Crisis MT: Developing A Cookbook for MT in Crisis Situations

William D. Lewis
Microsoft Research
Redmond, WA 98052
wilewis@microsoft.com

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

Stephan Vogel
Carnegie Mellon University
Pittsburgh, PA 15213
stephan.vogel@cmu.edu

Abstract

In this paper, we propose that MT is an important technology in crisis events, something that can and should be an integral part of a rapid-response infrastructure. By integrating MT services directly into a messaging infrastructure (whatever the type of messages being serviced, *e.g.*, text messages, Twitter feeds, blog postings, etc.), MT can be used to provide first pass translations into a majority language, which can be more effectively triaged and then routed to the appropriate aid agencies. If done right, MT can dramatically increase the speed by which relief can be provided. To ensure that MT is a standard tool in the arsenal of tools needed in crisis events, we propose a preliminary *Crisis Cookbook*, the contents of which could be translated into the relevant language(s) by volunteers immediately after a crisis event occurs. The resulting data could then be made available to relief groups on the ground, as well as to providers of MT services. We also note that there are significant contributions that our community can make to relief efforts through continued work on our research, especially that research which makes MT more viable for under-resourced languages.

1 Introduction

The connected world contains approximately 5000 languages – at least that is how many languages you could find at the other end of your phone right now. However, the majority of these languages are under-resourced, and they have few or no digital resources. In the event of a sudden onset crisis, people will immediately begin using their communication technologies – and their languages – to report their situations, request help, and seek out loved ones. Yet, in the event that such a crisis occurs in a region of the world where an under-resourced language is spoken, delivery of support or aid could be affected due to the inability to communicate. This was felt most strongly in the wake of the January 12, 2010 earthquake in Haiti. Local emergency response services were inoperable, but 70-80% of cell-towers were quickly restored. With 83% of men and 67% of women possessing cellphones, the nation remained largely connected. People within Haiti were texting, calling, and interacting with social media, primarily in Haitian Kreyòl (Munro, 2011). Yet, most of the aid that was being delivered to the country – initially, soley by the American Military – was being delivered by groups that did not communicate in Kreyòl. It was the first time that the world has seen a large-scale sudden onset crisis in a region with productive digital communications in an under-resourced language, but it certainly will not be the last.

We strongly believe that MT is an important technology to facilitate communication in crisis situations, crucially since it can make content in a language spoken or written by a local population accessible to those that do not know the language, in particular aid agencies. Multiple groups saw MT as a grand challenge in the Haitian crisis, and they set to work to make MT available as soon as possible after the crisis. Within two weeks of the crisis, the first two MT engines were built and were available to those who needed them. We believe that we can make MT available just as quickly in future crises, and, with the right preparation, tightly integrate MT into the communication infrastructure that is deployed (*e.g.*, the text messaging infrastructure). The challenge is doing the work now to make this vision possible.

In this paper, we describe the technologies that came to play in the Haitian crisis, how Haitian Kreyòl MT was developed, the problems of surprise languages and low resource MT, and detail the research and technologies, cast as a “Crisis MT Cookbook”, that will be essential for MT to form a core role in future crises. In Sections 2, 3, and 4 we discuss Mission 4636 and the technologies that came

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 501–511,
Edinburgh, Scotland, UK, July 30–31, 2011.

into play in Haiti and other recent crises, and the role that technologies can and should play in future crises. In Section 5, we discuss what made Haitian Kreyòl a special case of a "surprise language", and how MT was developed for the language. In Section 6, we review the NLP and MT research areas that will likely net big returns for under-resourced languages. In Section 7, we review the need for an MT Crisis Cookbook, and what the data and infrastructural components of the Cookbook should be. Finally, in Section 8 we review a sample crisis timeline, and how a crisis might play out with all the components of the Cookbook available. Section 9 wraps up the paper.

2 Mission 4636

In Haiti, crowdsourced translation enabled communications between the Kreyòl-speaking Haitian population and English-speaking emergency responders. A small group of international aid workers established a phone-number,'4636'[1] , that people were able to send text messages to for free within Haiti. The actual translations were made by about 2000 Kreyòl[2] and French speaking volunteers collaborating on an online microtasking platform that they used to translate, categorize, identify missing people and geolocate information on a map (Munro, 2010).[3] After a month, this work was gradually transferred to paid workers in Mirebalais, Haiti. These messages, about 80,000 in total, were used as part of the shared task for the *2011 Workshop on Machine Translation*. About 3,000 of the messages had the categories and coordinates refined by a third workforce working with the Ushahidi platform out of Boston.[4] They published this information on an online crisis map and worked directly with the main emergency responder, the American Military, to identify actionable information.

The strategy for translation was extremely effective - 80,000 messages equates to about 10 novels of information, translated in real-time, lifting a burden off people in Haiti. One high-ranking official described the translation process as a "perfect match" of social media and traditional emergency response (Anderson, 2010).

To meet the scale of translation needs, machine translation services were quickly shipped. A member of Mission 4636 built a high-precision, low-coverage dictionary-based system that was used by a number of translators. A couple of days later, the world's first publically accessible Stastical Machine Translation (SMT) engine for Kreyòl was developed by Microsoft Research, with Google Research following several days later with their own engine.[5] Although the statistical translation engines were not used directly in the SMS translation effort, there is evidence they were used by those who were involved in the relief effort, as determined by blog postings and a review of translation logs showing relief-centric translations. Although Kreyòl is not a high traffic language—it was not expected that it would be—about 5% of the traffic in the weeks and the months following the earthquake appeared to be relief-related, suggesting that machine translation was being used those who needed it most.[6] Had MT been integrated directly into the text messaging infrastructure used in Haiti, this percentage would have been significantly larger.

3 Translation and crisis response - a quickly changing field

To establish a ready-workforce to aid information processing in relief efforts an organization called the

[1] See the "Mission 4636" website at http://www.mission4636.org for more information about the organization and its efforts in Haiti.

[2] We use the term Kreyòl for the Creole spoken in Haiti to differentiate it from other Creoles. This is also in concord with customary usage in Haiti.

[3] An author of this paper, Robert Munro, coordinated this process and is a founding member and translation coordinator for the Standby Task Force, which is discussed later in the paper.

[4] For more information on Ushahidi, see http://www.ushahidi.org.

[5] A rough timeline of these developments can be seen in the commentary posted to the Language Log website (see specifically the archive at: http://languagelog.ldc.upenn.edu/nll/?p=2068).

[6] The logs output by Microsoft Translator's engine were examined, and categorized roughly into broad categories describing the type of content. These categories were: Relief Related (suspected), Colloquial or Common Expressions (which could, in fact, have been relief related), Chat, and Unknown. The analysis was done by hand on a random sample of 200 messages from the many thousands of messages received within a couple of months of the quake. There were a large number of strings that were difficult to categorize, including many partial strings, and a bias against Relief Related when it was not clear. Thus, the 5% estimate is likely a conservative one.

Standby Task Force was established in late 2010. Its founding members had worked together in the Haiti and/or subsequent Pakistan response efforts. It currently has several hundred members who specialize in tasks like report mapping, verification, media monitoring and translation. Of all the different tasks that volunteers can perform, translation is the *least* transferable from one crisis to the next.

Following from the lessons learned in Haiti, crowdsourced and machine translation have been combined for a number of aid efforts: vote monitoring for the referendem in Southern Sudan (Arabic); a UN-led earthquake simulation in Colombia (Spanish); and for crisis mapping following the tsunami in Japan (Japanese).

When information is immediately translated into a high resource language it can be quickly triaged by a greater number of people. The more time-intensive task of manually correcting any mistranslations can be performed in parallel. This workflow of combining machine and crowdsourced translation is largely a succesful one and is likely to become common practice in humanitarian information processing.

The combination of manual and machine-translation was found to be effective across unpredictable input:

> "An email came into the Sudan Vote Monitor platform in Indonesian - your plugin did a good job of translating it into English and Arabic"
>
> Helena Puig Larrauri, volunteer for Sudan Vote Monitor (P.C.)

But not without errors, especially across vital phrases like location names:

> "Names of neighborhoods such as Salitre or Puerta al Llano were not recognized as such and unnecessarily being translated."
>
> Marta Poblet, volunteer for Colombia earthquake simulation (P.C.)

When the uprisings hit Libya in early 2011 the United Nations did not have the capacity to collect vital ground-truth data in the lead up to their involvement. Information about refugee numbers and needs were on web-accessible articles and social media, as were reports about the movements of government and rebel troops and vunerable populations within the country. But there simply wasn't the workforce within the UN to aggregate and verify so much information. This was the first time the United Nations directly engaged a volunteer workforce for large-scale information processing, requesting the Standby Tasks Force's deployment. It was also the first time that so much information had come from social media, a potentially large but unstructured data source, but it gave the UN a huge head-start in their efforts (Verity, 2011). Crowdsourced and machine translation were also combined here, but in this case by directly engaging Arabic speakers in media monitoring and by using reports from *Meedan*.[7]

In a crisis, it will now be more common than not that the volume of available digital information will surpass the volume of information that aid-workers can collect directly from the ground. This rapid change is being quickly met by a rapid change in cloud-based and automated solutions to language processing, especially machine translation.

