Constructing a Scientific Blog Corpus for Information Credibility Analysis

Eric Nichols, Koji Murakami, Kentaro Inui, and Yuji Matsumoto
Nara Institute of Science and Technology
{eric-n,kmurakami,inui,matsu}@is.naist.jp

1 Introduction

Corpora are an essential resource for data-driven approaches to semantically demanding tasks such as recognizing textual entailment and multi-perspective question answering, however, they are expensive to construct and maintain, and existing corpora may not contain the information necessary to approach a new task.

In this paper we discuss the construction of an English corpus for use in evaluating the credibility of information on the web. Because the identification of conflicting opinions and their logical justifications is of great importance, we turn to scientific blogs as our primary source of data. By exploiting the format of the blog posts and the networked nature of the scientific blogging community, we minimize construction costs by automating many of the tasks necessary for data collection and annotation.

2 Motivation

The importance of the internet as a source of information cannot be disputed. A recent poll [8] by the Pew Research Center found that among Americans the internet has overtaken newspapers as a news outlet and rivaled television for those surveyed under the age of thirty. In this age of widespread internet access, anyone with a computer can put their ideas on the Web where they can be viewed by a large audience. However, this removal of the barriers to publication has also made it easier to spread disinformation.

2.1 The Anti-vax Movement: A Cautionary Tale

The anti-vaccination movement (hereafter "the anti-vax movement") is a good example of the danger of disinformation. In 1998, a group of researchers in the UK published a study implying a causal connection between Measles, Mumps, and Rubella (MMR) vaccinations and the development of autism in children [11]. Though further scrutiny of these initial results disproved the autism-vaccination link, culminating in the withdrawal of endorsements by 10 of the study's 12 authors, the damage had already been done.

The mainstream media picked up on the study, amplifying fears about the safety of vaccinations in an already nervous public. An anti-vaccination movement soon formed, fueled by celebrity activists. Online communities[1] developed, insulating their members against the medical evidence to the contrary. Vaccination rates plummeted despite the best efforts of public health organizations [3].

The result of the spread of the anti-vax movements was that in 2008, for the first time in over a decade, there was a resurgence in the number of reported cases of measles in both the United States [1] and Europe. The situation in the UK was serious enough to be elevated to an endemic [2]. Measles, which in the 1990s was considered a cured disease, was making a comeback.

2.2 The Importance of Evaluating Credibility

The case of the anti-vax movement causing a resurgence in measles is a sad one, but it could have been prevented. After all, Wakefield et al.'s study [11] was repeated numerous times in an attempt to verify the connection between MMR vaccinations and autism, and the results were overwhelmingly against such a causative connection[2]. But this information did not get to the very people concerned about the safety of vaccinations. Part of the blame belongs with the mainstream media which, both online and offline, was more interested in entertaining conspiracy theories than presenting the wealth of evidence disproving a vaccination-autism link, but the underlying problem that people did not know how to find trustworthy evidence to the contrary is illustrative of the importance of evaluating the credibility of information.

Clearly the problem of evaluating information credibility is important, but how can we help users decide what information to trust? One approach taken by several projects is to educate users about how to identify good information online. Services like snopes.com and factcheck.org debunk urban myths and provide fact checking to commonly made political claims. The Quackometer[3] uses language models to identifies pseudo-scientific language in webpages. Sense About Science[4] campaigns to educate users about the importance of the scientific method and peer review. Credibility Commons[5] provides tools to help users automatically evaluates the credibility of webpages. Finally, a number of professionals in fields ranging from science and medicine to history and economics share their expert opinions with the public through blogs.

The above projects are all invaluable, but users are often not aware of them, and there may not always be a dedicated resource for a given user's topic of interest. More needs

[1] http://www.ageofautism.com

[2] An updating list of studies can be found at http://en.wikipedia.org/wiki/MMR_vaccine_controversy

[3] http://www.quackometer.net/

[4] http://www.senseaboutscience.org.uk

[5] http://credibilitycommons.org

to be done to connect users with the good information out there on the Web. In order for a user to come to an informed opinion, he or she needs to be presented with all of the viewpoints on a topic and the justification or supporting evidence for each one.

