On Developing Extraction Rules for Mining Informal Scientific References from Altmetric Data Sources

Waqas Khawaja¹, Michael Taylor², and Brian Davis¹

1. National University of Ireland, Galway 2. Elsevier

Introduction

- Research impact measurement is through citation count (including more recent measures like H-Index)
- The measures do not provide for calculating impact on media, public discourse or government policies
- We work on developing Altmetrics that are non-traditional metrics aiming at measuring research impact from alternative data sources such as news or social media
- 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 sources includes: news, articles, blogs and government sources
- We search for mentions of scientists, research organizations, industry R&D departments, labs in our heterogeneous corpus
- We manually annotate these mentions and craft **Java Annotation Pattern Engine** (JAPE) grammar using the text analytics toolkit - **General Architecture for Text Engineering** (GATE)

Method

- Drawing a trigger list using Mimir queries.
 - 'said [1..10] (paper|study)'
 - 'quote [1..5] ({Person} | {Organization} | Dr) [1..5] questioned'
- Devised a list of trigger phrases.
 - Cite, Claim, Publish, Reason

Results

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

Future Work

- Improving Results
- A more fine grained annotation schema
- Develop gold standard
- Conduct inter-annotator agreement
- Augment rule based approach with machine learning

This project has been funded by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 and by targeted project funding from Elsevier.