Constructing a Scientific Blog Corpus for Information Credibility Analysis

Eric Nichols, Koji Murakami, Kentaro Inui, and Yuji Matsumoto
Computational Linguistics Laboratory,
Graduate School of Information Science,
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0192 JAPAN
{eric-n,kmurakami,inui,matsu}@is.naist.jp

Abstract

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 blog post metadata and the network structure of the scientific blogging community, we minimize construction costs by automating many of the tasks necessary for data collection and annotation. We propose a technique for gathering blog posts into multi-document *discussions* that share a common *source of interest* and evaluate a filtering method for reducing noise in the corpus data.

1 Motivation

The importance of the internet as a source of information cannot be disputed. A recent poll (Pew Research, 2008) 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 false information.

1.1 The Anti-vax Movement: A Cautionary Tale

The anti-vaccination movement (hereafter "the anti-vax movement") is a good example of the danger of misinformation. 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 (Wakefield et al., 1998). 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 (Finding Dulcenia, 2009).

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 (CDC, 2008) and Europe. The situation in the UK was serious enough to be elevated to an endemic (Eurosurveillance, 2008). Measles, which in the 1990s was considered a cured disease, was making a comeback.

1.2 The Importance of Evaluating Credibility

The case of the anti-vax movement causing a resurgence in measles is tragic, but it could have been prevented. After all, the study of Wakefield et al. (1998) 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 need to assist the evaluation of information credibility.

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

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

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 identify 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 evaluate the credibility of webpages. Finally, a number of professionals in fields ranging from science and medicine to economics and politics share their expert opinions 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 to be done to connect users with the good information out there on the Web. In order for users to come to informed opinions, they need 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* related 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 project (Murakami et al., 2009) is working on a system that automatically identifies and summarizes 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] http://www.quackometer.net/
[4] http://www.senseaboutscience.org.uk
[5] http://credibilitycommons.org

2 Scientific Blogs as a Corpus

Let us return to the example of the anti-vax movement to show the potential of scientific blogs as a corpus. In October of 2008, *Us Magazine* published an interview with celebrity and anti-vax activist Jenny McCarthy claiming that she had cured her autistic son by changing his diet (Us Magazine, 2008). The interview, which offered no evidence to support this claim, angered Phil Plait, a professional astronomer and blogger for the science news source *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[6], is not a medical doctor, but as a scientist he has a healthy respect for the scientific process. So he wrote an entry at his blog, *Bad Astronomy*, critical of the *Us Magazine* piece (Plait, 2008).

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*[7] 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 diseases like measles, 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*.

2.1 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 exposing bad science. They seek out examples of bad science (and bad science reporting) in the mainstream media and on the internet and refute them point-by-point, explaining the logical fallacies and other common pitfalls. When bad science appears, it is often surrounded by controversy: global warming denialism,

[6] http://www.randi.org
[7] "after therefore because of:" mistaking precedence for causality

[TOPICS] Antiscience, Autism, Debunking, Energy, General Science, Intelligent Design/Creationism, Jenny McCarthy, Kooks, Medicine, New Age Mysticism, Piece of mind, Pointless Words of Wisdom, Politics, Science, Skepticism, anti-vaccine lunacy, ...

[SOURCE] Us Magazine: Jenny McCarthy: My Son No Longer Has Autism
Despite criticism from the American Academy of Pediatrics, Jenny McCarthy says she helped her son, Evan, recover from autism. The actress —who believes the MMR vaccine was to blame for her son's diagnosis— 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 to a loving six-year-old.

[BLOGS]

- **Bad Astronomy**: ...but how do we recover from Jenny McCarthy?
  Ms. McCarthy has an autistic son. Or, according to her, he was autistic; now she's claiming her son has been cured of autism. She makes this claim in a Us magazine interview, saying changing his diet by removing wheat and dairy products has cured him.
- **Left Brain, Right Brain**: Jenny McCarthy and the Holy War
  Four months after that, under the headline 'Jenny McCarthy: My Son No Longer Has Autism' Jenny McCarthy says she helped her son, Evan, recover from autism.
- **MattMaroon.com**: An Open Letter To Parents Everywhere
  But don't take medical advice from Jenny McCarthy, or, for that matter, any Playboy Bunny, past or present. Take medical advice from doctors.
- **Skepacabra**: News From Around The Blogosphere 10.21.08
  Phil Plait goes off on Jenny McCarthy – Wow! Phil rarely covers the antivaccine movement in his blogs but that's 2 days in a row. Awesome!

