On developing extraction rules for mining informal scientific references from Altmetric data sources

Khawaja, Waqas; Taylor, Michael; Davis, Brian

2015-06-04

Khawaja, Waqas, Taylor, Michael, & Davis, Brian. (2015). On Developing Extraction Rules for Mining Informal Scientific References from Altmetric Data Sources. In Chris Biemann, Siegfried Handschuh, André Freitas, Farid Meziane & Elisabeth Métais (Eds.), Natural Language Processing and Information Systems: 20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015, Passau, Germany, June 17-19, 2015, Proceedings (pp. 443-447). Cham: Springer International Publishing.

Springer Verlag

http://dx.doi.org/10.1007/978-3-319-19581-0_44

http://hdl.handle.net/10379/6602

http://dx.doi.org/10.1007/978-3-319-19581-0_44

Some rights reserved. For more information, please see the item record link above.
On Developing Extraction Rules for Mining Informal Scientific References from Altmetric Data Sources

Waqas Khawaja\(^1\), Michael Taylor\(^2\), and Brian Davis\(^1\)

\(^1\) Insight Centre for Data Analytics, National University of Ireland, Germany
\{waqas.khawaja, brian.davis\}@insight-centre.org
http://www.insight-centre.org

\(^2\) Elsevier, Oxford, United Kingdom
\{Mi.Taylor\}@elsevier.com
http://www.elsevier.com

Abstract. Altmetrics measure scientific impact outside of traditional scientific literature. While different methods adopting citation counts measure the impact within scientific literature, altmetrics may look for research impact on government policies or public discourse by exploring alternative data sources such as news, articles, blogs, social media or government documents. We identify mentions of scientific research or entities like researchers, academic or research organizations. We collect a corpus containing blogs, articles, news items etc and manually analyse it for patterns of such informal mentions. We then apply text mining techniques by developing extraction rules for mining informal mentions. We apply them to our development corpus and present our results. This work takes us closer to developing concrete altmetrics for determining research impact on news, public discourse ultimately leading to government policies.

Keywords: Text Mining, Altmetrics, Informal Scientific References

1 Introduction

Citation count had been the foundation of measuring research impact for a long time until refined but similar measures were introduced like H-Index\(^{[14, 9]}\). While other factors are taken into account to measure a more careful impact, these approaches are based heavily on counting the citations a piece of research output has obtained. These methods do not provide ways to determine the impact research created on media, public discourse or even government policies. The advent of social media has immensely enhanced the diffusion of scientific impact and hence the increased interest in the study of altmetrics\(^{[10, 15]}\) which refers to the study of alternative measures for discovering research impact.
We obtain a heterogeneous corpus of sources such as news, articles, blogs and official government pages to work towards identifying existence of scientific research instances. We look for mentions of research itself or research related entities for example, scientists, research organizations, or research and development departments of commercial entities. We manually annotate these mentions and craft JAPE grammar rules to extract the same in General Architecture for Text Engineering (GATE) \[4\]. JAPE (Java Annotations Pattern Engine) is a pattern matching language over features and annotations implemented as a finite state transducer\[5\]. Our task is somewhat similar to scenario template extraction in Information Extraction but our intention is to convert the problem to sentence/relation classification task. We first look at the existing work done in the domain and then explain the types of mentions discovered. We then highlight the grammar development process and present our initial results before summarizing limitations and recommendations.

2 Altmetrics and related Work

Social media has attracted a lot of attention from scientists in search of altmetrics in the recent years. They looked at Linkedin, Facebook, Twitter, blogs and review websites etc and adopted different methods to develop and evaluate altmetrics\[16, 2\].

A Twitter study found that 6% of studied tweets contained first or second order link to research articles\[2\]. A deeper analysis of the content of tweets found that these mostly mentions summary or title of a paper\[2\]. Researchers concluded that Twitter is mostly using for making idea popular \[7\]. Young researchers with high presence on Twitter also tend to blog as well and have high networking with similar researchers\[11\]. Unlike Twitter, blogs are thought to be an effective medium for initiating discourse\[12\]. A collection of research paper samples from Delicious, Mendeley, Twitter and Wikipedia revealed that 62% of them were present of Mendeley with insignificant presence on other platforms\[16\].

Snowball Metrics also attempt to void the gap present in altmetrics \[3\]. It is a collection of 23 different methods described in a freely available recipe book.

While text mining has focused on traditional scientific databases and publications, methods from text mining/analytics\[6\] have seen adoption for analyzing opinions and sentiments as well as to determine the impact of scientific research on other research\[8\].

