The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence
Scholarly Big Data: AI Perspectives, Challenges, and Ideas:
Technical Report WS-16-13

Enabling Public Access to Non-Open Access Biomedical Literature via Idea-Expression Dichotomy and Fact Extraction

Xiaocheng Huang, Pauline C Ng*
Genome Institute of Singapore (A*STAR)
60 Biopolis Street, #02-01 Genome
Singapore 138672
{huangxc, ngpc4}@gis.a-star.edu.sg

Abstract

The general public shows great potential for utilizing scientific research. For example, a singer discovered her ectopic pregnancy by looking up clinical case reports. However, an exorbitant paywall impedes the public's access to scientific literature. Our case study on a social network demonstrates a growing need for non-open access publications, especially for biomedical literature. The challenge is that non-open access papers are protected by copyright licenses that bar free distribution. In this paper, we propose a technical framework that leverages the doctrine of "idea-expression dichotomy" to bring ideas across paywalls. Idea-expression dichotomy prevents copyright holders from monopolizing ideas, theories, facts, and concepts. Therefore *facts* may pass through paywalls unencumbered by copyright license restrictions. Existing fact extraction methods (such as information extraction) require either large training sets or domain knowledge, which is intractable for the diverse biomedical scope spanning from clinical findings to genomics. We therefore develop a rule-based system to represent and extract facts. Social networkers and academics validated the effectiveness of our approach. 7 out of 9 users rated the paper's information from the facts to be above average ($\geq$6/10). Only 7% of the extracted facts were rated misleading.

1 Introduction

The general public shows great potential for utilizing scientific research. Non-academic individuals have successfully utilized literature for health problems. For example, a pregnant singer discovered her back pain was a symptom of ectopic pregnancy by looking up PubMed case reports[1]. Similarly, an extreme athlete diagnosed her own genetic flaw (Yong 2014). Access to scientific research can also lead to scientific discoveries. A 15-year-old credits open access articles to his success in developing a method for detecting pancreatic cancer (Shockey 2013). These stories highlight the public's ability to utilize and benefit from scientific discoveries.

The need to access scientific papers is widespread. National Institute of Health's PubMed Central provides free access to some biomedical literature and has one million unique visitors daily (National Institute of Health 2015). 40% of its visitors are citizens rather than universities or companies (Swan and Unesco 2012).

However an exorbitant paywall impedes the public's access to some scientific literature, the price of which has outpaced inflation by more than 250%[2] over the past thirty years, with a biology journal costing $1583 on average[3]. 26% of people seeking health information online have hit a paywall (Fox and Duggan 2013) and must decide to pay for a publication which may not have the desired content.

To examine the public's interest in non-open access publications that are behind the paywall, we perform a fine-grained analysis on the social network *Reddit (Scholar)*[4], which is a forum where people submit requests for academic publications that they do not have access to. Our analysis in Section 2 reveals more than 700 Reddit requests per month for scientific papers and the number of new Redditors is growing 37% per month. As expected, the public's major concern is health (>40%). This analysis, for the first time, quantifies the growing need for non-open access publications, especially biomedical publications.

Efforts are being made to allow free or open access to scientific publications. Nonprofit organizations like Creative Commons[5] advocate new copyright licenses. Researchers can share their papers through social networks (such as Academia.edu[6]), but can encounter legal problems with major publishers (Howard 2013). In this paper, we aim to use artificial intelligence techniques to help *facts* bypass paywalls leveraging the concept of "*idea-expression dichotomy*".

Idea-expression dichotomy (Samuels 1988; Yen 1989) limits the scope of copyright protection by differentiating an idea from the expression or manifestation of that idea, and

*Corresponding author
DISCLAIMER: Nothing in this paper shall be construed as legal advice. The authors are not lawyers. Readers should consult qualified legal counsel if they require an opinion on legal matters.

[1] http://www.vanessacarlton.com/news/2013/11/5/first-four-november-tour-dates-canceled

[2] https://www.eff.org/issues/open-access

[3] http://www.lib.washington.edu/scholpub/facts/economics

[4] http://www.reddit.com/r/scholar

[5] http://creativecommons.org

[6] https://www.academia.edu/

thus prevents copyright holders from monopolizing ideas, theories, facts and concepts. In particular, facts are not copyrightable (Feist Publications, Inc., v. Rural Telephone Service Co., 499 U.S. 340 (1991)). We therefore propose a framework relying upon idea-expression dichotomy to extract and distribute facts as a substitute for the full-text of a paper. When someone (with access) views or downloads a paper, the main component of our framework, *Fact Extractor*, will extract facts from this paper locally. The facts are subsequently sent to our server that is searchable to the public. Client-side processing for fact extraction means no copies of the paper are distributed. Ideally, a user could get all the information of a paper by reading facts. At the very least, facts may help the user decide to purchase the paper.

Facts of a paper can be derived from named entities and relations. Existing information extraction techniques focus on general relations such as triples (person, birth date, nationality), and rely on large datasets (Grishman 2011). In addition, fact extraction needs to be domain-independent, since biomedical literature spans many subject areas from clinical findings to genomics. The most relevant work is (Schäfer et al. 2008) and (Blake 2010). The former extracts relations from scientific literature, but it relies on domain-dependent ontology (e.g. LT World ontology). Schäfer et al.'s work also stores all the relation components, which are almost all the words in a sentence, leading to a potential leakage of expressive elements in addition to facts when applied to non-open access papers. The latter identifies claims from full-text biomedical articles, which is a subset of facts. We borrow some results from the aforementioned work, which will be described later.

Example 1.1 *Consider the sentence below that compares 3 antihistamine drugs: diphenhydramine, desloratadine+verapamil and methylcellulose (MC):*

*"*Both diphenhydramine and desloratadine+verapamil treated animals performed significantly less well on the rotarod than the MC treated animals ($p<0.0001$)*" (Paper ID: PMID19356011).*

Fact Extractor's representation of the facts in this sentence is:

*"*diphenhydramine and desloratadine+verapamil treated animals significantly less on the rotarod than MC treated animals ($p<0.0001$)*"*.

We observe that our goal differs from information extraction in the sense that we extract facts for humans (i.e. the public) rather than machines. We therefore propose to keep only *the essence of a sentence* that is factual and important in order to make it comprehensible to humans. Another goal is to reveal all the details of the paper. Therefore we aim to capture the facts of each sentence.

We summarize ten general rules that define facts, such as comparisons (increase or decrease of numbers) and scientific terms (e.g. drugs or biomarkers). To identify domain-dependent scientific terms without relying on previously defined ontology, we observe that the important scientific terms are always repeated in a paper. We therefore use frequent n-grams to represent scientific terms. Example 1.1 presents a sentence and Fact Extractor's facts. A complete list of rules can be found in Section 4.

Note that our proposed framework is not restricted to non-open access biomedical publications. It could be used to extract facts for any biomedical publication.

Our contributions are summarized as follows.

- We provide a case study on the social network Reddit to demonstrate the public's need for non-open access papers. Our study shows a growing need for non-open access papers behind the paywall, especially for biomedical publications. To the best of our knowledge, it is the first study to quantify public's interest in non-open access papers.
- We aim to provide the public a method to bypass paper paywalls via the doctrine of idea-expression dichotomy. Our framework distributes facts, not full-texts. As facts cannot be copyrighted, their distribution is not bound by copyright licenses.
- We provide a rule-based definition and extraction method for facts. Two user studies on academics and social networkers validated the effectiveness of our approach.

The rest of the paper is organized as follows. Section 2 presents our case study on Reddit. Section 3 shows the system overview, and Section 4 describes the methodology to extract facts from sentences. Section 5 provides experiment results for the fact extraction method described in Section 4.

Related Work. Several different approaches could be used to bypass the paper paywall.

One approach is to summarize a scientific paper (Nenkova and McKeown 2012). A paper's abstract is typically the best summary of a paper and is usually available to the public. Users seeking full-text desire details. For these users, text summarization is not comprehensive and is inadequate.

