Building MT for a Severely Under-Resourced Language: White Hmong

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

Phong Yang
California State University Fresno
5245 N. Backer, M/S PB92
Fresno, CA 93740
pyang@csufresno.edu

Abstract

In this paper, we discuss the development of statistical machine translation for English to/from White Hmong (Language code: mww). White Hmong is a Hmong-Mien language, originally spoken mostly in Southeast Asia, but now predominantly spoken by a large diaspora throughout the world, with populations in the United States, Australia, France, Thailand and elsewhere. Building statistical translation systems for Hmong proved to be incredibly challenging since there are no known parallel or monolingual corpora for the language; in fact, finding data for Hmong proved to be one of the biggest challenges to getting the project off the ground. It was only through a close collaboration with the Hmong community, and active and tireless participation of Hmong speakers, that it became possible to build up a critical mass of data to make the translation project a reality. We see this effort as potentially replicable for other severely resource poor languages of the world, which is likely the case for the majority of the languages still spoken on the planet. Further, the work here suggests that research and work on other severely under-resourced languages can have significant positive impacts for the affected communities, both for accessibility and language preservation.

1 Introduction

Building Machine Translation for low-resource languages has recently come into favor in the MT research community. Languages without significant resources pose difficulties for statistically-biased approaches to MT, since one needs a critical mass of parallel and monolingual data to be able to build engines of reasonable quality. Much of the work on low resource MT has focused on languages that have large populations of speakers or have an official status in their countries of origin (*e.g.*, (Islam et al., 2010) for Bangla, (Somers, 2004) for Welsh, etc.). Little work has been done on the very low-end of the under-resourced languages spectrum—what we call "Severely Under-Resourced Languages", or SURLs—which likely represent the bulk of the 6,900 languages still spoken on the planet (Maxwell and Hughes, 2006). Some notable exceptions include the work by Chiang and Bird and colleagues on developing MT for language preservation[1], work by a number of research teams to build translation systems for Haitian Creole immediately following the earthquake in Haiti (*e.g.*, efforts at BBN, Google Research and Microsoft Research, with one effort discussed in (Lewis, 2010), etc.)[2], and ongoing research in "Crisis MT", based on the Haitian Creole model, as proposed by (Lewis et al., 2011). What we seek to do here is propose a model for building

[1] NSF IIS-1144167, focused initially on languages of the Eastern Highlands of Papua New Guinea. See http://nsf.gov/awardsearch/showAward.do?AwardNumber=1144167

[2] Although Haitian Creole is an official language of Haiti, and spoken by the majority of Haiti's inhabitants, it lacked significant available resources at the time, beyond those developed in the DIPLOMAT project (Frederking et al., 1997), and required considerable involvement from a crowd of native speakers to build up sufficient data for viable translation systems (*e.g.*, from disaster related SMS messages (Munro, 2010)). Due to these efforts, the authors no longer consider Haitian Creole to be a SURL.

Table 1: Sample Hmong Words

Word	English	Onset	Nucleus	Tone
hmoob	‘Hmong’	pre-aspirated bilabial nasal	nasalized back vowel	high
hmoov	‘destiny’	pre-aspirated bilabial nasal	nasalized back vowel	mid rising tone
ntawv	‘book’	pre-nasalized dental	diphthongized low vowel (“au”)	mid rising tone
tes	‘hand’	dental stop	mid-front vowel	low-tone
phom	‘rifle’	aspirated bilabial stop	back vowel	very low tone (“creaky voice”)

Table 2: Sample Hmong Classifiers

Classifier	Hmong Nouns	English Glosses	Used for
rab	rauj, phom, txiab	hammer, rifle, scissors	tools and weapons
daim	txiag, ntawv, pam	board, sheet of paper, blanket	flat things and surfaces
txoj	hlua, hmab	rope, vine	long, thin things
phau	nyiaj, ntawv	wad of money, book	piles of things
tawb	qaub ncauj, zis, quav	spit, urine, dung	bodily excretions

MT for the severely under-resourced languages of the world through strong community engagement, and to show the viability of the MT that is developed. Through MT, a community not only gains accessibility to content that might have otherwise been unavailable (*i.e.*, in another language), but they also have a viable method to preserve and continue to use their language.

2 The Hmong Languages

The Hmong Languages constitute a dialect spectrum spread across Southern China into Southeast Asia, with scattered populations throughout Vietnam and Laos. The Vietnam War proved to be calamitous for the Hmong peoples of Southeast Asia, forcing many to flee permanently from the area, with approximately 300,000 settling in the United States, and lesser numbers elsewhere (*e.g.*, Australia, France, Thailand). Spreading the people so thinly around the globe has had a cost: many children of the Hmong have stopped learning and using their ancestral language. For the first time in the history of the Hmong, the language may eventually become extinct.

The predominant Hmong dialects spoken in the United States are White Hmong (in native orthography, Hmoob Daw) and Green Hmong (Moob Njua, sometimes Blue Hmong). Our focus in this project is on the White Hmong dialect, which is mutually intelligible with Green Hmong (our plan is to extend the translation project to Green Hmong in the future). The project we describe here came about from a close collaboration between the greater Hmong community in the United States and our research team. One reason we sought to develop a Hmong translator was to help those Hmong who do not read English the ability to more comfortably use their native language to navigate public and private resources on the Web. We also saw it as a project that can help preserve the Hmong language, and encourage the youth to participate in learning and using their language.

White Hmong (language code: mww) is a monosyllabic, monomorphemic, tonal language (seven tones), and is strongly Subject-Verb-Object (SVO). The principal orthography for White Hmong (as well as for Green Hmong) is latin-based, and often referred to as the Romanized Populist Alphabet (RPA). Each word is written with a set of consonants representing syllable onset, which are then followed by the nucleic vowel. The final consonant, if present, is a representation of the word’s tone. See Table 1 for some sample words and descriptions.

Unlike languages with richer morphology, Hmong’s fairly reduced morphology can help counter data sparsity for MT (a serious problem for languages with richer morphology). However, Hmong does have a very rich classifier system, which increases sparsity. Classifiers are common in a number of Asian languages (and in a number of African and Amerindian languages). Classifier systems are roughly similar to gender systems seen in many languages, but rather than 2-3 genders, Hmong has approximately 70 nominal classes. The classes are basically semantic, as can be seen by the classifiers in Table 2 (from Jaisser (1987)).

Each noun in Hmong has a classifier associated with it, and classifiers are generally used in discourse with the nouns they attach to, with some exceptions. Note the distribution of phrases in Table 3, and how they differ from English sentences. In order to learn the correct word alignments, an aligner must learn that each noun (mostly) co-occurs with a specific classifier, thus aligning *a* or *the* with that classifier. Opposing this, the aligner must also recognize the contexts where the classifiers are used in Hmong but the equivalent determiners are *not* used in English (*e.g.*, with quantifiers). Further, since classifiers represent semantic classes, there are not necessarily 1:1 correspondences between each noun and each classifier; a different classifier can be used with a noun if the context requires it (*e.g.*, the meaning of *mov* can alternate between 'meal' and 'rice' depending on the classifier used with it).[3]

3 The Hmong Community Engagement

Engaging with the language community is essential for any project focused on the development of MT (or any NLP resources, for that matter) for SURLs. Data is essential for MT to work, and the community provides a means to *access* data by reviewing it, and also a means for generating it. They also are crucial for the eventual uptake of whatever the results of project are (they are the primary consumers, after all!). A community engagement involves two critical groups: (a) a community of native speakers who are willing to spend time on the project, and, (b) community leader(s) who can engage with and motivate the community, and who can also publicize the project, both as it is being developed, but also to solicit users within the community after the results of the project have been released. Alternatively, one can engage with the community directly, effectively bypassing (b), but other incentives would have to be provided to make community participation in data collection/generation possible, as well as to ensure translator uptake once developed. Community leaders reduce the costs associated with community engagement, and make it more likely that the project will be successful, since, effectively, they have a vested interest in its success.

