shagunsodhani / **OpenVocabularyNMT.md**

Created 5 years ago • Report abuse

Star

Code Revisions 1 Stars 1

Summary of "Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models" paper

`OpenVocabularyNMT.md`

Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models

Introduction

- The paper presents a novel open vocabulary NMT(Neural Machine Translation) system that translates mostly at word level and falls back to character level models for rare words.
- Advantages:
  - Faster and easier to train as compared to character models.
  - Does not produce unknown words in the translations which need to be removed using *unk replacement* techniques.
- Link to the paper

Unk Replacement Technique

- Most NMT operate on constrained vocabulary and represent unknown words with *unk* token.
- A post-processing step replaces *unk* tokens with actual words using alignment information.
- Disadvantages:
  - These systems treat words as independent entities while they are morphologically related.
  - Difficult to capture things like name translation.

Proposed Architecture

Word-level NMT

- Deep LSTM encoder-decoder.
- Global attention mechanism and bilinear attention scoring function.
- Similar to regular NMT system except in the way unknown words are handled.

Character-level NMT

- Deep LSTM model used to generate on-the-fly representation of rare words (using final hidden state from the top layer).
- Advantages:
  - Simplified architecture.
  - Efficiency through precomputation - representations for rare sources words can be computed at once before each mini-batch.
  - The model can be trained easily in an end-to-end fashion.

Hidden-state Initialization

- For source representation, layers of the LSTM are initialized with zero hidden states and cell values.
- For target representation, the same strategy is followed except for the hidden state of the first layer where one of the following approaches are used:
  - **same-path** target generation approach
    - Use the context vector just before softmax (of word-level NMT).
  - **seperate-path** target generation approach
    - Learn a new weight matrix **W** that will be used to generate the context vector.

Training Objective

- $J = J_w + \alpha J_c$
- J - total loss
- J_w - loss in a regular word-level NMT
- αJ_c - loss in the character-level NMT

Word Character Generation Strategy

- The final hidden state from character-level decoder could be interpreted as the representation of *unk* token but this approach would not be efficient.

- Instead, *unk* is fed to the word-level decoder as it is so as to decouple the execution for the character-level model as soon the word-level model finishes.
- During testing, a beam search decoder is run at the word level to find the best translation using the word NMT alone.
- Next, a character-level encoder is used to generate the words in place of *unk* to minimise the combined loss.

Experiments

Data

- WMT'15 translation task from English into Czech with newstest2013 (3000 sentences) as dev set and newstest2015 (2656 sentences) as a test set.

Metrics

- Case-sensitive NIST BLEU.
- chrF3

Models

- Purely word based
- Purely character based
- Hybrid (proposed model)

Observations

- Hybrid model surpasses all the other systems (neural/non-neural) and establishes a new state-of-the-art result for English-Czech translation in WMT'15 with 19.9 BLEU.
- Character-level models, when used as a replacement for the standard unk replacement technique in NMT, yields an improvement of up to +7.9 BLEU points.
- Attention is very important for character-based models as the non-attentional character models perform poorly.
- Character models with shorter time-step backpropagation perform inferior as compared to ones with longer backpropagation.
- Separate-path strategy outperforms same-path strategy.

Rare word embeddings

- Obtain representations for rare words.

- Compare the Spearman correlation between similarity scores assigned by humans and by the model.
- Outperforms the recursive neural network model (which also uses a morphological analyser) on this task.