A Corpus Study for Identifying Evidence on Microblogs

Paul Reisert[1] **Junta Mizuno**[2] **Miwa Kanno**[1] **Naoaki Okazaki**[1,3] **Kentaro Inui**[1]

[1] Gradute School of Information Sciences, Tohoku University / Miyagi, Japan

[2] Resilient ICT Research Center, NICT / Miyagi, Japan [3] Japan Science and Technology Agency (JST) / Tokyo, Japan

preisert@ecei.tohoku.ac.jp junta-m@nict.go.jp {meihe, okazaki, inui}@ecei.tohoku.ac.jp

Abstract

Microblogs are a popular way for users to communicate and have recently caught the attention of researchers in the natural language processing (NLP) field. However, regardless of their rising popularity, little attention has been given towards determining the properties of discourse relations for the rapid, large-scale microblog data. Therefore, given their importance for various NLP tasks, we begin a study of discourse relations on microblogs by focusing on evidence relations. As no annotated corpora for evidence relations on microblogs exist, we conduct a corpus study to identify such relations on Twitter, a popular microblogging service. We create annotation guidelines, conduct a large-scale annotation phase, and develop a corpus of annotated evidence relations. Finally, we report our observations, annotation difficulties, and data statistics.

1 Introduction

Microblogs have become a popular method for users to express their ideas and communicate with other users. Twitter[1], a popular microblogging service, has recently been the attraction of many natural language processing (NLP) tasks ranging from flu epidemic detection (Aramaki et al., 2011) to gender inference for its users (Ciot et al., 2013). While various tasks are available, despite its daily, rapid large-scale data, evidence relation studies have yet to be explored using Twitter data. Previous research exists for determining the credibility of information on Twitter (Castillo et al., 2011); however, the focus of this work is to determine and annotate evidence relations on microblogs.

Our primary motivation behind focusing on evidence relations includes the possibility of discovering support for a claim which can support the debunking of false information. During the March 2011 Great East Japan Earthquake and Tsunami disaster, victims turned to the Internet in order to obtain information on current conditions, such as family member whereabouts, refuge center information, and general information (Sakaki et al., 2011). However, false information, such as the popular *Cosmo Oil explosion causing toxic rain*, interfered with those looking to find correct information on the status of the disaster areas (Okazaki et al., 2013). This is a scenario in which identification of potentially false information is necessary in order to provide accurate information to victims and others relying on and trusting in the Internet. Therefore, as a start to find support for counterclaims for false information such as the Cosmo Oil explosion, we focus on dialogue between two individuals: a *topic starter*, or a post with no parent; and a *respondent* who provides either an agreeing or disagreeing claim and support for their claim. An example is provided in Figure 1.

We note that our task can appear similar to the field of Why-QA (Verberne, 2006; Oh et al., 2013; Mrozinski et al., 2008), which attempts to discover the answer for *Why* questions. Given our task of discovering agreeing or conflicting claims, and finding specific reasoning to support the claim, we end up with a *Why* question similar to *Why is it true/not true that X*, where *X* is the contents of the claim found in the parent post. However, we consider source mentions or hyperlinks, which can either stand alone or be contained in a statement, question, or request, as a way to answer the above question.

To the best of our knowledge, no corpora for evidence relations on microblogs currently exists. In terms of argumentation corpora, the Araucaria Argumentation Corpus [2] exists which utilizes various argumentation schemes (Walton, 1996; Katzav and Reed, 2004; Pollock, 1995). In this work, we

[1] https://twitter.com

[2] http://araucaria.computing.dundee.ac.uk/doku.php

This work is licenced under a Creative Commons Attribution 4.0 International License. Page numbers and proceedings footer are added by the organizers. License details: http://creativecommons.org/licenses/by/4.0/

LAW VIII - The 8th Linguistic Annotation Workshop, pages 70–74,
Dublin, Ireland, August 23-24 2014.

