Machine Translation from Inuktitut to English: Parsing Strategy

Christopher Liu, Brian Wang, Yao Yang
CS229 Project — Stanford University
{cwtliu, bwang16, yangyao}@stanford.edu

Motivation

- Inuktitut is one of the eight major Inuit languages of Canada. It is spoken by around 35,000 people. Unfortunately, translation tools do not yet exist for this language.
- Inuktitut is a polysynthetic language which means words are agglutinated to form bigger words.
  - ***Unmiaqjuaqmi* = *unmiaq* (boat) + *juaq* (big) + *mi* (in)**
- We wanted to explore pre-processing methods that increase translation quality for English and Inuit/Alaska Native languages.
- Data scarcity is often prohibiting translation to rare languages: we used parallel corpora of government proceedings.

Approach

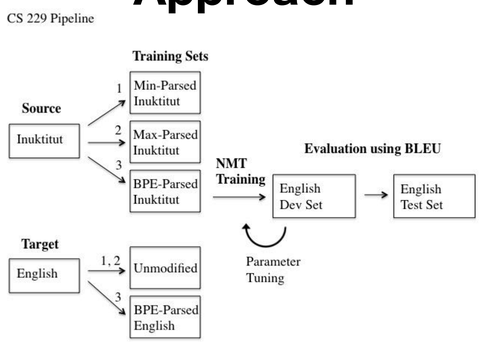

Figure 1: Pipeline of the project

1. Data tokenization (rule-based parsing and byte pair encoding).
2. Neural Machine Translation (LSTM Recurrent Neural Network) using *seq2seq* toolkit with attention.
3. Model evaluation using BLEU metric.

Data Preparation

Data:
Hansard: recordings from parliamentary proceedings (04/1999—11/2007) published by the legislative body of Nunavut
Pre-processing:

1. Modified Rule-Based Parser (Inuktitut Morphological Analyzer)
   a. Maximum and minimum parsing methods
2. BPE parsing with varying amount of token merges

Byte Pair Encoding (BPE)

- Neural machine translation typically uses finite vocabulary, but machine translation is an open vocabulary problem.
- BPE is an unsupervised tokenization method - initializes vocabulary with all individual characters and extends vocabulary repeatedly by merging the most frequent pair of existing vocabulary tokens.
- For our project, BPE was used to tokenize both training sets before model training. We used various vocabulary sizes in order to explore the level of tokenization that produces the highest accuracy.

Neural Machine Translation

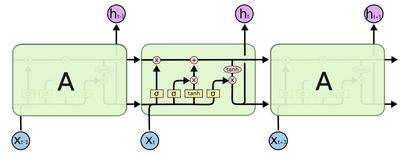

Figure 2: Representation of a LSTM RNN

- Recurrent neural networks are state-of-the-art for machine translation tasks; however, optimal neural network architecture for performance is an open question. Our method applied bidirectional models with attention.
- As part of parameter tuning, we explored the performance time trade-off of adding additional layers, units, and batch size to the model.

Experiments

We did parameter tuning using the shortest 50,000 Inuktitut lines. The BLEU results did not vary significantly so we chose a parameter set for our full runs that optimized speed. Our main analysis included three experiments using the following datasets:

1. Minimum-parsed Inuktitut → Unmodified English
2. Maximum-parsed Inuktitut → Unmodified English
3. BPE-parsed Inuktitut → BPE-parsed English (joint parsed)

Results: BLEU Scores

Exp	Dev Set	Test Set
min_parser	2.17	2.09
max_parser	4.79	4.92

Figure 3: Rule-Based parsing results at maximum BLEU

Exp	Dev Set	Test Set
BPE 10k	2.22	2.55
BPE 32k	2.32	2.95

Figure 4: BPE number of merges comparison at step with highest BLEU, from English to Inuktitut

Results: BLEU Graphs

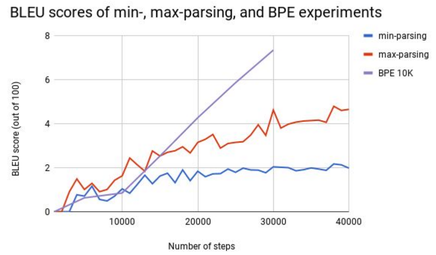

Figure 5: Min/max rule-based and BPE 10K parser results

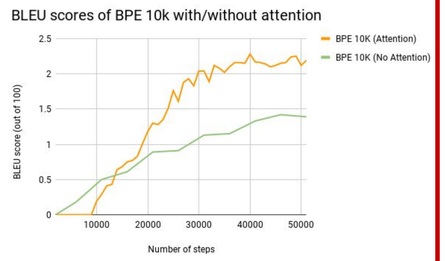

Figure 6: BPE results comparing 10k and 32k (Eng -> Inuk)

Analysis

- Conclusions
  - Full Experiments
    - The max-parsing strategy outperformed the min-parsing strategy
    - So far, BPE-based parsing returns higher BLEU than rule-based parsing
  - BPE 10K vs 32K
    - BPE at 10K merges performed similarly to the BPE at 32K merges
- Challenges
  - Determining optimal vocabulary size specific to Inuk/Eng and dataset size
- Next Steps
  - More analyses to report in the final paper (more BPE sizes, tuning batch)
- Future work
  - Gather more training data
  - Training on standardized post-base conjugations
  - Targeted BPE script specific to the language pair and dataset amount

Complementary Project

- How might selective data augmentation improve translator performance?
- We added (1) the full Bible or (2) only conversational Bible verses to our dataset.
- Adding the full Bible significantly decreased prediction BLEU, while adding only conversational Bible verses had a similar or marginally higher prediction BLEU.

Acknowledgments & References

Professors Andrew Ng and Dan Boneh, TA: Ziang Xie

[1] The UQAILAUT Project. http://www.inuktitutcomputing.ca/Uqailaut/info.php
[2] Sennrich, R.; Haddow, B.; Birch A. (2016). Neural Machine Translation of Rare Words with Subword Units.
[3] Bahdanau, D.; Cho, K. & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate.
[4] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation.
[5] Yonghui W. et al. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation