Short message communications: users, topics, and in-language processing [*]

Robert Munro
Department of Linguistics
Stanford University
Stanford, CA 94305
rmunro@stanford.edu

Christopher D. Manning
Department of Computer Science
Stanford University
Stanford, CA 94305
manning@stanford.edu

ABSTRACT

This paper investigates three dimensions of cross-domain analysis for humanitarian information processing: citizen reporting vs organizational reporting; Twitter vs SMS; and English vs non-English communications. Short messages sent during the response to the recent earthquake in Haiti and floods in Pakistan are analyzed. It is clear that SMS and Twitter were used very differently at the time, by different groups of people. SMS was primarily used by individuals on the ground while Twitter was primarily used by the international community. Turning to semi-automated strategies that employ natural language processing, it is found that English-optimal strategies do not carry over to Urdu or Kreyol, especially with regards to subword variation. Looking at machine-learning models that attempt to combine both Twitter and SMS, it is found that the cross-domain prediction accuracy is very poor, but some loss in accuracy can be overcome by learning prior distributions over the sources. It is concluded that there is only limited utility in treating SMS and Twitter as equivalent information sources – perhaps much less than the relatively large number of recent Twitter-focused papers would indicate.

1. INTRODUCTION

Short-message systems have risen to recent prominence in emergency response communications, especially in a growing body of work that is looking at how aid agencies can directly engage crisis-affected communities in the wake of sudden onset disasters.

We compare Twitter and SMS communications from two recent crises, the 2010 earthquake in Haiti and the 2010 floods in Pakistan. They are both the largest disasters of their kind in living memory, and both feature prominent use of both SMS and Twitter. We investigate the application of natural language processing to identify novel information in the messages, comparing systems built on the original languages (Haitian Kreyol and Urdu) to systems built on the English translations of those messages.

The first author ran the crowdsourced information processing system for the Haitian Kreyol messages, *Mission 4636*, and adapted the same platform for processing the messages in Urdu for the second system, *Pakreport*. Contrary to many media and some academic reports [11, 30], Twitter was not widely used in either case, and was not considered a high-value source of information at the time by those of us running large information processing systems. This study is part of a larger body of actual deployments and critical analysis focusing on automated and semi-automated approaches to processing communications in development contexts [22, 20, 21, 18].

From analyzing the data, several conclusions are drawn:

(1) Twitter and SMS filled very different functions in these crises: Twitter was primarily used by the aid community and SMS by the crisis-affected community.

(2) Because of the differences, systems built on Twitter alone cannot be extrapolated to large scale projects that engage the crisis-affected population.

(3) Systems that are optimal for one language are unlikely to be optimal for all, with English likely to be an outlier with respect to modeling subword variation.

(4) Systems optimized on one type of short message (SMS or Twitter) do not perform well when applied to the other, but learning a prior over the type of message can help reduce the error.

Given that the *majority* of recent research has looked only at Twitter in English (see section 3), the conclusions raise some questions about the appropriate directions for research and development.

DEV '12, March 11-12, Atlanta, GA
Copyright © 2012 ACM 978-1-4503-1262-2/12/03... $10.00

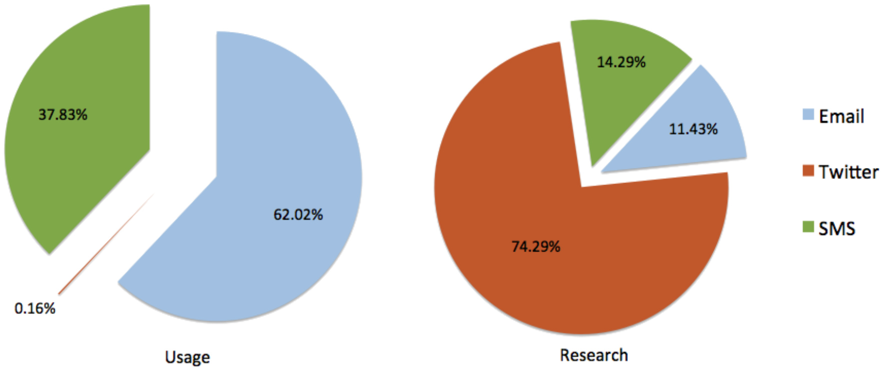

Figure 1: The Twitter bias. Left, the relative amount of communications globally, comparing SMS, Twitter and email. Right, the same comparison, but comparing the relative amount of papers written between June 2010 and June 2011 that looked at automatic filtering or information extraction from each of SMS, Twitter and email. While Twitter accounts for much less than 1% of the world's communications, it accounts for almost three-quarters of the research. The utility of Twitter-based studies to broader communication therefore heavily depends on the extent to which Twitter-communications are similar to SMS or email.

The paper is in four parts. First, we look at the disparity between usage and research of short message systems. Second, we investigate who actually used short message platforms in Haiti and Pakistan. Third, we compare the topics of messages. Finally, we look at classification through machine-learning, investigating the domain dependence of the results and comparing subword models in Urdu and Kreyol with their English translations.

2. BACKGROUND

Text messaging (SMS) is a popular communication technology in social development, supporting health [16], banking [24], access to market information [14], literacy [12] and emergency response [20]. Moreover, it is the dominant form of remote communication in many regions, especially less-resourced parts of the world, meaning that it often has the broadest coverage for communications with the people that development technology supports.

To show that there really *is* a bias in recent research, we conducted a review of all peer-reviewed IEEE, ACM, and ACL publications between June 2010 and June 2011 that focused automated or semi-automated processing of Twitter, email or SMS, and compared them to the actual global usage of the technologies, as reported by [28, 13, 7].

