What Kind of #Conversation is Twitter? Mining #Psycholinguistic Cues for Emergency Coordination

Hemant Purohit[a,c], Andrew Hampton[b,c], Valerie L. Shalin[b,c], Amit P. Sheth[a], John Flach[b], Shreyansh Bhatt[a]

Ohio Center of Excellence In Knowledge-enabled Computing (Kno.e.sis), USA

[a]*Department of Computer Science & Engineering, Wright State University, USA*
[b]*Department of Psychology, Wright State University, USA*

[c]*Corresponding Authors:*
(Hemant Purohit, Andrew Hampton, Valerie L. Shalin)
{hemant,andrew,valerie}@knoesis.org
+1-937-775-5217 (work)
377 Joshi Research Center,
3640 Colonel Glenn Highway,
Dayton, OH, USA - 45435

1. Introduction

Social media (e.g., Twitter) promise to facilitate coordinated response to disasters such as the Haiti (2010) and Japan (2011) earthquakes, or Hurricane Irene (2011). Citizens can participate in coordinated emergency response in at least two ways. First, they may serve a passive role, reporting on the changing state of affairs, like citizen sensors [1]. Second, they may mobilize and direct their own resources, in hopes of supplementing or improving the resources of the formal emergency response system. However, in the absence of an established link to the formal emergency response system, a relevant citizen response is difficult to identify and incorporate into the formal system's planning process.

The exploitation of social media for emergency response requires an approach to filtering the large volume of message traffic in near real time. We address this requirement with the development of a model for detecting coordination [2], defined as the harmonious functioning of parts for effective results[2]. Goodwin & Heritage [3] and others claim that processes of social interaction lead to shared meaning, mutual understanding, and the coordination of human conduct. Thus, the present paper models coordination in language as a means to filter useful segments of social media traffic.

Certainly, anecdotal evidence supports a relationship between general message traffic and public action. For example, in 2009 unusually substantial Twitter traffic preceded the resignation of the controversial political figure Van Jones. In the US, Twitter traffic following Joe Wilson's insult to the US president during a State of the Union address (2009) correlated with an increase in donations to Wilson's political opponent. And considerable message traffic preceded a conservative talk show host's 2010 rally in Washington DC.

The present research aims to exploit domain-independent linguistic features of coordination in conversation [2, 3, 4] as the first step in narrowing the candidate set of messages for domain-dependent and computationally intensive analysis of coordination content [5]. In the remaining subsections of this introduction we describe the platform properties of the Twitter medium before turning to the methodological and theoretical problem of identifying a corpus of conversation. We conclude the introduction with a summary

[2]Dictionary definition for coordination: http://www.merriam-webster.com/dictionary/coordination

of related work in linguistics that grounds our hypotheses for detecting and modeling the coordination in Twitter conversation.

1.1. Twitter Social Media Platform

Twitter is a microblogging service that provides a social network structure and a medium for information flow, where users post updates and subscribe to (referred to as 'following') other users to receive updates (microblogs).

- ***Tweet***: A short message/post/status/microblog from a user on Twitter, spanning a maximum of 140 characters. Tweets include updates about user activities, sharing useful information, forwarding other users' statuses, conversing with others, etc. The 140 character limit influences expression.

- ***Hashtag***: Denoted by a word with preceding '#' symbol (e.g., #Japan-Earthquake), the hashtag is a platform convention for user-defined topics, invented to identify a topic of communication using minimal characters. It is also an important tool to provide a basis to group conversations by topic.

- ***Short URLs***: Tweets may contain links to web-pages, blogs, etc. To avoid lengthy URLs, Twitter users employ condensed versions of those URLs, shortened by external services (e.g., http://bit.ly/IyBgIO).

- ***Reply***: Reply is a platform-provided feature to communicate with tweet author by clicking on Twitter's 'Reply' button in response to a tweet. For example, user *hemant_pt* tweets *"today's discussion on linguistic coordination was just brilliant!"*, while user U uses the built-in Reply button to indicate *"@hemant_pt I was excited too about today's discussion"*. The Reply syntax automatically inserts the originator's user name.

- ***Retweet***: Retweet forwards a tweet from users to their followers, similar to e-mail forwarding. In so doing, the writer credits the source using the built-in 'Retweet' feature resulting in *'RT @USER_NAME'*. For example, *"RT @hemant_pt: it is not enough to depend on platform provided indicators for conversations #coordination #psycholinguistic"*. Here a new user retweeted a tweet from *hemant_pt*.

- ***Mention***: Mention acknowledges a user with the symbolic '@' sign, but without using the 'Reply' platform feature. For example, *"Thanks @hemant_pt, we hope to see you in next year's conference too for further discussion on #coordination"*.

1.2. Potential Conversation in Twitter

Despite the constraint of 140 characters, a growing research base supports the claim that some Twitter posts constitute a conversation. Danescu-Niculescu-Mizil et al. [6] showed that Twitter exchanges reflect the psycholinguistic concept of communication accommodation, where participants in conversations tend to converge to one another's communicative behavior; they coordinate using a variety of dimensions including choice of words, syntax, utterance length, pitch and gestures. Gouws et al. [7] analyzed the effects of user demographics, context and modes of information sources (web vs. mobile clients) on lexical usage in the Twitter medium. Their study showed a convergence in the adoption of unusual vocabulary terms, a potential culturally relevant behavior. Further, the authors found that contextual indicators including geographic location account for lexical variance from the standard English language. This phenomenon of lexical transformation supports our conceptualization of some Twitter exchange as a kind of conversation.

To identify the diagnostic features for a classification model of conversation, we require positive instances of messages that likely reflect conversation. Most of the relevant work on Twitter focused on a corpus incorporating the 'Reply' feature. As discussed below, we examine Reply, Retweet, and Mention separately and together. Figures 1,2, and 3 below provide examples for each function, illustrating both positive and negative examples of conversation. The negative examples support our claim that platform features alone do not assure conversation.