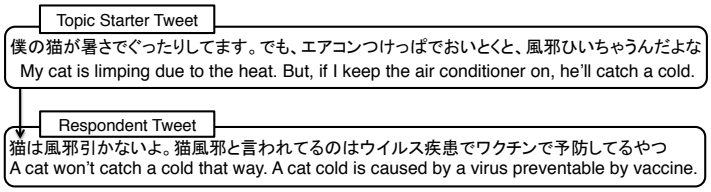

Figure 1: topic starter post and respondent post on the microbloging service Twitter.

manually annotate evidence relation claim and support. We conduct a corpus study that uses both current data and March 2011 data from Twitter, manually observing its structure and evidence, and devising guidelines based on our findings. We utilize these guidelines for conducting a large-scale annotation stage and develop a corpus with our results. We present our findings, challenges in annotation, and also the result statistics in the later sections. The corpora and annotation guidelines are currently available at: http://www.cl.ecei.tohoku.ac.jp/index.php?Open%20Resources%2FEvidence%20Relation%20Corpusw

2 Annotation Method

In this section, we describe evidence relation structure, target data, and our annotation method outline.

Evidence relations, defined by Mann and Thompson (1988) consist of a **claim**, or something that an author wishes for a reader to believe, and **support**, or something that increases the believability of the claim, and it can be understood by the following: *The program as published for calendar year 1980 really works. In only a few minutes, I entered all the figures from my 1980 tax return and got a result which agreed with my hand calculations to the penny.*, where the latter is support to the former claim.

With this in mind, we aim to explore what type of claims and support units exist on microblogs. Our microblog choice is Twitter, where users post *tweets* containing up to 140 characters. Tweets may then be replied to by other users. Each pair in our corpus consists of, what we refer to as, a *topic starter's* tweet and all of its direct reply tweets, or *respondent's* tweets. The topic starter's tweet is a top-level tweet not in response to another tweet, and the respondent tweet consists of a tweet directly in reply to the topic starter's tweet. We then discover respondent claims that agree or disagree with the topic starter. In addition, we target only pairs which contain an evidence relation.

The outline for annotation is as follows: 1) Given two tweets (topic starter and respondent), detect relation at agreeing or disagreeing level 2) Mark the claim and support in the respondent tweet

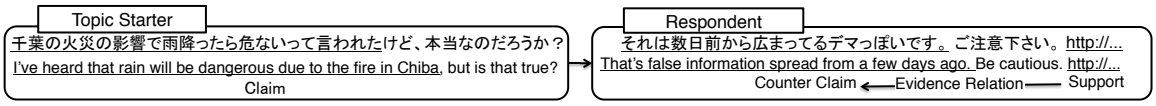

Figure 2: Evidence relation within a response.

We target this scenario because: 1) we assume that a topic will be presented in a topic starter's tweet and a respondent's direct reply will be responding to the topic content, and 2) it is possible to lose important information, such as topic keywords, for a reply tweet not in reply to a topic starter tweet.

Using this outline, we utilized Japanese Twitter data from the 2011 Great Eastern Tohoku Earthquake, specifically hottolink[3] data, and created a list of guidelines to use for our large-scale annotation in the next section by manually observing roughly 6,000 tweet pairs.

3 Large-scale Annotation

In order to obtain and observe more evidence relations, we conducted a large-scale annotation stage. We discuss the data and our statistics and observations.

3.1 Data

Using the guidelines in the previous section, we composed a list of 56,033 filtered tweets from various time periods, shown in the table below.

[3]http://www.hottolink.co.jp/press/936

Table 1: Data for large-scale annotation phase (A = Agree, D = Disagree, P = Partly A/D, O = Other)

#	Set	Pairs	Evidence	A	D	P	O
1	3-11 False Rumor topic starter Data	5753	1029	177	637	74	141
2	Togetter Controversial Category Data	2410	283	164	105	12	2
3	Togetter Negative Tag Data	1233	129	51	71	7	0
4	Twitter Random Controversial Topic Filtered Data	6918	277	168	94	14	1
5	3-11 Random Data	13064	381	241	115	21	4
6	3-11 Negative respondent Keyword Data	26655	1543	836	521	126	60
	Total	**56033**	**3642**	**1637**	**1543**	**254**	**208**

For Sets 1, 5, 6, we utilize the hottolink corpus mentioned briefly in the previous section. Set 1 consists of filtered pairs containing a well-known rumored topic from a list of 10 topics, such as Cosmo Oil Toxic Rain and Drinking Isodine for Radiation Prevention, and also contained a negative keyword in the respondent's tweet. We also included all other direct replies for the topic starter's tweet. Similarly, Set 5 contains random data from the hottolink corpus, unfiltered, and Set 6 contains pairs filtered via a negative keyword in the reply only.

