An Investigation of Evidence Relations within Social Media Conversations

Paul Reisert[†]　*Junta Mizuno*[‡]　*Miwa Kanno*[†]　*Naoaki Okazaki*[†§]　*Kentaro Inui*[†]

Tohoku University[†]　*National Institute of Information and Communications Technology (NICT)*[‡]

Japan Science and Technology Agency (JST)[§]

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

1 Introduction

In the field of discourse analysis, the importance of recognizing relations between segments can assist with understanding how text is structured. Standard approaches for recognizing implicit and explicit discourse relations include utilizing popular corpora such as the Penn Discourse Treebank [6], which includes relations for over 1 million Wall Street Journal corpus. However, for corpora such as the PDTB, we miss out on various properties of conversations such as personal requests, desires, suggestions, and the ability to directly agree or disagree with a claim made by another individual. Due to their ability to support the debunking of false information, we begin the investigation of discourse relations for social media conversations by focusing primarily on evidence relations. Overall, this paper reports our initial findings for evidence relations in social media conversations after observing two data sets and briefly discusses our future annotation methods.

Evidence relations, as defined by Mann et al. [9], are composed of a claim and its supporting segment(s). To further elaborate, the claim's role, authored by an individual, is to provide information in which an individual wishes for another to believe. The supporting segment's role is to increase the believability of the stated claim. Our primary motivation for focusing on evidence relations as a start includes their ability to provide reasoning as to why a specific claim is either true or false. Therefore, as we are focusing on the discovery of discourse relations at a conversational level, we focus on conversations between two individuals: a *topic starter* who makes a specific claim, and a *respondent* who provides either an agreeing or disagreeing claim and its supporting segments. In Figure 1, we show an example of this scenario by showing both a respondent's counter claim, which does not contain any specific topic keyword and is in response to a topic starter's original claim, and its supporting segment, in this case, a hyperlink as a source.

False information such as the one in Figure 1, which have the ability to cause major, unnecessary panic, surfaced during the 2011 Great Eastern Tohoku Earthquake and Tsunami Japan period. During this devastating disaster, several victims whom were affected, both directly and indirectly, turned towards the Internet in order to find information regarding location status, family whereabouts, and other information related to the disaster. When determining how to easily identify evidence relations within a respondent's text, we utilize data from the 2011 Great Eastern Tohoku Earthquake and Tsunami Japan period for our first stage and furthermore investigate evidence relations on current, non-disaster specific data at our second stage.

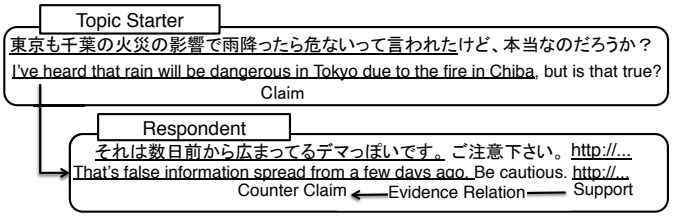

Figure 1: Evidence Relation within a Respondent's Tweet

At this time, as there exists no corpora for annotated evidence relations for social media conversations, we begin the investigation and development of a corpus composed of annotated evidence relations for replies which contain either an agreeing or disagreeing response claim and their respective support segment(s). We filter social media data consisting of pairs of conversations between a topic starter and their respective respondents. Finally, we report our annotation method, including encountered findings and challenges, and discuss our future involvement in discourse relations for conversations.

2 Related Work

The framework for annotating discourse relations for arguments consisting of agreeing and disagreeing claims can be thought of as a combination of previous researches.

Nichols et al. [3] provides a fundamental strategy for discovering evidence. Given a topic query such as *Milk is good for the body*, their system searched for sentences in the form of *Because milk* ***X****, it is* ***Y*** *for the body*, in which **X** provides reason, or evidence, as to why milk is **Y** for the body, where **Y** can either agree (e.g. *good, etc*) or disagree (e.g. *bad, etc*) with the original query. For our paper, we expand upon their work by investigating evidence relations which are not confined to only one sentence which contains an explicit contextual cue such as ので, から, ため, but also those which contain implicit evidence relations in adjacent or spanning texts. We take a similar approach where we consider a claim within a topic starter's tweet to act as the query and discover claims within a respondent's tweet which either agree or disagree with the topic starter's claim. Using such a direct conversational setting, we gain the ability to discover respondent claims which may or may not contain an original topic keyword.

In terms of creating a corpus for annotated discourse re-

lations outside of the Penn Discourse Tree Bank, Tonelli et al. [8] created a corpus for relations within a spoken conversational dialogue setting. However, for our purposes of determining evidence which may contain support in the form of hyperlinks, quotes, and other forms, we focus primarily on conversational dialogue on social networking.