Focusing exclusively on postings with 'Reply', Ritter et al. [8] analyzed content dependent and language dependent vocabulary in a computationally intensive model of structuring conversation element sequences and disentangling dialogues on Twitter. While their distinction between content and language dependent vocabulary is similar to our distinction between domain dependent and independent analyses, we advocate reliance on the domain independent cues as a computationally inexpensive way of screening the Twitter corpus prior to domain dependent analysis.

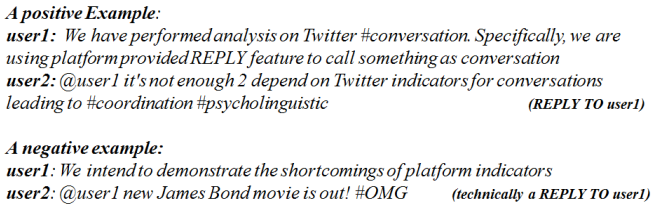

Figure 1: ***'Reply' feature based conversation***

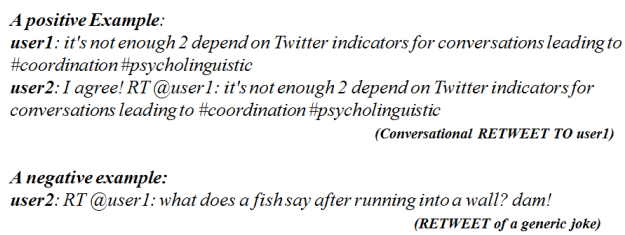

Figure 2: ***'Retweet' feature based conversation***

While Twitter's Retweet practice seems like a means simply to disseminate information, it also potentially functions as a type of conversation where multiple recipients comprise listeners for the original author. Three observations support our claim for conceptualizing retweets as conversation. Boyd et al. [9] noted three forms of potential conversational properties of Twitter replies, retweets, and messages that included hashtags. Their extensive study of Retweet behavior indicated that this type of conversation is distributed across a non-cohesive network in which the recipients of each message change depending on the sender, often missing conversational structures. Further, as shown in Figure 2, users in the Retweet diffusion chain sometimes prefix their opinion to the forwarded message. This represents a localized conversation between the followee and her immediate followers based on the action of the follower. Finally, the action of retweeting bears some similarity to backchanneling in verbal exchange, in which the listener confirms continued attention and comprehension with action [10].

Similarly, Mention-based tweeting can form a conversation where one user addresses another user rather than simply referring to him (e.g., *"@user1 it's not enough 2 depend on Twitter indicators for conversations leading to coordination psycholinguistic"*) without using the Reply feature of Twitter. Honeycutt et al. [11] focused on the coherence of exchanges involving the

A positive Example:
user1: *We have performed analysis on Twitter #conversation. Specifically, we are using platform provided REPLY feature to call something as conversation*
user2: *I kind of agree with @user1 that it's not enough 2 depend on Twitter indicators for conversations #coordination #psycholinguistic*
(MENTION OF user1, but not using REPLY feature)

A negative example:
user: *felt like @MichaelJordan at last basketball game!*

Figure 3: ***'Mention' feature based conversation***

'@' sign. They observed a surprising degree of conversationality using lexical patterns particularly when using '@' as a marker of addressivity.

We identify two implications of our focus on platform-driven subsets as linked to conversation. First, each subset is more likely to exhibit coordination features relative to the remainder tweets. Everything else is less likely to be a conversation. Therefore we should see relatively more coordination indicators in the platform-driven subsets than the remainder. Second, the prevalence of coordination language may decline with the type of platform indicators. 'Reply' should have the most coordination language, as it is the most explicit indicator of conversational intent.

User@user1:
*@**user2** goodnight I am about out of here also :)*

User@user2:
*@**user1** I adore charlie sheen everyday He is Kewl I pray he puts on a Telethon for Japan & HELPS them out!charlie sheen follows me*

Figure 4: ***Platform indicators do not ensure coherent conversation.***

We specifically deny the stronger claim that platform indicators alone determine coordination. For example, using the Reply feature may simply reflect a convenient way to distribute a message. Blind retweeting to a broader network need not reflect concurrence or endorsement consistent with a kind of conversation. And including the name of another Twitter user in a message need not invoke a response.

Just as platform indicators do not guarantee conversation, the absence of platform indicators does not guarantee the absence of conversation. We are particularly concerned with posts that contain hidden conversation, without platform indicators, e.g., *"what's going on with that city? How many people escaped? Please tell me!"* by a user *@JT800*.

We do claim that platform indicators, relative to the remaining subset of tweets, are more likely to reflect the properties of conversational coordination. By identifying a reliable set of theoretically based indicators of conversational coordination, we obtain a bootstrapped model for classifying any message as reflecting linguistic coordination and we can potentially identify the features that reflect coordinated effort in any individual posting, independent of platform indicators. A final justification of the search for coordination indicators independent of platform indicators is that compliance with convention often fails under stressful circumstances of disaster. We suspect that recommendations for coordination that hinge on imposing low-level communication templates on informal social media communities will fail under stressful and non-standard circumstances [12]. Therefore, the ability to mine conversation provides a robust alternative to brittle user compliance and we assert a clear need to first understand the whole communication landscape of Twitter and then perform systematic study for conversations leading to coordination.

1.3. Candidate Coordination Features in Twitter Conversation

Before identifying the specific features we will examine in Twitter posts, we note two methodological concerns. First, the traditional linguistics literature focuses on positive cases of conversation. This strategy does not identify whether the properties of conversation are diagnostic because it does not examine non-conversation for the reduced prevalence of these features. Moreover, the oversight betrays the lack of an operational definition for non-conversation. Second, the space constraint in Twitter potentially alters coordination practices. For example, one might imagine a reduction in dialogue management relative to face-to-face conversation. Thus, a successful model of diagnostic conversational features in Twitter supports the claim that fundamental patterns transcend communication media.

Social scientists have been investigating the role of linguistic patterns in coordination for decades. Notably, Clark et al. [2] showed that users follow certain linguistic patterns while communicating for coordination. Properties of an exchange, including opening and closing phrases, anaphora, and deixis, reveal the existence of coordination. Goodwin & Heritage [3] analyzed various facets of the conversation landscape. Similarly, Mark [4] showed conventions followed in the collaborative environment.

Table 1 presents the features we examine in the tweets. The examination of articles (h1 and h2) follows Chafe [13], who asserted that "the" assumes a previously established topic. A set of dialogue management items

Heuristic	Heuristic Description
h1	Determiners (the)
h2	Determiners (a, an)
h3	Subject Pronouns (she, he, we, they)
h4	Mixed Subject/ Object pronouns but centered on individual (my, I, me)
h5	Relative Pronouns (that, this, these, those)
h6	Possessive Pronouns (mine, yours, his, hers, ours, theirs)
h7	Relative Pronouns (who, what, which, whom, whose)
h8	Intensive/ Reflexive Pronouns (myself, yourself, himself, herself, itself, ourselves, themselves, yourselves)
h9	Dialogue management indicators (thanks, yes, ok, sorry, hi, hello, bye, anyway, how about, so, what do you mean, please, {could, would, should, can, will} followed by pronoun)
h10	Word Counts
h11	Hedge Words (kinda,sorta)
h12	Ambiguous Pronoun (you)
h13	Ambiguous Pronoun (it)
h14	Object Pronouns (us,them,him,her)

Table 1: ***Candidate Heuristic Features for Identifying Conversation***

(h9) capture the typical conversational openings and closings and requests for clarification. The preponderance of hypotheses related to pronouns captures anaphora (reference to a previous exchange) and deixis (grounding in a physical setting). We anticipate more of these words when participants share common ground. We identified separate hypotheses by grammatical part of speech and person. First and second person pronouns should appear in a coordinated exchange. However, first person pronouns also appear in the personal status reports that pervade Twitter, and may therefore not diagnose conversation. Other pronoun forms (possessives, relatives, reflexives) could obtain grounding within the post itself, rather than a previous post.

We now identify our hypotheses and research questions:

- ***H1***: Linguistic coordination indicators distinguish replies, retweets and mentions from other tweets.
- ***R1***: The degree of success in separating replies, retweets and mentions from non-conversation reflects the degree to which these platform indicators function as conversation.
- ***R2***: The degree of success in separating replies, retweets and mentions

from non-conversation depends on the extent to which the surrounding context promotes coordination.

- ***R3***: The diagnostic features of conversation transcend platform indicator.

- ***H2***: Coordination indicators correlate with information density.

2. Method

We first describe data collection for the proposed study, followed by our approach for testing the hypotheses mentioned above, using fine grained conversation categorization, conversation features, and modeling.

2.1. Data Collection

The Twitter *Streaming* API provides real-time tweet collection. Alternatively, the Twitter *Search* API provides keyword based search query, returning the 1500 most recent tweets in one response and excluding tweets from users who opt for privacy. The query provides tweet text and metadata, such as timestamp, location, and author information. The API access rate depends on the role of service authorization, with 350 requests/hour for non-whitelisted access (no special service authorization) or 20000 requests/hour for whitelisted access (special service authorization).

To study tweet events, we created a crawler in the Twitris system [14][15][16] that queried the Twitter Search API every 30 seconds for event-related keywords (e.g., "hurricane" for the event "Hurricane Irene storm 2011") for the duration of the event period. We initiated the keyword set with seed keywords and hashtags. We then expanded the initial set by extracting its top key phrases and adding them to the crawler. We maintained human oversight for seed keyword selection to maintain relevance to the event context. One can also utilize a sophisticated computation method, such as Continuous Semantics framework from our prior work [17], to model the evolving knowledge and to find highly relevant keywords for an event, but that is not the focus in this paper. We collected tweets for six different events. To reflect language behavior in response to a disaster, we examined the Haitian and Japanese earthquakes and hurricane Irene. For the purposes of comparison with non-disaster events, we examined the debt ceiling debate, the Skype Microsoft deal, and the Glenn Beck rally (described later in the Table 2).

2.2. Algorithm to construct data corpuses for conversation types

As argued above, Twitter provides three mechanisms 'Reply', 'Retweet' and 'Mention' that potentially enable conversation. We constructed our separate corpuses as follows:

1. Collect the event-centric tweet corpus for an event, denoted as A.
2. Extract all tweets that were part of Twitter's Reply-based conversation feature. Append remaining tweets in those conversation threads which were not present in our corpus in step 1 and call the extracted (and appended) set the Reply-based conversation set, denoted as RP.
3. From remaining corpus in step 2, now extract tweets with Retweet (RT) usage and call this set the RT-based conversation set.
4. From remaining corpus in step 3, extract tweets where "@" is used and call it the Mention-based conversation set, denoted as M.

The remaining corpus from step 4 {A- RP-RT-M}, is the *Non-conversation set*, denoted by NC.

2.3. Classification model and Feature Ranking

- ***Data Sampling***

  We examined the presence or absence of conversation indicators across a chunk of three tweets, the average size of a Reply based sequence on Twitter at the time of our crawls. We created balanced equal sized data samples for both the positive conversation sample and the negative conversation sample, where we consider positive samples of conversation as samples belonging to any of the RP, RT, or M conversation type corpuses. Negative conversation samples consist of those belonging to NC corpus. Thus, our question is whether we can detect language patterns in these three-tweet chunks that distinguish them from three-tweet chunks in the NC corpus.

- ***Classification Model***

  We conducted classification modeling to establish the degree of conversationality shown by a potential conversation text sample, each one characterized with a value for each of our candidate linguistic features, including variants of the feature words in the social media space to compensate (e.g., 'you' as 'u'). A classifier is a mathematical function that combines feature values to judge the class membership of a data

sample, in this case, as a conversation or not. We used Decision Tree classifiers [18] for our analysis, which provides an interpretable classification tree of nodes as linguistic features (from root node to leaf node) and the leaf node as the conversation class (decision).

We used machine learning techniques to develop the conversation classifier. We therefore created training sets (to learn from the data) and testing sets (to test on the new data and make a more robust classifier) of the data samples. We created balanced (equal number of positive and negative samples) training sets and test sets using data samples corresponding to each of the conversation type class (RP, RT or M) and non-conversation class (NC). We used the Weka [19] data mining tool to perform modeling and experimentation.

- ***Feature Ranking***

  We ranked the linguistic features which reflect significant alignment with the conversation class suggested by any of RP, RT or M corpuses as compared to non conversation class (NC). We used a chi^2 test for feature ranking. In a separate analysis that examined only correctly classified tweet segments (hits and correct rejections) we checked for the direction of the relationship between feature and class.

- ***Evaluation***

  Using ten-fold cross validation to understand unbiased accuracy of conversation classifiers and feature ranking models, the system generates partitions of samples of data into complementary subsets, performing the analysis on nine subsets (the training set), and validating the analysis on the remaining one subset (the testing set). This process will repeat the procedure ten times and then generate average accuracy statistics for classification such as the area under the Receiver Operating Characteristic (ROC) curve.

3. Experiments and Results

3.1. Data Summary

We collected a set of six events for analyzing conversation characteristics where the set represents diversity of study by spanning different time periods of different length and covering varied social significance. We set the end of

the event period as when the volume of information flow dropped steeply. Table 2 shows a summary of the corpus:

EVENTS	DURATION	DAYS	TWEETS	AUTHORS	RP	RT	M	NC
Japan Earthquake 2011	2011-03-11 – 2011-03-30	20	609853	26916	60223	240090	41234	268306
Haiti Earthquake 2010	2010-01-13 – 2010-03-10	57	583747	26460	56896	200955	54171	271725
Hurricane Irene Storm 2011	2011-08-26 – 2011-09-10	16	181871	8335	14345	72441	12146	82939
Debt Ceiling Debate 2011	2011-07-25 – 2011-07-30	6	75788	4068	8294	29761	6490	31243
Skype Microsoft Deal 2011	2011-05-10 – 2011-05-15	6	19331	959	1158	5709	3640	8824
Glenn Beck Rally 2010	2010-08-30 – 2011-09-05	7	3848	386	594	1173	372	1709
TOTAL			1474438	67124	141510	550129	118053	664746

Table 2: ***Statistics about the event-centric data sets and for various conversational corpuses- Reply (RP), Retweet (RT), Mention (M) and Non-conversation (NC)***

The first three events in Table 2 show the nature of a disaster situation, involving many tweets and are likely to correlate with higher coordination. The remaining three events are generic in nature and represent general conversations. The choice of events with varied amounts of data allows us to demonstrate generalized usage of linguistic cues on the conversation in the new communication paradigms of social media. Nevertheless, we create conversation classifier models on normalized values of the features to account for generalization of the model.

3.2. Conversation Classifiers

We developed a separate decision tree classification model and feature ranking on the data sets of each event. First, the data sets were partitioned into corpuses of conversation types, followed by creation of data samples from each of the conversation types (RP, RT, M) with non-conversation type corpus (NC), as described in the previous section. Table 3 summarizes the results for learned models of conversation classifiers. The table includes accuracy for the classifier (ability to distinguish between the platform-indicated conversation and NC) for each of the platform indicators in the first three columns as well as ROC area values in the subsequent three columns. Higher accuracy and ROC values indicate a better classifier.

Each row in the Table 3 shows the performance for classification ability for an event, with accuracy and ROC measures for each of the three platform based subsets. Accuracy measures range from 62 - 78. ROC measures range from 0.63 to 0.84. These measures suggest fair to good accuracy in general, with relatively superior scores for the case of disaster events relative

to the non-disasters events, replies relative to retweets and retweets relative to mentions. Across all events, the ROC values are 0.8, 0.77 and 0.69 for distinguishing replies, retweets, and mentions from NC using a common model of heuristics. The conversation classifier suggests the elimination of 23 percent of the replies, 30 percent of the retweets and 33 percent of the Mention-based tweets from further analysis, despite the presence of platform indicators. However, the conversation classifier also promotes an average of 31 percent of the tweets that are not marked with these platform indicators as exemplifying the characteristics of conversation. Given the distribution of tweets in our corpus of approximately 1.5 million, the conversational classifier focuses further analysis on approximately 570,000 tweets with platform indicators, and a potential 200,000 tweets not marked with platform indicators.

	Accuracy %			ROC Value		
EVENTS	REPLY based	RT based	MENTION based	REPLY based	RT based	MENTION based
Japan Earthquake 2011	**78.06**	71.33	**66.49**	**0.84**	**0.78**	**0.71**
Haiti Earthquake 2010	70.54	71.07	64.9	0.75	0.77	0.69
Hurricane Irene Storm 2011	75.2	**71.46**	59.45	0.8	0.77	0.61
Debt Ceiling Debate 2011	68.9	68.77	62.51	0.73	0.74	0.66
Skype Microsoft Deal 2011	65.54	67.76	56.38	0.67	0.72	0.57
Glenn Beck Rally 2010	62.37	64.32	64.92	0.63	0.67	0.65
Common/Mixed dataset	***74.27***	***70.9***	***64.35***	***0.8***	***0.77***	***0.69***
Common/Mixed dataset for Disasters	***74.61***	***71.24***	***64.72***	***0.8***	***0.78***	***0.69***
Common/Mixed dataset for Non-Disasters	***68.96***	***68.15***	***61.53***	***0.73***	***0.73***	***0.65***

Table 3: ***Classification Model performance for various types of conversations based on the linguistics cues***

3.3. Feature Ranking

Table 4 shows the features in the models ranked from left (best) to right column (worst) for classification, for each of the event data sets and for each of the conversation type corpuses - RP (Reply), RT (Retweet), M (Mention). As in Table 3, the last rows in Table 4 provide results for the entire event data set, and disaster and non-disaster events. Figure 5, 6, and 7 provide graphical summaries for the top four features omitting the highly influential heuristic "you" [h12] to preserve a readable effectiveness scale on the remaining heuristics. In general h3, h4, h9, and h12 appear in the top 5 across the platform indicators and types of events. RT-based exchanges are identified by h10 and h1 as well. Mention-based exchanges are identified by h10 and h1.