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

safety concerns regarding the Large Hadron Collider, and alternative medicine are often-addressed topics by science bloggers. Furthermore, the blog 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 blog post metadata and the network structure of the blogging community facilitate this task: category tags are aggregated from blog posts identifying the *discussion*'s topic, and trackbacks and blogrolls are used to identify other sources of related posts.

Blog posts contain a *source of interest*: a link to the original mainstream media news article, event, or other blog post that inspired authors to respond with their own opinions. In this example, the *source of interest* is the *Us Magazine* article. Once *discussions* are formed, we will identify *statements of focus*– opinions, facts, or justification relevant to the topic of discussion– and annotate them with logical relations.

2.2 Logical Relations for Annotation

The Statement Map Project targets a wide variety of logical relations, described in detail in (Murakami et al., 2009). 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 from the *discussion* 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 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 parents 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 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 *discussions*. 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.

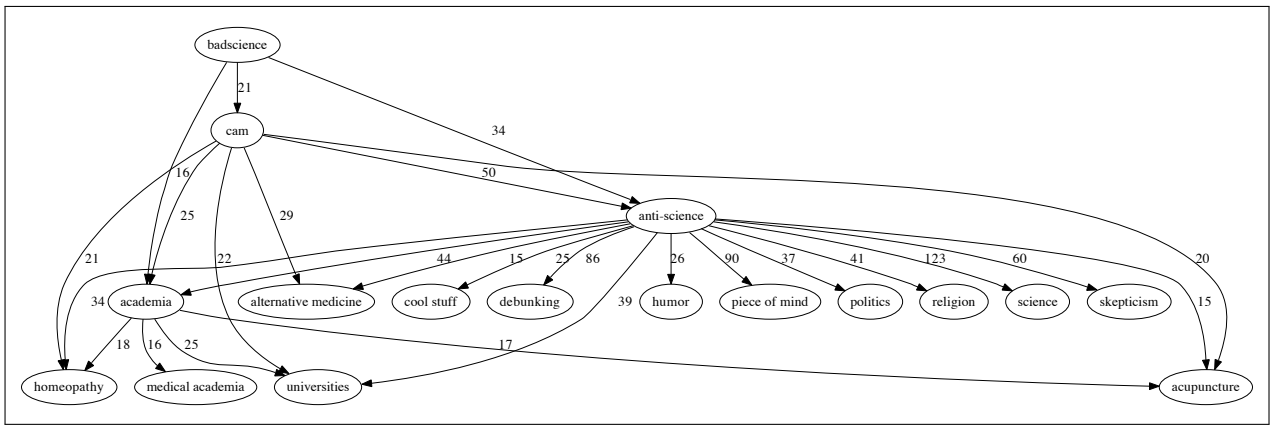

Figure 2: A sample of blog categories that co-occur with *badscience*

3.1 MPQA Corpus

The Multi-Perspective Question Answering Corpus (Wiebe et al., 2005) is composed of news articles annotated with opinions. Because the data is primarily indirect reports of opinions, it focuses on identifying an opinion's source, target, and intensity at the subsentential level. In contrast, our corpus is composed primarily of first-person reports of opinions, annotated with logical relations between statements.

3.2 Large, Unannotated Blog Collections

The TREC Blog Track data (Macdonald and Ounis, 2006) and the Spinn3r Blog Dataset[8] 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 be annotated with logical relations.

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 annotate the logical relations between these statements. In this section, we propose algorithms for finding scientific blog data and grouping posts into *discussions*

Scientific blogs feeds are discovered and stored in an RSS reader. New posts are downloaded and converted into text. Blog posts that share topics are organized into *discussions*.

1. **Subscribe to scientific blog RSS feeds**
   - Gather feeds in *Google Reader*[9]
2. **Expand subscriptions**
   - Search blogrolls for related blogs
   - Follow trackbacks in blog posts
   - Get suggestions from Google Reader's *Recommendations* service[10]
3. **Download blog posts**
   - Use screen-scraping tools to convert posts to plain text
4. **Create a *discussion* from linked posts**
   - Group blog posts that link to the same *source of interest*