4 Translation and low-resource languages

We were fortunate that Arabic, Spanish and Japanese are high resource languages for which online machine translation services already exist. Speakers of low resource languages cannot currently benefit from this kind of translation service and yet low resource languages are disproportionally spoken by the world's most vunerable populations. Over the last 12 months many problems have been solved regarding the workflow of managing crisis data, but one of the biggest remaining problems is the ability to quickly deploy machine-translation systems to augment relief efforts.

While translation is not widely discussed aspect of crisis response, it is "a perennial hidden issue" (Disaster 2.0, 2011):

> "Go and look at any evaluation from the last ten or fifteen years. 'Recommendation: make effective information available

[7] *Meedan* is an NGO that seeks to create greater understanding between the Arabic and English speaking world by translating media reports and blogs between the languages, combining quick machine-translation with corrections by a volunteer community.

> to the government and the population in their own language.' We didn't do it . . . It is a consistent thing across emergencies."
>
> Brendan McDonald, UN OCHA in (Disaster 2.0, 2011)

Beyond the particular use case of small-to-medium scale emergency information processing, machine translation can also contribute to aid efforts when the scale of information is beyond any manual processing. In addition to the Libya deployment, a recent Red Cross survey (2010) found that nearly half the respondents would use social media to report emergencies. It simply would not be possible to translate all real-time reports when expressed through social media, but translation into a high resource language could aid semi-automated methods for discovering and prioritizing information.

There is, therefore, a great need to explore methods for rapid deployment of machine-translation systems into minority languages. The questions that we seek to address in this paper is how we as a community can prepare for the eventuality of the next crisis, can draw from the lessons we learned in the Haitian crisis, and might significantly impact the aid effort in the next and future crises.

5 Surprise Languages: What Made Haiti Different?

On January 19th, 2010, the Microsoft Research Translator team received an e-mail from the field requesting that they develop an MT engine for Haitian Kreyòl to assist in the relief effort. At the time, no publically available MT engine existed for Kreyòl. In less than five days, the Microsoft Translator site was supporting the language. Given that it can take weeks to months to develop an MT engine for a new language, it would not seem possible that an engine could be developed so quickly, especially for a low-resource, minority language. The reasons this was possible are varied, and are in some ways unique to Kreyòl.

Haitian Kreyòl, as it turns out, has proven to be an exceptional case for a surprise language. Unlike the languages in Surprise Language Exercises of nearly a decade ago (Oard, 2003; Oard and Och, 2003), in which participants were given a month to collect data and build language technologies for previously unknown languages, including Machine Translation systems, there was a surprising amount of data for Kreyòl at the start of the Haitian crisis, and it became available relatively quickly. Partly, this is due to the growth of the Web, which has proven to be a surpisingly diverse multi-lingual resource. But it also stems crucially from work that had been done in the past on Kreyòl, specifically, the work that was done in the DIPLOMAT and NESPOLE! projects at CMU (Frederking et al., 1997). It was possible to assemble a reasonable sample of data for the language in very short order (*i.e.*, days). Further, since the language itself is fairly reduced morphologically, it is an easier target for SMT. In contrast, if one were to sample a language at random from the set of the 7,000 languages spoken on the earth, one is more likely to find a language that is morphologically richer (*e.g.*, fusional, aggutinating, polysynthetic). Morphological richness compounds the data sparsity problem, reducing the quality of the resulting SMT engines.

In other words, a combination of a simple morphology combined with reasonably accessible sources of data made the rapid deployment of MT for Kreyòl far more likely. That is not to say that there weren't problems. First, Kreyòl is fairly "young" as a written language[8], and is still in the early stages of orthographic standardization and normalization (Allen, 1998). This has led to inconsistencies in the orthography that increases data sparseness and noise. Further, Kreyòl has multiple registers in its written form: a "high" register that uses full forms for pronouns and a set of function words, and a "low" register that corresponds more closely to its spoken form, and is written with many contractions. For example, the Kreyòl word for the first person pronoun is *mwen*. It can be written as *mwen* (the high register), or contracted to *m'* (the low register). The form can either be attached to the succeeding word or written with a following space. Likewise, the first person possessive is also *mwen* which is written following the word that is possessed. This

[8] Although Haitian Kreyòl in written form goes back as far as the late 18th century (see Lefebvre (1998) for material on some of these texts), Kreyòl as a written language did not become more commonplace until the 20th century, not achieving official status in Haiti until 1961.

can be written as *'m*, and can be attached to the word or delimited by a space. Both *m'* and *'m* appear in some texts as just *m*. The same patterns hold for all pronouns, and some function words as well. See Table 1 for a list of these reductions.

Table 1: Sample Pronouns and Reductions

Pronoun	Gloss	Appears as
mwen	I, me, mine	m, 'm, m'
nou	you (pl), us	n, 'n, n'
ou	you	w, w'
li	he, she, it	l, l', 'l

Additionally, writers of Kreyòl use a large number of abbreviated forms for common expressions, a kind of shorthand. For example, *avèn* can be used to represent *avèk nou*, *mandem* can be used for *mande mwen*, etc. Overall, the number of alternations and multi-way ambiguities also increases the level of noise and data sparsity. [9]

So, even with a morphologically reduced language like Kreyòl, one has issues with data sparsity beyond the mere lack of availability of data. This compounds the low-data aspect of the language. Adding in a multitude of morphological variants, as one might encounter in a Turkic language, or worse, in an Inuit language, would only make the problem more severe. The big challenge for Crisis MT is not only to deal with the data availability problem, but once one has the data in hand, to deal with the reduction in the utility of that data caused by noise and the multiplication of word forms. These pose major challenges to our community, which can be countered through additional research, a motivated and active community, and scores of rapidly applied heuristics and data repairs.

6 Research Areas to Counter Data Sparsity

As noted, the major problems with low-resource MT is the lack of data and various data issues that increase the sparsity of data already in short supply. What are the research challenges? How can we make MT viable quickly for low-resource and simultaneously morphologically rich languages?

The following constitutes a rough list of solutions, many of which map to very interesting research problems:

- Crowdsourcing – Beyond the use of crowdsourcing in the crisis context itself (*e.g.*, to translate or process text messages, much as what was done by Mission 4636), novel techniques for tapping the crowd could also be used to add or repair data:
  - Repairing and evaluation – In this scenario, the crowd would be used to repair data that is obviously noisy, evaluate problems with particular data points, or even make simple determinations as to whether the data in question is actually in the language(s) of interest or too noisy to use.
  - Translating content, generating new data – Given crowd sourced, micro-tasking platforms such as Amazon's Mechanical Turk and Crowdflower, one can now easily tap the crowd to generate new data. The major challenge will be identifying if speakers of the target language(s) are available on the desired platform, and if not, if they could be motivated to particpate.[10] Likewise, infrastructure and resources will be needed to evaluate the quality of the resulting translations (Zaidan and Callison-Burch, 2011).
  - Active Crowd Translation – This method combines active learning with crowdsourcing for annotation of parallel data in comparable resources, and can be used to increase the amount of data that is found (Ambati et al., 2011). Active learning might be applicable to other crowdsourcing tasks as well, such as being used in crowdsourcing for translating content or repairing translated content.
- Tapping non-traditional sources – Critical to traditional approaches of SMT is parallel training data. Parallel data is difficult to impossible to come by for a large number of the world's

[9] For more details of the Haitian Kreyòl translation systems developed at Microsoft Research, please see Lewis (2010).

[10] Based on the results of an informal survey, there may be speakers of a hundred or more languages on Mechanical Turk. See http://www.junglelightspeed.com/amt_language/ for a list of the languages that may be available on Turk.

languages. Tapping non-traditional sources of data can help increase the supply of ever valuable training data for a language:

  - Mining comparable sources of data – mining comparable data for parallel data has a long history, including mining comparable sources for named entities (Udupa et al., 2009; Irvine et al., 2010; Hewavitharana and Vogel, 2008; Hewavitharana and Vogel, 2011), mining Wikipedia for parallel content, including sentences (Smith et al., 2010), and many more too numerous to list. There is always room for improvement and hybridization in this space, as well as tapping additional sources of data, such as the volumes of noisy comparable data on the Web.
  - Monolingual – More recent work has focused on mining monolingual sources of data, treating MT as a decipherment problem (Ravi and Knight, 2011), rather than a source-target mapping problem.
  - Dictionary bootstraps and backoffs – Despite the absence of context, dictionaries can be useful, especially for resolving out-of-vocabulary items (OOVs). Many bilingual dictionaries also contain example sentences, which can be harvested and used in training.
  - Field data from linguists – Given that linguists have variously studied a large percentage of the world's languages, tapping the supply of data that they have accumulated could prove quite fruitful. Some recent work tapping annotated bitexts (at this time, for over 1,200 languages) produced by linguists may prove useful in the future (Lewis and Xia, 2010), if for nothing more than to provide information about linguistic structure (*e.g.*, morphological complexity or divergences, potential distortion rates, and structural divergence (*a la* Fox (2002))). Engaging with the documentary linguistic community and providing tools to facilitate the collection of data might produce additional data, especially data where alignment is assisted through human input (Monson et al., 2008).