Finally, the idea of recognizing agreement or disagreement arguments within replies on social media has also been researched by Misra et al. [1]. Similarly, for this work, we construct a list of keywords signifying disagreement for reply claims.

3 Experiment

For both stages in our experiment, we utilize the popular social networking website Twitter for our data set. Twitter users have the ability to create a post, hereby referred to as a *tweet*, which can then be replied to by other users. For this study, we hereby refer to such users as *topic-starters* and *respondents*, respectively. As mentioned in Section 1, a goal of creating a corpus specifically for evidence relations in microblogs is to discover such evidence relations which could assist in the debunking of false rumors. Assuming the topic starter creates a topic to which a respondent replies to, we focus on discovering evidence relations within a respondent's tweet that is composed of both a claim, which either agrees or disagrees with a topic starter's original claim, and the claim's support. In preparation of utilizing our corpus towards the detection of false information, we put an emphasis on discovering evidence relations composed of a disagreeing claim and its support.

3.1 Annotation Method

For our current annotation method, we signify the agreeing or disagreeing claim within a respondent's tweet data using a blue text color. Likewise, for our support, we label its range with a red text color. An example of this can be seen in Figure 2. For a list of the claims and support we discovered during both stages, please see Sections 4.1 and 4.2.

3.2 First Stage

For our initial data set, we utilized Twitter data collected around March 11th, 2011, the time period of the 2011 Great Eastern Tohoku Earthquake and Tsunami Japan. Our motivation for using such tweets originated from the heavy amount of now known false rumors which originated during this time period. After receiving the initial data during this period, in order to confine to our topic starter-respondent structure, we create *(topic starter,respondent)* pairs consisting of a topic starter's tweet and a respondent's tweet. As a topic starter's tweet may have multiple replies, one pair may share the same topic-starting tweet with another. Applying this filter results in nearly 449,000 unique pairs consisting of all specified top-level tweets and their replies.

Next, we created a simple list of disagreeing keywords in Japanese which have the meaning of either *false information* or *rumor*, such as デマ, 嘘, and ガセ. Our initial assumption was that we would be able to discover respondent claims with such a disagreeing keyword along with a supporting segment to complete our evidence relation in the form of the example provided in Section 1. We were able to quickly discover around 300 evidence relations within a respondent's tweet consisting of a claim with a disagreeing keyword and its supporting segment. For each of these 300 relations, we manually mark the agreeing or disagreeing claim in blue and mark its associated support in red, as shown in Figure 2.

In addition to our initial assumption of finding evidence relations consisting of a claim such as *That information is false* and either an associated hyperlink or supportive claim, however, we discovered various, unexpected ways of presenting third-party source mentions for support segments such as the use of commands and suggestions. This encourages us to focus existing methods for labeling such utterances, as mentioned in Section 5.2.

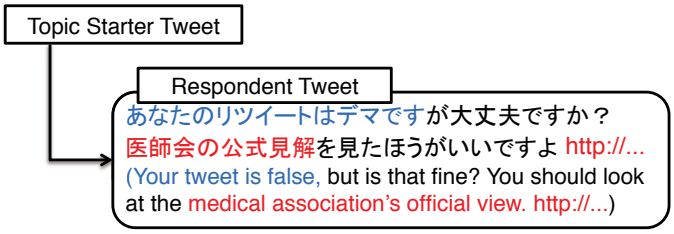

Figure 2: Annotated Evidence Relation for Unexpected Respondent tweets. Claims are labeled in blue and support is labeled in red.

In Figure 2, we discover that a suggestion is provided containing the entity for which the claim was most likely concluded from. Therefore, we annotate the argument within the suggestion as a supporting segment. In addition to suggestions, we also observed commands containing third party entity mentions. A comprehensive summary of the observations we have encountered thus far for both claims and support can be found in Section 4.1 and Section 4.2, respectively.

3.3 Second Stage

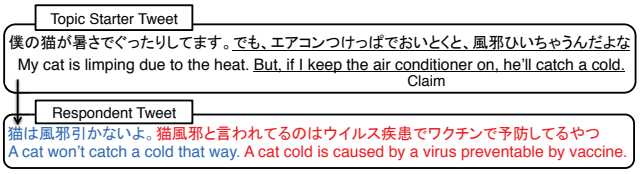

Figure 3: Evidence Relation discovered in Stage 2

Following up on our observations for data around the 3-11 disaster time period, we observed current data which is not specific to a major disaster. For this stage, we do not observe only replies with a negative keyword, but instead, filter a topic starter's tweet by a list composed of around 300 controversial topics. Our motivation for using such a filter arose during the first observation of the data and the difficulty in finding evidence relations due to the overwhelming amount of basic greetings between two individuals. Applying this filter of controversial topics resulted in roughly 12,500 pairs. We then choose 100 random samples to observe other ways of presenting agreeing and disagreeing claims which may or may not contain a negative keyword as in Section 3.2. Overall, we discovered 22 evidence relations for our random sampling stage.

Respondent Claim Examples
同意。/ 賛成。(I agree.)
そんなことないです！ (That's not true!)
RT しないで下さい！(Don't retweet that information!)
それほんと？ (Is that true?)
スマホじゃなくてスマフォだよ (It's not Sumafo, but Sumaho)
どうやら過去の話のようです。 (Apparently, that's past information)

Table 1: Observed Respondent claims

Respondent Tweets
それはデマです。 http://... That information is false. http://...
コスモ石油の HP によると、デマみたいです。 According to the Cosmo Oil company's website, that appears to be false.
デマです。医師会の公式見解見て！ That's false information. Look at the medical association's official opinion!
ここまでいわき市関係のツイートが多いんだからデマではないと思います。 Because there's several tweets regarding Iwaki-shi, I don't think that it's false information.
その情報はデマです！存在しない住所だそうです！ That information is false! It appears that address doesn't exist!

Table 2: Respondent Evidence Relations containing both Claim and Source

4 Investigated Findings

4.1 Claim Units

As seen from Table 4.1, we encounter various types of claims either agreeing or disagreeing with a topic starter. With phenomena such as questions and commands, we are making the underlying assumption that the reason for requesting someone to remove information or asking them if the information is actually correct is that they indirectly disagree with the topic starter's tweet, hence we consider this to be an implicit disagreeing claim.

4.2 Support Units

After a claim has been provided, we observed its support consists of mainly either a *source* or a *supportive claim*. We describe both below and provide examples along with our current annotation. For our support results, we note that our observed segments consists of a source mention, mainly in the form of a third-party entity to the respondent, a quote, and a hyperlink. We provide examples in Table 4.2 for our findings. For simplicity sake, we specify the claim as *That is false information.* for some examples. Originally, we expected to find mainly hyperlinks or claim and support signified by *because* contextual cues, similar to Nichols et al. [3]; however, as signified in the table, we discover the presentation of third-party entities via commands and other contextual cues such as *According to*. In order to properly label such findings, we consider the integration of speech act and argumentation scheme labeling for our evidence relations, as denoted in the next section.

4.3 Stage Comparison

When comparing both the first and second stage, a notable difference includes the frequency of evidence relations within the disaster data opposed to the non-disaster specific data. Therefore, prior to filtering the non-disaster specific data by controversial topic, the ability to discover evidence relations within the respondent's tweet proved to be a challenge. Even after filtering the non-disaster specific data by false information topics, we had difficulty in searching for samples using negative keywords similar to the preliminary stage, hence our decision to choose 100 random samples to annotate.

Furthermore, we discovered in our second stage that it was more difficult to find claims or support in the form of a command, or support in the form of a suggestion. When observing our preliminary stage data, we determine that most claims in the form of commands were provided in response to false information. This includes commands such as *Do not post false information* and *Do not retweet such information.* We equate this to the fact that there were less well-known rumors being spread during our second stage and thus less knowledge of external resources to link to a topic starter.

5 Discussion

During our annotation process, we encountered many difficulties which are discussed below.

5.1 Difficulties

Claim or Support only in respondent's text
Occasionally, we encountered cases where a support segment could also be interpreted as a claim if presented in a different context. To avoid confusion, we label only instances in which the claim and support are both present.

Label Range
In terms of range to label, we ignore labeling any speech disfluencies, emoticons, or other jargons used. Therefore, in the following example, we label only the underlined portion: いやいや、危ないって。笑.

Topic Starter Hyperlinks
When determining the relationship between the respondent's claim and the topic starter's claim, we occasionally encountered situations in which the respondent's claim was not in direct response to the topic starter's, but directly opposing a claim within the content of the topic starter's provided hyperlink. Future work will include extracting relevant information from hyperlinks in order to determine the original claim. The same will be applied for hyperlinks that are provided as support to a given claim. In other words, in the event a known rumor is posted by a topic starter and the respondent posts only a hyperlink which contains information which contradicts the topic starter's information, than we can make an underlying assumption that their lies an implicit claim within the respondent's tweet.

Ambiguous topic-starter claims
We ignored instances in which it was too difficult for us

to determine the claim within the topic starter's tweet. In doing so, we plan to compose a corpus containing pairs with both a clear topic starter claim and a clear evidence relation within the respondent's tweet.

Interchangeable relation

We encountered cases in which two segments could either have the structure Claim-Support or Support-Claim; in other words, the first segment could be considered the claim and the second could be considered support, or vise versa. An example includes the following: 有害物質は確認されていませんよ*?(The dangerous material hasn't been confirmed?)* and 悪質なデマと言われていますけど。*(It's being said that it's a dishonest false information.)*. For cases such as this, we note the possibility of interchangeable relation and choose the claim in which we feel is most likely to agree/disagree with the topic starter's claim and also be supported by the other claim.

Tweet length

In terms of tweet length, we observed that in the event a topic starter's tweet length was small, evidence relations were less likely to be found within a respondent's tweet. Through personal observation, several topic starter tweets with a small amount of characters mainly consisted of personal, first-person tweets such as greetings or personal conditions. Therefore, we found a challenge in discovering evidence relations with such a limited amount of characters in the reply. In future work, we plan to integrate filters into our data in order to make annotation easier.

5.2 Further Annotation

As a result of our findings and difficulties for evidence detection, we plan to integrate the following into our framework:

Speech Acts

Towards a more fine-grained classification of our data, we plan to annotate speech acts, or ways to classify utterances, for our evidence detection system. Work by Zhang et al. [7] has already been done for detecting speech acts on Twitter for topic summarization purposes. Likewise we plan to take a similar approach for our work. In labeling various speech acts with our data, we hope to add more structure into our system for identifying evidence. Currently, we have already started segmenting our tweets and detecting speech acts by utilizing the dependency graph querying language DGrep [4] to categorize commands, questions, suggestions and normal statements.

Argumentation Schemes

Considering the respondent's claim as a conclusion and its support as a premise, we have the ability to assess the quality of our evidence relations via integration of Argumentation Schemes which describe various types of arguments, in our case, types of evidence relations. Walton et al. [2] composes a list of around 60 schemes, such as *Argument from Expert Opinion* and *Argument from Position to Know*, which we can already utilize for relations discovered in Table 4.2.

6 Conclusion and Future Work

In this paper, we investigate evidence relations in social media conversations between a topic starter and the respondent. Given the importance of evidence relations, especially when debunking false information, we filter a collection of tweets by well-known false information topics. We annotate roughly 300 pairs for data around the 2011 Great Eastern Tohoku Earthquake and Tsunami Japan, each containing an evidence relation in a respondent's tweet consisting of a disagreeing claim and its support. We also randomly sample 100 instances of current data and discover 22 out of 100 evidence relations within a respondent's tweet.

In our future work, we focus on expanding our corpus for evidence relation detection using a supervised machine learning model. In addition, with the presence of commands, requests, and questions within a social media conversation, we plan to annotate using speech acts and argumentation schemes in order to create more structure for our work.

Finally, we plan to utilize the Internet Argument Corpus which consists of around 11,000 discussion threads in English on controversial topics such as Evolution, Abortion, and Gun Control, to name a few [5]. Such a corpus can assist in detecting a new variety of evidence relations which can then be applicable in our Twitter environment, especially when we begin detection of evidence in other languages such as English.

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

[1] Amita Misra and Marilyn Walker. Topic independent identification of agreement and disagreement in social media dialogue. In *Proceedings of SIGDIAL2013*, pages 41–50, 2013.

[2] Douglas Walton and Chris Reed and Fabrizio Macagno. *Argumentation Schemes*. 2008.

[3] Eric Nichols and Junta Mizuno and Yotaro Watanabe and Kentaro Inui. Towards evidence search. In *In the Proceedings of the Seventeenth Annual Meeting of the Association for Natural Language Processing.*, pages 880–883, 2011.

[4] Eric Nichols and Paul Reisert. Dgrep: A pattern-matching tool for dependency trees. In *Proceedings of the Twentieth Annual Meeting of the Association for Natural Language Processing. (to appear)*, 2014.

[5] Marilyn Walker and Jean Fox Tree and Pranav Anand and Rob Abbott and Joseph King. A corpus for research on deliberation and debate. In *Proceedings of LREC2012*, pages 812–817, 2012.

[6] Rashmi Prasad and Nikhil Dinesh and Alan Lee and Eleni Miltsakaki and Livio Robaldo and Aravind Joshi and Bonnie Webber. The penn discourse treebank 2.0. In *Proceedings of LREC2008*, pages 2961–2968, 2008.

[7] Renxian Zhang and Wenjie Li and Dehong Gao and Ouyang You. Automatic twitter topic summarization with speech acts. *IEEE Transactions on Audio, Speech & Language Processing*, 21(3):649–658, 2013.

[8] Sara Tonelli and Giuseppe Riccardi and Rashmi Prasad and Aravind K. Joshi. Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of LREC2010*, pages 2084–2090, 2010.

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