An ideal technological solution would mine the Web for opinions on a user's topic of interest, aggregate and summarize these opinions, and show them to the user together with the support for each opinion. The user would be told *who* holds the opinion (i.e. their qualifications); *what* the opinion is; *when* it was held (to insure its relevance); and, finally, *why* they hold that opinion (i.e. the factual or logical justification for the opinion). In order to create such a solution, we need to be able to: *(i)* find statements of focus to the user's topic of interest on the Web; *(ii)* classify these statements into different opinions; and *(iii)* identify logical relations between the statements (i.e. agreement, conflict, support, etc.).

The Statement Map[6] project [7] is working on a system that automatically identifies and summaries opinions for Japanese internet users. This work represents the first step to expand the project to handle English, starting with the acquisition of the resources necessary to build tools capable of detecting the statements of focus and the logical relations of interest to users. We do so by exploiting a previously-untapped data source: *scientific blogs*.

3 Scientific Blogs as a Corpus

Let us return to the example of the anti-vax movement to show the potential of using scientific blog data as a corpus. In October of 2008, and anti-vax activist Jenny McCarthy claiming that she had cured her autistic son by changing his diet [10]. The interview, which offered no evidence at all to support this claim, angered Phil Plait, a professional astronomer and blogger for the scientific publication *Discover Magazine*.

The author of *Bad Astronomy*, whose other pursuits include debunking the claims of moon landing skeptics and presiding over the James Randi Educational Foundation[7], is not a medical doctor, but as a scientist he has a healthy respect for the scientific process: the verification of testable hypotheses through repeatable experimentation. So he wrote an entry at his blog, *Bad Astronomy* [9], critical of the *Us Magazine* piece.

Bad Astronomy's author pointed out that medical doctors have not verified the claimed recovery of Jenny McCarthy's son, and explained the logical fallacy of *post hoc ergo propter hoc*[8] present in both her claims of a vaccination-autism link and her son's cure through a change in diet. He reminded his readers that failure to get their children vaccinated facilitates the spread of infectious diseases like measels, and ended with a plea not to buy in to the anti-vaxxers' groundless claims.

Soon, other members of the scientific blogging community had noticed the *Bad Astronomy* post, and weighted in with their own opinions. One blogger pointed readers to stopjenny.com, a website dedicated to refuting the arguments of the anti-vax movement and its spokeswoman. An entire discussion about the credibility problems in the *Us Magazine* article was sparked by the blog post at *Bad Astronomy*.

3.1 A Collection of Discussions

This kind of linked discussion on the same topic, participated in by many members of the blogging community is what makes the construction of our corpus possible. The authors of scientific blogs share a common goal of celebrating good science while tearing down bad science. They seek out examples of bad science (and bad science reporting) in the mainstream media and on the internet and refute it point-by-point, explaining the logical fallacies and other common pitfalls. When bad science appears, it is often surrounded by controversy: global warming denialism, safety concerns regarding the Large Hadron Collider, and alternative medicine are often-addressed topics by science bloggers. Furthermore, the blogs posts are written for a general audience in an informal, easy-to-understand manner, instead of the terse, jargon-laden prose common to scientific publications.

We construct our corpus by forming *discussions* – collections of posts from different blogs discussing and organized around a single topic or article. The *discussion* created from Phil Plait's blog post on Jenny McCarthy is shown in Figure 1. The structure of the blogs and the networked nature of the blogging community facilitate this task. Tags in each blog post make it easy to identify the topic of discussion. Blog posts contain a link to the *source of interest* – the original mainstream media news article, event, or other blog post that inspired authors to respond with their own opinions. Once *discussion*s are formed, we identify *statements of focus* – opinions, facts or justification pertinent to the topic of discussion – and annotate the logical relations between them.

3.2 Logical Relations for Annotation

The Statement Map Project targets a wide variety of logical relations, described in detail in [5, 7]. To simplify the construction of our corpus, we focus on a small number of relations that are plentiful in the scientific blog data and that are most important to the task of detecting and reporting a variety of viewpoints: contrasting opinions, logical refutations, elaboration, and agreement. We give some examples below taken from the *discussion* shown in Figure 1.