In the Hmong project, through community leaders we were able to engage with a wide spectrum of members across the Hmong community, including:

- college students, many of whom are taking classes in Hmong,
- school teachers, many of whom teach Hmong in elementary and high schools,
- school administrators, deans and professors at local universities and colleges,
- business people, such as publishers of Hmong texts and dictionaries, as well as Hmong-language TV and radio broadcasters,
- elders, who have significant respect throughout the community, and have more time to review and correct, and are motivated by the strong desire to see their language preserved, and,
- high school students, many of whom are semi-literate in their language, are encouraged by their parents and other family members to learn and become literate (and who see MT as a "cool" way to get there).

4 The Hunt for Data

As a SURL, it was exceptionally difficult to find data in Hmong. Since Hmong is not the official language of any government or country, it was impossible to turn to government sources of data (as has been possible with other under-resourced MT projects, such as for Welsh (Somers, 2004) or Inuktitut (Martin et al., 2003)). There are also no readily available corpora in the language, neither bilingual nor monolingual. It was thus necessary for us to cobble together resources opportunistically through engagement with the community, and off the Web. Our first resource was the Bible, for which we were able to locate a Hmong translation in electronic form (use of the Bible for under-resourced NLP has a long history, for example, see (Resnik et al., 1999). However, pairing the Hmong translation with several different simple Bible translations proved to generate rather low quality alignments, and was aban-

[3] English has classifiers, they are just not used as frequently. Examples of English classifiers include "*herd* of cows", "*flock* of birds", "*pride* of lions", "*pack* of dogs", etc.

Table 3: Sample Phrases with Classifiers

Hmong	English	Notes
lub tsev	the house	
ib lub tsev	a (one) house	A/one used interchangeably
ib lub tsev tshiab	a new house	
kuv lub tsev tshiab	my new house	
yim lub tsev tshiab	eight new houses	Classifiers used with quantifiers
kuv him lub tsev tshiab	my eight new houses	
tus tsov	the tiger	
ib tus tsov	a tiger	
kuv ntshai tsov	I fear tigers	No classifier (indefinites)
ib pluag mov	a meal	Classifier for 'meals'
ib taig mov	a bowl of rice	Classifier for 'bowls'
ib taig zaub	a bowl of vegetables	

doned later in the project.[4] Subsequently, community members were able to provide some resources. One crucial resource was the Hmong dictionary at HmongDictionary.com, provided by the publisher, which had a rich inventory of vocabulary in Hmong language (approximately 6000 words, with some contexts), a set of parallel sentences used in classroom instruction (approximately 3200 sentences), and a small set of phrases used on a mobile phone app for Hmong (approximately 300 phrases, created by the developer of the app).[5] With all of this data, however, we still had less than 5000 bilingual sentences/translation units we could use, much too little to build viable translation systems. We thus had to look farther afield.

One common source of data for MT is the Web, which contains large amounts of parallel content for many of the world's languages, including surprisingly, a growing cache of content for SURLs. Notable efforts by Scannell and colleagues (Scannell, 2007) to collect monolingual content for many SURLs (to date Scannell has collected corpora for over 1,000 languages) gave us hope that content for Hmong could be found and added to our small supply of data.

The difficulty of locating resources for Hmong, however, as with any SURL, is that no search engine indexes the language, making it difficult to query existing search APIs to find pages in the language. To locate data we started by finding a few high quality Hmong pages on the Web, which we found via existing search engines using a couple of very simple and unique Hmong strings, namely, *xov xwm hmoob*, which means 'Hmong news', and *dab neeg hmoob*, which means 'Stories of the Hmong people'. Using the smallish corpus of monolingual Hmong content distilled from these pages, we then identified additional common n-gram sequences (1-4 grams) in Hmong, which were then used to do large-scale queries against the Bing index. This allowed us to identify a much larger sample of Hmong pages; ultimately, we were able to locate approximately 16K pages that likely contained Hmong data. With the aid of community leaders, we then worked with a small number of college students who were conversant in Hmong to do two things:

1. Review the pages that had been identified as being in Hmong, and verify that they were in fact in Hmong (specifically White Hmong), and,
2. Identify additional pages on the relevant sites that might be source English pages from which the Hmong pages were translated.

