Haitian Creole: How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes

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

Abstract

We describe the effort of the Microsoft Translator team to develop a Haitian Creole statistical machine translation engine from scratch in a matter of days. Haitian Creole presents a number of difficulties for devleoping an SMT system, principal among these is the lack of significant amounts of parallel training data and an inconsistent orthography, both of which lead to data sparseness. We demonstrate, however, that it is possible to build a translation engine of reasonable quality over very little data by engaging with the native language community and reducing data sparseness in creative ways. As such, we show that MT as a technology and as a service can be deployed rapidly in crisis situations.

1 Credits

Credit is due to all members of the Microsoft Translator team who spent many sleepless nights putting together and shipping the Haitian Creole Translators. I am proud and humbled to call you colleagues. Our whole team in turn is indebted to the researchers and data providers who helped us locate or directly provided us with training data for our systems, to the Butler Hill Group and their employees who donated many, many hours to the effort, to Moravia Worldwide and Welocalize who helped translate content, and most especially to the volunteers involved in the relief effort, especially those at Ushahidi and Mission 4636, who offered crucial advice and provided us with data for our engine. This entire effort would not have succeeded without your help.

© 2010 European Association for Machine Translation.

2 Introduction

The unprecedented disaster in Haiti triggered a massive relief effort from around the world. Since a significant portion of the population of Haiti speaks Haitian Creole—approximately seven out of every eight people in Haiti speak the language[1]— and since Haitian Creole is not widely spoken outside of Haiti, it became obvious that Machine Translation might be a useful tool for the relief effort in Haiti, where it could be applied to translate emergency relief documents, medical documents, SMS text messages, and even common phrases and expressions. Unfortunately, no readily available Machine Translation engine existed for Haitian Creole at the time. Our team received an e-mail from colleagues on the ground in Haiti asking if we could develop a translation engine to aid in the relief effort. Here we relate this effort and the daunting challenges we faced to produce such an engine from scratch in just a few days.

3 Motivation

Haitian Creole is a resource poor language spoken principally on the western portion of the Island of Hispaniola. On January 12th, 2010 a devastating earthquake struck the island, with an epicenter in

[1]This calculation based on the number of speakers in Haiti of 6,960,000 in 2001 as listed Ethnologue (http://www.ethnologue.com/show_language.asp?code=hat), and an extrapolation of the United Nations' estimate of Haiti's total population of 8,326,000 in 2003, reduced by the country's estimated annual growth rate of 1.32% per year from 2000-2005 (http://www.nationsencyclopedia.com/Americas/Haiti-POPULATION.html), to a figure of 8,107,644 for 2001 .

the neighborhood of Port-au-Prince. Due to the strength of the quake, and generally poor building standards in the city, many tens of thousands of people were killed during the quake and in the days and weeks that followed; many more were left without shelter, food or water.[2] The humanitarian crisis prompted relief organizations to come to the aid of the Haitian people. It became evident early in the crisis that medical and relief documentation in the native Haitian Creole language would be of significant utility to relief workers on the ground in the country. Efforts were started to locate medical and relief related documents in Creole and make these freely available.[3] Further, since aid organizations such as Usahidi and Mission 4636 had built up an infrastructure to receive text messages from those seeking aid, translating the many thousands of the Haitian Creole text messages into English in real time became critical. Based on these and other potential uses, a Haitian Creole Machine Translation engine might be of significant value to the relief effort.

Developing NLP tools and resources for low density or resource poor languages is not without precedent. A notable example of this was the *Surprise Language Exercise* (SLE) sponsored by the DARPA Translingual Information Detection Extraction and Summarization (TIDES) program (Oard, 2003). The idea behind this initiative was to "address the need for technologies in less common languages through experiments in rapid technology porting where data collection, resource creation, and technology development take place simultaneously within a very short time period (i.e., one month)" (Strassel et al., 2003). Sixteen research teams from around the globe participated in the initiative, along with several additional organizations that participated as data providers. (Strassel et al., 2003) discusses in detail the Linguistic Data Consortium's efforts to cobble together sources of data for Cebruano and Hindi in a compressed timeframes. (Oard and Och, 2003) present the results of a rapidly deployed Cebuano Statistical Machine Translation system, built over resources collected as part of the SLE.

4 The Challenge

On January 19th, 2010, one week after the earthquake, we received an e-mail from a colleague who was involved in the aid effort asking us if it would be possible for us to develop a translation engine for Haitian Creole. Within a few hours, we rallied a small team of developers, testers, and computational linguists to decide on how best we could develop such an engine. No one on our team had any knowledge of Creole: no native speakers, no linguistic background in the language (other than trivial knowledge learned in college), no idea about the grammatical structure, encoding, orthography, registers (e.g., differences in training data between high and low registers), degree of literacy in the speaker population, etc. Further, we had no Haitian Creole data of any kind, nor any readily available documents or other materials in Haitian Creole or about Haitian Creole. In effect, we were starting at zero.