REPLY-based	Feature Rank							
EVENTS	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8
Japan Earthquake 2011	h4	h12	h9	h13	h3	h10	h5	h2
Haiti Earthquake 2010	h12	h4	h13	h10	h9	h3	h5	h2
Hurricane Irene Storm 2011	h4	h12	h9	h3	h13	h10	h5	h2
Debt Ceiling Debate 2011	h12	h10	h4	h3	h13	h5	h9	h2
Skype Microsoft Deal 2011	h4	h12	h10	h13	h5	h9	h3	h2
Glenn Beck Rally 2010	h4	h12	h13	h3	h10	h5	h1	h2
Common/Mixed dataset	**h4**	**h12**	**h9**	**h13**	**h3**	**h5**	**h10**	**h2**
Common/Mixed dataset for Disasters	**h4**	**h12**	**h9**	**h13**	**h3**	**h5**	**h10**	**h2**
Common/Mixed dataset for Non-Disasters	**h4**	**h12**	**h9**	**h13**	**h3**	**h5**	**h10**	**h2**
RT-based	Feature Rank							
EVENTS	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8
Japan Earthquake 2011	h10	h9	h12	h6	h3	h7	h1	h4
Haiti Earthquake 2010	h10	h4	h9	h12	h1	h2	h13	h14
Hurricane Irene Storm 2011	h10	h4	h14	h9	h6	h12	h7	h13
Debt Ceiling Debate 2011	h10	h3	h4	h5	h14	h2	h8	h12
Skype Microsoft Deal 2011	h10	h3	h13	h14	h12	h6	h5	h7
Glenn Beck Rally 2010	h10	h1	h5	h4	h7	h2	h3	h14
Common/Mixed dataset	**h10**	**h9**	**h12**	**h4**	**h1**	**h3**	**h14**	**h6**
Common/Mixed dataset for Disasters	**h10**	**h9**	**h12**	**h4**	**h1**	**h3**	**h14**	**h6**
Common/Mixed dataset for Non-Disasters	**h10**	**h9**	**h4**	**h12**	**h1**	**h6**	**h3**	**h14**
MENTION-based	Feature Rank							
EVENTS	Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7	Top-8
Japan Earthquake 2011	h9	h10	h12	h4	h14	h3	h6	h5
Haiti Earthquake 2010	h12	h10	h3	h9	h2	h4	h13	h1
Hurricane Irene Storm 2011	h10	h4	h12	h9	h3	h6	h1	h2
Debt Ceiling Debate 2011	h10	h12	h9	h1	h4	h5	h2	h3
Skype Microsoft Deal 2011	h5	h6	h3	h7	h1	h2	h4	h12
Glenn Beck Rally 2010	h1	h9	h5	h6	h2	h4	h12	h3
Common/Mixed dataset	**h12**	**h10**	**h9**	**h3**	**h4**	**h2**	**h6**	**h14**
Common/Mixed dataset for Disasters	**h12**	**h10**	**h9**	**h3**	**h4**	**h2**	**h6**	**h14**
Common/Mixed dataset for Non-Disasters	**h10**	**h12**	**h9**	**h3**	**h4**	**h2**	**h6**	**h13**

Table 4: ***Feature ranking for classification for conversation types for various events***

3.4. Correlation study for features in the correctly classified sample set

Table 5 shows the correlation coefficients for correctly classified data only. While the magnitude is meaningless because of the restricted sample, the direction of the relationship is always positive for the most highly ranked features. Thus the presence of the features we tested discriminate between positive and negative instances of conversation, according to the platform indicators.

3.5. Information Density

A domain-dependent analysis of tweet information content is beyond the scope of the present paper. However, we provide a generic indication of tweet information density using Pennebaker's Linguistic Inquiry Word Count (LIWC) software (http://www.liwc.net/), developed to provide percentanges for the presence of various pre-defined categories of words. Here we report analyses using measures of communication, sensed experience, and social

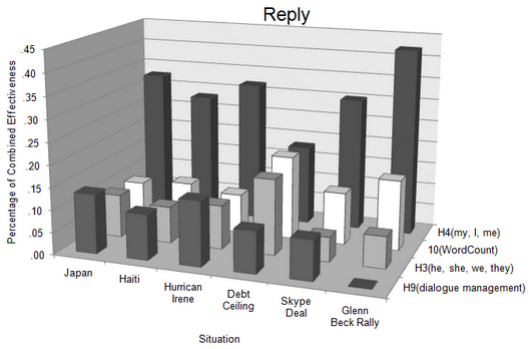

Figure 5: ***Top heuristics within the Reply framework***

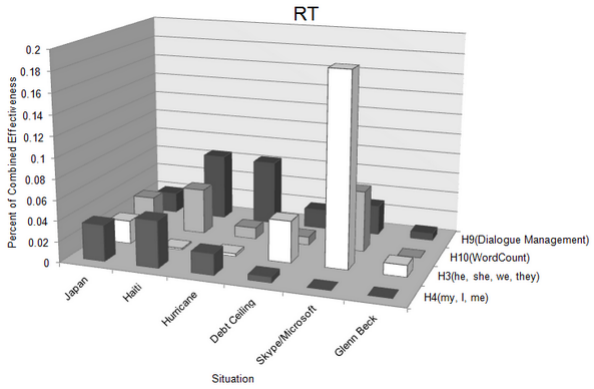

Figure 6: ***Top heuristics within the RT framework***

interaction. Measures of communication include 130 words such as "call", "speak", and "listen". Measures of sensed experience include 112 words, such as "drink", "eat", and "look". Measures of social interaction include 325 words such as "rumour", "secret", and "aunt". Although LIWC provides separate tallies for these categories, we note some degree of content overlap. For example, the word "ask" appears in the LIWC dictionaries for all three categories. However we edited the social interaction measure to exclude the words we used to build our conversation classifier.