Identifying parallel pages often consisted of very simple STRAND-like (Resnik and Smith, 2003) pattern matching, but sometimes required more involved traversing of target pages. (1) alone gave us monolingual data in Hmong that could be useful for building target language models in the language for

[4] We are considering reintroducing the Bible into the data, given that the much stronger Model 1's built over our currently much larger data sets could act as a filter to remove weakly aligned sentences in the sentence alignment phase.

[5] The Hmong Translator App, developed by Joel Fries, is available for download from the ITunes store or the Android marketplace.

an English>Hmong system, and could also be used for building a robust language identifier in the future. Paired with (2), however, we are able to identify pages in Hmong and English that could be distilled into parallel training sentences crucial for developing MT. After several months of concerted effort, we distilled the 16K pages into a reliable set of 2,700 Hmong pages and documents, of which 1,000 were parallel with pages and documents in English. This gave us a core training set of over 30,000 sentences/600,000 words. Not huge, as compared to available training data for other languages, but certainly a nice core to start with.

To facilitate the engagement with the Hmong Community around data, we used the newly developed Microsoft Translator Hub infrastructure.[6] The Hub allows community members to upload data, pick and choose data that they would like to use to train MT models, train the models, verify quality against test data (test data that they either provided or that was auto-selected), and engage other community members to review the quality of translated output and even to repair translations over "elicitation data", which is then placed back into training. Further, community members could iterate over all of these steps freely until finally converging on a set of models they felt had the quality for deployment to the official Microsoft Translator site.[7]

5 Description of the Data and Training Pipeline

Following is our data and training pipeline:

1. Extract Data from native file formats: Since our data came in a variety of formats, it was first necessary to extract the sentences we needed for training from these formats. For extracting from PDF, we used the TET tool.[8] For html documents, we used our own custom tools to extract text from the html body. Text documents were used in their native format, and any Word documents were saved to text. All documents were converted to Unicode so that the encoding was consistent throughout.
2. Extracting sentences: Sentences were extracted from source and target documents, and saved in files, one sentence per line. To extract sentences, we broke on typical sentence end characters (*e.g.*, ".!?"), using heuristics to decide when there was a sentence break.
3. Sentence alignment: We used a derivative of the Moore aligner[9] to align source and target data for all our parallel data. Whenever new data was added, we used the model 1's from previous alignments as a prior, in order to improve alignment.
4. Data cleaning: We subjected the data to a rigorous set of data cleaning filters in order to remove noise, badly encoded characters, etc. Further, we applied filters to remove or normalize HTML tags, and applied others to ensure a reasonable ratio between alpha and non-alpha characters (in order to remove obvious noise). The length of source and target sentences were also examined, and any alignments that showed a highly skewed ratio between source and target were removed, as were overly long sentences (since, even if good quality, word alignment would likely suffer).
5. Dev/Test/Train: After collecting a sizable amount of data, we split the parallel sentences into dev, test, and train. To ensure an adequate and random sample, we used a "shuffle" step in each random sample to protect against accidental clumping in the samples. 1500 sentences were ultimately extracted from the training data for dev and test, with 1000 sentence used for dev, and 500 for test.
6. Training models: We used custom-built phrasal and tree-to-string (T2S) systems for training the models for our engines: English-Hmong was trained using a source-side parser and T2S, and Hmong-English with a phrasal system.[10]

[6] http://hub.microsofttranslator.com/

[7] www.microsofttranslator.com

[8] http://www.pdflib.com/products/tet-pdf-ifilter/

[9] See (Moore, 2002) for details, or download from here: http://research.microsoft.com/en-us/downloads/aafd5dcf-4dcc-49b2-8a22-f7055113e656/

[10] The T2S system that we developed uses technology described in (Quirk and Menezes, 2006), and requires a source-side dependency parser. For Hmong-English, since we have no source-side parser, we used our custom-built phrasal decoder, similar in many ways to the decoder in Moses, and discussed in more detail in (Quirk and Moore, 2007). Further details about the decoders is beyond the scope of this paper. The reader is

For the English-Hmong system, we trained a 4-gram LM over all monolingual data, including the target side of the parallel data. For the Hmong-English system, we trained a 4-gram LM in the same manner, but also trained a second 5-gram English language model trained over a much larger corpus of English language data, including Web crawled content, licensed corpora (such as LDC's Gigaword), etc. We used Minimum Error Rate Training (MERT) (Och, 2003) for tuning the lambda values for both systems.

