CMU Haitian Creole-English Translation System for WMT 2011

Sanjika Hewavitharana, Nguyen Bach, Qin Gao, Vambhi Ambati, Stephan Vogel
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{sanjika,nbach,qing,vamshi,vogel+}@cs.cmu.edu

Abstract

This paper describes the statistical machine translation system submitted to the WMT11 Featured Translation Task, which involves translating Haitian Creole SMS messages into English. In our experiments we try to address the issue of noise in the training data, as well as the lack of parallel training data. Spelling normalization is applied to reduce out-of-vocabulary words in the corpus. Using Semantic Role Labeling rules we expand the available training corpus. Additionally we investigate extracting parallel sentences from comparable data to enhance the available parallel data.

1 Introduction

In this paper we describe the CMU-SMT Haitian Creole-English translation system that was built as part of the Featured Translation Task of the WMT11. The task involved translating text (SMS) messages that were collected during the humanitarian operations in the aftermath of the earthquake in Haiti in 2010.

Due to the circumstances of this situation, the SMS messages were often noisy, and contained incomplete information. Additionally they sometimes contained text from other languages (e.g. French). As is typical in SMS messages, abbreviated text (as well as misspelled words) were present. Further, since the Haitian Creole orthography is not fully standardized (Allen, 1998), the text inherently contained several different spelling variants.

These messages were translated into English by a group of volunteers during the disaster response. The background and the details of this crowdsourcing translation effort is discussed in Munro (2010). Some translations contain additional annotations which are not part of the original SMS, possibly added by the translators to clarify certain issues with the original message. Along with the noise, spelling variants, and fragmented nature of the SMS messages, the annotations contribute to the overall difficulty in building a machine translation system with this type of data. We aim to address some of these issues in out effort.

Another challenge with building a Haitian Creole-English translation system is the lack of parallel data. As Haitian Creole is a less commonly spoken language, the available resources are limited. Other than the manually translated SMS messages, the available Haitian Creole-English parallel data is about 2 million tokens, which is considerably smaller than the parallel data available for the Standard Translation Task of the WMT11.

Lewis (2010) details the effort quickly put forth by the Microsoft Translator team in building a Haitian Creole-English translation system from scratch, as part of the relief effort in Haiti. We took a similar approach to this shared task: rapidly building a translation system to a new language pair utilizing available resources. Within a short span (of about one week), we built a baseline translation system, identified the problems with the system, and exploited several approaches to rectify them and improve its overall performance. We addressed the issues above (namely: noise in the data and sparsity of parallel data) when building our translation system for Haitian Creole-English task. We also normalized

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 386–392,
Edinburgh, Scotland, UK, July 30–31, 2011. ©2011 Association for Computational Linguistics

different spelling variations to reduce the number of out-of-vocabulary (OOV) tokens in the corpus. We used Semantic Role Labeling to expand the available training corpus. Additionally we exploited other resources, such as comparable corpora, to extract parallel data to enhance the limited amount of available parallel data.

The paper is organized as follows: Section 2 presents the baseline system used, along with a description of training and testing data used. Section 3 explains different preprocessing schemes that were tested for SMS data, and their effect on the translation performance. Corpus expansion approach is given in Section 4. Parallel data extraction from comparable corpora is presented in section 5. We present our concluding remarks in Section 6.

2 System Architecture

The WMT11 has provided a collection of Haitian Creole-English parallel data from a variety of sources, including data from CMU[1]. A summary of the data is given in Table 1. The primary in-domain data comprises the translated (noisy) SMS messages. The additional data contains newswire text, medical dialogs, the Bible, several bilingual dictionaries, and parallel sentences from Wikipedia.

Corpus	Sentences	Tokens (HT/EN)
SMS messages	16,676	351K / 324K
Newswire text	13,517	336K / 292K
Medical dialog	1,619	10K / 10K
Dictionaries	42,178	97K / 92K
Other	41,872	939K / 865K
Wikipedia	8,476	77K / 90K
Total	124,338	1.81M / 1.67M

Table 1: Haitian Creole (HT) and English (EN) parallel data provide by WMT11

We preprocessed the data by separating the punctuations, and converting both sides into lower case. SMS data was further processed to normalize quotations and other punctuation marks, and to remove all markups.