3 Implementation and Experiment

3.1 Mentions

We first considered the question that what qualifies as a mention. We are looking for appearances of scientific research or entities. The scientific entities may be researchers, academic or research organizations, as well as research related operations of commercial companies. The entities also include government health departments and officials. A few examples of possible candidates are given below.
1. In relation to the cost of R&D reported by industry an article in the journal BioSocieties (Feb. 2011), a publication of the London School of Economics (LSE), argued that the real cost of R&D is, in fact, a fraction of the commonly cited estimates.

2. According to the Centers for Disease Control and Prevention (CDC), approximately 40,000 persons are infected with HIV each year.

3. GlaxoWellcome has embarked on a major direct-to-consumer advertising barrage.

The first one (1.) is a mention of scientific research from an academic institute and the second (2.) is a figure stated by an official health department. The third (3.) however, is an action and is not considered as a mention. There are many other variations of mentions found later in the corpus. More observations indicate that papers are not explicitly cited in resources such as the case with conventional bibliometrics.

3.2 Corpus Analysis

We collected a corpus of around 500 documents reaching to 130 MB. These documents were indexed from the web against the keyword Tamiflu. The corpus includes news reports, articles, and reports among others. We indexed the corpus in Mimir [13] which is a semantic search platform. Mimir allows searching over patterns of semantic/linguistic annotations. Our queries looked for mentions of entities using a variety of combinations. We manually annotated this corpus with 232 mentions of scientific research or entities.

We assumed certain words that were likely to appear in the target mentions. We then issued queries in Mimir to look for such instances. Some of the Mimir queries we issued are:

\begin{verbatim}
'said [1..10] (paper|study)'
'quote [1..5] ({Person} | {Organization} | Dr) [1..5] questioned'
\end{verbatim}

We further refined our Mimir queries based on the results we obtained. Further analysis of results returned by Mimir and manual study of the corpus for similar patterns enabled us to devise of list of 30 trigger phrases that frequently appear in these mentions.

3.3 JAPE Grammar Development

We compiled all our triggers in finite state custom gazetteer. We included the default VP chunker (rule based) in our pipeline to account for different forms of verb phrases. We first developed a JAPE preprocessor to annotate all words in our trigger list as triggers using gazetteer lookup. The following JAPE grammar is of the preprocessor.
We then crafted JAPE grammars initially taking inspiration from the queries we wrote for Mimir. We noted the possible appearance of entities in beginning, middle or end of sentences. The following text is an example of a person entity appearing in the middle of a mention.

However, the shortage doesn’t include the capsule form of Tamiflu, which remains in good supply, said Dr. Michael Jhung, a medical officer with the U.S. Centers of Disease Control and Prevention’s Influenza Division.

We created a corpus pipeline in GATE which consists of some default shallow linguistic processing resources from the standard ANNIE information extraction pipeline that includes a Sentence Splitter and POS Tagger. We replaced the ANNIE NER with Stanford NER based on a simple comparison experiment and added Verb Group (VG) chunker and finally two custom JAPE grammars to identify mentions in our corpus based on the trigger words and their lexico-syntactic/semantic context.

The following JAPE grammar rule looks for mentions in text where an entity appears first that may be followed by one or more organizations and finally a trigger phrase before the end of the sentence.

Rule:Reference3

\[
\begin{align*}
\text{Rule:Reference3} & \quad (\text{({Person} | {Organization})+ {Trigger} {Split}}) \\
\text{ bind} & \quad --> \quad \text{ bind.TempMention} = \{\text{rule=Reference3, type=reference}\}
\end{align*}
\]

Our entities and trigger phrases can appear anywhere in a sentence. Consider for example the mention, Tamiflu is an antiviral medication that blocks the actions of the influenza virus in the body, says Dr. Sterkel. The trigger phrase and entity are the last two in the sentence. In order to capture the complete sentence as a mention, we first create a TempMention and finally the JAPE grammar given below creates the annotates the complete sentence as a mention.

\[\text{http://goo.gl/bwpqTG}\]
4 Results

Our manually annotated corpus was used for analysis of trigger phrases as well as for developing the JAPE grammars. Our results presented here are obtained from testing the JAPE grammars on our development corpus described in Section 3.2.