6 Experiments and Analysis

Data was the fundamental question in this project — could we get enough of it to make a viable set of translation engines for the language? Table 4 gives the results by various sources for English<>Hmong experiments, and Table 4 shows similar experiments for the opposing direction. Both charts show the BLEU score for each experiment (against the test data sampled from all training data, as described in Section 5), and the difference from the preceding experiment and the baseline. Table 4 also shows the number of parallel and monolingual sentences in each iteration. Each experiment is labeled with a unique letter ((a) for the Baseline), with the letters roughly aligning between the two directions where relevant or possible.

Some analysis of the experiments, by experiment number (as enumerated in the tables), follows:

(a) The baseline consisted of all of the base data described in Section 4, that is, the set of data first provided by the Hmong community, minus the Bible data, classroom data, and Hmong Dictionary data, but including a small subset of parallel data from the Web. For English-Hmong (EH), no monolingual data was included in the baseline. For Hmong-English (HE), the baseline used only our default English LM (as noted above).

(b) For both systems, we altered the LMs that were being used. For EH, we added all monolingual Hmong PDF data that we had crawled. For HE, we included the English LM (as in step (a)), but also built a second LM over the English side of the parallel data (this was done for all subsequent experiments.)

(c) The classroom data of approximately 3,000 sentences was added. The reader will note that this data had a significant impact on BLEU, clearly adding valuable vocabulary and contexts.

(d) The entirety of the Hmong dictionary was added—after filtration, approximately 3,100 items. These data were doubled in order to increase their weight in the models that were built. The two systems showed contrary results: EH dropped by about 1/2 point,and HE increased by 0.64. Clearly, additional vocabulary was added by the dictionary. However, since there was no contextual information, and most importantly, almost no classifiers, the EH direction was adversely affected. This is consistent with similar experiments that interact with data sparsity, where the direction interacts with the morphological richness of the source. [11] Nonetheless, given the enriched vocabulary coverage, it was decided to keep the entirety of the Hmong dictionary data.[12]

(e) The first pass of crowd collected data was added, consisting of 500+ sentences. These data were created by Hmong community members over elicitation data. Surprisingly, the crowd data had a minor positive impact on EH, but a larger negative impact on HE. By examining the OOV rates between (d) and (e), HE showed that there was a reduction in OOVs, so it was decided to keep the data (and all subsequent crowd sourced data that was added after the experimental period).

(f) (EH only) We included the Matthew chapter of the Bible as a test of the Bible data. As shown, results were not promising.

(g) (EH only) The Hmong LM was increased dramatically by including all monolingual Hmong data that had been collected. Despite the fact

encouraged to refer to the sources provided for additional information.

[11] See (Lewis, 2010) for a discussion about how the normalization of the "richer" Haitian Creole improved the Haitian Creole>English system, but could not be employed in the opposing direction. Similar experiments from morphologically richer Bangla to English are discussed in (Islam et al., 2010).

[12] For SURLs, dictionary data is one of the broad coverage resources that are available, so consuming quality dictionary data, even sans contexts, can significantly help vocabulary coverage.

Table 4: English-Hmong (EH) Experiments

Description	BLEU	Diff/Prev	Diff/Baseline	# parallel snts	# Mono snts
a. Baseline (no mono)	20.11			17,159	17,159
b. Include all available PDF data for LM	20.70	0.59	0.59	17,159	126,520
c. Include full classroom data	22.03	1.33	1.92	17,624	
d. Include full Hmong Dictionary (double)	21.52	-0.51	1.41	21,802	
e. Include crowd data	21.80	0.28	1.69	22,366	
f. Include Matthew	21.23	-0.57	1.12	23,220	
g. Include all available HTML data for LM	21.62	0.39	1.51	23,220	761,996
h. Include new parallel Web data (& Bible)	22.56	0.94	2.45	49,548	
i. Exclude Bible data	23.72	1.16	3.61	45,448	
j. Clean LM data	23.81	0.09	3.70		274,336