As Figure 1 shows, while Twitter makes up much less than 1% of the world's communications, it accounts for almost 75% of recent research. Most of the papers stated that Twitter was used as a stand-in for social media and/or SMS more generally.[1] The utility of Twitter-based studies to broader communication therefore very heavily depends on the extent to which Twitter-communications are similar to SMS or email.

For all three, the percentages for use indicate the number of messages sent, not the number read (excluding SPAM). Emails and SMS sent to multiple recipients and Tweets read by multiple followers are only counted once. Email and Twitter are the most likely platforms to have multiple readers per message. On the other hand, Tweets and (especially) emails may be ignored and not read at all, while this is much less likely with SMS. All these generalizations are very context dependent – Pakistan in particular makes productive use of group SMS messaging – but the exact numbers could vary greatly here and the disparity would remain.

2.1 Evaluation Data

Four sets of communications are used here: Mission 4636 text-messages from within Haiti; Tweets sent from within or about Haiti during the same time period; Pakreport text-messages from Pakistan and Tweets sent within or about Pakistan during the same time period.

The Mission 4636 messages were sent predominantly in Haitian Kreyol. At the time, every message was translated into English in near-time by crowdsourced volunteers and workers, giving us a parallel corpus of Haitian Kreyol/English messages. We use 40,811 of these messages here.

The Tweets about Haiti were taken directly from the Twitter API. Part of Mission 4636 was a mapping initiative called 'Ushahidi @ Tufts'/'Ushahidi Haiti', who imported about 3,000 of the text messages and published them online. They also looked for relevant information in social media and so the Tweets used here were also incorporated.

The Pakreport messages were sent in either Urdu, Sindhi, Pashto or English. Some of the non-English messages were translated into English by crowdsourced volunteers at the time. The system was smaller in scale, as, unlike Haiti, most existing channels of communication remained open. There were only 400 messages sent to the system in total, so in order not to have too-small a data set we retranslated every message into both English and Urdu. We paid microtaskers to do this, giving the task to multiple workers to ensure data quality and creating a separate evaluation task to pick the

[1] Interestingly, not a single paper says that it uses Twitter as a stand-in for email, despite email still being our most frequent form of remote communication.

Nou tigwav,nou pa gen manje nou pa gen kay. m. ‘We are Petit Goave, we don’t have food, we don’t have a house. Thanks.’
RT wyclef Haiti is in need of immediate AID please text Yele to 510 510 and donate $5 toward earthquake relief. Please Help! *(no translation)*
sukkur mien thambu , dhavaieyan ki zaroorath hein.or dhood powder ka zakth zaroorath hein. ‘in sukkur there is desperate need of tents, clothes and medicines, even a strong need of powder milk
Karachi: Project Madad: Need Volunteers for data entry for relief inventory ‘Karachi: Project Madad: rahat soochi keliye atha pravishti karne keliyae svayanasevak ka zaroorat hai.’

Table 1: Examples of the four kinds of communications: the first is an SMS sent from within Haiti, the second is a Tweet about Haiti, the third is an SMS sent from within Pakistan, the fourth is a Tweet about Pakistan.

most accurate translations. We did this for *all* the messages, not just the ones without a prior translation, in order to ensure that it was a consistent data set.

The Tweets about Pakistan were taken directly from the Twitter API and were imported into PakReport at the time. There were only 40 tweets that ultimately made it into the PakReport. While this number is small we increased it by similarly translating them all into Urdu (they were all English) thus allowing comparisons with the SMS in both languages. They were augmented by a further 500 tweets that contained the hashtag ‘#pkfloods’, giving us a large enough set for evaluation.

Who was actually using Twitter and SMS during the crises in Pakistan and Haiti? Despite the many reports about both to-date, no-one has yet tried to calculate the percentage of SMS and Twitter that was by individuals who were actually within the crisis-affected regions.

We undertook a simple analysis to label the data by source: was it sent by an individual within the affected region (citizen reporting), or was it sent by an aid organization or international person. We did this for all the Pakistan data and a random subset of 1,000 SMS and 1,000 Tweets from Haiti.

As with the translation of the Pakistan messages, we used paid microtaskers to annotate the messages as being citizen reporting from within Haiti/Pakistan or not (ie, being from an aid organization or a person outside of Haiti). The task was given to multiple workers to ensure reliability. It was a three-way choice between ‘from an individual within the crisis-affected region’, ‘from an organization and/or someone outside the crisis-affect region’, or ‘unambiguous’. A surprisingly low percentage of messages across all sets were ambiguous. People from organizations typically clearly identified themselves as such, despite the character limits, and the content and choice of pronouns made it clear when it was a person from with a region (“I/we need food”) a person outside the region (“they need food”) or an organization (“we need volunteers to help with packing”). However, when the tweets reported second-hand information it was often impossible to judge whether the actual source was from on the ground or via the media. For this reason we didn’t attempt to encode the messages for evidentiality.

3. WHO WAS USING SMS AND TWITTER?

The results are in Figure 2. It is clear that there is a stark difference between Twitter and SMS. The former was predominantly used by international community and the latter by the crisis-affected community, with very little overlap. There is a real danger that the crisis-affected community and international aid community, while both utilizing short-message systems, ultimately end up talking past each other.

Having established that the users of Twitter and SMS were largely separate groups, it should be clearer why the topics of the messages are also different.