- Novel ways of countering data sparsity
  - Systematizing data cleaning heuristics – Undoubtedly, the same kinds of filtration and data cleaning heuristics used for Kreyòl could prove useful for speeding up the processing of data for new languages. Applying Machine Learning techniques to data filtration and data cleaning could aid and generalize the process, thus decreasing overall latency from acquisition to training.
  - Strategies to make the source look more like the target (or vice versa) – A corollary to data sparsity is faulty word alignment, where low frequency words fail to get good alignments because there is not enough data to reinforce fairly weak hypotheses, or where source-target distortion is high. Both problems disfavor what alignments do exist. If the source and target are reordered so that one side more closely matches the other, or one side is "enriched" to be more like the other, one can reduce distortion related effects, and might also counter the large number of forms in morphologically rich languages (*e.g.*, (Yeniterzi and Oflazer, 2010; Genzel, 2010), and many others).

- Strategies to systematically deal with complex morphology – this is one on-going area of research that could still net large returns, since, even with some relatively high-data languages, such as Finnish, data is made sparser due to the multiplication of possible forms. There is too long a literature to really do justice here, but some recent work includes discrimitative lexicons (Jeong et al., 2010), sub-word alignment strategies (Bodrumlu et al., 2009), learning the morphological variants in a language (Oflazer and El-kahlout, 2007), using off-the-shelf morphological tools, *e.g.*, Morfessor [11], etc.

- Use syntax or linguistic knowledge in the translation task – By reducing the hypothesis space for possible alignments, syntax-based

[11] http://www.cis.hut.fi/projects/morpho/

approaches can do better in lower-data situations and can handle source-target discontinuities better than straight phrase-based systems (*e.g.,* (Quirk and Menezes, 2006; Li et al., 2010)).

7 The MT Crisis Cookbook

Given the relatively narrow domain context of Crisis MT—generally the needed vocabulary and data should be centered on relief work, medical interactions, and communicating with the affected populations—it may be possible to approach Crisis MT as we would MT for any domain (*e.g.*, news, government, etc.). With enough data relevant to a particular domain or sub-domain (*e.g.*, earthquake, tsunami, nuclear disaster, flooding, etc.), it would be possible to build the relevant translation memories (TMs) and train highly domain-specific MT engines to produce translations of reasonable quality and utility. Even with highly inflected languages, a domain-specific approach may get around many of the data sparsity issues.

It is also crucial that no data be thrown out. Relief specific content that was relevant to an earlier crisis can certainly contribute to subsequent crises. Among these data are difficult to replicate sources of data, such as SMS messages. This data would constitute a highly domain specific set of data which would only grow over time.

7.1 Outline of the Cookbook

The recipe for the MT Crisis Cookbook consists of two parts:

1. The **content** that would be most useful in crisis situations. This consists of relief-centric vocabulary, phrases, sentences, and other material. It should be in some common "source" language, likely English (English is a reasonable "pivot" in and out of many other languages, given the ubiquity of English-to-X content).
2. The **infrastructure** to support relief workers, aid agencies, and the affected population. As made obvious in Haiti, an SMS messaging infrastructure integrated into a crowd-sourced translation infrastructure, proved to be crucial. For future crises, this infrastructure should be streamlined and have public MT APIs integrated directly into it (to support first pass MT).

7.2 Cookbook Data

As noted in Section 5, one way to counter the data sparsity problem is to build domain specific engines, with a set of data ready-to-go in the event of a crisis. This data, which would exist in English and possibly other languages, would be translated into the target language (if needed), distributed to to aid organizations (as needed), and used to train MT engines and other language processing resources. The following list constitutes a set of possible sources. It is by no means complete (for instance, some resources specific to particular crisis types, *e.g.*, floods, nuclear disasters, etc. are not included), but it does represent a good central core of resources that should be part of any Crisis Cookbook[12] :

- Where There is No Doctor – This is one of the most recognized and widely used and useful references in under-resourced regions around the world. The publisher of the text, the Hesperian Foundation, has already had the text translated into 75 languages, and it is available in PDF as a free download from their website.[13]
- CMU Medical Domain Phrases, Sentences, and Glossary – Collected under the jointly NSF/EU funded NESPOLE! and DIPLOMAT projects (Frederking et al., 1997), this data consists of common phrases and sentences that would be useful in a crisis medical scenario, and would be quite useful for training MT, as it was for training the Kreyòl engines. Only the English side of this data would be relevant to future crises.
- Anonymized Crisis-related SMS Messages – Relief-related SMS messages may be particularly useful in future crises, since those collected in a crisis scenario are likely to contain content that transfers readily to similar crises. A selected sample of the 80,000+ messages resulting from the Haitian crisis could constitute

[12] Some of the resources listed here are under copyright. There may need to be some negotiation with the copyright owners to ensure that the texts can be used, and how they can be used (*e.g.*, to train MT, to be used in TMs, to be distributed in hardcopy form, etc.).

[13] http://hesperian.org/

a reasonable core of SMS messages that could be added to over time.

- Red Cross Emergency Multilingual Phrasebook – A small, but highly focused, set of phrases and questions useful in an emergency medical context. Available in multiple languages.
- Emergency and Crisis Communication Vocabulary – An example bilingual set was prepared by the Canadian Government in both French and English[14] , consisting of a small list of "official" terms needed in crisis situations, and their associated descriptions. Although the terms on the Canadian site are translated and defined only in English and French and have a bias to the Canadian government nomeclature, having such a list of terms from multiple government agencies and their definitions could prove useful for relief vocabulary as well as for vocabulary needed for official announcements.
- High Frequency Wikipedia Disaster Content – This would consist of vocabulary that recurs across multiple related crisis pages on wikipedia. The idea is to harvest those terms that repeat across multiple pages of the same "sub-domain" (*e.g.*, those that cover events with floods, earthquakes, nuclear disasters, etc.), but document disasters in different locales, where cross-page repeated vocabulary is favored (substracting out high-frequency vocabulary that occurs elsewhere). This vocabulary could be distilled automatically from a set of relevant pages, and would likely contain core vocabulary for specific crisis and disaster contexts. For instance, shared vocabulary between the Japanese, Indonesian, Pakistani, and Haitian Earthquake pages might contain a reasonable set of vocabulary relevant to earthquake crises as a whole.

7.3 Cookbook Infrastructure

The Cookbook infrastructure draws directly on what was found to be useful in the Haitian Crisis. Here are the infrastructural components we see as crucial:

- A crowd sourced microtasking infrastructure to translate and route messages from the field. This proved to be essential in Haiti. Having such an infrastructure ready-to-go for future crises would shave days off implementation and likely have profound effects on the rapidity of the response.
- Integration of the APIs for the publically available MT services, such as Microsoft Translator and Google Translate, into the microtasking and messaging infrastructure, enabling processing of SMS messages, Twitter feeds, etc. In this way, when any of these services deploy MT for a given crisis language, the switch can be flipped and first-pass can be MT activated at a moment's notice.
- A ready-to-go smart phone app that acts as a crisis Translation Memory, which can be populated with Cookbook content as it becomes available. In this manner, rather than relying on the distribution of paper copies of Cookbook materials, relief workers on the ground could just sync-up their mobile devices to get the latest content. This is particularly important in crisis locales where "data plan" access is limited, and phones will thus not necessarily have online access to cloud based resources on a regular basis.

8 A Sample Crisis Timeline

The following timeline is only meant to demonstrate what might be possible with the right infrastructure in place and the community fully engaged. The mantra of "every crisis is different" applies, and this timeline should not be interpreted as a "cookbook" for a future event. All place and entity names are intended to add realism; there was no intention to leave anyone in or out.

Day 0 – A massive earthquake hits the island nation of Palladi.

Day 1 – The first aid organizations arrive on the island with food and humanitarian aid, although only the two major cities are directly accessible. Thousands of Palladians are not reachable by aid organizations, and the exact numbers that are affected and their locations are not known.

[14] http://www.btb.gc.ca/publications/documents/crise-crisis.pdf

The native population of Palladians is nearly 80% monolingual. There is a dire need for Palladian interpreters, but also of translated Palladian content. Notified of the need for Palladian translations, MT community volunteers begin efforts to collect and license data in Palladian. The relief community responds by activating the crowd sourcing infrastructure used in other relief scenarios. Researchers and disaster response teams are notified at Microsoft Research and Google Research of the critical need for crisis content to be translated into Palladian. Native Palladian speakers are being looked for by all parties.

Day 2 – As with the Haitian crisis, a text messaging infrastructure is put in place such that text messages can be received from the population and routed to a crowd of rapidly assembling volunteers. Since there is some internet access, including via mobile phones, twitter feeds are monitored. Until messages start arriving, a small crowd of Palladian speakers begin translating content into Palladian, focused specifically on the Cookbook and off-the-shelf SMS content.

The first text messages start arriving by late afternoon. These text messages are routed directly to the text messaging and microtasking infrastructure. The small but growing crowd of Palladian translators begin translating this growing tide of messages.