Another approach may be to paraphrase each sentence of a scientific paper. Existing text paraphrase approaches (Quirk, Brockett, and Dolan 2004; Bhagat and Hovy 2013) train a paraphrase model on parallel corpora, typically news articles. However, we observed that when applied to scientific sentences (which tend to be factually dense and minimally expressive), the paraphrases are almost identical to the original sentences. We are concerned that the similarity between the input and output of such paraphrasers may lead to a leakage of expressive elements across the idea-expression barrier; therefore we opt for fact extraction.

Named entities (NE) and their relations are facts and can be mined from massive datasets using information extraction technologies (Sarawagi 2008; Grishman 2011). However, extracting NEs and relations from one individual document requires either domain knowledge (Schäfer et al. 2008), hand-crafted patterns, or (semi-)supervised models for each NE/relation. Since relations in biomedical publications can be quite diverse, it is intractable to enumerate all possible ontologies and relations. One particular relation that we consider is comparative relations. (Liu, Hu, and Cheng 2005; Jindal and Liu 2006b) proposes a supervised-learning approach for extracting comparisons including objects being compared (e.g. Cannon, Nikon) and features (e.g. optics) from product reviews. However, it assumes the objects are simple nouns, which in general is not true for scientific lit-

erature. In Example 1.1, one object in the sentence is a complex phrase "*MC treated animals*". Thus, their method is not applicable for comparison extraction on scientific literature.

(Blake 2010) identifies claims from full-text biomedical articles. Claims are a subset of facts, and according to (Blake 2010), only 22% of sentences contain claims. We will include claims in our future work.

(Seifert et al. 2013) proposes to extract facts from the tables in scientific publications. Their work complements our work on fact extraction from texts and we incorporate some of their work.

2 Case Study

In this section, we perform a fine-grained analysis on a social network and find a growing need for non-open access scientific publications, especially for biomedical publications.

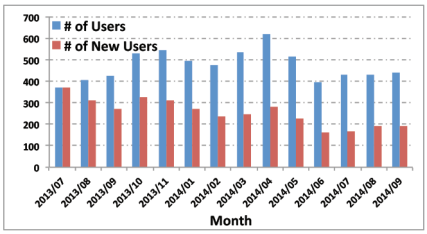
Data Collection. Reddit Scholar is a forum where people submit requests for academic papers that they do not have access to. Other Redditors that have access to a requested paper will share it through file-sharing websites. We downloaded the requests for a 15-month span between 2013-07-01 and 2014-09-30 through the Python Reddit API[7]. Overall, 3,566 distinct users submitted 11,369 paper requests averaging 758 paper requests monthly. A small fraction of submissions requested multiple papers (<1%). We managed to obtain ∼65% (7426) full-texts of the requested papers.

User Analysis. The average number of Redditors requesting papers monthly is 472 (Figure 1a). We further seek to identify the number of **new** users in each month. We define a user as new the first time he/she makes a paper request during the 15-month period. Thus the users in July 2013 are all new. The number of new users decreases each month and eventually stabilizes in June 2014 (Figure 1a). Henceforth, the average number of new users monthly is 177 (June 2014-September 2014). Thus, a substantial fraction of Redditors are new to the Scholar forum (37.5%, 177/472).

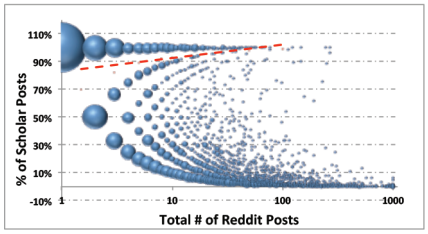
We also analyze if the users are non-academics by looking at their last 1000 posts in Reddit[8], which is the parent host of Reddit Scholar (Figure 1b). The bubble size at a point (x, y) represents the total number of users who submitted x posts to Reddit in the past, and y is the percentage of Reddit Scholar posts. The lower the y is, the more probable the user is a non-academic. It is clear from Figure 1b that the users can be classified into two clusters (separated by the red line). One cluster of users submitted most of their posts to Reddit Scholar, and thus may be academics (above the red line). In contrast, a substantial number of users (72.3%) are not on Reddit for purely academic reasons (% of Reddit Scholar posts $< 100\%$).