- **Contrasting Opinions**
  A: Jenny McCarthy says she helped her son, Evan, recover from autism.
  B: We have not seen any diagnoses of her son.
- **Logical Refutation**
  A: The actress believes the MMR vaccine was to blame for her son's diagnosis of autism.
  B: Doctors have made very careful studies of this, and there is no link between vaccines and the onset of

[6] 言論マップ *genron map* in Japanese
[7] http://www.randi.org
[8] "after therefore because of:" mistaking precedence for causality

```
<TOPICS: Autism, vaccinations, cures>
    ...                                  ...
[Us Magazine]            Jenny McCarthy: My Son No Longer Has Autism
[Bad Astronomy]          ...but how do we recover from Jenny McCarthy?
[stopjenny.com]          Stop Jenny McCarthy
[Pharyngula]             What an excellent name for a website
[Skeptico]               Stop Jenny McCarthy!
[Bad Astronomy]          Antivaxxers must be stopped! NOW.
[Matt Maroon]            An Open Letter to Parents Everywhere
[Bad Astronomy]          Antivaxxers are doing real damage to society
    ...                                  ...
```

Figure 1: An example *discussion* on the topic of **autism**

autism.

- **Elaboration**
  A: Ms. McCarthy is engaging in a mistaken way of thinking called post hoc ergo propter hoc.
  B: Vaccinations are given around the same time children can be first diagnosed with autism! So it makes a link, a false link in a parent's mind.
- **Agreement**
  A: Ms. McCarthy and the antivaxxers have lots of anecdotes, but the real evidence is totally against them.
  B: There is no evidence vaccines cause autism. Jenny McCarthy is wrong.

3.3 Comparison to Existing Corpora

In constructing a corpus from scientific blog data, we focus on making *coarse-grained* annotations of *statements of focus* in multi-document *discussion*s. As the project progresses, we will increase the level of annotation information necessary to create tools for the Statement Map project. In this section, we compare our annotation approach to similar resources.

3.3.1 MPQA Corpus

The Multi-Perspective Question Answering Corpus [12] is composed of news articles with text indicating opinions annotated. Because the data is primarily indirect reports of opinions, it focuses on identifying an opinions source, target, and intensity at the sub-sentential level. In contrast, our corpus is composed primarily of first-person reports of opinions, with annotated logical relations between statements.

3.3.2 Large, Unannotated Blog Collections

The TREC Blog Track data [4] and the Spinn3r Blog Dataset[9] used at ICWSM 2009 are both large (> 100GB) collections of RSS feeds, article texts, and comments. Data contains no markup and may include spam. Our corpus is also constructed using RSS feeds, but it is smaller in scope, organized into multi-document *discussion*s, and will contain annotated logical relations between statements.

4 Constructing the Corpus

Our goal is to minimize annotation cost by exploiting the structure of the blogs to automatically identify and download promising data, extract statements of focus, and eventually annotate the logical relations between these statements. In this section, we propose algorithms for finding scientific blog data, grouping posts into *discussion*s, and identifying statements of focus for annotation.

4.1 Finding the Data

Scientific blogs feeds are discovered and stored in an RSS reader. New posts are downloaded and converted into text.

1. **Search for science blogs**
   Look for patterns in blog names (e.g. *BAD*, *SKEPTIC*):

   ```
                BAD Astronomy
                BADchemist
   Good Math,   BAD Math
                BAD Science
                ...
       Action   SKEPTICs
      Secular   SKEPTIC
                SKEPTchick
                SKEPTICo
                SKEPToblot
                ...
   ```

2. **Subscribe to RSS feeds**
   Gather feeds in *Google Reader*[10] for search and archival.
3. **Find similar blogs**
   - Search blogrolls for related blogs
   - Follow trackbacks in blog posts
   - Get suggestions from Google Reader's *feed discovery service*[11]
4. **Download blog posts**
   - Use screen-scraping tools to convert posts to text

4.2 Annotating the Data

Blog posts are organized into *discussion*s using links in posts together with category tag information. *Statements of focus* are identified in the *discussion* and marked for annotation.

1. **Get topics from blog post's category tags**
   Topics help identify *statements of focus* for annotation. E.g. *autism*, *vaccines*, *Jenny McCarthy*, *antivaxxers*, ...