Day 3 – The humanitarian information processing community, with the support of many organizations and volunteers, releases the first sections of the Crisis Cookbook. The Crisis Cookbook is transmitted directly to aid organizations on the ground in Palladi, and soft- and hard-copies are distributed to aid workers as quickly as feasible.

AT&T puts into place several cell towers with satellite connectivity for areas that do not have cell coverage. Within hours, text and twitter messages from the field increase dramatically.

Day 4 – Microsoft and Google release the first versions of their Palladian-English translators, with ready access via their public APIs. Since the text messaging infrastructure already has both APIs integrated directly into the microtasking and message processing infrastructure, both engines are activated immediately, and all messages are translated first by one or the other engine, and the MT'd content along with the original message are handed to volunteers.[15] Translations are repaired, and routed directly to aid organizations, and to the Google and Microsoft teams (for retraining models).

Day 5 – Additional cookbook materials are translated. Researchers at Johns Hopkins locate a stash of Palladian data at the Palladian Central University. This data is posted at the CMU site, and is immediately consumed by all parties working on the MT problem.

Day 6 – Researchers at University of Edinburgh develop a novel algorithm for dealing with Palladian vowel harmony, which has been a major problem with Palladian MT, since data sparsity is exacerbated by the problem. The Edinburgh researchers publish the algorithm immediately to their Web site, and notify both Microsoft and Google.

Day 10 – Armed with algorithmic improvements and an increasing volume of data, machine translated content is now achieving sufficient quality to warrant passing it directly to aid organizations. Palladian volunteers now work principally on the hard to translate cases (those with high OOVs), and on post-response data clean-up. The fruits of their labor result in iterative improvements on the various MT engines that have been deployed.

Day 11+ – The deployment of language technologies, specifically MT, in the Palladian crisis results in saving untold thousands of lives. The lessons learned in the Palladian earthquake will be applied to future crises, and the translated content produced by volunteers will be added to the cookbook for use in the next crisis.

[15] Determining which engine to send translations to is a problem that should be resolved in advance. A combination of either random selection or on-the-fly OOV calculations could be used to determine routing.

9 Conclusion

In this paper, we propose that MT is an important technology in crisis events, something that can and should be an integral part of the rapid-response infrastructure. By integrating MT services directly into a messaging infrastructure (whatever the type of messages being serviced, *e.g.*, text messages, Twitter feeds, blog postings, etc.), MT can be used to provide first pass translations into a majority language, which can assist in triaging messages and routing them to appropriate aid agencies. If done right, MT can dramatically increase the speed by which relief can be provided. To ensure that MT is a standard tool in the arsenal of tools used in crisis events, we propose a preliminary *Crisis Cookbook*, the data contents of which could be translated into the relevant language(s) by volunteers immediately after a crisis event takes place. The resulting data can then be made available to relief groups on the ground, as well as to providers of MT services. We also note that there are significant contributions that our community can make to relief efforts through continued work on our research, especially that research which makes MT more viable for under-resourced languages.

Credits

This paper is dedicated to the thousands of volunteers who worked selflessly for many, many hours in aid of the people of Haiti. Without their help, many hundreds more would have perished. We also wish to express our deepest appreciation to all those who have devoted their lives to aid people in need, especially the first responders in crisis events. It is our sincerest hope that that the small measures our community can take to assist in relief efforts will help make your jobs more effective, and that our efforts will ultimately assist you and those you strive to help.

References

Jeffrey Allen. 1998. Lexical variation in Haitian Creole and orthographic issues for Machine Translation (MT) and Optical Character Recognition (OCR) applications. In *Association for Machine Translation in the Americas (AMTA) Workshop on Embedded MT Systems: Design, Construction, and Evaluation of Systems with an MT Component*, Langhorne, Pennsylvania.

Vamshi Ambati, Sanjika Hewavitharana, Stephan Vogel, and Jaime Carbonell. 2011. Active Learning with Multiple Annotations for Comparable Data Classification Task. In *Proceedings of ACL 2011*, Portland, Oregon, June.

Sharon Anderson. 2010. Talking with Adm. James G. Stavridis Supreme Allied Commander, Europe Commander, U.S. European Command. *CHIPS - The Department of the Navy Information Technology Magazine*, 28.

Tugba Bodrumlu, Kevin Knight, and Sujith Ravi. 2009. A New Objective Function for Word Alignment. In *Proceedings of the NAACL/HLT Workshop on Integer Programming for Natural Language Processing*, Boulder, Colorado.

Disaster 2.0. 2011. *Disaster Relief 2.0: The Future of Information Sharing in Humanitarian Emergencies*. United Nations Foundation, UN Office for the Coordination of Humanitarian Affairs (UN OCHA), Vodafone Foundation, Harvard Humanitarian Initiative.