We therefore conclude that the public's need for academic papers is growing.

Impact Factor (IF) of Publications. To examine types of publications of interest, Zotero[9] is used to obtain the meta-info of publications, including title, type and journal. Overall, we obtain meta-information for 5980 (out of 7426) publications. 96.8% (5729/5980) are journal articles. We further analyze the impact factor[10] of the journals. The journals' impact factors for 3021 articles are obtained (Figure 2a). Surprisingly, most of the articles (62%) are not from high-profile journals (IF≤5).



(a) Monthly Users on Reddit Scholar



(b) User History on Reddit

Figure 1: User Analysis

Topics of Interest. To examine Redditors' topics of interest, AlchemyAPI[11] is used to classify the papers into their most probable topic category. For papers without full-text, the paper title is used to obtain the most probable topic. To avoid uncertain classification, we throw out the papers that are assigned topics with low probability (probability<0.4).

Of the 11,369 submissions, topics for 8531 papers are retrieved (5448 from full-texts and 3083 from paper titles). The topic distribution from full-texts and paper titles have similar trends, with "Health and Fitness" and "Science" being the top 2 topics. The overall distribution is in Figure 2b. People are mainly interested in "Health and Fitness" (41%). Moreover, within the "Science" category, 27% of the papers belong to the "Medicine" subcategory (Figure 2c). In total, 45.9% (3916/8531) of all requested papers are biomedical (categories Health & Fitness, Science/Medicine). This underscores the public's need to access information from biomedical publications.

3 System Overview

In this section, we present the framework for extracting facts from biomedical papers, thus enabling us to invoke the doctrine of idea-expression dichotomy. Recall that facts are not copyrightable, and thus not bound by copyright license.

Figure 3 shows the overview of our system. Academics (e.g. students, scientists, and librarians) have access to non-open access publications while the public is behind a paywall. The academic will locally install our browser exten-

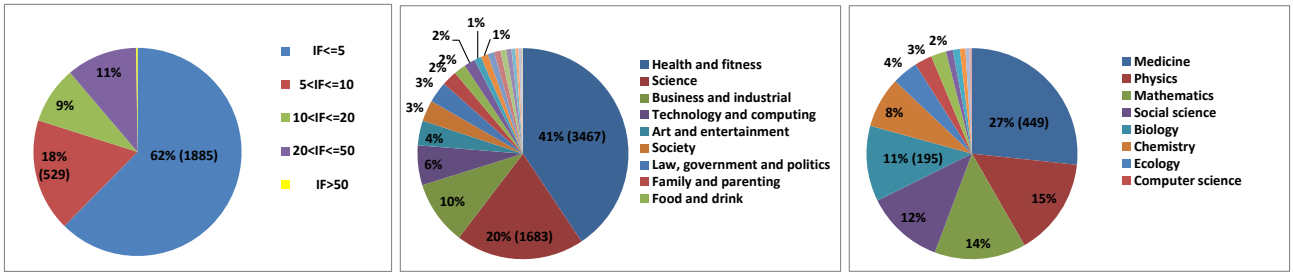
[7] https://praw.readthedocs.org/en/v2.1.21/

[8] Reddit restricts the maximum downloads to 1000.

[9] https://www.zotero.org/support/retrieve_pdf_metadata

[10] http://www.uco.es/servicios/biblioteca/indpacto/jcr_2013.pdf

[11] http://www.alchemyapi.com/



(a) Journals' Impact Factor (IF) of the Requested Articles

(b) Top-level Topics of Interest

(c) Science Subtopics

Figure 2: Requested Publications

sion that can extract facts on his/her computer. If the extension detects that a scientific publication is being downloaded or viewed, the browser extension will call Fact Extractor to extract the facts from the paper, and then send these facts to our server. The extraction process is done locally on the academic's computer to assure that the paper is never distributed. In theory, the doctrine of idea-expression dichotomy means that the facts stored on the main server are not subject to copyright and thus free for the public to search and access.