Posts	Blog name	Field
1,657	Pharyngula	Biology
1,390	Badscienceblogs	General science
1,019	Freakonomics	Economics
1,014	Bad astronomy	Astronomy
903	A blog around the clock	Biology
806	Skepchick	Skepticism
778	Sandwalk	Biochemistry
740	Language log	Linguistics
653	Change.org's autism blog	Autism
622	Skepacabra	Skepticism
503	Debunking christianity	Religion
463	Respectful insolence	Medicine
330	The panda's thumb	Biology
306	Be lambic or green	Skepticism
275	New urban legends	Skepticism

Table 2: Names & fields of most productive blogs

5 Evaluation

We are currently monitoring the RSS feeds of over 70 science blogs. These blogs cover a wide variety of

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

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

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

# of posts per discussion	Unfiltered discussions	Unfiltered precision				Filtered discussions	Filtered precision			
2	2,498	11	/	25	(44.0%)	2,335	11	/	23	(**47.8%**)
3	571	7	/	25	(28.0%)	520	7	/	22	(**31.8%**)
4	213	16	/	25	(64.0%)	170	14	/	21	(**66.7%**)
5	103	10	/	25	(40.0%)	77	9	/	17	(**52.9%**)
6	71	10	/	25	(40.0%)	30	9	/	11	(**81.8%**)
7	48	16	/	48	(33.3%)	29	15	/	29	(**51.7%**)
8	31	8	/	31	(25.8%)	23	8	/	23	(**34.8%**)
9	16	5	/	16	(31.2%)	12	5	/	12	(**41.7%**)
10	7	2	/	7	(28.6%)	4	2	/	4	(**50.0%**)
11	13	4	/	13	(30.8%)	8	4	/	8	(**50.0%**)
12	13	2	/	13	(15.4%)	9	2	/	9	(**22.2%**)
13	10	2	/	10	(20.0%)	5	2	/	5	(**40.0%**)
14	5	1	/	5	(20.0%)	3	1	/	3	(**33.3%**)
15	1	1	/	1	(100%)	1	1	/	1	(100%)
16	3	1	/	3	(33.3%)	2	1	/	2	(**50.0%**)
17	3	1	/	3	(33.3%)	1	1	/	1	(**100%**)
18	1	1	/	1	(100%)	1	1	/	1	(**100%**)
>18	23	0	/	23	(0.0%)	6	0	/	6	(0.0%)
total	3,630	98	/	299	(32.8%)	3,236	93	/	198	(**47.0%**)

Table 1: Hand-verified *discussions* ordered by number of posts containing *source of interest*

scientific disciplines: biology, linguistics, economics, and information science are but a few of the fields represented. Table 2 gives a breakdown of the 15 blogs with the greatest number of posts and their fields.

Figure 2 gives a sample of the categories shared by blog posts labeled as *badscience*. The numbers on the graph edges represent the number of times the two categories co-occur. Only categories co-occurring at least 15 times are shown, including categories that share an intermediary with *badscience*. Categories like *badscience*, *anti-science*, and *skepticism* may be useful for building *discussions* because they indicate that blogs posts in those categories are likely to contain detailed refutals of pseudoscience.

Over a period of 9 months, we have collected over 14,000 blog posts and automatically organized them into over 3,200 *discussions*. A breakdowns of the *discussions* by size in blog posts are given in Table 1. Figure 3 shows a graphical distribution for *discussions* between 3 and 19 posts in size with power series trendlines for comparison. This is more data than we can annotate, so we focus on finding methods of detecting the most promising blog posts and gathering them into *discussions*. We begin by evaluating a method of filtering out invalid *discussions*.

First, we manually verify the acceptability of a small sample of *discussions*. We evaluated all *discussions* composed of at least 7 links, because they were few in number. For *discussions* with less than 7 links, we evaluated a random sample of 25. We used the following evaluation criteria:

- *discussions* must contain at least two different blogs
- the *source of interest* must be relevant to the linked posts

Our evaluation showed that *discussions* were often composed solely of links from a single blog, or that the *source of interest* was the top-level domain of the news source, rather than the link to the actual ısource of interest. We also found that a large portion of *discussions* judged unacceptable were for blog community activities, such as offline meetings; announcements of new blogs or upcoming linkfests for blogs on similar topics; or links to blog aggregator websites. While some of these discussions may be useful for extending our collection of scientific blogs, they are not useful as corpus data.

We developed a short blacklist of less than 25 blog category tags and urls that often appeared in invalid *discussions*. With this blacklist, we were able to filter out close to 400 discussions with a minimal loss of desirable *discussions*, as shown in Table 1. On re-evaluation, the precision of the filtered *discussions* had increased from less than a third to 47 percent.