Heidi Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of EMNLP 2002*, Philadelphia, Pennsylvania.

Robert Frederking, Alexander Rudnicky, and Christopher Hogan. 1997. Interactive speech translation in the diplomat project. In *Workshop on Spoken Language Translation at ACL-97*, Madrid.

Dmitriy Genzel. 2010. Automatically Learning Source-side Reordering Rules for Large Scale Machine Translation. In *Proceedings of COLING 2010*, Beijing, August.

Sanjika Hewavitharana and Stephan Vogel. 2008. Enhancing a Statistical Machine Translation System by using an Automatically Extracted Parallel Corpus from Comparable Sources. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Workshop on Comparable Corpora*, Marrakech, Morocco, May.

Sanjika Hewavitharana and Stephan Vogel. 2011. Extracting Parallel Phrases from Comparable Data. In *Proceedings of ACL 2011*, Portland, Oregon, June.

Ann Irvine, Chris Callison-Burch, and Alexandre Klementiev. 2010. Transliterating from all languages. In

Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010), Denver.

Minwoo Jeong, Kristina Toutanova, Hisami Suzuki, and Chris Quirk. 2010. A Discriminative Lexicon Model for Complex Morphology. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*, Denver.

Claire Lefebvre. 1998. *Creole Genesis and the Acquisition of Grammar: The case of Haitian Creole*. Cambridge University Press, Cambridge, England.

William D. Lewis and Fei Xia. 2010. Developing ODIN: A Multilingual Repository of Annotated Language Data for Hundreds of the World's Languages. *Literary and Linguistic Computing*. See: http://research.microsoft.com/apps/pubs/default.aspx?id=138757.

William D. Lewis. 2010. Haitian Creole: How to Build and Ship an MT Engine from Scratch in 4 Days, 17 Hours, & 30 Minutes. In *EAMT 2010: Proceedings of the 14th Annual conference of the European Association for Machine Translation*, Saint RaphaeÏ, France, May.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Ann Irvine, Lane Schwartz, Wren N. G. Thornton, Ziyuan Wang, Jonathan Weese, and Omar F. Zaidan. 2010. Joshua 2.0: A Toolkit for Parsing-Based Machine Translation with Syntax, Semirings, Discriminative Training and Other Goodies. In *In Proceedings of Workshop on Statistical Machine Translation (WMT10)*, Uppsala, Sweden.

Christian Monson, Ariadna Font Llitjos, Vamshi Ambati, Lori Levin, Alon Lavie, Alison Alvarez, Robert Frederking Roberto Aranovich, Jaime Carbonell, Erik Peterson, and Katharina Probst. 2008. Linguistic Structure and Bilingual Informants Help Induce Machine Translation of Lesser-Resourced Languages. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

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

Robert Munro. 2011. Subword and spatiotemporal models for identifying actionable information in Haitian Kreyol. In *Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011)*, Portland.

Douglas W. Oard and Franz Josef Och. 2003. Rapid-Response Machine Translation for Unexpected Languages. In *MT Summit IX*, New Orleans.

Douglas W. Oard. 2003. The Surprise Language Exercises. *ACM Transactions on Asian Language Information Processing - TALIP*, 2(2):79–84.

Kemal Oflazer and Ilknur Durgar El-kahlout. 2007. Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation. In *In Proceedings of the Statistical Machine Translation Workshop, ACL 2007*, Prague.

Chris Quirk and Arul Menezes. 2006. Dependency Treelet Translation: The convergence of statistical and example-based machine translation? *Machine Translation*, 20:43–65.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of ACL 2011*, Portland, Oregon, June.

RC. 2010. The American Red Cross: Social Media in Disasters and Emergencies. Presentation.

Jason Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, Los Angeles.

Raghavendra Udupa, K Saravanan, A Kumaran, and Jagadeesh Jagarlamudi. 2009. MINT: A Method for Effective and Scalable Mining of Named Entity Transliterations from Large Comparable Corpora. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, Athens, Greece.

Andrej Verity. 2011. What the UN could not have done without the Volunteer Technical Community. In *United Nations Dispatch*. The Disaster Relief 2.0 Blog Series.

Reyyan Yeniterzi and Kemal Oflazer. 2010. Syntax-to-Morphology Mapping in Factored Phrase-Based Statistical Machine Translation from English to Turkish. In *Proceedings of the ACL 2010*, Uppsala, Sweden.

Omar Zaidan and Chris Callison-Burch. 2011. Crowdsourcing Translation: Professional Quality from Non-Professionals. In *Proceedings of ACL 2011*, Portland, Oregon, June.