The browser extension calls the fact extraction package. The fact extraction package is in Java and thus is platform-independent. Most non-open access papers are in PDF (Portable Document Format) format. We use the state-of-the-art work by (Klampfl et al. 2014) for converting PDF to section texts (including section titles, paragraphs). Fact Extractor processes the section texts, and extracts facts from each sentence. Our code can be found on GitHub[12]. The facts are displayed on our public web server http://factpub.org.

We now describe Fact Extractor in detail. The technical solution on fact extraction is designed to be light-weight, because it is intended to run on computers locally. Fact Extractor therefore uses a rule-based system to extract facts.

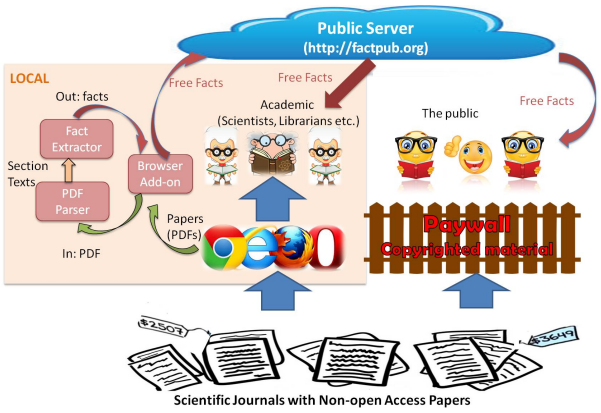


Figure 3: System Overview

[12]https://github.com/happybelly

4 Fact Extractor

In this section, we describe the main component of our system, Fact Extractor. Fact Extractor processes the section texts, and extracts facts from each sentence.

Fact Extractor takes as input a publication's text. A publication's text is broken down into section titles (e.g. Results) and their corresponding paragraphs (Klampfl et al. 2014). Paragraphs are further broken down into sentences using Stanford CoreNLP (Manning et al. 2014). Fact Extractor processes sentences and then outputs sentence-level facts for each section (Figure 3).

Given a sentence from a publication, there could be multiple ways to represent the sentence's facts. Here, we define a sentence's facts as a word sequence (possibly with gaps) extracted from the sentence that conveys the *factual* and *important information*.

Example 1.1 (in Introduction) gives an example of Fact Extractor's representation of facts. Our goal is for a reader to deduce the true meaning of the original sentence by looking at Fact Extractor's representation.

We define ten rules for extracting facts from a sentence. These rules are applied to each sentence from a publication to identify a word sequence of factual terms. We retain the original word order in the sentence and extract the important terms that provide factual information. Punctuation is retained from the original sentence for the sake of readability. After applying to all sentences in a publication, a list of sentence-level facts is generated. For brevity, Fact Extractor's sentence-level facts will be referred to as FE facts.

The fact extraction process consists of rule matching and result composition. Prior to rule matching, Stanford CoreNLP is used to tokenize words and obtain Part-Of-Speech tags (POS tags) and word stems. The methodology for each rule match is described below:

Rule 1. Noun identification. Nouns (e.g. "diphenhydramine") are important in the biomedical field because they commonly describe a method or a treatment. A word is designated as a noun if its POS tag is NN, NNS, NNP or NNPS.

Rule 2. Scientific terms. Using nouns only can miss more complex scientific terms (e.g. "desloratadine+verapamil", "treated animals"). We observe that important concepts and scientific terms are repeated within a paper. Therefore, we choose *frequent n-grams* to represent scientific terms. *N-*

grams of a sentence, for n=1 to total number of tokenized words in the sentence, are word subsequences (without gaps) of length n of the sentence that contain at least one noun and do not start or end with propositions. In this paper, we define a *frequent* n-gram to be an n-gram that appears more than once among all the sentences. Note that, frequent n-grams may overlap with each other; we will handle this case at the final step. Given a sentence, any subsequence (without gaps) of the sentence that is identical to any frequent n-gram is deemed a scientific term.