6 Conclusion and Future Work

While our corpus is off to a promising start, there is still much work to do. Now that we have collected a sufficient number of interesting *discussions*, we will focus on annotating *statements of focus*. This entails evaluating the coverage and accuracy of category tag based retrieval. Automatic detection of logical relations between statements would facilitate the construction of our corpus. Mishne (2007) observed that

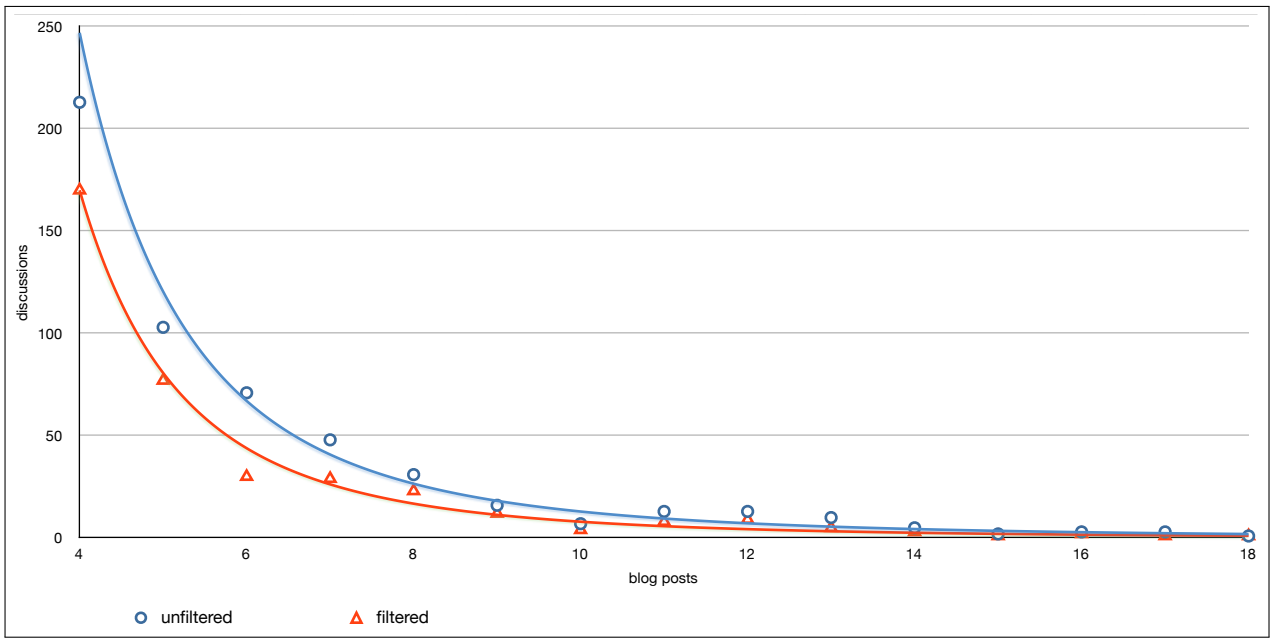

Figure 3: A comparison of unfiltered and filtered *discussion* sizes for $4 <= posts <= 18$

temporal cues and comment information were both useful in improving blog search. Using that information together with contextual cues like the proximity of links in text may help to reduce the search space. Finally, it will also be necessary to correctly attribute opinions to their authors. The work by Siddharthan and Teufel (2007) on scientific attribution may provide some valuable insights in this area.

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 provided invaluable feedback. Finally, we thank all of the bloggers, especially Ben Goldacre of badscience.net, for the data.

References

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

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

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

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

Gilad Mishne. 2007. Using blog properties to improve retrieval. In *Proceedings of ICWSM 2007*.

Koji Murakami, Eric Nichols, Suguru Matsuyoshi, Asuka Sumida, Shouko Masuda, Kentaro Inui, and Yuji Matsumoto. 2009. Statement map: assisting information crediblity analysis by visualizing arguments. In *WICOW '09: Proceedings of the 3rd workshop on Information credibility on the web*, pages 43–50. ACM, New York, NY, USA.

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

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

Advaith Siddharthan and Simone Teufel. 2007. Whose idea was this, and why does it matter? attributing scientific work to citations. In Candace L. Sidner, Tanja Schultz, Matthew Stone, and ChengXiang Zhai, editors, *HLT-NAACL*, pages 316–323. The Association for Computational Linguistics.

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

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. 1998. Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, 351(9103).

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