BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 7 2010, pages 982–984
doi:10.1093/bioinformatics/btq060

Data and text mining

Advance Access publication March 3, 2010

PaperMaker: validation of biomedical scientific publications

D. Rebholz-Schuhmann*, S. Kavaliauskas and P. Pezik

European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The automatic analysis of scientific literature can support authors in writing their manuscripts.

Implementation: PaperMaker is a novel IT solution that receives a scientific manuscript via a Web interface, automatically analyses the publication, evaluates consistency parameters and interactively delivers feedback to the author. It analyses the proper use of acronyms and their definitions, and the use of specialized terminology. It provides Gene Ontology (GO) and Medline Subject Headings (MeSH) categorization of text passages, the retrieval of relevant publications from public scientific literature repositories, and the identification of missing or unused references.

Result: The author receives a summary of findings, the manuscript in its corrected form and a digital abstract containing the GO and MeSH annotations in the NLM/PubMed format.

Availability: http://www.ebi.ac.uk/Rebholz-srv/PaperMaker

Contact: rebholz@ebi.ac.uk

Received on October 5, 2009; revised on January 25, 2010; accepted on February 9, 2010

1 INTRODUCTION

The primary purpose of scientific publications is to report on new scientific findings and to embed them in prior research work. The author receives best acceptance if a large audience accurately perceives what he had in mind, and manuscripts of good quality have a higher likelihood to pass the review process.

Scientists write their manuscripts in loosely structured natural language but have to comply with standards concerning the document format, the use of language and the citation of prior work. The availability of electronic data resources such as ontologies, reference databases and electronic literature in the biomedical scientific community exposes the scientists to new requirements: the use of domain-specific terminology has to follow standards from scientific databases and the author has to support submission of data to the reference database as part of the manuscript submission process (Leitner and Valencia, 2008). Furthermore, the author has to avoid duplication of the existing work. Reliable automatic feedback on any of these parameters will improve the speed and efficiency of the manuscript preparation phase and the review phase for authors, publisher and reviewers. Existing solutions for the analysis of the scientific literature focus on improving the search of relevant information (Kim and Rebholz-Schuhmann, 2008) but cannot be applied during the manuscript preparation phase.

2 METHODS

2.1 Text analysis (Whatizit) and evaluation

PaperMaker uses the modular infrastructure of Whatizit (Kirsch *et al.*, 2006; Rebholz-Schuhmann *et al.*, 2008). Modules are described in the order as they are applied (Fig. 1). Candidates for acronyms are all tokens with 2–11 characters in length having an initial uppercase letter. Tokens at the beginning of a sentence were ignored as well as stop words including highly frequent English function words (e.g. In, If). The co-occurrence of the acronym with the long form was identified with syntactic patterns (Schwartz and Hearst, 2003).

Terminologies for the named entity recognition and normalization comply with public resources [e.g. UniProtKb/Swiss-Prot, Gene Ontology (GO), DrugBank, Medline Subject Headings (MeSH), see below]. Complete Medline has been analyzed to generate a terminological resource of all known biomedical terms.

The annotation of the gene and protein named entities is based on the BioLexicon (prec/rec: 94/63; Rebholz-Schuhmann *et al.*, 2009). ChEBI was the reference data resource for the identification of chemical entities. For the annotation of medical entities we used the terms for diseases and syndromes from UMLS (Jimeno *et al.*, 2008) in combination with drug names (DrugBank).

The annotation of text passages with GO concepts uses the solution by (Gaudan *et al.*, 2008): F-measure close to 40% for the terms at Rank 1. The categorization of manuscripts with MeSH terms relies on k-nearest neighbor clustering to attribute text passages to MeSH categories similar to the ones delivered with Medline abstracts (Trieschnigg *et al.*, 2009).

The references in the document were identified based on syntactic patterns and compared with the list of citations in the reference section. The syntactical patterns are defined according to the citation standard in *Nucleic Acids Research* (*NAR*). Missing publications in the reference section as well as unused citations are recognized by the automatic analysis. The identification of related work is achieved by translating a given passage into a Lucene query, which is then run against the Medline and Pubmed Central repositories. PaperMaker's analysis was evaluated on 50 randomly selected publications from NAR.

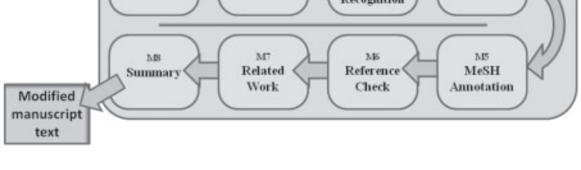

Fig. 1. The diagram gives an overview of the different processing steps for the manuscript. The results from each step are presented to the author.

*To whom correspondence should be addressed.

2.2 Processing of proprietary document formats

The Web application PaperMaker uses JODConverter to convert files into the OpenDocument Format (ODF), an XML format. It integrates OpenOffice.org/Writer functionality that enables import and export of OpenDocument and Microsoft documents (DOC and RTF). Formatting details in the document are preserved through special XML tags in the document structure. Once the text has been analyzed and prepared to be returned to the author, PaperMaker converts it back to the original XML format, which can be used to generate the original and alternative file formats (JODConverter, OpenOffice.org).

3 RESULTS

3.1 Use of PaperMaker

The user interface enables the upload of manuscripts in different file formats (see Section 2.2). None of the data will be stored on site. The interface provides an overview on the availability of required services (top right corner) and a button for help.

The uploaded manuscript will be processed step by step by the modules M1–M8 (Fig. 1). Each module is embedded in its own Web page with its own Help page and the 'GoTo'-butting indicates the index of the next module. The last page summarizes the results of the analysis.