Rule 3. Acronyms. Acronyms are important because long scientific terms are often abbreviated into a shortened form that is used consistently within a paper (e.g. "MC" stands for "methylcellulose"). We use the algorithm described in (Schwartz and Hearst 2003) to obtain acronyms.

Rule 4. Comparative and superlative terms. When two nouns are identified, their relationship becomes meaningful by including any terms that describe the comparative relation. Therefore, comparative and superlative terms are important (e.g. X "less than" Y). A word is designated as a comparative or superlative term if it has one of the following POS tags: JJR, RBR, JJS, and RBS.

Rule 5: Terms expressing comparative relations. Comparative relations of what "increases" or "decreases" are also part of biological facts. We manually compiled a list of 24 words by finding synonyms for the words "increase", "decrease" and included the 79 comparison keywords from (Jindal and Liu 2006a). A word's stem is checked to see if it belongs to the list of 103 words.

Rule 6: Terms expressing discovery or importance. Discovery terms such as "reveal" indicate important scientific findings. A list of 35 words were manually compiled by finding synonyms for "discover" and "significant". A word's stem is checked against this list.

Rule 7: Negative and quantitative terms. Negative terms (e.g. "not") and quantitative terms (e.g. "any") are important because they can change the meaning of a sentence. We built a word list of such terms and Fact Extractor identifies and keeps such words when observed in a sentence. 4,784 negative opinion terms were obtained from sentiment analysis (Hu and Liu 2004). An additional six terms ("few", "all", "some", "any", "other", "except") were manually added to the list. A word's stem is checked against this list.

Rule 8. Numbers. Numbers (and their units) may form the basis of a fact, and are identified using regular expressions.

Rule 9. Units. Units of measurement are identified using JNumWu library[13].

Rule 10. P-values. P-values also bear information of the significance of fact (e.g. "p<0.0001"), and are identified using regular expressions.

After applying the ten rules to a sentence, the words potentially containing factual information are combined into a FE fact. To resolve the overlaps of different rules, we maintain a boolean marker for each word in the sentence. The marker is set to 1 if the word matches a rule. All the words in the sentence that are marked as 1 constitute the FE fact. The sentences from a paper are compiled as a list of FE facts.

[13]http://sourceforge.net/projects/jnumwu/

Our method is neither perfect nor complete. Take Example 1.1 in Introduction and its corresponding FE facts. Removing word "Both" might mislead readers to think "diphenhydramine and desloratadine+verapamil" is one treatment if readers are unaware that the paper compares three treatments: "diphenhydramine", "desloratadine+verapamil" and "MC". We will discuss this in further detail in the experiment section.

5 Experiments

In this section, we evaluate the performance of our Fact Extractor by user studies. We conducted two user studies targeting two different types of users, scientists and nonacademics. Scientists were recruited from our research group and nonacademics were from Reddit who posted a biomedical paper request in the past. In both cases, users selected their own papers, and were asked to read and rate FE facts from a biomedical paper.

Experimental design

User recruitment process. Two groups completed Fact Extractor: Redditors and fellow scientists. We invited Redditors who posted a biomedical paper request (and received the full-text) on the Reddit Scholar forum to evaluate extracted FE facts, and offered $5 for participation. Recruitment started on June 1, 2015 and 100 invitations were sent out. Nineteen Redditors expressed willingness to participate and were provided the FE facts of their requested paper. Four (out of 19) Redditors eventually provided their feedback.

Five scientists from our own research group selected papers for journal club and were also provided with FE facts. We refer the four Redditors as R1-R4 and the five academics from our research group as A1-A5.

Validation process. To evaluate the performance of Fact Extractor independently of text extraction from the PDF, we manually copied and pasted the sections (section titles and paragraphs without tables or figures) from full-text papers into Fact Extractor. Fact Extractor outputs FE facts, which are presented in an Excel spreadsheet, where each row is either a section title, or a FE fact.

The user was then asked to 1) give an overall evaluation of facts and 2) rate the FE fact for each sentence exclusively to be *misleading* or *meaningless* and give feedback in words if possible. Otherwise, we assume the fact to be correct. The overall evaluation consists of the following two questions:

Q1 Amount of information: How much information did you get (rate 0-10)?