A resource poor language presents a daunting challenge for developing a Statistical Machine Translation (SMT) engine. Since, at the time, we knew of no corpora of any size in the language, including no known annotated corpora nor parallel corpora, developing an initial set of tools to ramp up the effort would prove to be exceedingly difficult. Although there are public and government web sites with Haitian Creole content on the Web, much of this content is in idiosyncratic formats (such as PDF) which are not easily consumed.

Fortunately, Carnegie Mellon University released a small database of parallel Haitian Creole and English data shortly after the quake[4], which was made freely available to the public with limited restrictions on use. Much of this data was developed for a Speech to Speech Translation project called DIPLOMAT which had been abandoned in the 1990's (Frederking et al., 1997). CrisisCommons made some data available to developers [5], although much of this data overlapped with readily available data on the Web (such as the Haitian Constitution) [6]. Also available was the Haitian Creole

[2] See http://en.wikipedia.org/wiki/2010_Haiti_earthquake for overview and references about the earthquake and the relief effort.

[3] E.g., Hesperian's distribution of *Where There is No Doctor* and *Where Women Have No Doctor* through free downloads. See http://www.kron.com/News/ArticleView/tabid/298/smid/1126/ArticleID/4775/reftab/610/t/Berkeley%20Nonprofit%20Helping%20Quake%20Victims%20With%20Medical%20Materials%20Translated%20into%20Haitian%20Creole/Default.aspx.

[4] See http://www.speech.cs.cmu.edu/haitian/ for the materials, and http://www.cmu.edu/news/archive/2010/January/jan27_haitiancreoletranslation.shtml for the initial press release.

[5] http://wiki.crisiscommons.org/wiki/Machine_Translation_System

[6] http://pdba.georgetown.edu/constitutions/haiti/haiti1987.html

Bible, although the language in the Bible is somewhat "stilted" and archaic, especially on the English side.

5 The Plan

We rapidly developed a plan for building the engine:

- First, we needed to quickly identify available data sources and start processing them. The resources from CMU and the Bible were the first to be processed. We identified additional publicly available data, such as documents provided by government agencies, which were assigned to a small cadre of developers to process. Many of these were in idiosyncratic formats (such as PDF), and so were often one-off conversion and clean-up projects, sometimes cleanable only through some manual effort. As soon as data had been extracted and cleaned, it was passed thorough our sentence aligner and then added to our slowly growing data store.

- Since we had no expertise in Haitian Creole at all, we needed to find individuals who were either native speakers of the language or who had some linguistic training on the language, or both. Fortunately, with in a couple of days of starting the project, we found an individual who was fluent in the language and was also linguistically trained. He proved invaluable for answering questions about the structure of the language, spelling conventions, orthographic norms, etc.

- We also needed native speakers to help us translate documents, verify translations, or correct output when it was corrupted or unusuable. We found individuals with these skills within a couple of days of starting the project.

- We engaged with relief community to determine what they might need translations of, and what data sources they might know of. A crucial contact was Ushahidi, an organization that had set up an infrastructure for receiving text messages to a phone number in Haiti, namely 4636. Ushahidi later turned the management of 4636 to Crowdflower, which renamed the effort to *Mission 4636*. Crowdflower has been providing us dumps of the 4636 text messages, in both Creole and English, with the former being translated into the latter by human translators from around the world. We used bilingual speakers to clean up this content (since SMS messages tend to be quite noisy, and the translations are done quickly and therefore can be of mixed quality), and added it to our training data. Table 1 shows some sample SMS messages and the various kinds of noise associated with them.[7]

6 Some of the Challenges Creole Presented

Beyond being a "low data" language, there were a number of challenges that Creole presented that made creating a translation engine more difficult. Since Creole is fairly "young" as a written language[8], and is still in the early stages of orthographic standardization and normalization (Allen, 1998), inconsistencies in the orthography increase data sparseness and noise.

Creole 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 Haitian 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 possessive is also *mwen* which is written following the word that is possessed. This 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 2 for a list of these reductions. Unfortunately, the number of alternations make it difficult to train on Creole texts, since the patterns are inconsistent, and with very little training data, data sparseness is increased and errors are more frequent. We coun-

[7] Please note: As we clean up and translate the SMS messages we return them to Mission 4636 so that they can use them for their efforts. We are in the process of anonymizing the SMS messages so that they can be distributed to the much larger community involved in the relief effort.