Table 5: Hmong-English (HE) Experiments

Description	BLEU	Diff/Prev	Diff/Baseline	# OOVs
a. Baseline (uses only ENU LM)	16.82			
b. Include LM over local data & ENU LM	19.23	2.41	2.41	
c. Include full classroom data	20.90	1.67	0.79	
d. Include full Hmong Dictionary (double)	21.54	0.64	1.43	84
e. Include crowd data	20.94	-0.60	0.83	79
h. Include new parallel Web data (& Bible)	21.51	0.57	1.40	
i. Exclude Bible data	23.41	1.90	3.30	

that the LM quintupled in size, the BLEU score went up only by 0.39. This problem was addressed in step (j).

(h) All parallel Web data was included, as was the Bible data, resulting in nearly a point gain in EH, and about 1/2 point in HE.

(i) After some analysis by Hmong community members as to the quality of the aligned text from the Bible, it was decided to exclude the data from training. The result was significant in both directions.

(j) Since the Hmong monolingual data in (g) had not undergone rigorous data cleaning (as described in Section 5), it was subjected to it in this step. The result was a dramatic drop in the amount of data, but almost no effect in BLEU. This clearly speaks to the necessity of data cleaning, even at the sacrifice of data.

Community participants have been very helpful in testing the engine and in offering suggestions to improve the engine. As expected, one of the biggest noted problems has been with getting classifier-noun combinations correct. As with any statistical MT system, overcoming sparsity-related effects is best solved with more data.

Out of Vocabulary items (OOVs) also proved to be an issue. However, since Hmong does not possess vocabulary for a number of technical vocabulary items—*e.g.*, IT terms, legal terms, terms for modern machinery, terms for objects that do not exist in Southeast Asia, etc.—a certain number of OOVs is actually acceptable in Hmong, much as borrowed vocabulary in other languages is also acceptable. Although the floor for the number of OOVs in Hmong may be equal to or higher than other languages, classifiers are still necessary even for the borrowed vocabulary. Unfortunately, since classifiers represent semantic classes, there is no "neutral" classifier that can be used universally for all new vocabulary. Table 6 shows examples of acceptable OOVs and their classifiers. We often get these wrong, either by leaving off the classifier, or by attaching the wrong one.

7 Community and Crowd Engagement After Release

Community engagement did not end with locating data and resources for building the initial MT systems; it continued after the MT engines went into production. Community members improved the quality of MT output, which was done in two ways:

Table 6: Sample 'Acceptable' OOVs and their Classifiers

Hmong	**English**	**Notes**
U.S. Department of Agriculture	lub	round, bulky things, places where people live or work
Wyoming	lub	ditto
computer	lub	ditto
vehicle	lub	ditto
carpet	daim	flat item
wipes	daim	flat item
sanitizer	hom	liquids, gels
gel	hom	liquids, gels
Fresno	nroog	city

1. By running a set of "elicitation" sentences through an engine, and offering corrections to the output generated by that engine. Initial elicitation sentences consisted of sentences constructed from words in from the Hmong Dictionary (HmongDictionary.com), recent news content harvested from news sites, and medically related questions and answers.[13] Figure 1 shows a sample elicitation session in the Hub where the user supplies repaired Hmong translations for the English translations on the left.

2. By using the translator to generate content on community relevant English websites, and then subsequently correcting Hmong translations of these sites. In addition to providing training data for the Hmong<>English engines, users of the affected websites get the immediate benefit of high quality human translations (mixed with unrepaired MT'd content).

Data from both (1) and (2) were, and continue to be, iteratively added back into training for the engines. The community continues to provided improved translations for Hmong<>English content, with some community members providing 20 or more corrections per day.

