the candidate set has a total of 1212 and 339 sentences for training and testing, respectively. The test set includes 69 answers to only 15 of the 24 queries, while the training set covers all queries with 190 correct claims. Table 1 shows the performance results of the claim selection component. The authors in [1] relied on lexical similarity and a *Z-score* that computes the sentence relevance, with respect to the distribution of similarity scores of other sentences across the dataset. However, While this scoring function contributes to precision, it also affects the recall performance metric. The robustness of our proposed answer detection component relies on the combination of semantic and context features, with an effective ranking algorithm that ranks the sentences according to relevancy, instead of only classifying them as relevant/irrelevant.

[4] https://spacy.io/.

Table 1. Claim extraction results.

	Precision		Recall		F1	
	Answer	Non-answer	Answer	Non-answer	Answer	Non-answer
AlAmri [1]	0.56	0.92	0.57	0.92	0.56	0.92
Model (general features)	0.92	1	1	0.67	0.96	0.80
Model (general & domain-based features)	0.94	1	1	0.75	0.96	0.86

5.2 Contradiction Detection Results

Performance comparison between models is a non-trivial task, therefore we deploy the same evaluation metrics as in [3]. Since there is a bias in the distribution of *YES/NO* classes in the corpus, the results are best reported through precision, recall and F1. The baseline performance is measured by annotating all claims with the majority class *YES*. All evaluation results are shown in Table 2. Our model was able to improve the accuracy of detecting contradictions, namely the *NO* category, and still maintain good results regarding the entailment. The achieved improvement is due to the enhanced negation detection through the NegEx framework, and the inclusion of antonyms.

Table 2. Contradiction detection results.

	Precision		Recall		F1	
	Entailment	Contradiction	Entailment	Contradiction	Entailment	Contradiction
Baseline [3]	0.69	0.0	1.0	0	0.82	0.0
AlAmri [3]	0.85	0.80	0.94	0.60	0.89	0.69
Proposed model	0.95	0.85	0.93	0.89	0.94	0.87

6 Conclusions and Future Work

In this paper, we are interested in identifying conflicting findings reported in biomedical literature. We focus on information found in the abstracts as it summarizes all research methodology and conclusion, and conveys important findings without redundancy. It divides the extraction process into two phases, finding the relevant sentences and detecting contradiction. The model combines both semantic and domain-bases features, to enhance the claim detection process. It relies on an SVM classifier that integrates negation, antonyms and alignment scoring to detect conflicting statements. The evaluation results are very promising, specifically in the contradiction detection component, achieving better performance than other systems.

The results may be influenced by the small size of *ManConCorpus*, and hence further investigations are needed by scaling up the evaluation of the model on much larger corpora. Furthermore, a numeric mismatch in-between sentences is not regarded as a contradiction in the proposed system, since there are no contradiction instances of that type available in the corpus. However, when comparing

clinical evidence found in biomedical literature, specifically when reporting recommended doses, considering variations in numeric values is important. Other possible extensions to the proposed model include incorporating domain knowledge resources, such as *UMLS*, Unified Medical Language System, and possibly integrating contrasting word embeddings [16].

References

1. Alamri, A., Stevenson, M.: Automatic detection of answers to research questions from medline abstracts. In: Proceedings of BioNLP, vol. 15, pp. 141–146 (2015)
2. Alamri, A., Stevensony, M.: Automatic identification of potentially contradictory claims to support systematic reviews. In: Proceedings of IEEE International Conference Bioinformatics and Biomedicine (BIBM), pp. 930–937, November 2015. https://doi.org/10.1109/BIBM.2015.7359808
3. Alamri, A.: The detection of contradictory claims in biomedical abstracts. Ph.D. thesis, University of Sheffield (2016)
4. Alamri, A., Stevenson, M.: A corpus of potentially contradictory research claims from cardiovascular research abstracts. J. Biomed. Semant. **7**(1), 36 (2016)
5. Burges, C.J.: From ranknet to lambdarank to lambdamart: an overview. Learning **11**(23–581), 81 (2010)
6. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. J. Biomed. Inform. **34**(5), 301–310 (2001)
7. Chen, R.C., Spina, D., Croft, W.B., Sanderson, M., Scholer, F.: Harnessing semantics for answer sentence retrieval. In: Proceedings of the Eighth Workshop on Exploiting Semantic Annotations in Information Retrieval, pp. 21–27. ACM (2015)
8. Chklovski, T., Pantel, P.: VerbOcean: mining the web for fine-grained semantic verb relations. In: EMNLP, vol. 4, pp. 33–40 (2004)
9. De Marneffe, M.C., Rafferty, A.N., Manning, C.D.: Finding contradictions in text. In: ACL, vol. 8, pp. 1039–1047 (2008)
10. Del Corro, L., Gemulla, R.: Clausie: clause-based open information extraction. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 355–366. ACM (2013)
11. Harabagiu, S., Hickl, A., Lacatusu, F.: Negation, contrast and contradiction in text processing. In: AAAI, vol. 6, pp. 755–762 (2006)
12. Ioannidis, J.P.: Why most published research findings are false. PLoS Med. **2**(8), e124 (2005)
13. Pavlopoulos, I., Aris Kosmopoulos, I.A.: Continuous space word vectors obtained by applying word2vec to abstracts of biomedical articles. Technical report, NLP Group, Department of Informatics, Athens University of Economics and Business, Greece Institute of Informatics and Telecommunications, NCRS Demokritos, Greece (2014)
14. Jameson, J.L., Longo, D.L.: Precision medicine – personalized, problematic, and promising. Obstet. Gynecol. Surv. **70**(10), 612–614 (2015)
15. Järvelin, K., Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 41–48. ACM (2000)

16. Li, L., Qin, B., Liu, T.: Contradiction detection with contradiction-specific word embedding. Algorithms **10**(2), 59 (2017)
17. Liu, T.Y., et al.: Learning to rank for information retrieval. Found. Trends® Inf. Retrieval **3**(3), 225–331 (2009)
18. Metzler, D., Kanungo, T.: Machine learned sentence selection strategies for query-biased summarization. In: SIGIR Learning to Rank Workshop, pp. 40–47 (2008)
19. Miller, G.A.: Wordnet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
20. Padó, S., de Marneffe, M.C., MacCartney, B., Rafferty, A.N., Yeh, E., Manning, C.D.: Deciding entailment and contradiction with stochastic and edit distance-based alignment. In: TAC (2008)
21. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
22. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980). Google Scholar
23. Prasad, V., Cifu, A., Ioannidis, J.P.: Reversals of established medical practices: evidence to abandon ship. Jama **307**(1), 37–38 (2012)
24. Prasad, V., Vandross, A., Toomey, C., Cheung, M., Rho, J., Quinn, S., Chacko, S.J., Borkar, D., Gall, V., Selvaraj, S., et al.: A decade of reversal: an analysis of 146 contradicted medical practices. In: Mayo Clinic Proceedings, vol. 88, pp. 790–798. Elsevier (2013)
25. Preum, S.M., Mondol, A.S., Ma, M., Wang, H., Stankovic, J.A.: Preclude: conflict detection in textual health advice. In: 2017 IEEE International Conference on Pervasive Computing and Communications (PerCom), pp. 286–296. IEEE (2017)
26. Ritter, A., Downey, D., Soderland, S., Etzioni, O.: It's a contradiction–no, it's not: a case study using functional relations. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 11–20. Association for Computational Linguistics (2008)
27. Sarafraz, F.: Finding conflicting statements in the biomedical literature. Ph.D. thesis, University of Manchester (2012)
28. de Silva, N., Dou, D., Huang, J.: Discovering inconsistencies in pubmed abstracts through ontology-based information extraction. In: ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB) (2017, p. to appear)
29. Sox, H.C., Greenfield, S.: Comparative effectiveness research: a report from the institute of medicine. Ann. Internal Med. **151**(3), 203–205 (2009)
30. Yang, L., Ai, Q., Spina, D., Chen, R.-C., Pang, L., Croft, W.B., Guo, J., Scholer, F.: Beyond factoid QA: effective methods for non-factoid answer sentence retrieval. In: Ferro, N., et al. (eds.) ECIR 2016. LNCS, vol. 9626, pp. 115–128. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30671-1_9