Q2 Satisfaction: How satisfied are you with the information (rate 0-10)?

Three reading tips were provided to our users: 1) FE facts are represented as a word sequence extracted from each sentence; 2) assume positivity if something is missing (adjectives or adverbs or verbs may not be stated); and 3) figures and tables are missing.

Evaluation

The overall evaluation from the Redditors and scientists is presented in Table 1. The FE facts are helpful in most cases.

Table 1: Overall Evaluation
(Q1-Amount of information; Q2-Satisfaction)

User	Q1	Q2	User	Q1	Q2
R1	6	5	A1	8.5	6.5
R2	7	7	A2	7	7
R3	6	5	A3	9	9
R4	4	3	A4	7	6
			A5	3.5	5
Avg	5.75	5	Avg	7	6.7

7 out of 9 users rated the paper's information from the facts to be above average (Q1, rating ≥ 6). Yet the satisfaction level is slightly lower (Q2). We could not assess inter-rater agreement, because our readers were not interested in the same paper. However the performance rates were similar among all readers (6.44 ± 1.83, Q1; 5.94 ± 1.70, Q2).

The scientists were in general more satisfied than Redditors with FE facts. This may be because scientists have background knowledge on the paper, such as fundamental scientific concepts and mechanisms, so it is easier for them to infer the true meaning of FE facts. Readers might find it helpful to be given a set of basic scientific concepts prior to reading the facts; this will be explored in future work.

It is worth noting that a closer look at Redditor R4's feedback actually shows that R4 may not have understood the exercise. His ratings were low, and when we looked into more detail on his feedback, approximately 26% did not seem reasonable. For example, R4 rated Fact Extractor's output "Patients with KIT exon 9 mutations showed best efficacy." to be meaningless when the original sentence was "Patients with KIT exon 9 mutations showed the best efficacy."

The four Redditors also commented on whether each FE fact was misleading or meaningless. (The scientists skipped this exercise.) The misleading (or meaningless) rate is defined as the proportion of misleading (or meaningless) facts, and is shown in Table 2. The last column in Table 3 gives the total error rate, which is the sum of the misleading rate and the meaningless rate. The average misleading rate is much lower than the meaningless rate (7.15% vs. 22.40%), which is satisfying because we would not want users to be misled by FE facts. The total error rate is 19.7% on average. Therefore, readers found 80% of FE facts to be meaningful.

We also analyzed the error rate of different sections of papers. Fact Extractor performed better for the Results sections than Introduction and Method sections with error rates 18.90%, 30.45% and 35.89%, respectively.

After examining Redditors' feedback, we find several causes for error. First, approximately 33% of erroneous FE facts lacked important verbs, including general verbs (e.g. "*suspend*") and scientific verbs (e.g. "*phagocytose*"). Second, approximately 30% lacked essential adjectives (e.g. "*gastrointestinal* tumors"). In addition, approximately 14% lacked meaning-changing words (e.g. "*avoid*", "*until*", and "*despite*"). We also observed that some sentences were not split correctly since periods were placed before citations, which calls for a new sentence detection model. Some important prepositions involved with scientific processes were removed (e.g. "*after* treatment"). A small fraction of erroneous FE facts might be recovered by coreference resolution (Recasens, Can, and Jurafsky 2013), since misinterpretation was due to missing pronouns (e.g. "*their* study").

Table 2: Fraction of Misleading and Meaningless Facts

User	# of FE Facts	Misleading Rate (%)	Meaningless Rate (%)	Error Rate (%, sum)
R1	162	0	2.47	2.47
R2	167	17.37	2.40	19.77
R3	109	2.75	11.01	13.76
R4	118	8.47	73.73	82.20
Avg	139	7.15	22.40	19.70

6 Conclusions and Future Work

In this paper, we demonstrate the public's growing need for non-open access biomedical publications and propose a framework to extract facts as substitutes for full-texts. Fact Extractor represents facts as word sequences (with gaps) extracted from each sentence using ten general rules. User studies on social networkers and academics validate the promise of our approach. Future work includes 1) improving the readability of FE facts by providing additional information such as core scientific concepts, figures and tables; and 2) reducing FE facts' error rate by incorporating important verbs and adjectives, resolving coreferences among pronouns, and using crowdsourcing to refine the facts.

