CMU Haitian Creole-English Translation System for WMT 2011

Sanjika Hewavitharana, Nguyen Bach, Qin Gao, Vamshi Ambati, Stephan Vogel

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA.

Introduction

The WMT 11 Featured task involved translating Haitian Creole text (SMS) messages into English. These messages were collected during the humanitarian operation in the aftermath of the earthquake in Haiti in 2010.

- Challenges involved in translating SMS messages
  - SMS messages are noisy
    - Contains fragments and incomplete information
    - Spelling variants due to non-standard orthography
    - Manually translated messages contain additional annotations
  - Lack of parallel data
    - Limited resource because Creole is a less-commonly spoken language
- Our approach was to rapidly build a statistical translation system utilizing available resources and techniques
- The effort involved
  - Clean the SMS messages by collapsing of OOV words and spelling normalization
  - Obtain additional parallel data by corpus expansion with SRL rules and by automatic extraction from comparable resources
- System Architecture
  - Translation systems were built using standard SMT tools
  - Baseline system was trained with all provided training data (124K sentence pairs)
  - One development set (dev clean) and others unseen test sets
  - Results reported in case-insensitive BLEU

Collapsing of OOV Words

- Some words in the raw SMS data contain special characters such as asterisk
  - Possibly due to processing errors in the pipeline
- Collapse special word *s* into the closest in-vocabulary entry
  - We build :
    - A dictionary using the full provided corpus
    - A lexicon by cross-referencing cleaned dev and raw dev
  - If a closet match *c* (single edit distance) exists in the dictionary, collapse *s* into the most frequent *c*
  - If no close match is found, look up in the lexicon and return a potential substitution if exists
- This approach collapses about 80% of the words with special characters

Spelling Normalization

- Performs spelling correction on misspelled Creole words
  - Single edit-distance operation on the misspelled word
    - For a word with *k* characters, there are 66*k*+31 possible corrections
  - If the correction appears in a French dictionary (Gigaword), select the French word with the highest probability as the desired correction
    - Large part of Hatian Creole lexicon contains French words
  - If the correction appears in an English dictionary (Gigaword), ignore it
- Examples: *tropikal:tropical*, *economiques:economique,* irjan:iran, idantifie:identifie
- Percentage of OOV tokens and types in test sets before/after spelling normalization

	dev (clean)	devtest (clean)	test (clean)
Before	2.6 ; 16.00	2.7 ; 16.00	2.6 ; 16.00
After	2.2; 13.63	2.3 ; 13.95	2.2 ; 14.30

- Build translation systems with spelling corrected corpus

  SN-All: Spelling normalization applied to all words (11.5%)

  SN-LFW: Only Normalize words that have a frequency < 2 (4.5%)

Corpus Expansion using SRL

- Expand the training corpus using semantic role label (SRL) substitution rules (Gao & Vogel, 2011)
  1. Parse and label semantic roles of the English side using the ASERT labeler
  2. Using the word alignment models of the parallel corpus, extract SRL substitution rules
     - A rule consists of: source & target phrases covering the semantic role, the verb frame they belong to, and role labels of the constituents
  3. For each sentence in the corpus, replace embedded SRL substitution rules with equivalent rules that have the same verb frame and the role label
  4. Filter out grammatically incorrect sentences using an SVM Classifier
     - Skipped, due to the lack of manually labeled samples for Haitian Creole to training the SVM classifier
- Training corpus with 124K sentences expanded into 505K sentences
- Build a translation system combining the baseline with the expanded corpus (B + Expanded)

Automatic Extraction of Parallel Data

- Because the available Creole-English parallel data is limited, automatically extract parallel data from the web
- We found several hundred Creole medical articles, which were linked to comparable English articles
  - Two main sources: www.rhin.org and www.nlm.nih.gov
  - Converting the pdfs into text breaks the document structure, leading to fragmented text. Also, structural differences in article pairs lead to difference in the order of the sentences
- A maximum entropy classifier similar to (Munteanu & Marcu, 2005) to extract parallel sentence pairs
  - ME score for a phrase pair (*S,T*)

$$p(c \mid S,T) = \frac{\exp(\lambda \cdot f(c,S,T))}{Z(S,T)}$$

  - Two classes: C=0 (*non-parallel*) and C=1(*parallel*)
  - Feature vector f is defined based on lexical probabilities
    - No explicit sentence alignment
  - Classifier train/test sets obtained using 175/100 parallel sentences
  - (*S,T*) is parallel if $p(c=1 \mid S,T) > \varepsilon$
  - Performance: F-1 score of 79.5% on the test set
- Detect parallel sentences in medical articles
  - 220 articles pairs with 20K/18K source/target sentences
  - Classifier selected 10K as parallel
- Build a translation system combining the baseline with the extracted data (B + Extracted)

Results

	dev (clean)	devtest (clean)	devtest (raw)
Baseline (B)	32.28	33.49	29.95
SN-All	32.18	30.22	25.45
SN-LFW	28.90	31.06	27.69
B + Expanded	31.79	32.98	30.10
B + Extracted	32.29	33.29	29.89

- The systems with spelling normalization does not outperform the baseline
  - Restricting it for low frequency words only helped in devtest
- Small improvement on raw devtest, but decrease in performance with the other two test sets
  - Low quality of SRL parsing on the SMS corpus
- No significant difference in performance with the smaller set of automatically extracted parallel data