Table 6 presents the three measures for the three conversation models we constructed (replies, mentions, and retweets) using different randomly selected 400,000 tweet samples of the data we used to build our models. Analyses for each of the nine combinations of measure and model appear

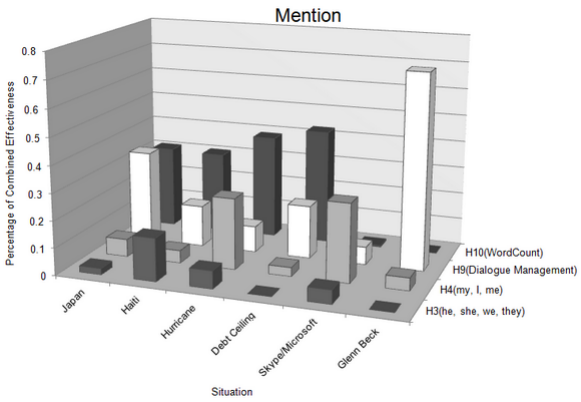

Figure 7: ***Top heuristics within the Mention framework***

REPLY-based		***RT-based***		***MENTION-based***	
CORR(C,h4)	0.5379435089	CORR(C,h10)	0.6890163385	CORR(C,h12)	0.4951287188
CORR(C,h12)	0.4933621573	CORR(C,h9)	0.1427287627	CORR(C,h9)	0.4317517945
CORR(C,h13)	0.3608575332	CORR(C,h12)	0.1373937122	CORR(C,h10)	0.418519636
CORR(C,h9)	0.3603677094	CORR(C,h3)	0.0990999819	CORR(C,h4)	0.2772313278
CORR(C,h3)	0.3298950158	CORR(C,h6)	0.0635809668	CORR(C,h3)	0.2370879435
CORR(C,h5)	0.2904841023	CORR(C,h7)	0.0511771093	CORR(C,h2)	0.1006701597
CORR(C,h10)	0.1302279126	CORR(C,h5)	0.0506748247	CORR(C,h13)	0.1000650892
CORR(C,h2)	0.1295144148	CORR(C,h14)	0.023213349	CORR(C,h5)	0.0941631627
CORR(C,h7)	0.0615535592	CORR(C,h8)	0.001804053	CORR(C,h6)	0.0596707947
CORR(C,h8)	0.0468423027	CORR(C,h13)	-0.0074445414	CORR(C,h14)	0.0381844937
CORR(C,h6)	0.0467575461	CORR(C,h11)	-0.0152468614	CORR(C,h8)	0.0084248154
CORR(C,h11)	0.0425604815	CORR(C,h2)	-0.029237837	CORR(C,h7)	0.0049817937
CORR(C,h14)	0.0313333062	CORR(C,h1)	-0.0708892543	CORR(C,h11)	0.0044300814
CORR(C,h1)	0.0298269769	CORR(C,h4)	-0.1322003669	CORR(C,h1)	-0.0446478387

Table 5: ***Correlation of features (h_i) with predicted class C in the true classified instances, for conversation types for common dataset for Disasters***

in separate two-by-two contingency tables. Rows of the contingency tables indicate the presence or absence of a platform indicator for the tweet type in question. Columns of the contingency table indicate material classified as conversation and material classified as non-conversation. Thus the cell in the upper left hand corner of the table represents the social interaction measure for replies. The data in the first row of cells correspond to the percentage of social interaction words for replies classified as conversation and for replies classified as non-conversation. The data in the second row of cells correspond to the percentage of social interaction for tweets that are not replies, mentions, or retweets, classified as conversation and non-conversation. Row totals correspond to the percentage of the social metric for replies and the non-replies, non-mentions, and non-retweets, that is platform-based NC. Each cell in each table represents the content of 100,000 tweets.

Each row and each column represent the content of 200,000 tweets.

REPLY			
Social			
	Conv	Non-Conv	
RP	3.58	3.47	3.56
NC	3.79	3.07	3.25
	3.64	3.17	
Senses			
	Conv	Non-Conv	
RP	2.08	1.6	1.96
NC	1.72	1.37	1.46
	1.99	1.42	
Comm			
	Conv	Non-Conv	
RP	1.55	1.29	1.49
NC	1.3	1.12	1.17
	1.49	1.16	

RT			
Social			
	Conv	Non-Conv	
RT	3.7	3.31	3.52
NC	3.48	3.04	3.29
	3.58	3.17	
Senses			
	Conv	Non-Conv	
RT	1.62	1.54	1.58
NC	1.43	1.31	1.38
	1.52	1.42	
Comm			
	Conv	Non-Conv	
RT	1.48	1.31	1.41
NC	1.23	1.13	1.19
	1.35	1.22	

MENTION			
Social			
	Conv	Non-Conv	
M	4.17	3.53	3.91
NC	3.7	3.22	3.37
	4.01	3.33	
Senses			
	Conv	Non-Conv	
M	1.61	1.58	1.59
NC	1.54	1.42	1.46
	1.58	1.48	
Comm			
	Conv	Non-Conv	
M	1.44	1.33	1.39
NC	1.29	1.14	1.19
	1.39	1.21	

Table 6: ***Three LIWC analysis features- Senses, Social Interaction and Communication, for the three conversation models (Reply, RT, Mention)***