3.2 Identification of unknown terms

In the first step (module M1), we determine the number of unknown, and thus potentially non-standard or misspelled terms in the scientific publications. A potentially unknown term is one which is neither mentioned in the British National Corpus (BNC, general English) nor in any Medline abstract. Both resources combined contain 3 015 437 distinct single-word token 'terms' in total: 89% from Medline only, 5% from both resources and 6% from BNC only. All 50 documents contained unknown terms: 29.5 terms per document and 1474 in total. Most unknown terms turned out to be identifiers (64%, 935 terms): nomenclature identifier, public database identifier and other identifier, and author names.

The remaining 36% of unknown terms (539) are either novel hyphenations of known terms (57%, in 47 publications) or truly novel terms. Removing the hyphenation leads either to composite terms or to single words (e.g. fig leaf, pigeonhole) that conform to dictionary resources (*Shorter Oxford English Dictionary*). Amongst the truly novel terms, we identified algorithms, methods, brand names and gene names. A few terms are neologisms derived from other English words (e.g. organismally). All remaining 22 novel terms (4%) have been misspelled variants of known terms.

3.3 Categorization of known terms

3.3.1 Acronyms resolution (M2) PaperMaker identified 1445 occurrences of a known acronym without a definition in all 50 documents (Gaudan *et al.*, 2005). Of the occurrences, 49.6% (717) were correct. False positive results included short sequences of nucleotides, general English words and Roman numerals (e.g. AC, AS and II). Three hundred and sixteen acronym mentions are novel and would require a definition: 11% (45) of the mentions represent gene/protein named entities and were resolved at a later stage through the biological entity tagger.

In contrast, the majority of acronyms were correctly defined in the manuscript. However, a small subset uses a less frequent definition. The acronym resolution filter also identified acronyms that have been defined at least twice (14) and that have been used before its proper definition.

3.3.2 Identification of named entities (M3) The average document contained 3877 words (reference section not included), 1242 unique words, 29 unique acronyms, 6 unique gene/protein mentions, 18.5 chemical terms and 7 medical terms. Medical terms were identified at 78.7% precision, protein/gene names at 71.2% and chemical entities at 61.7% according to our manual analysis. The low precision of the last module results from the fact that this set is very heterogeneous (DNA is a chemical entity as well as glutamate). Preferred names proposed by the reference data resources were used at low frequency: 19.3% for medical terms, 33% for protein/gene names and 46.3% for chemical entities.

3.3.3 GO and MeSH recognition (M4, M5) PaperMaker assigns first GO terms and then MeSH taking the whole publication into consideration. Forty percent (391 of 973) of the MeSH concepts matched the manually attributed MeSH terms and 20% of manually assigned MeSH headings were not identified.

3.3.4 Reference check module (M6) A total of 4738 publications from PubMed Central were analyzed. Five hundred and seventy-one publications had inconsistencies in the reference section. A random selection of 54 publications was manually inspected to find 12 (22%) true positive results where the authors were mentioned in the bibliographic section but not referenced in the main text. False positive results were due to authors being mentioned in tables or pictures but not in the main text.

3.3.5 Related work and summary (M7, M8) In the last two steps, the author receives an overview on related work and the summary of all findings.

4 CONCLUSIONS

The integration of scientific publications into the biomedical data resources is ongoing work (Rebholz-Schuhmann *et al.*, 2005). PaperMaker supports authors in this integration work without putting efforts to curation teams (Leitner and Valencia, 2008). The final result is the production of structured abstracts in the NLM/PubMed format. Furthermore, the manuscript generation process is well embedded into the literature search and the referencing of relevant prior research.

ACKNOWLEDGEMENTS

The scientific full-text documents were made available by the journal *Nucleic Acids Research*, published by Oxford Journals, Oxford University Press.

Funding: EC Strep project 'BOOTStrep' (FP6-028099).

Conflict of Interest: none declared.

REFERENCES

Gaudan,S. *et al.* (2008) Combining evidence, specificity, and proximity towards the normalization of Gene Ontology terms in text. *EURASIP J. Bioinform. Syst. Biol.*, 342746.

Jimeno Yepes,A. *et al.* (2008) Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinform.*, **9**(Suppl. 3), Article S3.

Kim,J.J. and Rebholz-Schuhmann,D. (2008) Categorization of services for seeking information in biomedical literature: a typology for improvement of practice. *Brief Bioinform.*, **9**, 452–465.

Kirsch,H. *et al.* (2006) Distributed modules for text annotation and IE applied to the biomedical domain. *Int. J. Med. Info.*, **75**, 496–500.

Leitner,F. and Valencia,A. (2008) A text-mining perspective on the requirements for electronically annotated abstracts. *FEBS Lett.*, **582**, 1178–1181.

Rebholz-Schuhmann,D. *et al.* (2008) Text processing through web services: calling Whatizit. *Bioinformatics* **24**, 296–298.

Rebholz-Schuhmann,D. *et al.* (2009) Measuring prediction capacity of individual verbs for the identification of protein interactions. *J. Biomed. Inform.* [doi:10.1016/j.jbi.2009.09.00, October 8].

Rebholz-Schuhmann,D. *et al.* (2005) Facts from text—-is text mining ready to deliver? *PLoS Biol.* **3**, e65.

Schwartz,A.S. and Hearst,M.A. (2003) A simple algorithm for identifying abbreviation definitions in biomedical text. *Pac. Symp. Biocomput.*, 451–462.

Trieschnigg,D. *et al.* (2009) MeSH up: effective MeSH text classification for improved document retrieval. *Bioinformatics* **25**, 1412–1418.