[9] http://www.icwsm.org/2009/data/

[10] http://www.google.com/reader/

[11] http://www.google.com/support/reader/bin/answer.py?hl=en&answer=80468

2. **Expand set of topics**
   Cluster topics with tf-idf and find synonyms in WordNet
3. **Create a** *discussion* **from linked posts**
   - Follow links from initial blog post
   - Count number of links back to scientific blog community to identify *source of interest*
   - Group related blog posts with *source of interest* to create a collection of opinions with context
4. **Find** *statements of focus*
   - Simple search for terms in expanded topic set
5. **Annotate** *statements of focus* **in the** *discussion*
   - Use proximity of links in text to help narrow down annotation search space

5 Current Progress

We are currently monitoring the RSS feeds of over 40 science blogs and have collected over a thousand blog posts. We are creating *discussion*s on topics as diverse as autism and vaccinations, the Large Hadron Collider, omega acid dietary supplements, fruit fly research, and electronic voting.

Here are some examples of *statements of focus* that cannot be picked up using topic information alone.

- **Pronoun resolution**
  The actress (= Jenny McCarthy) believes the MMR vaccine was to blame for her son's diagnosis.
- **Loose paraphrasing**
  [Jenny McCarthy] says a strict no wheat-and-dairy-free diet has changed her son from *a quiet little boy who used to flail his arms around* (= autistic) to *a loving six-year-old* (= cured).

6 Future Work

While our corpus is off to a promising start, there is still much work to do. First, we plan to evaluate the usefulness of links in detecting *sources of interest* and grouping blog posts into *discussion*s.

Once we have collected a sufficient number of interesting *discussion*s, we will focus on annotating *statements of focus*. This entails evaluating the coverage and accuracy of category tag based retrieval. As we have seen, there are *statements* that cannot be picked up by shallow methods.

Finally, automatic detection of logical relations between statements would facilitate the construction of our corpus. Mishne observed that temporal cues and comment information were both useful in improving blog search [6]. Using that information together with contextual cues like the proximity of links in text may help to reduce the search space.

7 Acknowledgments

This work is supported by the National Institute of Information and Communications Technology Japan. Francisco Dalla Rosa Soares, Hiram Calvo, Francis Bond, and Michael Goodman also provided invaluable feedback. Finally, we would like to thank all of the bloggers, especially Ben Goldacre of badscience.net for the data.

References

[1] CDC. Update: Measles outbreaks continue in U.S. Website for Centers for Disease Control and Prevention, 2008. Available at: http://www.cdc.gov/Features/MeaslesUpdate/.

[2] Eurosurvelliance. Measels once again endemic in the United Kingdom. Eurosurvelliance, 13(27), 2008. Available at: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=18919.

[3] Finding Dulcenia. European health officials cope with measles outbreaks, lower vaccination rates. Finding Dulcenia: Librarian of the Internet, 2009. Available at: http://www.findingdulcinea.com/news/health/2009/jan/European-Health-Officials-Cope-With-Measles-Outbreaks--Lower-Vaccination-Rates.html.

[4] Craig Macdonald and Iahd Ounis. The TREC Blogs06 collection : Creating and analysing a blog test collection. DCS Technical Report Series, 2006.

[5] Suguru Matsuyoshi, Koji Murakami, Yuji Matsumoto, and Kentaro Inui. A database of relations between predicate argument structures for recognizing textual entailment and contradiction. In Proceedings of the Second International Symposium on Universal Communication, pages 366–373, December 2008.

[6] Gilad Mishne. Using blog properties to improve retrieval. In Proceedings of ICWSM 2007, 2007.

[7] Koji Murakami, Suguru Matsuyoshi, Asuka Sumida, Hiraku Morita, Chitose Sao, Shoko Masuda, Yuji Matsumoto, and Kentaro Inui. Generating statement maps for capturing supportive and contrastive relations between statements. In The Technical Report of IEICE, volume NL186, pages 55–60, July 2008. (in Japanese).

[8] Pew Research. Internet overtakes newspapers as news outlet. Website for the Pew Research Center for the People & the Press, 2008. Available at: http://people-press.org/report/479/internet-overtakes-newspapers-as-news-source.

[9] Phil Plait. ...but how do we recover from Jenny McCarthy? Bad Astronomy, 2008. Available at: http://blogs.discovermagazine.com/badastronomy/2008/10/20/but-how-do-we-recover-from-jenny-mccarthy/.

[10] Us Magazine. Jenny McCarthy: My son no longer has autism. Website for Us Magazine, 2008. Available at: http://www.usmagazine.com/news/jenny-mccarthy-my-son-is-no-longer-autistic/.

[11] A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, and J A Walker-Smith. Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. The Lancet, 351(9103), 1998.

[12] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation (formerly Computers and the Humanities), 39(2/3):164–210, 2005.