To build a baseline translation system we followed the recommended steps: generate word alignments using GIZA++ (Och and Ney, 2003) and phrase extraction using Moses (Koehn et al., 2007). We built a 4-gram language model with the SRI LM toolkit (Stolcke, 2002) using English side of the training corpus. Model parameters for the language model, phrase table, and lexicalized reordering model were optimized via minimum error-rate (MER) training (Och, 2003).

The SMS test sets were provided in two formats: raw (r) and cleaned (cl), where the latter had been manually cleaned. We used the *SMS dev clean* to optimize the decoder parameters and the *SMS devtest clean* and *SMS devtest raw* as held-out evaluation sets. Each set contains 900 sentences. A separate *SMS test*, with 1274 sentences, was used as the unseen test set in the final evaluation. For each experiment we report the case-insensitive BLEU (Papineni et al., 2002) score.

Using the available training data we built several baseline systems: The first system (Parallel-OOD), uses all the out-of-domain parallel data except the Wikipedia sentences. The second system, in addition, includes Wikipedia data. The third system uses all available parallel training data (including both the out-of-domain data as well as in-domain SMS data). We used the third system as the baseline for later experiments.

	dev (cl)	devtest (cl)	devtest (r)
Parallel-OOD	23.84	22.28	17.32
+Wikipedia	23.89	22.42	17.37
+SMS	32.28	33.49	29.95

Table 2: Translation results in BLEU for different corpora

Translation results for different test sets using the three systems are presented in Table 2. No significant difference in BLEU was observed with the addition of Wikipedia data. However, a significant improvement in performance can be seen when in-domain SMS data is added, despite the fact that this is noisy data. Because of this, we paid special attention to clean the noisy SMS data.

3 Preprocessing of SMS Data

In this section we explain two approaches that we explored to reduce the noise in the SMS data.

[1] www.speech.cs.cmu.edu/haitian/

3.1 Lexicon-based Collapsing of OOV Words

We observed that a number of words in the raw SMS data consisted of asterisks or special character symbols. This seems to occur because either users had to type with a phone-based keyboard or simply due to processing errors in the pipeline. Our aim, therefore, was to collapse these incorrectly spelled words to their closest vocabulary entires from the rest of the data.

We first built a lexicon of words using the entire data provided for the Featured Task. We then built a second probabilistic lexicon by cross-referencing *SMS dev raw* with the cleaned-up *SMS dev clean*. The first resource can be treated as a dictionary while the second is a look-up table. We processed incoming text by first selecting all the words with special characters in the text, and then computing an edit distance with each of the words in the first lexicon. We return the most frequent word that is the closest match as a substitute. For all words that don't have a closest match, we looked them up in the probabilistic dictionary and return a potential substitution if it exists. As the probabilistic dictionary is constructed using a very small amount of data, the two-level lookup helps to place less trust in it and use it only as a back-off option for a missing match in the larger lexicon.

This approach only collapses words with special characters to their closest in-vocabulary words. It does not make a significant difference to the OOV ratios, but reduces the number of tokens in the dataset. Using this approach we were able to collapse about 80% of the words with special characters to existing vocabulary entries.

3.2 Spelling Normalization

One of the most problematic issues in Haitian Creole SMS translation system is misspelled words. When training data contains misspelled words, the translation system performance will be affected at several levels, such as word alignment, phrase/rule extractions, and tuning parameters (Bertoldi et al., 2010). Therefore, it is desirable to perform spelling correction on the data. Spelling correction based on the noisy channel model has been explored in (Kernighan et al., 1990; Brill and Moore, 2000; Toutanova and Moore, 2002). The model is generally presented in the following form:

$$p(\hat{c}|h) = \arg\max_{\forall c} p(h|c)p(c) \qquad (1)$$

where h is the Haitian Creole word, and c is a possible correction. $p(c)$ is a source model which is a prior of word probabilities. $p(h|c)$ is an error model or noisy channel model that accounts for spelling transformations on letter sequences.