At the time, the SMS and Twitter messages were categorized according to a standard set of UN-defined labels (“request for food”, “request for water”, “response taken”, etc). The same categories, with only a few changes, were used for both Mission 4636 and Pakreport, so we can therefore run the same comparison across both. Many of the categories had few or no messages, so these were conflated into larger top-level categories. There were also large volumes of messages for which there was no pre-defined category that fit the description, like “request for volunteer laborers”. Given the findings in the previous section, there is also an important distinction between topics depending on the sender. An aid organization requesting food for distribution is very different, from an operational standpoint, than an individual requesting food within a crisis-affected region, but in the coding schemes they are treated equally (perhaps because citizen-reporting simply never made it into reports in the past). Here, these categories are separated in an attempt to capture broad trends in the actual topics, not their mapping to information requirements of specific relief agencies in the given context. Each message was given zero or more categories:

1. Ground report - a direct report about the conditions within the affected region.

2. Request for workers/volunteers - a request for help in some service (e.g.: loading items into trucks).

3. Forwarding news - passing on information that was already second-hand (e.g.: linking to a news article, retweeting, etc).

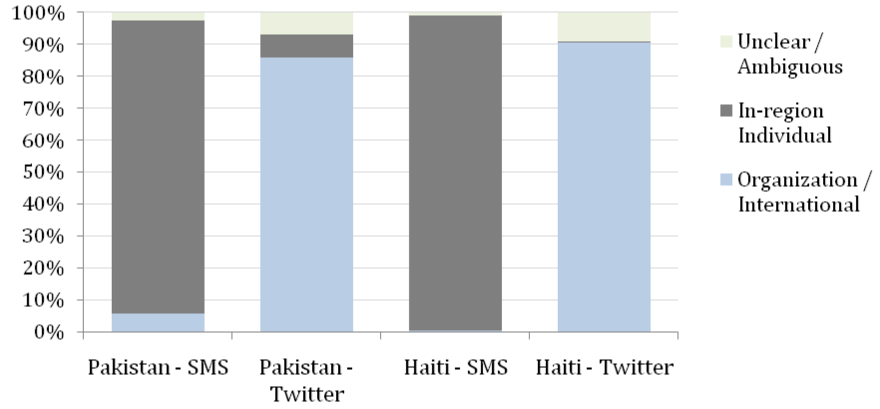

Figure 2: The users of Twitter and SMS, showing that in both cases SMS was primarily used by people within the crisis-affected region while Twitter was used by the international community.

4. Request for food/aid - a request for aid (made from anywhere).
5. Complaint - a criticism of the relief efforts.

4. THE TOPICS OF SHORT MESSAGE SYSTEMS

The breakdown of relative frequency of topics is given in Figure 3. As with the source data, it is clear that the topics of the messages differs greatly between SMS and Twitter.

However, the distribution of topics across the crises is very similar. In both cases, most of the SMS were a request for food/aid and/or a report about the conditions on the ground. The biggest topic for Twitter was simply relaying messages – mostly sharing links to online articles or retweets.

The biggest difference between Pakistan and Haiti can be traced back to the relative lack of resources that went into the Pakistan relief efforts. Pakistan had a greater number of people displaced but received a much smaller response from the international aid community, meaning that the Pakistan response required a larger logistical component with fewer resources. This is reflected in the Twitter messages, where many aid organizations were requesting people to help with simple tasks like loading containers with food. It is also reflected in the SMS 'complaint' messages where people were complaining about not having received aid or the national id card that would enable them to receive aid.

5. CLASSIFYING MESSAGES

The final and largest focus of this paper is comparing strategies for classifying messages with natural language processing. We extend earlier studies that looked at subword variation in text message classification in Chichewa, a language of Malawi, and Kreyol [21, 22].

5.1 Domain dependence

Domain dependence is a well-known problem in natural language processing – models trained on one domain, mode, register or genre will often perform badly when applied to another. For example, a system trained to identify spam in emails might not work as well when applied to spam-detection in social media. From the analysis of Twitter and SMS so far, while both are short message systems it is likely that they are different enough that models trained on one do not adapt well. By looking at the cross-domain accuracy, and the extent to which domain dependence is a problem, this gives us another lens for investigating the difference between SMS and Twitter, this time in the context of potential operational effectiveness.

We compared five different strategies to gauge the extent of the domain dependence and potential mitigating strategies. First we looked at models trained only on each type of message, that is, models trained and tested only on SMS, and models trained and tested only on Twitter. Second, we looked at cross-domain accuracy, for example, calculating the error in classification in Twitter using models trained only on SMS. Third, we built combined models, training and testing on both Twitter and SMS, treating both sources as equal. Finally, we extended the combined model by explicitly modeling the source ('Twitter' or 'SMS'), allowing the model to independently derive an optimal prior probability per source while using feature vectors across both.

5.2 Features and Subword variation

The simplest feature vector to implement is to just model the words that occur in the messages. We modeled the words, bigrams, and every pair of words in the message. We were able to model every pair because they were short messages. With $O(N^2)$ cost, this would not be as viable a strategy for longer texts.

We also modeled subword variation. By 'subword' variation, we mean any spelling variant of a given word. Broadly, this fits into three categories, morphology, phonological/ orthographic variation, and alternate spellings.

English has a very high level of standardization, even when the spelling is historical/arbitrary (e.g., 'tongue'), supported by extensive literacy education of most first and second language speakers. In contrast, many other languages have incomplete or inconsistent spelling standards, where even very fluent speakers will still have limited literacy, and so spelling is often ad hoc and inconsistent, roughly reflecting the pronunciation of words (as indeed it was in English at the time of Chaucer, when written English was first becoming common place).

The overwhelming majority of languages also have more complex morphology than English. Morphology refers to the fundamental components that make up words, most often combined through prefixing, suffixing or compounding. In English, 'unthinkingly' would have four morphemes, 'un', 'think' (the stem), 'ing', and 'ly'. Morphemes that can occur alone, like 'think', are called free morphemes, while those that must be attached to a free morpheme (like most suffixes and prefixes) are called bound morphemes. It is easy to see how word-based features might fail to capture this variation. For example, should a search for 'think' also turn up 'unthinkingly'? English has one of the world's least complex morphological systems, averaging close to just one morpheme per word – two or three per word is more typical and fully formed one-word sentences are common. Compounding is another frequent method for combining morphemes, but in this case with two or more free morphemes, like 'notebook'.

Phonological/orthographic variation will often result from more or less phonetic spellings of words. For example, the English plural has both the 'z' and 's' sound, and so a more phonetically accurate transcription for 'cats and dogs' would be 'cat*s* and dog*z*'. This can also result from morphology. For example, when we combine 'go' with 'ing' in English, we also pronounce the glide 'w' between the vowels. Someone not familiar with the language might write the 'w', 'gowing'. In some cases, a character that is not pronounced may be included or omitted. For example, the 'e' in 'bite' is not pronounced – it simply changes the vowel-quality of the 'i'. In Urdu, the same is true for the 'h' in 'th'. It indicates that the 't' is the aspirated variant of 't' (much like it does in German) but not a separate phoneme as it often does in English. Similarly, a repeated vowel can either mean a different phoneme or a lengthened vowel. This variation will most often occur when the writer has limited literacy, they are writing in a language that does not have established spelling conventions (which is most languages), or they are writing in a script that is not typical for the language, as with the Urdu speakers here favoring Roman script for text messages.

Alternate spellings are simply when there are multiple accepted spellings for a word, often the result of geopolitical divisions, like with 'reali*z*e' and 'reali*s*e'. They will also often reflect more or less phonetic pronunciations.

Urdu is an Indo-Aryan language that is almost completely co-intelligible with Hindi. The two are, grammatically, the same language, but traditionally use different scripts and have different borrowing patterns from other languages. With the Urdu messages analyzed here, it is easy to find examples of these kinds of variation. For example, *zaroorat*, 'need' is spelled at least four different ways (*zaroorath, zaroorat, zarorat, zarorath*). A typical former creole, Haitian Kreyol has very simple morphology but the text message-language produces many compounds and reductions. For example, *fanmi mwen* 'my family', is spelled at least six ways (*fanmi mwen, fanmwen, fanmi m, fanmi'm, fanmim', fanmim*).[2]

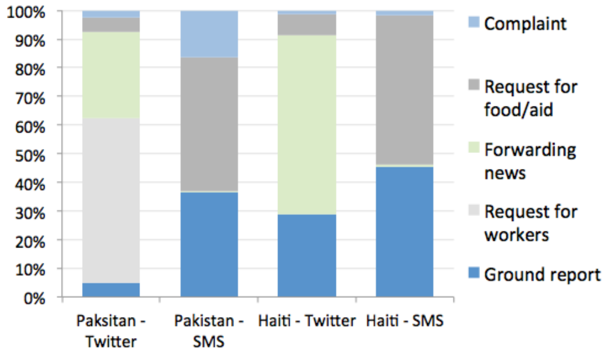

Figure 3: The topics of Twitter and SMS, showing that they were being used to disseminate very different kinds of information. This can be understood in terms of the results in the previous section – very different groups of people were using the two technologies. The distribution of topics for short messages is actually more similar within the type of message across crises than across the type of messages within each crisis. Twitter in Pakistan has a relatively large number of messages requesting help for logistics (e.g., loading food onto trucks). This may reflect a new use for Twitter in this later context or simply the context itself: the logistical operations in Pakistan were greater than in Haiti but undertaken by a smaller number of people who therefore needed to reach out more broadly for volunteer labor.

Most of the variations above have one thing in common: they follow predictable patterns. This means that they can be discovered and modeled through statistical analysis and unsupervised machine-learning. The common alternations like 's' and 'z' have the same place of articulation and differ only in voicing (that is, your mouth is in the same configuration and differ only by whether you vibrate your vocal chords). Is is therefore not a random alternation, and will follow patterns of usage that can discovered by statistical modeling. The same is true for other common alternations like 't'/'d' and 'p','b'. Similarly, the affixes like 'un', 'ing', and 'ly' will occur across multiple words an can be identified as common subword patterns. When an automated system that knows that 'zarootat' means 'need', it can therefore normalize spelling variants and separate morphemes to make informed predictions about the meaning 'zarootath', 'sarootat', and 'zarootathai' (note two of the spellings in table 1). If an automated system did *not* attempt to model unseen subword variations, then any novel forms of a word would not be able to contribute to classification predictions.

5.3 Testing framework

[2] These are typical examples in the messages here, with this kind of variation productive across both the Urdu and Kreyol messages. These particular examples are chosen because it is easy to imagine that 'my family' and 'need' are important forms to recognize when processing emergency messages.

We compared systems that used subword models to ones that only used word-level features.

For phonological/orthographic variation and common spelling alternates, we implemented a set of heuristics that represented forms for phonemes with the same potential place of articulation ('c/k'), forms with an adjacent place-of-articulation that are common phonological alternates ('l/r', 'e,i'), voicing alternations ('s/z'), limiting the potential alternations to *only* these. The messages were normalized such that the heuristic H were applied to a word w in the corpus C resulting in an alternate w', if, and only if, $w' \in C$. This method limits the alternates to only those whose existence is supported by the data.