A common pattern emerges across all nine analyses, described here for the analysis in the upper left hand corner of the table. We do not provide statistical analyses for these data due to the very large sample sizes in question, which shrink the standard errors and trivialize testing. The density of social interaction information (3.56% of the words) in the platform-based indicator data (replies, in the present case) is greater than the density of information in the platform-based NC (3.25%). The density of information in the content classified as conversation (3.64%) is greater than the density of information in the content classified as non-conversation (3.17%). Within the content marked as a Reply, the subset classified as conversation (3.58%) has a higher percentage of content than the subset classified as non-conversation (3.47%). Within the content of platform-based NC, the subset classified as conversation has a higher information density (3.79%) than the subset classified as non-conversation (3.07%). This pattern holds for all nine analyses, albeit with varying magnitudes. The only anomaly is the higher information density in the Reply analysis for the NC-conversations relative to the platform-based conversations. We also note that in many cases (the social interaction metric in particular), among the tweets that are not marked with platform indicators, the information density of classified conversation exceeds the information density of the tweets marked with platform indicators that are not classified as conversation.

Thus, we demonstrate that our conversation-based sampling heuristic for tweets correlates with higher densities of information content.

4. Discussion

- Classification Ability

  Our goal was to separate the Twitter stream into subsets more and less likely to contain citizen coordination revealed in conversation. We modeled conversation indicators in a conversation classifier for three types of Twitter postings assumed to contain a high proportion of conversation. Using simple heuristics based on pronouns, dialogue management, and word count, we demonstrated the ability to classify tweets as instances of replies, retweets, and mentions versus none of these with accuracy up to 78% and ROC area values up to 0.84. These generally good values support the claim that social media platform indicators reflect the coordination inherent in conversation conventions.

  Certainly our ability to classify declines with the type of Twitter exchange, but in an interpretable fashion. We do best at classifying replies, which should rely most heavily on coordination indicators because the intended purpose of Reply is conversation. Similarly, we do better with the disaster corpus than the non-disaster corpus as shown in Table 3. This supports an association between linguistic indicators of coordination and the actual coordination that the disaster invokes.

  Despite relative success in distinguishing different types of tweets from non-conversation, our discrimination statistics are not perfect. This is in part due to the expected contamination of replies, retweets, and mentions with non-conversation, and the presence of otherwise undetected conversation in the non-conversation subset. Indeed we seek the model of heuristics that enables us to transcend platform indicators in the detection of coordination.

- Psycholinguistic Theory

  We know of no other studies that attempt to test an account of conversation against a control corpus, in part because of the challenge of defining such a corpus. The bulk of linguistic theory hinges on the analysis of positive instances of conversation. Thus, we had not been able to test the diagnosticity of conversation indicators.

The models generally depend on a common set of highly effective heuristics, across individual events, types of events, and types of conversation. Personal pronouns, relative pronouns, and dialogue indicators play major roles in discriminating conversation types from non-conversations. Consistent with psycholinguistic theory, the preponderance of pronouns reflects the prior common grounding of important entities (agents and objects) in previous exchange. However, the length of conversation plays a greater role in retweets. Crediting the original source and adding opinion prefixes necessarily extend the length of tweets, unless already at the 140 character limit. Thus the length heuristic is likely an artifact of the Twitter medium. However, denser diffusion networks result from retweets with a credited source, reinforcing their retention as observed by Nagarajan et al. [20] as well.

In addition to demonstrating the diagnosticity of conversational indicators relative to a control condition of non-conversation, we also have demonstrated a greater density of information content in tweets that reflect conversation. Twitter traffic marked with platform indicators (replies, retweets, and mentions) classified as conversation has a greater density of information content. Twitter traffic marked with platform indicators that does not get classified as conversation appears to have less content. This theoretically relevant association between conversational indicators and content has practical merit. We cannot assume that all platform marked traffic is actually information rich conversation, providing a basis for trimming an otherwise unwieldy volume of message traffic.

- Limitations/ Future Work

  Alternative machine learning approaches such as boosting and bagging could improve the performance of the conversation classifier. However, our goal here is to present an existence proof for a conversation classifier as the foundation for the study of coordination. Although linguistic theory would anticipate a universal need for cooperation in conversation, our heuristics are limited to English and could require revision as we extend them to other languages. Finally, the space constraint in Twitter leads to unconventional English and the emergence of new writing conventions, such as hashtags. Therefore, space constraints potentially override the tacit concern for coordination in ordinary con-

versation.

The conversational filter we have developed serves as the first, domain independent step in the extraction of nuggets of coordination. One risk of the approach is that we ignore something important simply because it does not appear conversational. On the other hand, important content that is buried in slow processing is functionally unavailable. Furthermore, subsequent semantic analysis must mine the conversations for actionable content. In this paper, we relied on generic semantic metrics (for communication, sensed experience, and social interaction) simply to demonstrate the potential information gain in the conversational subsets. Although encouraging, this is no substitute for the semantic analysis that identifies actionable nuggets. Our ongoing efforts focus on the semantic models, both domain independent and domain specific, to further mine, sort, and aggregate actionable content. We are also investigating display methods for providing actionable nuggets to the emergency response community, balancing the tradeoff between effortful information search and passive information overload, via our data management tool Twitris (http://www.twitris.org).

- Implications

  By combining the existing platform-based indicators with our linguistic model, we can sort the voluminous Twitter traffic into subsets more likely to contain coordination. This is of potential interest to the emergency response community. Using the conversational classifier on tweets marked with platform indicators cuts the corpus by nearly two thirds. Admitting tweets classified as conversation without platform indicators still cuts the corpus in half. When combined with semantic analysis we hope to quickly and inexpensively identify the most salient messages among millions of noisy transmissions. The current method of achieving this goal is too computationally intensive, slow, and expensive to be practical in dynamic emergency management situations. Creating a smaller, targeted data set can direct information flow to those with the power to help the most.