8 Tools

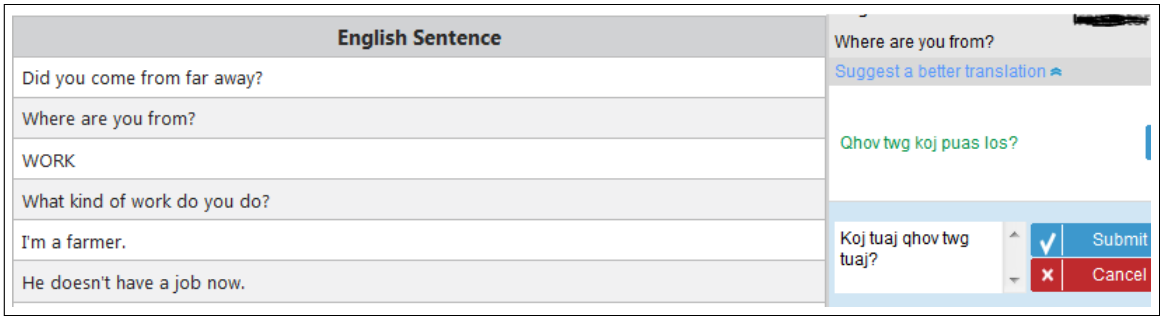
Since Hmong is treated no differently than any other language we ship with Microsoft Translator, all of the tools and resources that have already been developed for other languages are available for Hmong.[14] Notably, our API allows software developers to create Hmong specific tools and apps that generate translations through a simple call, supporting the translation of strings to and from Hmong 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). It is through CTF that many Hmong community members continue to contribute training data. They do this by repairing translations on web pages where the widget is installed[15]

Finally, the Translation Bot (TBot) can be added to Messenger IM sessions, whereby IM messages between users can be translated into and out of Hmong in real time. Students at California State

[13] An alternative strategy, to improve the utility of the data that was provided to the engines subsequently, would be to use an Active Learning strategy, *á la*(Ambati et al., 2010). We are considering such an approach in the future.

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

[15] For example, see the use of the widget on http://go2fresnostate.com/hmongtranslator/, http://www.fresnounified.org/dept/parentuniversity, and http://www.lwsd.org/school/muir/Pages/default.aspx. On any of the pages, use the widget to translate the source content (which is in English) into Hmong. Many of the Hmong translations were provided by Hmong community members and override those provided by the automatic translator (selecting *Improve Translation* for any sentence on the pages will show the alternative translations that were provided).



English Sentence
Did you come from far away?
Where are you from?
WORK
What kind of work do you do?
I'm a farmer.
He doesn't have a job now.

[redacted]

Where are you from?

Suggest a better translation

Qhov twg koj puas los?

Koj tuaj qhov twg tuaj?

Figure 1: Sample Session Correcting Elicitation Data from English to Hmong

Fresno and Fresno Unified have found this a particularly intriguing feature since they can communicate with students through IM who do not speak Hmong, translating their IM from and to Hmong in real-time. Using IM tools in their native language has made Hmong “cool”, and can help in language preservation.

9 The Real Effects of the Hmong Translator within the Hmong Community

The release of a Hmong translator had a significant impact on the Hmong community, and has been widely publicized throughout the community, through Hmong radio broadcasts in markets where there are sizable Hmong populations, such as in Fresno, California and Minneapolis, Minnesota, and on internationally broadcast Hmong TV. It has instilled a sense of pride in their language, which many feel had been lost, especially among the younger population.

Community members continue to add Hmong translations to English only sites (*e.g.*, using the Widget and CTF), and engage with other community members to repair these translations. The Fresno Unified and Lake Washington School Districts both with sizable Hmong student populations, have adopted the translator on some or all of their sites to better serve their communities. California State University Fresno serves a community with a large Hmong population, and is planning on using the translator to localize content for their Web sites, and engage with the community to help repair translations.[16] All the data that is collected through these engagements has been or will be added into training data to improve the Hmong translators going forward.

Finally, one anecdote from a Fresno Unified that is using the translator speaks to its utility. Shortly after translating their Web pages to Hmong, and making them available to community members, calls for support to navigate the site dropped precipitously. One of the administrators in the office noted the sudden drop in phone support calls, and after speaking with several community members, discovered that native Hmong speakers could reliably navigate what was once an English-only site, and could find the documents and information they needed without calling for assistance. Thus, the utility of machine translated content, even when sometimes poor, was better than no access at all.