For the morphological segmentation we used a *Hierarchical Dirichlet Process* (HDP) [34] adapted from Goldwater et al. [8], where every morpheme in the corpus m_i is drawn from a distribution G which consists of possible morphemes (the affixes and stems) and probabilities associated with each morpheme. G is generated from a Dirichlet Process (DP) distribution $DP(\alpha_0, P_0)$, with morphemes sampled from P_0 and their probabilities determined by a concentration parameter α_0. The context-sensitive model where H_m is the DP for a specific morpheme is:

$$\begin{array}{rll} m_i|m_{i-1} = m, H_m \sim H_m & & \forall m \\ H_m|\alpha_1, G & \sim DP(\alpha_1, G) & \forall m \\ G|\alpha_0, P & \sim DP(\alpha_0, P_0) & \end{array}$$

As with [21], we included *both* the segmented and unsegmented forms in the final supervised model, which allows the learner to arrive at the optimal distribution of weights between the segmented and non-segmented words. This was shown to be more accurate than earlier work that learned models post-normalization [22].

We use the Maximum Entropy model for the supervised learning component of the systems (the *the Stanford Classifier*). Without going into the full technical detail, it can be thought of as a system that calculates the optimal weight to apply to every feature (in our case, all the words and subwords), which is sensitive to changes in information value resulting from direct correlations between features, but is insensitive to exclusive-or-like correlations.

With a relatively small number of messages to learn models over we simplified the division of categories for prediction, combining the categories to a binary division in all sets: *Ground reports* and *News reports*.

The first, 'Ground reports', includes all reports from within a crisis-affected region that reports information about that region. The latter, 'News reports', includes all reports that are derived from existing media sources. We removed any that were ambiguous. While a more fine-grained analysis would be interesting, there simply wasn't a high enough volume of category types across both SMS and Twitter to allow this. This particular binary division is to reflect a realistic use cases: separating novel from known information.

Accuracy was calculated as the *F-value* for identifying novel information, which is the harmonic mean of the precision and recall. For the cross-domain evaluations we simply trained on one of SMS/Twitter and tested on the other. For the Pakistan data, we reduced the training set size for cross-domain evaluation in order to avoid an imbalance in model size. For all other models, a leave-one-out cross-validation was used, where accuracy was calculated by leaving out one data item (one message) as the evaluation, while training on all other items, and repeating over all items. This ensured that we trained on as much data as possible (necessary with the relatively small data sets) but that accuracy was still calculated over unseen messages.

5.4 Results - domain dependence

The results in table 2 confirm that there is domain dependence. The most striking figures are the very low accuracies for Twitter in Haiti. This mostly reflects the fact that ground reports messages from the ground were very rare in Twitter, relative to reports taken from the news, meaning that the baseline accuracy is much lower (about $F = 0.1$ here, as compared to about $F = 0.9$ for SMS). The lower accuracy is therefore expected in this context, but from an operational standpoint it is still clear that identifying novel reports in Twitter would be much less accurate in similar contexts.

Another reason that Twitter was so hard to classify in Haiti is because of the very high number of reports that are difficult to identify as news-based without external knowledge. For example, there are more than 20 different Twitter messages from Haiti about the Caribbean Market from *after* the search and rescue operations had ended (the number of retweets/quotes number in the 1000s). They are written in a manner that is close to indistinguishable from the ground reports, eg: "There are people still alive in the Caribbean Market", but at the time they were written they could not have been from the ground. This is actually the *most* reported event in the Tweets. By contrast, it only occurs once in the SMS messages. It gives some empirical backing towards the widespread concern of the reliability of social media during a crisis, and it certainly rules out cross-validated verification through counting the number of independent reports about an event.

For Haiti in particular, the accuracy in identifying messages that contain information from the ground drops from $F = 0.912$ to $F = 0.206$. This is much *less* than the baseline, meaning that cross-domain prediction is worse than a zero-information system that simply guessed every message was a ground report. By training on both and explicitly modeling the source, though, the accuracy is increased to $F = 0.954$, which is probably at the level of inter-annotator agreement. One of the biggest problems here is the prior probability. Because there were relatively few ground reports in Twitter the model is biased towards predicting fewer ground reports in the SMS. The precision/recall reflects this: while the precision is still high at 0.899, meaning that the system was accurate when it *did* predict that a message was a ground report, the low recall of 0.116 indicates that it made very few predictions, missing almost 9 out of 10. This is why a system was tested that also modeled the prior probability of each source – it allows the one system to be used across both sources, assigning separate priors to each.

	Haiti			**Pakistan**		
	Precision	Recall	F-value	Precision	Recall	F-value
SMS only	0.898	0.926	0.912	0.864	0.969	0.913
SMS trained on Twitter	0.899	0.116	0.206	0.893	0.878	0.885
SMS modeling source	0.930	0.979	0.954	0.667	0.842	0.744
Twitter only	0.398	0.243	0.301	0.690	0.851	0.762
Twitter trained on SMS	0.176	0.882	0.293	0.633	0.974	0.768
Twitter modeling source	0.331	0.249	0.284	0.900	0.983	0.940
Combined	0.882	0.895	0.888	0.864	0.967	0.913
Combined modeling source	0.846	0.820	0.833	0.866	0.964	0.912

Table 2: A comparison of accuracy for the four sets of data, comparing different combinations of source data and training techniques. Note that all the results in this table are restricted to English only in order to maximize the size of the data sets used and to avoid conflating the cross-linguistic factors that we investigate later in the paper. The combined models include both Twitter and SMS as training/test. The results show that cross-domain adaption is generally poor, and that modeling the source of the message generally improves the results, but that there is no one architecture that will produce the most accurate results in all cases.