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

Table 1: Sample SMS Messages and Problems

Original SMS Message	Translation	Corrected Translation	problem
Mwen rele FIRST LAST mwen se yon bòs mason kay mwen kraze mwen gen kat pitit numero mwen se 99999999	My name is FIRST LAST. I am a construction worker and I have four children. Please call me at that number 99999999.	My name is FIRST LAST. I work in construction, and I have four children. My number is 99999999.	Original SMS contains no sentence breaks, and the original is slightly incorrect
Ki sa pou nou f? ak timoun yo kos?nan lekol la e pui kile moun duval nan croi des bouket ap jwen manje pou met nan vant yo	What can we do with the children regarding school and when will the people of duval in croix des bouquets get food to put in their bellies?		Encoding problems in the original SMS lead to no accented characters in source (realized as "?")
Alo mwen ap viv andedan delma 2 ak tout moun men nou pifo sinistre nou bezwen ?d avek tant manje	Hi, I am living in Delmas 2, most of the people there lost evrything, we need help and food please.	Hi, I live in Delmas 2 with everyone but most of us are destitute. We need help with tents, food	Original SMS has encoding problem, accents missing (e.g., pifò), and translation is inaccurate
Voye kÃk konsÃy pou mwen.	Send me some advice.	Send me some advice.	Encoding problems in the original SMS, where è shows up as Ä.
Je suis stephanie douyon agee de 20 ans vivant a delmas 33 rue charbonniere imp cala.mes parents et moi ont tout perdu nous a besoin daide.on vous attend.merci.			SMS not in Creole.

tered this problem by normalizing all of the contracted forms in our training data to the full forms, doing the same for all input we receive at runtime. Further, we trained our word breaker to look for accented forms (for those we might have missed) in order to ensure that there were no stranded apostrophes.

Table 2: 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

Writers of Creole also 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. Here too, we normalize to the higher register forms in order counter data sparseness, and also to ensure more consistent translations. See Table 3 for a small list of these shorthand forms.

There are three accented characters in Creole, è, ò and à, with the last being somewhat uncommon. Unfortunately, Creole is written inconsistently, especially in SMS messages. The Creole word for *thank you*, for instance, can be written as *mesi* or *mèsi*, with the latter being correct. Likewise, the Creole equivalent of *do* in English can be written as either *èske* or *eske*, with the former being cor-

Table 3: Sample Shorthand Notations

Abbreviated Form	Full Form
s'on	se yon
avèn	avèk nou
relem	rele mwen
wap	ou ap
map	mwen ap
zanmim	zanmi mwen
lavel	lave li
fanmim	fanmi mwen
ekriw	ekri you
mandew	mande ou
fèw	fè ou
poum	pou mwen
lèm	lè mwen
edem	ede mwen
lèw	lè ou
santim	santi mwen
lavim	lavi mwen
dim	di mwen
peyem	peye mwen
voyel	voye li
frèm	frè mwen
bliyem	bliye mwen
frèw	frè ou
wale	ou ale
beniw	beni ou
achtel	achte li
SVP	silvouple

rect. For our Creole to English translation engine, we normalized the forms to their unaccented counterparts, since the cost in that direction was minimal (there were no minimal pairs that we knew of that would have any consequence). However, we could not do the same in the other direction since accents occurred correctly in most of our data. We had no consistent way to rectify this problem in English to Creole, and had to live with the sparseness created by this inconsistency.

We also apply fairly standard normalizations to the data which also counteract data sparseness, such as normalizing quotes and apostrophes so that they are consistent throughout the data, and normalizing some encoding problems related to misencoded content (such as fixing UTF-8 encoded data misidentified as latin1).

An additional strategy we are exploring for countering data sparseness in our Creole to English system involves creating a backoff dictionary of "faux" Creole words. To construct this dictionary, we are mapping a large vocabulary of existing French words drawn from our French to English SMT system to their phonetic forms using an Automatic Speech Recognition (ASR) dictionary. These phonetic representations are then mapped to the equivalent Creole phones which we derived from the ASR dictionary provided by CMU.[9] We are generating a faux Creole vocabulary from these phonetic forms. Although not all forms in the dictionary will be valid Creole words; where they are, we will have a valid Creole to English mapping that we would not have otherwise had.[10]

7 The Result

After several days of work, we were able to release our first Creole to English and English to Creole engines. In fact, the delivery of the engine took about four and one-half days from concept to delivery. In the months that have followed we have made a large number of additional changes to the engine and have released several updates. Our system has around 150,000 segments of training data at this time, and the latest version has a BLEU score of 29.89 on a held out set of CMU and SMS data for Creole to English, and 18.30 on English to Creole over the same set.[11] [12] The engine has been used by relief workers in the field, and we are discussing plans to integrate it into the the Mission 4636 SMS process, where it can be used to provide first pass translations, or used to feed classifiers built over the English text which can route messages automatically to the correct aid organizations. We have a continued relationship with Mission 4636 to this day and continue to help them translate content.

