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, Galway, Ireland
{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. We identify mentions of scientific research or entities like researchers, academic or research organizations in a corpus containing blogs, articles, news items etc. We manually analysed the corpus for patterns of such informal mentions and then applied text mining techniques by developing extraction rules for mining informal mentions. We applied them to our development corpus and present our results. This work takes us closer to developing concrete altmetrics for determining research impact on news and public discourse ultimately leading to measuring impact of scientific research on 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. The more recent measures like H-Index[10, 7] still rely on citation count to measure impact. While more recent, they do not provide ways to determine the impact research created on media, public discourse or even government policies.

We look for mentions of research itself or research related entities for example, scientists, research organizations, or research and development departments of commercial entities in a heterogeneous corpus of sources such as news, articles, blogs and official government pages. We manually annotated these mentions and crafted JAPE grammar rules to extract the same in General Architecture for Text Engineering (GATE)[2]. JAPE(Java Annotations Pattern Engine) is a pattern matching language over features and annotations implemented as a finite state transducer[3]. 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.

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[11, 1].

A Twitter study found that 6% of studied tweets contained first or second order link to research articles[1]. Researchers concluded that Twitter is mostly using for making idea popular [5]. Unlike Twitter, blogs are thought to be an effective medium for initiating discourse[8].

While text mining has focused on traditional scientific databases and publications, methods from text mining/analytics[4] have seen adoption for analyzing opinions and sentiments as well as to determine the impact of scientific research on other research[6].

3 Implementation and Experiment

We collected a corpus of around 500 documents reaching to 130 MB. As a use case for scholarly discourse, we choose the anti virual drug *Tamiflu* and indexed the documents from the web against this keyword. The corpus includes news reports, articles, and reports among others. We indexed the corpus in Mimir [9] which is a semantic search platform. 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 also devised a list of 30 trigger phrases that frequently appear in these mentions.

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 process lookup triggers that overlap only with verbs for simple disambiguation. We put this all together in 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[12] based on a simple comparison experiment[3] 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
(
  ({Person} | {Organization})+ {Trigger} {Split}
):bind
-->
:bind.TempMention = {rule=Reference3, type=reference}
```

[3] http://goo.gl/bvpqTG

Our entities and trigger phrases can appear anywhere in a sentence. Consider for example the mentiond, *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 highlighted with green and red respectively. 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.

```
Rule: Reference1
(
{Sentence contains TempMention}
):bind
-->
:bind.Mention = {rule=Reference2, type=reference}
```

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. The number of annotations and precision, recall, and F-measure information can be seen in Tables 1 and 2 respectively.

Table 1 Annotation Results of Pipeline

	Key	System
Total	321	329
Match	162	
Only Key	76	
Only System	84	
Overlap	83	

Table 2 Precision, Recall & F-Measure

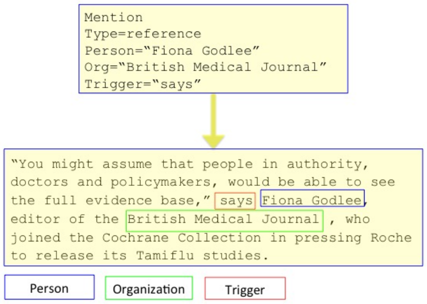
Recall	Strict	0.50
	Lenient	0.76
	Average	0.63
Precision	Strict	0.49
	Lenient	0.74
	Average	0.61
F measure	Strict	0.49
	Lenient	0.75
	Average	0.62

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 annotation 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 although subject verb dependency is clearly incorrect as the rule is too relaxed. The trigger phrase is highlighted in green and the organization is highlighted in red:

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.

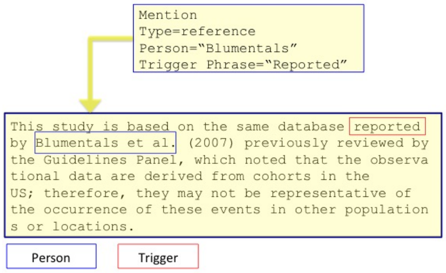
Further example of correct mentions annotated by our JAPE grammars can be seen in Figures 1 & 2.



Mention
Type=reference
Person="Fiona Godlee"
Org="British Medical Journal"
Trigger="says"

"You might assume that people in authority, doctors and policymakers, would be able to see the full evidence base," says Fiona Godlee, editor of the British Medical Journal , who joined the Cochrane Collection in pressing Roche to release its Tamiflu studies.

Fig. 1 Example of a mention with Person, Organization and Trigger Phrase



Mention
Type=reference
Person="Blumentals"
Trigger Phrase="Reported"

This study is based on the same database reported by Blumentals et al. (2007) previously reviewed by the Guidelines Panel, which noted that the observa tional data are derived from cohorts in the US; therefore, they may not be representative of the occurrence of these events in other population s or locations.

Fig. 2 Another Captured Mention with Person and Trigger

5 Future Work

Our goal is 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 tune our rules and 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[4].

Acknowledgments

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

[4] http://www.scopus.com

References

1. 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.
2. H. Cunningham. Gate, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254, 2002.
3. 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.
4. J. Elder IV and T. Hill. *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press, 2012.
5. K. Holmberg and M. Thelwall. Disciplinary differences in twitter scholarly communication. *Scientometrics*, 101(2):1027–1042, 2014.
6. R. N. Kostoff, M. Temixco, M. M. J. A. Humenik, M. Rockville, and M. L. A. M. Ramírez. Citation mining.
7. H. F. Moed. *Citation analysis in research evaluation*, volume 9. Springer Science & Business Media, 2006.
8. 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.
9. V. Tablan, K. Bontcheva, I. Roberts, and H. Cunningham. Mímir: An open-source semantic search framework for interactive information seeking and discovery. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2014.
10. M. Thelwall. Bibliometrics to webometrics. *Journal of information science*, 2008.
11. 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.
12. 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