Table 1: JAPE Grammar Results

|                  | Key  | System |
|------------------|------|--------|
| Total            | 321  | 329    |
| Match            | 162  |        |
| Only Key         | 76   |        |
| Only System      | 84   |        |
| Overlap          | 83   |        |
| Recall           |      |        |
|                  | Strict | 0.50  |
|                  | Lenient | 0.76  |
|                  | Average | 0.63  |
| Precision        |      |        |
|                  | Strict | 0.49  |
|                  | Lenient | 0.74  |
|                  | Average | 0.61  |
| F measure        |      |        |
|                  | Strict | 0.49  |
|                  | Lenient | 0.75  |
|                  | Average | 0.62  |

We expanded our basic custom gazetteer using synonyms from Wordnet lookup. While this enabled capture of more mentions, it also introduced some false positives which needed to be corrected.

There are some erroneous mentions as well that are annotated by our JAPE grammars. Consider the following mention annotations that has been annotated because R&D has been annotated as an organization which is to be expected from an off the shelf NER and a trigger phrase reported is found though in an incorrect syntax.

However, these figures should be taken with caution as they are usually taken from pharmaceutical industry reports which are known for the lack of transparency in relation to the cost of R&D and there are difficulties for verifying the figures reported.

Figure 1 & 2 present mentions annotated by our JAPE grammars.
5 Conclusions and Future Work

Our work does not take co-reference into account yet. There are many instances of mentions where a statement from a person is followed by a few other statements where the person is referred to as ‘he’ or ‘she’. Our system does not resolve nominal coreference yet. So the mentions of scientific research that have nouns like ‘the study concluded’ or ‘the report classified’ or ‘the researchers said’ are not captured.

We ultimately want to measure the impact of scientific research in government literature and policies. We have already gathered a corpus of government documents indexed from government health related websites. A more fine grained annotation schema must be developed that would later be used in development of a gold standard model corpus with inter annotator agreement and involving at minimum three annotators. The rule based extraction offers high precision over recall that is more suited for boot strapping machine learning in the absence of a training corpus. We also plan to augment the rule based extraction patterns with machine learning to enhance our recall. Finally, we will look at linking the extracted mentions of entities to unique identifiers in to scientific databases such as Scopus.

6 Acknowledgement

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

References

1. S. Atdag and V. Labatut. A comparison of named entity recognition tools applied to biographical texts. In Systems and Computer Science (ICSCS), 2013 2nd International Conference on, pages 228–233. IEEE, 2013.

4 http://www.scopus.com
2. J. Bollen, H. Van de Sompel, A. Hagberg, and R. Chute. A principal component analysis of 39 scientific impact measures. *PloS one*, 4(6):e6022, 2009.

3. L. Colledge. Snowball metrics recipe book. *Amsterdam, the Netherlands: Snowball Metrics program partners*, 2014.

4. H. Cunningham. Gate, a general architecture for text engineering, *Computers and the Humanities*, 36(2):223–254, 2002.

5. H. Cunningham, D. Maynard, and V. Tablan. JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield, Nov. 2000.

6. J. Elder IV and T. Hill. *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press, 2012.

7. K. Holmberg and M. Thelwall. Disciplinary differences in twitter scholarly communication. *Scientometrics*, 101(2):1027–1042, 2014.

8. R. N. Kostoff, M. Temixco, M. M. J. A. Humenik, M. Rockville, and M. L. A. M. Ramirez. Citations mining.

9. H. F. Moed. *Citation analysis in research evaluation*, volume 9. Springer Science & Business Media, 2006.

10. J. Priem, P. Groth, and D. Taraborelli. The altmetrics collection. *PloS one*, 7(11):e48753, 2012.

11. H. Shema, J. Bar-Ilan, and M. Thelwall. Research blogs and the discussion of scholarly information. *PloS one*, 7(5):e35869, 2012.

12. H. Shema, J. Bar-Ilan, and M. Thelwall. How is research blogged? a content analysis approach. *Journal of the Association for Information Science and Technology*, 2014.

13. V. Tablan, K. Bontcheva, I. Roberts, and H. Cunningham. Mimir: An open-source semantic search framework for interactive information seeking and discovery. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2014.

14. M. Thelwall. Bibliometrics to webometrics. *Journal of information science*, 2008.

15. M. Thelwall, S. Haustein, V. Larivière, and C. R. Sugimoto. Do altmetrics work. twitter and ten other social web services. *PloS one*, 8(5):e64841, 2013.

16. Z. Zahedi, R. Costas, and P. Wouters. How well developed are altmetrics? a cross-disciplinary analysis of the presence of alternative metrics in scientific publications. *Scientometrics*, 101(2):1491–1513, 2014.

17. J. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005.