8 Tools

Since Haitian Creole is treated no differently than any other language we ship with Microsoft Translator, all of the tools and resources that have already been developed for other languages are available for Haitian Creole.[13] Notably, our API allows software developers to create Haitian Creole specific tools and apps that generate translations through a simple call, supporting the translation of strings to and from Creole into and out of any of the other languages we support using a variety of interfaces, including AJAX, HTTP, and SOAP. Likewise, our widget can be activated on web pages by inserting a simple java script snippet, which enables real-time, in-place translations, and also enables the Community Translation Framework (CTF). CTF allows users and Web developers to contribute alternative translations which can override Machine Translated content when "published" to the page (these alternatives are also available to the community of users through a translation memory). Finally, the Translation Bot (TBot) can be added to Messenger IM sessions, whereby IM messages between users can be translated into and out of Haitian Creole. For instance, one user can be messaging in Creole and another in English, and the messages will be translated between the two users by TBot as they are entered. A sample session using TBot between English and Haitian Creole is shown in Figure 1.

[9]The ASR mappings were provided as part of the Haiti data CMU made available to the public to help with the relief effort, as described in Section 4.

[10]Where they are not, the words are effectively ignored, since they will generally not occur in source text. Unfortunately, this strategy does not work in the English to Creole system, since the faux Creole words will show up in cases when an Out of Vocabulary (OOV) item is encountered.

[11]The held out evaluation data consisted of a random sample of 550 sentences from the CMU data, and 36 sentences from translated SMS sentences. A much larger portion of the translated SMS sentences (around 1,500 sentences at the time of this writing) were used in training.

[12]We are currently conducting a human evaluation on the Creole to English and English to Creole translators, but the results of this evaluation were not available at the time of this writing.

[13]See http://www.microsofttranslator.com/Tools/ for a complete set of Microsoft Translator tools and documentation.

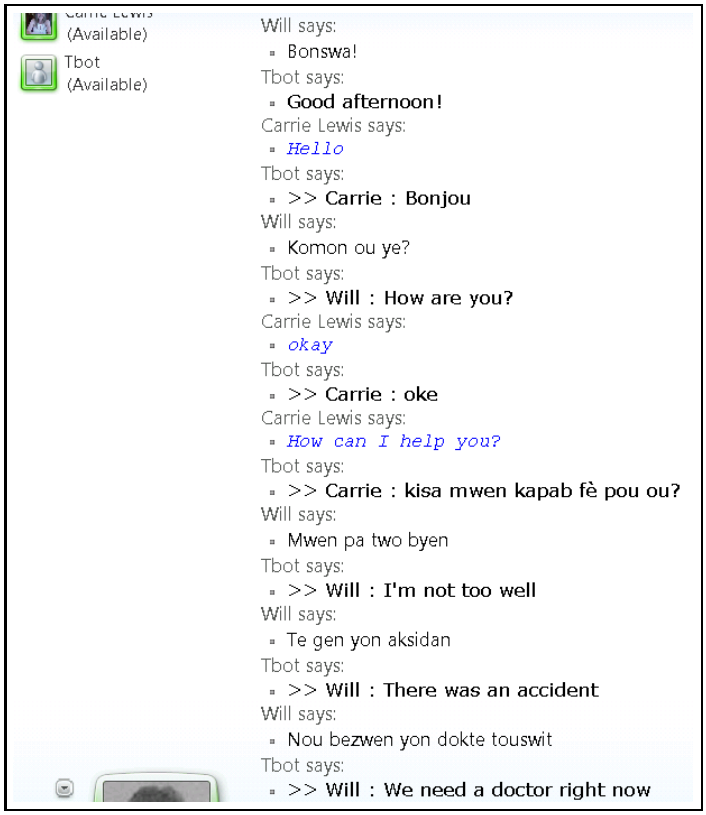

Figure 1: Sample Session In Haitian Creole and English using TBot

9 Conclusion

Delivering the two Haitian Creole translation systems (Creole to English and English to Creole) was a significant achievement, and proves that it is possible to build a translation engine of reasonable quality over very little data and in an extremely compressed timeframe. We hope that our effort demonstrates two things: (1) in crisis situations, MT can be a crucial component, and can be deployed rapidly, and (2) statistical MT systems can be developed and deployed for resource poor languages, crucially by cobbling together data from a variety of sources, contending with data sparseness in creative ways, and by engaging native speakers and a broader community to assist in the effort.

References

Allen, Jeffrey. 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.

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

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

Oard, Douglas W. and Franz Josef Och. 2003. Rapid-response machine translation for unexpected languages. In *In Proceedings of MT Summit IX*, September.

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

Strassel, Stephanie, Mike Maxwell, and Christopher Cieri. 2003. Linguistic resource creation for research and technology development: A recent experiment. *ACM Transactions on Asian Language Information Processing - TALIP*, 2(2):101–117.