  The ability to detect coordinated response inherent in conversation has relevance to other domains in addition to emergency response. We envision augmenting standard methods of capturing public attitude with conversation-based coordination metrics, potentially correlated

with action, such as attending a new movie. Similar analyses apply to political opinion and likelihood of action, e.g., casting a ballot for state elections.

5. Conclusion

We have presented an extensive analysis on communication characteristics of the Twitter-sphere in this study. We show that theory-driven linguistic features are present in the message traffic of new communication paradigms in social media, which can help to locate linguistic coordination via conversations. The study also grounds the new communication paradigm of social media in fundamental properties of human communication, and as such can be further explored on adaptive human behavior in communication. Our specific use-case of this study is to improve crisis response coordination by helping locate meaningful information from the voluminous message traffic.

6. Acknowledgement

This work is supported by NSF (IIS-1111182, 09/01/2011 - 08/31/2014) SoCS program under the grant titled "Social Media Enhanced Organizational Sensemaking in Emergency Response". We also thank our colleagues for great support and valuable comments on our preliminary investigations, especially Dr. Christopher Thomas (alumnus) and Dr. Meenakshi Nagarajan (alumnus).

References

[1] A. P. Sheth, Citizen sensing, social signals, and enriching human experience., IEEE Internet Computing 13 (4) (2009) 87–92.

[2] H. H. Clark, D. Wilkes-Gibbs, Referring as a collaborative process, Cognition 22 (1) (1986) 1–39.

[3] C. Goodwin, J. Heritage, Conversation analysis, Annual Review of Anthropology [CPL] (1990) 283–307+.

[4] G. Mark, Extreme collaboration, Communications of the ACM 45 (2002) 89–93.

[5] H. Purohit, A. Hampton, V. Shalin, A. Sheth, J. Flach, Framework for the analysis of coordination in crisis response, in: Proceedings of the CSCW workshop on Collaboration & Crisis Informatics, 2012.

[6] C. Danescu-Niculescu-Mizil, M. Gamon, S. Dumais, Mark my words!: linguistic style accommodation in social media, in: Proceedings of the 20th international conference on World wide web, WWW '11, ACM, New York, NY, USA, 2011, pp. 745–754. doi:10.1145/1963405.1963509.

[7] S. Gouws, D. Metzler, C. Cai, E. Hovy, Contextual bearing on linguistic variation in social media, in: Proceedings of the Workshop on Languages in Social Media, LSM '11, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 20–29.
URL http://dl.acm.org/citation.cfm?id=2021109.2021113

[8] A. Ritter, C. Cherry, B. Dolan, Unsupervised modeling of twitter conversations, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010), Proceedings of the Main Conference, June 24, 2010, Los Angeles, California, Association for Computational Linguistics, ACL, Stroudsburg, PA, USA, 2010, pp. 172–180.

[9] D. Boyd, S. Golder, G. Lotan, Tweet, tweet, retweet: Conversational aspects of retweeting on twitter, in: hicss, IEEE Computer Society, 1899, pp. 1–10.

[10] D. Clarke, The Sequential Analysis of Action Structure, European Studies in Social Psychology, Cambridge University Press, 1982.
URL http://books.google.com/books?id=nxs4AAAAIAAJ

[11] C. Honeycutt, S. C. Herring, Beyond microblogging: Conversation and collaboration via twitter., in: HICSS, IEEE Computer Society, 2009, pp. 1–10.

[12] R. Dietrich, T. von Meltzer, Communication in High Risk Environments, Linguistische Berichte: Sonderheft, Buske, 2003.
URL http://books.google.com/books?id=YupoAAAAIAAJ

[13] W. Chafe, B. C. S. P. University of California, Cognitive Constraints on Information Flow, Berkeley cognitive science report, Cognitive Science Program, Institute of Cognitive Studies, University of California at

Berkeley, 1984.
URL http://books.google.com/books?id=W-5uHQAACAAJ

[14] M. Nagarajan, A. P. Sheth, S. Velmurugan, Citizen sensor data mining, social media analytics and development centric web applications., in: S. Srinivasan, K. Ramamritham, A. Kumar, M. P. Ravindra, E. Bertino, R. Kumar (Eds.), WWW (Companion Volume), ACM, 2011, pp. 289–290.
URL http://dblp.uni-trier.de/db/conf/www/www2011c.htmlNagarajanSV11

[15] M. Nagarajan, K. Gomadam, A. P. Sheth, A. Ranabahu, R. Mutharaju, A. Jadhav, Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences., in: G. Vossen, D. D. E. Long, J. X. Yu (Eds.), WISE, Vol. 5802 of Lecture Notes in Computer Science, Springer, 2009, pp. 539–553.
URL http://dblp.uni-trier.de/db/conf/wise/wise2009.htmlNagarajanGSRMJ09

[16] A. Smith, A. Sheth, A. Jadhav, H. Purohit, L. Chen, M. Cooney, P. Kapanipathi, P. Anantharam, P. Koneru, W. Wang, Twitris+: Social media analytics platform for effective coordination., NSF SoCS Symposium, 2012.

[17] A. P. Sheth, C. Thomas, P. Mehra, Continuous semantics to analyze real-time data., IEEE Internet Computing 14 (6) (2010) 84–89.
URL http://dblp.uni-trier.de/db/journals/internet/internet14.htmlShethTM10

[18] S. Safavian, D. Landgrebe, A survey of decision tree classifier methodology, Systems, Man and Cybernetics, IEEE Transactions on 21 (3) (1991) 660–674.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The weka data mining software: an update, SIGKDD Explor. Newsl. 11 (1) (2009) 10–18. doi:10.1145/1656274.1656278.
URL http://dx.doi.org/10.1145/1656274.1656278

[20] M. Nagarajan, H. Purohit, A. Sheth, A qualitative examination of topical tweet and retweet practices (2010).
URL http://aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1484