The Twitter messages for Haiti are more robust when classified on models trained on SMS, falling from $F = 0.301$ to $F = 0.293$, but modeling the source actually hurts the performance further, reducing the accuracy to $F = 0.284$.

One notable outlier has the accuracy *increased* for the Twitter messages in Pakistan when trained on SMS, jumping from $F = 0.762$ to $F = 0.768$ with most of the gain in recall, an increase that is trending but not quite significant ($\rho < 0.1$, χ^2 with Yates' Correction). From investigating the models and results, it is difficult to work out exactly why there was an increase here, and we suspect that it might simply be an artifact of the bias that we already identified in the data – the SMS had a greater number and more varied ground reports than the Tweets, and so it is possible that this increased richness alone helps overcome the change of domain. As a single data point from a relatively small amount of training data (less than 500 items) we are careful not to draw too broad a conclusion, especially as the small gain is overshadowed by the much more accurate model combining training data sources, giving $F = 0.940$. If a larger study *did* confirm this trend to be significant, then there are some interesting implications for creating training data. While this study has emphasized the importance of in-domain training data, it may be as important in some contexts to collect as broad a range of training examples as possible, specifically targeting out-of-domain examples to make up gaps in knowledge.

Despite the fact that SMS and Twitter are both short message systems and the actual messages are taken from the same time and events, the results here show that we cannot assume to use one as a stand-in for the other. The loss in accuracy resulting from domain dependence can be mitigated by modeling the source, but this does not necessarily correct all the errors – note the overall Haiti accuracy drops from $F = 0.888$ to $F = 0.833$. The solution to domain dependence therefore seems to be context specific. Other solutions to domain adaptation might be more appropriate, and in some cases it might be easier to just keep separate models per-source.

5.5 Results - subword models

Finally, we compared the results across languages, looking at the relative impact of subword models on accuracy in Kreyol, Urdu and English.

The results are given in Figure 3, and show improvements for the Urdu and Kreyol messages. The English data does improve with significance in the Pakreport data ($\rho < 0.05$, χ^2 with Yates' Correction) but by investigating the features that received the greatest weight in the models it was clear that most of this was simply removing the hashtag '#' from the beginning of words like '#pakistan' in Tweets. The same was true of some earlier results on a smaller set of Haiti Twitter data. While this is certainly a valuable and useful segmentation, it would also be somewhat trivial to hard-code into a Twitter-filtering system. However there is probably still some value in applying subword models to Twitter in this context.

Even if people are using controlled vocabularies, like '#need #food', it is difficult to imagine that it will be easy to ensure that they are widely adopted during the chaos of a crisis, or that they would be used consistently. What is the relative weight that should be given, then, between the sequences '#need #food' and 'need food', or even the much simpler '#pakistan' and 'pakistan'? Should it be the same weight for all tag/non-tag pairs? Probably not. To what extent should the context of each tag/non-tag sequence be taken into account when arriving at a weight? It is difficult to image that this could be hard-coded into any system with human-tuned weights. For the subword models proposed here, as the system has access to both the tag and non-tag equivalents in a given message and a history of the distribution across past messages, it will arrive at an optimal weight for each pair/context according to past data. This might be one way to evaluate the effectiveness of controlled vocabulary systems: how much does the accuracy degrade when we strip out all the '#'s?

The results from the previous section showed that SMS and Twitter were used for substantially different topics by very different groups of people. The advantage here is that it therefore means that we can conclude that the need for subword modeling is not a simple quirk of the particular platform but a general property of the Urdu and Kreyol languages. If this study was first conducted in English alone, it would have seemed the correct choice to discard subword models for SMS. In that case, none of the gains in Figure 3 would have been discovered. We conclude that it is necessary to build and evaluate natural language processing systems in the languages of the crisis-affected populations.

6. RELATED WORK

Despite the prevalence of Twitter and SMS in recent development discussion and research, there is no prior study that compares the two when both were used in the same context (which is, of course, one of the main reasons for this paper).[3]

At least in part, the bias is towards availability of data and not just that social media studies are a current trend. Twitter's open API has made it much easier for people to scrape data, relative to more typically closed communication channels like SMS, email, and even the more private social media of Facebook. We can see evidence of this by looking at a recent workshop focused on social media, *'#SocialMedia: Computational Linguistics in a World of Social Media'*, where there are plenty of Twitter papers, but no Facebook ones.

However, the availability of data is not the whole story, as much of the data in this study has been freely available during this same time period. Several thousand of the SMS in Haitian Kreyol and all but a handful of the SMS in Urdu were accessible on public websites. With the exception of the people who built the first Machine-Translation system for Kreyol [17] and ran the translation/categorization platform [21], the Kreyol messages have not been studied. The Urdu messages have not been studied at all.

Caragea et al. came close, outlining a system that scrapes Twitter and classifies Tweets [3]. In order to evaluate their system, they used some of the SMS from Haiti instead of Tweets (an inversion of most research that used Twitter as a stand-in for SMS). However, they only used the English translations of those messages. Therefore, they evaluated their Twitter classifier on the English translations of Kreyol SMS messages, making both the language and platform equivalent assumptions that are challenged here. They report (p.c.) that this choice was made because of familiarity with English. Their proposed architecture relies on machine-translation, which might also be problematic as most languages do not have machine-translation services, and such systems take some time to build [17, 18].