7 Acknowledgements

We thank the authors of (Klampfl et al. 2014) for providing their code. We thank Reddit volunteers and our colleagues for validating our results.

References

Bhagat, R., and Hovy, E. H. 2013. What is a paraphrase? *Computational Linguistics* 39(3):463–472.

Blake, C. 2010. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *J. of Biomedical Informatics* 43(2):173–189.

Fox, S., and Duggan, M. 2013. Health online 2013.

Grishman, R. 2011. Information extraction: Capabilities and challenges.

Howard, J. 2013. Posting your latest article? You might have to take it down. *The Chronicle of Higher Education.* Retrieved from http://chronicle.com/blogs/wiredcampus/posting–your–latest–article–you–might–have–to–take–it–down/48865.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *SIGKDD*, 168–177.

Jindal, N., and Liu, B. 2006a. Identifying comparative sentences in text documents. In *SIGIR*, 244–251.

Jindal, N., and Liu, B. 2006b. Mining comparative sentences and relations. In *AAAI*, 1331–1336.

Klampfl, S.; Granitzer, M.; Jack, K.; and Kern, R. 2014. Unsupervised document structure analysis of digital scientific articles. *Int. J. on Digital Libraries* 14(3-4):83–99.

Liu, B.; Hu, M.; and Cheng, J. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *WWW*, 342–351.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J. R.; Bethard, S.; and McClosky, D. 2014. The stanford corenlp natural language processing toolkit. In *ACL*, 55–60.

National Institute of Health. 2015. Plan for increasing access to scientific publications and digital scientific data from NIH funded scientific research.

Nenkova, A., and McKeown, K. 2012. A survey of text summarization techniques. In *Mining Text Data*. 43–76.

Quirk, C.; Brockett, C.; and Dolan, W. B. 2004. Monolingual machine translation for paraphrase generation. In *EMNLP*, 142–149.

Recasens, M.; Can, M.; and Jurafsky, D. 2013. Same referent, different words: Unsupervised mining of opaque coreferent mentions. In *NAACL HLT*, 897–906.

Samuels, E. 1988. The idea-expression dichotomy in copyright law. *Tenn. L. Rev.* 56:321.

Sarawagi, S. 2008. Information extraction. *Foundations and Trends in Databases* 1(3):261–377.

Schäfer, U.; Uszkoreit, H.; Federmann, C.; Marek, T.; and Zhang, Y. 2008. Extracting and querying relations in scientific papers on language technology. In *LREC*.

Schwartz, A. S., and Hearst, M. A. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *PSB*, 451–462.

Seifert, C.; Granitzer, M.; Höfler, P.; Mutlu, B.; Sabol, V.; Schlegel, K.; Bayerl, S.; Stegmaier, F.; Zwicklbauer, S.; and Kern, R. 2013. Crowdsourcing fact extraction from scientific literature. In *HCI-KDD*, 160–172.

Shockey, N. 2013. Open access empowers 16-year-old Jack Andraka to create breakthrough cancer diagnostic. *Right to Research Coalition.* Retrieved from http://www.righttoresearch.org/blog/open–access–empowers–16–year–old–to–create–breakth.shtml.

Swan, A., and Unesco. 2012. *Policy guidelines for the development and promotion of open access*. Paris: United Nations Educational, Scientific, and Cultural Organization.

Yen, A. C. 1989. First amendment perspective on the idea/expression dichotomy and copyright in a work's total concept and feel, A. *Emory LJ* 38:393.

Yong, E. 2014. DIY Diagnosis: How an extreme athlete uncovered her genetic flaw. *Mosaic Science.* Retrieved from http://mosaicscience.com/story/diy–diagnosis–how–extreme–athlete–uncovered–her–genetic–flaw.