Unfortunately, in the case of Haitian Creole SMS we do not have sufficient data to estimate $p(h|c)$ and $p(c)$. However, we can assume $p(c|h) \approx p(c)$ and c is in the French vocabulary and is not an English word. The rationale for this, from linguistic point of view, is that Haitian Creole developed from the 18th century French. As a result, an important part of the Haitian Creole lexicon is directly derived from French. Furthermore, SMS messages sometimes were mixed with English words. Therefore, we ignore c if it appears in an English dictionary.

Given h, how do we get a list of possible normalization c and estimate $p(c)$? We use edit distance of 1 between h and c. An edit can be a deletion, transposition, substitution, or insertion. If a word has l characters, there will be 66l+31 possible corrections[2]. It may result in a large list. However, we only keep possible normalizations which appear in a French dictionary and do not appear in an English dictionary[3]. To approximate $p(c)$, we use the French parallel Giga training data from the Shared Task of the WMT11. $p(c)$ is estimated by MLE. Finally, our system chooses the French word with the highest probability.

	dev (cl)	devtest (cl)	test (cl)
Before	2.6 ; 16	2.7 ; 16	2.6 ; 16
After	2.2 ; 13.63	2.3 ; 13.95	2.2 ; 14.3

Table 3: Percentage of OOV tokens and types in test sets before and after performing spelling normalization.

Table 3 shows that spelling normalization helps to bring down the percentage of OOV tokens and types by 0.4% and 2% respectively on the three test

[2] l deletions, l-1 transpositions, 32l substitutions, and 32(l+1) insertions; Haitian Creole orthography has 32 forms.

[3] The English dictionary was created from the English Gigaword corpus.

sets. Some examples of Haitian Creole words and their French normalization are (*tropikal:tropical*), (*economiques:economique*), (*irjan:iran*), (*identifie:identifie*).

	dev (cl)	devtest (cl)	devtest (r)
Baseline	32.28	33.49	29.95
S1	32.18	30.22	25.45
S2	28.9	31.06	27.69

Table 4: Translation results in BLEU with/without spelling correction

Given the encouraging OOV reductions, we applied the spelling normalization for the full corpus, and built new translation systems. Our baseline system has no spelling correction (for the training corpus or the test sets); in S1, the spelling corrections is applied to all words; in S2, the spelling correction is only applied to Haitian Creole words that occur only once or twice in the data. In S1, 11.5% of Haitian Creole words had been mapped to French, including high frequency words. Meanwhile, 4.5% Haitian Creole words on training data were mapped to French words in S2. Table 4 presents a comparison of translation performance of the baseline, S1 and S2 for the SMS test sets. Unfortunately, none of systems with spelling normalization outperformed the system trained on the original data. Restricting the spelling correction only to infrequent words (S2) performed better for the devtest sets, but not for the dev set, although all the test sets come from the same domain.

4 Corpus Expansion using Semantic Role Labeling

To address the problem of limited resources, we tried to expand the training corpus by applying the corpus expansion method described in (Gao and Vogel, 2011). First, we parsed and labeled the semantic roles of the English side of the corpus, using the ASSERT labeler (Pradhan et al., 2004). Next, using the word alignment models of the parallel corpus, we extracted Semantic Role Label (SRL) substitution rules. SRL rules consist of source and target phrases that cover whole constituents of semantic roles, the verb frames they belong to, and the role labels of the constituents. The source and target phrases must comply with the restrictions detailed in (Gao and Vogel, 2011). Third, for each sentence, we replaced one of embedded SRL substitution rules with equivalent rules that have the same verb frame and the same role label.