Set 2 consists of crawled data from Togetter[4]. Togetter offers a summarization of popular, and potentially controversial, tweets for various categories, such as news, society, and sports. We first crawled all popular categories around January 2014 and obtained unique tweet IDs. We then used the Twitter API[5] to extract the tweet information from its ID in order to determine if it was a direct respondent tweet. If so, we obtained its topic starter tweet and thus created our pairs.

For Set 3, we appended negative keywords to the Togetter hyperlink (e.g. *http://togetter.com/t/デマ)* in order to obtain tweets that had been tagged with a negative keyword. We then used the same procedure as Set 2 in order to obtain topic starter and respondent tweet pairs.

Finally, Set 4 consists of 6,918 tweet pairs randomly selected from a collection of tweet pairs from Togetter, where each topic starter tweet is filtered by a topic from a list of around 300 controversial topics.

3.2 Statistics and Observations

In this section, we summarize the results of the annotated large-scale corpus by first providing information on the discovered evidence relations. Of 56,033 pairs, 3,642, or roughly 6.5%, were labeled as containing an evidence relation. Shown in Table 1 are the specific amount of evidence relations found in each set, along with the exact amount of claims that either agree, disagree, partly agree and disagree, and other. Also shown in Table 1 is the number of agreeing, disagreeing, partly agreeing/disagreeing, and other statistics for pairs labeled as evidence for each set.

3.2.1 Type Distribution

Since an important goal of this paper was to determine *what types* of claims and support we would discover, we classified random annotated tweet pairs by claim type and support type.

attitude Claim contains only reply user attitude (e.g., "I agree with you" or "It's false information")

request Claim requests some action (e.g., "Please delete and correct your tweet immediately")

question Claim is a question regarding the original tweet (e.g., "Why do you think so?" or only "?")

statement Claim is an opinion of a reply user (e.g., "Radiation cannot be reduced by a normal filter.")

Table 2: Type distribution results

Claim Type	Support Type	Disaster	General
attitude	causality	36	27
	elaboration	45	40
	source	35	7
request	causality	13	9
	elaboration	4	8
	source	15	0
statement	causality	21	22
	elaboration	49	31
	source	12	7
question	causality	1	2
	elaboration	3	9
	source	2	0
summary		**236**	**162**

[4]http://togetter.com

[5]https://dev.twitter.com/docs/api

Each of the three types of support (below) are in square brackets.
causality Support is a reason of a claim (e.g., "Isodine is no good because [it will ruin your health]")
elaboration Support is not a reason of a claim, but an elaboration (e.g., "topic starter: I definitely do not ride side by side with a car when I'm on my bicycle. respondent: Me too. [I do not ride side by side even when I ride a motorbike]")
source Support contains source information of the claim, such as hyperlink and name of the media (e.g., "Please read this web site [URL]" or "I saw it on the TV")

From Table 2, we found many source samples during disaster times but not for non-disaster periods. For our second finding, we discovered that attitude and statement were tweeted with support, while request and question were not. This indicates that people require some action without any support. For our third finding, we found that there were many replies which contain a statement and its support, while Twitter allows only 140 characters. This indicates many informative support segments on Twitter.

3.2.2 Annotation Issues

Below we enumerate issues that were encountered during our annotation process.