Within the natural language processing community more broadly, there are probably too few researchers with the interests/skills to tackle non-English materials. There is only one previous study looking at cross-linguistic SMS in the literature, which is our previous work on the Chichewa language of Malawi [22], working with communications from *Medic Mobile*. We modeled the subword variation and built NLP systems to classifying messages according to medical labels, finding substantial differences when comparing the system the results to a system that used the English translations of the same messages. We add the results here to these, emphasizing the need for subword modeling in a wide variety of languages.

Building on this work, we also looked at realistic deployment scenarios for identifying actionable items within a constant stream of data [21]. This was building the system I wish existed during Mission 4636 (first author). The work also compared Twitter and SMS, but without labeled Twitter data it was simply trying to identify actionable SMS among Twitter data. While there were good results for Twitter, we suggested that this may have been the learner identifying the domain as much as the 'actionability', and the results here seem to confirm this. It is worth noting, though, that while the cross-validation methods used here give the clearest picture of the differences between types of short messages and across languages, the streaming architecture in [21] gives a more accurate idea of the potential accuracy in a sudden-onset deployment scenario.

There has been some more remotely related recent work in normalizing abbreviations in text messages [4, 15, 5, 25, 2, 33, 19]. All were evaluated on English (and one in French), but some systems normalized without pre-labeled English data, meaning that there is the potential to extend their methods to other languages. However, the most common SMS data set used for evaluations were not actual messages, but a dictionary of 303 common pairs of abbreviations/actual words [4, 5, 19], meaning that it was a very different task to the one investigated here. Earlier work on SMS is limited to identifying SPAM [9, 10, 6].

Xue et al. looked at both Twitter and SMS data for normalization [36]. Like Munro [21], they found that the linguistic structures of Twitter and SMS were independent enough that the optimal models for normalizing the respective messages were significantly different.

Vieweg, Hughes, Starbird and Palen looked at Twitter during two disasters in the United States [35]. This is an interesting counter-point to the papers here and some citizen reporting, but it is not clear what the most common use-case is. Starbird and Palen looked at Twitter in the context of the earthquake in Haiti [30]. Similar to the work here, they identified only four original users of Twitter from within Haiti. The analysis of the interaction on Twitter is an interesting dimension to processing information that is not covered in our work. However, with only four people sharing information from within Haiti, we do not share their conclusions that this necessarily represents a new method for crisis information processing, especially in light of the evidence here that the nature of Twitter communications is very different to other channels, even other short message systems.

Much recent work on Twitter has been in sentiment analysis [23, 1, 29], unsupervised event detection [27], and controlled vocabularies of Twitter hashtags [31, 32].

[3] Moreover, it looks like no one has *ever* looked at automated processing of email in a social development context, despite email being our single largest form of remote communication and having the longest history in automated filtering with a relatively large body of work on spam-detection and more recent work looking at the ENRON corpus [26]. We regret not also attempting to address this imbalance in this paper.

	Precision	Recall	F-value
Haiti			
Kreyol	0.882	0.957	0.918
Kreyol with Subword Models	0.925	0.997	0.960
English	0.921	0.998	0.958
English with Subword Models	0.921	0.998	0.958
Pakistan			
Urdu	0.643	0.857	0.735
Urdu with Subword Models	0.692	0.900	0.783
English	0.654	0.895	0.756
English with Subword Models	0.663	0.982	0.791

Table 3: A comparison of the effects of subword models on accuracy, showing significant increases in accuracy in both Kreyol and Urdu. While the results for English in Haiti are identical in accuracy, the actual predictions differed slightly.

7. CONCLUSIONS

Twitter is not a stand-in for SMS, or at least, the two have been used in very different ways to-date, meaning that lessons learned from one do not necessarily carry over the other. The same is true for languages. Automated systems that are evaluated on English alone will not necessarily arrive at the optimal strategies for the languages of a crisis-affected population.

However, the lack of overlap between SMS and Twitter may be encouraging from an operational standpoint. If they are used by different actors in different ways, then they are likely to contain complementary information about an emerging crises. From a qualitative analysis, this is true of the communications studied here – the most frequent use-case for SMS in both contexts were reports about conditions on the ground. The most frequent use-case for Twitter in Haiti was to inform the international community how they could help. The most frequent use-case for Twitter in Pakistan was to request help from people within Pakistan.

While the SMS were taken in whole or randomly sampled, the Twitter data was limited to that which response organizations considered important, so if anything it should be biased *in favor* of ground-based reports, but nonetheless they are very rare – at a best guess much less than 1 in 10,000 about each event. However, there is no single user-group for any platform and there were individual reports from both Twitter and SMS, organizational reports from both, and English, Kreyol and Urdu from both. So while the majority use of Twitter has not been for citizen reporting from within a crisis, that use is still attested in the data.

For natural language processing, while there was not one solution that produced the most accurate results, there was a consistent increase in accuracy when employing subword models and learning priors over the source. Most importantly, the results here highlight the need for research into Natural Language Processing for low resource languages. If there is not a way to efficiently filter large volumes of communications at short notice, then an important information source is lost.

Acknowledgements

With thanks to the anonymous reviewer who suggested that topic-divergence was not necessarily evidence that different models were required – our domain dependence research grew from this. The first author was supported by a Stanford Graduate Fellowship.

8. REFERENCES

[1] P. P. Alexander Pak. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceeding of the 2010 International Conference on Language Resources and Evaluation (LREC 2010)*, 2010.

[2] R. Beaufort, S. Roekhaut, L. Cougnon, and C. Fairon. A hybrid rule/model-based finite-state framework for normalizing sms messages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 770–779, 2010.