The original method includes an additional but crucial step of filtering out the grammatically incorrect sentences using an SVM classifier, trained with labeled samples. However, we were unable to find Haitian Creole speakers who could manually label training data for the filtering step. Therefore, we were forced to skip this filtering step. We expanded the full training corpus which contained 124K sentence pairs, resulting in an expanded corpus with 505K sentences. The expanded corpus was force-aligned using the word alignment models trained on the original unexpanded corpus. A new translation system was built using the original plus the expanded corpus. As seen in Table 5, we observed a small improvement with the expanded corpus for the raw devtest. This method did not improve performance for the other two test sets.

	dev (cl)	devtest (cl)	devtest (r)
Baseline	32.28	33.49	29.95
+Expanded	31.79	32.98	30.1

Table 5: Translation results in BLEU with/without corpus expansion

A possible explanation for this, in addition to the missing component of filtering, is the low quality of SRL parsing on the SMS corpus. We observed a very small ratio of expansions in the Haitian Creole-English data, when compared to the Chinese-English experiment shown in (Gao and Vogel, 2011). The latter used a high quality corpus for the expansion and the expanded corpus was 20 times larger than the original one. Due to the noisy nature of the available parallel data, only 61K of the 124K sentences were successfully parsed and SRL-labeled by the labeler.

5 Extracting Parallel Data from Comparable Data

As we only have a limited amount of parallel data, we focused on automatically extracting additional parallel data from other available resources, such as comparable corpora. We were not able to find comparable news articles in Haitian Creole and English. However, we found several hundred Haitian Creole medical articles on the Web which were linked to comparable English articles[4]. Although some of the medical articles seemed to be direct translations of each other, converting the original pdf formats into text did not produce sentence aligned parallel articles. Rather, it produced sentence fragments (sometimes in different orders) due to the structural differences in the article pair. Hence a parallel sentence detection technique was necessary to process the data. Because the SMS messages are related to the disaster relief effort, which may include many words in the medical domain, we believe the newly extracted data may help improve translation performance.

Following Munteanu and Marcu (2005), we used a Maximum Entropy classifier to identify comparable sentence. To avoid the problem of having different sentence orderings in the article pair, we take every source-target sentence pair in the two articles, and apply the classifier to detect if they are parallel. The classifier approach is appealing to a low-resource language such as Haitian Creole, because the features for the classifier can be generated with minimal translation resources (i.e. a translation lexicon).

5.1 Maximum Entropy Classifier

The classifier probability can be defined as:

$$Pr(c_i|S,T) = \frac{exp\left(\sum_{j=1}^{n} \lambda_j f_{ij}(c_i,S,T)\right)}{Z(S,T)} \quad (2)$$

where (S,T) is a sentence pair, c_i is the class, f_{ij} are feature functions and $Z(S)$ is a normalizing factor. The parameters λ_i are the weights for the feature functions and are estimated by optimizing on a training data set. For the task of classifying a sentence pair, there are two classes, $c_0 = non - parallel$ and $c_1 = parallel$. A value closer to one for $Pr(c_1|S,T)$ indicates that (S,T) are parallel.

The features are defined primarily based on translation lexicon probabilities. Rather than computing word alignment between the two sentences, we use lexical probabilities to determine alignment points as follows: a source word s is aligned to a target word t if $p(s|t) > 0.5$. Target word alignment is computed similarly. We defined a feature set which includes: length ratio and length difference between source and target sentences, lexical probability scores similar to IBM model 1 (Brown et al., 1993), number of aligned/unaligned words and the length of the longest aligned word sequence. Lexical probability score, and alignment features generate two sets of features based on translation lexica obtained by training in both directions. Features are normalized with respect to the sentence length.

5.2 Training and Testing the Classifier

To train the model we need training examples that belong to each of the two classes: parallel and non-parallel. Initially we used a subset of the available parallel data as training examples for the classifier. This data was primarily sourced from medical conversations and newswire text, whereas the comparable data was found in medical articles. This mismatch in domain resulted in poor classification performance. Therefore we manually aligned a set of 250 Haitian Creole-English sentence pairs from the medical articles and divided them in to a training set (175 sentences) and a test set (100 sentences).