Reliability For determining annotation reliability, we had 10% of random samples from Set 1 annotated by another annotator and found that the inter-annotator agreement Cohen's kappa value was only .476. Both annotators marked 45 of the same pairs as evidence. Annotator A marked 60 other pairs as evidence, while Annotator B marked 15 other pairs as evidence. We believe this statistic is because tweets with evidence were infrequent and that many examples contained implicit relations, opposed to containing a discourse marker. From Annotator A's results, we found that only 9 examples contained an explicit discourse marker and 96 did not. Prasad et al. (2008) has already recognized that it is difficult to annotate relations when no discourse marker is present. We plan to automatically annotate evidence relations via machine learning and provide a probability that a pair is evidence to help manual annotation.

Multiple Claims With Twitter's character constraints, we expected to discover only one claim per reply with multiple support segments. However, we found that a few of our annotated segments contained multiple claim and multiple support segments.

Range Annotation range was a problem we discovered after observing our annotated data. Although such annotated cases were small (only 2 respondent tweets), most likely due to annotators avoiding such annotations, we still believe this type of annotation is important for future work. The example below was labeled as *unsure*:
{コスモ石油の件は}$_{\text{CLAIM}}${本社HPで}$_{\text{SUPPORT}}${デマだと公表されています。}$_{\text{CLAIM}}$ ({The Cosmo Oil case,}$_{\text{CLAIM}}$ {on the official HP,}$_{\text{SUPPORT}}$ {is publicly announced false.}$_{\text{CLAIM}}$)

4 Conclusion and Future Work

As no corpora exists for evidence relations on microblogs, we conducted a corpus study using the popular microblogging service, Twitter. We created a list of guidelines for evidence relation annotation by observing roughly 6,000 tweet pairs from March 2011 Twitter data, or disaster-specific data. Next, we conducted a large-scale annotation stage, consisting of 56,033 tweets, and discovered 3,642 contained a type of evidence relation. Our annotated data set is available at: http://www.cl.ecei.tohoku.ac.jp/index.php?Open%20Resources%2FEvidence%20Relation%20Corpus

We manually observed that the presence of evidence relations do indeed exist on microblogs; however, their existence is rather infrequent. To address this sparsity issue for future annotation, we plan to increase the number of pairs containing an evidence relation per data set by constructing a model that can automatically annotate evidence relations and provide a probability that a pair contains an evidence relation. In this work, we did not analyze the quality of evidence we discovered. Therefore, we aim towards determining the factuality, or degree of certainty, for a given claim and support in order to determine the evidence relation's overall quality.

Acknowledgments

We would like to acknowledge MEXT (Ministry of Education, Culture, Sports, Science and Technology) for their generous financial support via the Research Student Scholarship. This study was partly supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant No. 23240018 and Japan Science and Technology Agency (JST). Furthermore, we would like to also thank Eric Nichols (Honda Research Institute Japan Co., Ltd.) for his discussions on the topic of evidence relations.

References

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1568–1576.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684.

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of Twitter users in non-English contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145.

Joel Katzav and Chris Reed. 2004. A classification system for arguments. *Technical Report*.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Joanna Mrozinski, Edward Whittaker, and Sadaoki Furui. 2008. Collecting a why-question corpus for development and evaluation of an automatic QA-system. In *Proceedings of ACL-08: HLT*, pages 443–451.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. 2013. Why-question answering using intra- and inter-sentential causal relations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1733–1743.

Naoaki Okazaki, Keita Nabeshima, Kento Watanabe, Junta Mizuno, and Kentaro Inui. 2013. Extracting and aggregating false information from microblogs. In *Proceedings of the Workshop on Language Processing and Crisis Information*, pages 36–43.

John L. Pollock. 1995. *Cognitive Carpentry: A Blueprint for How to Build a Person*. MIT Press.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Takeshi Sakaki, Fujio Toriumi, and Yutaka Matsuo. 2011. Tweet trend analysis in an emergency situation. In *Proceedings of the Special Workshop on Internet and Disasters*, SWID '11, pages 3:1–3:8.

Suzan Verberne. 2006. Developing an approach for why-question answering. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, EACL '06, pages 39–46.

Douglas N. Walton. 1996. *Argumentation Schemes for Presumptive Reasoning*. Psychology Press.