[3] C. Caragea, N. McNeese, A. Jaiswal, G. Traylor, H.-W. Kim, P. Mitra, D. Wu, A. H. Tapia, L. Giles, B. J. Jansen, and J. Yen. Classifying text messages for the haiti earthquake. In *Proceedings of the 8th International Conference on Information Systems for Crisis Response and Management (ISCRAM2011)*, Lisbon, Portugal, 2011.

[4] M. Choudhury, R. Saraf, V. Jain, A. Mukherjee, S. Sarkar, and A. Basu. Investigation and modeling of the structure of texting language. *International Journal on Document Analysis and Recognition*, 10(3):157–174, 2007.

[5] P. Cook and S. Stevenson. An unsupervised model for text message normalization. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 71–78. Association for Computational Linguistics, 2009.

[6] G. V. Cormack, J. M. G. Hidalgo, and E. P. Sánz. Feature engineering for mobile (SMS) spam filtering. In *The 30th annual international ACM SIGIR conference on research and development in information retrieval*, 2007.

[7] S. Garrett. Big goals, big game, big records. In *Twitter Blog (http://blog.twitter.com/2010/06/big-goals-big-game-big-records.html)*,

2010.

[8] S. Goldwater, T. L. Griffiths, and M. Johnson. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54, 2009.

[9] M. Healy, S. J. Delany, and A. Zamolotskikh. An assessment of case-based reasoning for Short Text Message Classification. In *The 16th Irish Conference on Artificial Intelligence & Cognitive Science*, 2005.

[10] J. M. G. Hidalgo, G. C. Bringas, E. P. Sánz, and F. C. García. Content based SMS spam filtering. In *ACM symposium on Document engineering*, 2006.

[11] N. Hodge. Texts, Tweets Saving Haitians from the Rubble. *Wired Magazine*, 2010.

[12] S. Isbrandt. Cell Phones in West Africa: improving literacy and agricultural market information systems in Niger. White paper: Projet Alphabétisation de Base par Cellulaire, 2009.

[13] ITU. The world in 2010 - the rise of 3G. In *International Telecommunication Union*, 2011.

[14] A. Jagun, R. Heeks, and J. Whalley. The impact of mobile telephony on developing country micro-enterprise: A Nigerian case study. *Information Technologies and International Development*, 4, 2008.

[15] C. Kobus, F. Yvon, and G. Damnati. Normalizing SMS: are two metaphors better than one? In *The 22nd International Conference on Computational Linguistics*, 2008.

[16] C. Leach-Lemens. Using mobile phones in HIV care and prevention. *HIV and AIDS Treatment in Practice*, 137, 2009.

[17] W. Lewis. Haitian Creole: How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes. In *14th Annual Conference of the European Association for Machine Translation*, 2010.

[18] W. Lewis, R. Munro, and S. Vogel. Crisis MT: Developing A Cookbook for MT in Crisis Situations. In *Annual Workshop on Machine Translation, EMNLP*, Edinburgh, 2011.

[19] F. Liu, F. Weng, B. Wang, and Y. Liu. Insertion, deletion, or substitution? normalizing text messages without pre-categorization nor supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011.

[20] R. Munro. Crowdsourced translation for emergency response in Haiti: the global collaboration of local knowledge. In *AMTA Workshop on Collaborative Crowdsourcing for Translation*, 2010.

[21] R. Munro. Subword and spatiotemporal models for identifying actionable information in haitian kreyol. In *Fifteenth Conference on Natural Language Learning (CoNLL)*, Portland, OR, 2011.

[22] R. Munro and C. D. Manning. Subword variation in text message classification. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010)*, Los Angeles, CA, 2010.

[23] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[24] G. Peevers, G. Douglas, and M. A. Jack. A usability comparison of three alternative message formats for an SMS banking service. *International Journal of Human-Computer Studies*, 66, 2008.

[25] D. Pennell and Y. Liu. Normalization of text messages for text-to-speech. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4842–4845. IEEE, 2010.

[26] K. Peterson, M. Hohensee, and F. Xia. Email formality in the workplace: A case study on the enron corpus. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics ACL 2011*, page 86, 2011.

[27] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010)*, 2010.

[28] Pingdom. Internet 2010 in numberse. In *Royal Pingdom Blog (http://royal.pingdom.com/2011/01/12/internet-2010-in-numbers/)*, 2011.

[29] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. Short text classification in twitter to improve information filtering. In *Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval*, 2010.

[30] K. Starbird and L. Palen. Voluntweeters: Self-organizing by digital volunteers in times of crisis. In *ACM CHI Conference on Human Factors in Computing Systems*, Vancouver, CA, 2011.

[31] K. Starbird and J. Stamberger. Tweak the tweet: Leveraging microblogging proliferation with a prescriptive syntax to support citizen reporting. In *Proceedings of the seventh international ISCRAM Conference*. ACM, 2010.

[32] K. Starbird and J. Stamberger. Tweak the Tweet: Leveraging Microblogging Proliferation with a Prescriptive Syntax to Support Citizen Reporting. In *Proceedings of the 7th International ISCRAM Conference*, 2010.

[33] S. Stenner, K. Johnson, and J. Denny. Paste: patient-centered sms text tagging in a medication management system. *Journal of the American Medical Informatics Association*, 2011.

[34] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *In Advances in Neural Information Processing Systems*, 17, 2005.

[35] S. Vieweg, A. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1079–1088. ACM, 2010.

[36] Z. Xue, D. Yin, and B. D. Davison. Normalizing microtext. In *Proceedings of the AAAI Workshop on Analyzing Microtext*, 2011.