The parallel sentence pairs were directly used as positive examples. In selecting negative examples, we followed the same approach as in (Munteanu and Marcu, 2005): pairing all source phrases with all target phrases, but filter out the parallel pairs and those that have high length difference or a low lexical overlap, and then randomly select a subset of phrase pairs as the negative training set. The test set was generated in a similar manner. The model parameters were estimated using the GIS algorithm. We used the trained ME model to classify the sentences in the test set into the two classes, and notice how many instances are classified correctly.

Classification results are as given in Table 6. We notice that even with a smaller training set, the classifier produces results with high precision. Using

[4] Two main sources were: www.rhin.org and www.nlm.nih.gov

	Precision	Recall	F-1 Score
Training Set	93.90	77.00	84.61
Test Set	85.53	74.29	79.52

Table 6: Performance of the Classifier

the trained classifier, we processed 220 article pairs which contained a total of 20K source sentences and 18K target sentences. The classifier selected about 10K sentences as parallel. From these, we selected sentences where $pr(c_1|S,T) > 0.7$ for translation experiments. The extracted data expanded the source vocabulary by about 5%.

We built a second translation system by combining the baseline parallel corpus and the extracted corpus. Table 7 shows the translation results for this system.

	dev (cl)	devtest (cl)	devtest (r)
Baseline	32.28	33.49	29.95
+Extracted	32.29	33.29	29.89

Table 7: Translation results in BLEU with/without extracted data

The results indicate that there is no significant performance difference in using the extracted data. This may be due to the relatively small size of the comparable corpus we used when extract the data.

6 Conclusion

Building an MT system to translate Haitian Creole SMS messages involved several challenges. There was only a limited amount of parallel data to train the models. The SMS messages tend to be quite noisy. After building a baseline MT system, we investigated several approaches to improve its performance. In particular, we tried collapsing OOV words using a lexicon generated with clean data, and normalize different variations in spelling. However, these methods did not results in improved translation performance.

We tried to address the data sparseness problem with two approaches: expanding the corpus using SRL rules, and extracting parallel sentences from a collection of comparable documents. Corpus expansion showed a small improvement for the raw devtest. Both corpus expansion and parallel data extraction did not have a positive impact on other test sets. Both these methods have shown significant performance improvement in the past in large data scenarios (for Chinese-English and Arabic-English), but failed to show improvements in the current low-data scenario. Thus, we need further investigations in handling noisy data, especially in low-resource scenarios.

Acknowledgment

We thank Julianne Mentzer for assisting with editing and proofreading the final version of the paper. We also thank the anonymous reviewers for their valuable comments.

References

Jeff Allen. 1998. Lexical variation in haitian creole and orthographic issues for machine translation (MT) and optical character recognition (OCR) applications. In *Proceedings of the First Workshop on Embedded Machine Translation systems of AMTA conference*, Philadelphia, Pennsylvania, USA, October.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2010. Statistical machine translation of texts with misspelled words. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, California, June.

Eric Brill and Robert C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL 2000)*, pages 286–293.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Qin Gao and Stephan Vogel. 2011. Corpus expansion for statistical machine translation with semantic role label substitution rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, June.

Mark D. Kernighan, Kenneth W. Church, and William A. Gale. 1990. A spelling correction program based on a noisy channel model. In *Proceedings of the 13th conference on Computational linguistics - Volume 2*, COLING '90, pages 205–210.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, June.

William 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 Annual conference of the European Association for Machine Translation (EAMT)*, Saint-Raphaël, France, May.

Robert Munro. 2010. Crowdsourced translation for emergency response in haiti: the global collaboration of local knowledge. In *AMTA Workshop on Collaborative Crowdsourcing for Translation*, Denver, Colorado, USA, October-November.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

Sameer S. Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Daniel Jurafsky. 2004. Shallow semantic parsing using support vector machines. In *Proceedings of the Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting (HLT/NAACL-2004)*.

Andreas Stolcke. 2002. An extensible language modeling toolkit. In *Proc. of International Conference on Spoken Language Processing*, volume 2, pages 901–904, Denver, CO, September.

Kristina Toutanova and Robert Moore. 2002. Pronunciation modeling for improved spelling correction. In *40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*.