10 Conclusion and Future Directions

We demonstrated that it is possible to build a statistical machine translation system for a severely under-resourced language, and put it to practical use. The project required very close collaboration with the native speaking language community, and required active participation and buy-in from that community to succeed. We see the collaborative model here as viable for developing machine translation for other severely under-resourced communities.

In the near term, we see attacking the classifier-noun data sparsity problem in Hmong as a critical change needed to improve the quality of the Hmong translation systems. In that vein, we have conducted some preliminary experiments to generate content with correct classifier-noun combinations,

[16] The 2010 census shows that 3.6% of Fresno's population is Hmong, the largest Asian minority in the area.

and added that data back into training. The results are inconclusive, in that we have seen improvements in quality for the targeted feature (that is, classifier and noun agreement), but have seen other translations adversely affected.

More broadly, we plan to extend our work to the mutually intelligible, but orthographically distinct Green Hmong dialect, and possibly to other underserved communities around the globe.

Acknowledgments

Without the support and tireless work of the Hmong community, the Hmong translation engines would never have become a reality. Our special thanks go to all members of the Hmong community who helped, most especially Maika Yang from Stone Soup Fresno, David Lee and Huaj Lis from 3 Hmong Publishing, Jay Xiong from HmongDictionary.com, Xee Yang from Fresno Unified, Julie Lee and Mai Ong from CSU Fresno, and Chue Her and all other Hmong elders who put their support behind the translation project. We'd also like to extend thanks to the College of Arts and Humanities at CSU Fresno who provided facilities for numerous project meetings. Further, Kristin Tolle with Microsoft Research Connections requires special mention since she devoted many hours to the project, and she and her team were responsible for providing the seed funding for the essential Microsoft Translator Hub. Also, we would like to thank David Harrison for the time he devoted to the project while he was a visiting researcher at Microsoft Research.

References

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2010. Active Learning and Crowd-Sourcing for Machine Translation. In *Proceedings of the 7th LREC*, Vallatta, Malta, May.

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.

Md. Zahurul Islam, Joïg Tiedemann, and Andreas Eisele. 2010. English to Bangla Phrase-Based Machine Translation. In *Proceedings of the 14th EAMT*, Saint Raphaël, France, May.

Annie Jaisser. 1987. Hmong Classifiers. *Linguistics of the Tibeto-Burman Area*, 10(2):169–176.

William D. Lewis, Robert Munro, and Stephan Vogel. 2011. Crisis MT: Developing A Cookbook for MT in Crisis Situations. In *Proceedings of the Sixth WMT*, Edinburgh, Scotland, July.

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

Joel Martin, Howard Johnson, Benoit Farley, and Anna Maclachlan. 2003. Aligning and Using an English-Inuktitut Parallel Corpus. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, June.

Mike Maxwell and Baden Hughes. 2006. Frontiers in linguistic annotation for lower-density languages. In *Proceedings of COLING/ACL2006 Workshop on Frontiers in Linguistically Annotated Corpora*.

Robert C. Moore. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. In *Proceedings of the 5th AMTA Conference*, pages 135–144, Tiburon.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st ACL*, Sapporo, Japan.

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

Chris Quirk and Robert C. Moore. 2007. Faster Beam-Search Decoding for Phrasal Statistical Machine Translation. In *Proceedings of MT Summit XI*.

Philip Resnik and Noah Smith. 2003. The Web as a Parallel Corpus. *Computational Linguisitics*, 29(3):349–380, September.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues'. *Computers and the Humanities*, 33(1):129–153.

Kevin P. Scannell. 2007. The Crúbadán project: Corpus building for under-resourced languages. In C. Fairon, H. Naets, A. Kilgarriff, and G-M de Schryver, editors, *Proceedings of the 3rd Web as Corpus Workshop*, Louvain-la-Neuve, Belgium, September.

Harold Somers. 2004. Machine Translation and Welsh: The Way Forward. Technical report, The Welsh Language Board.