Karlsruhe Institute of Technology
Dr. Eunah Cho
Institute for Anthropomatics and Robotics, Interactive Systems Labs
eunah.cho@kit.edu

MT Praktikum - NMT

14. Juli 2017

Um die Aufgaben auszuführen, können Sie Ihre Daten in folgendem Verzeichnis speichern: `/project/smtstud/ss17/systems/USERNAME/`

We are going to use a test data in German. Copy this file into your directory.

`/project/smtstud/ss17/data/nmt/test.de`

Later you are going to preprocess them. Copy it into your directory.

`/project/smtstud/ss17/bin/Preprocess.de.sh`

NMT: Preprocess and sub-word units

1. We are going to load a translation model and translate a test set from German to English. Preprocess the test data using the script.

   - Check which files are created. What is the difference?

   The test data is tokenized and true-casing is applied.

2. Copy the sub-word operation script into your directory and run it for the preprocessed file.

   `/project/smtstud/ss17/bin/BPE.sh`

   - Which words are split and how?

   Less frequent words are split.

3. Now you have the preprocessed file to translate. Log into rg3hpc1, and then to i13hpc28 or i13hpc29. Once you are in either i13hpc28 or i13hpc29, Move back into your working directory. Copy the script for translation into your directory, and run the translation. It may take some minutes to load the model.

   `/project/smtstud/ss17/bin/Translate.sh`

Word	How it is split
offenem	offen@@ em
Musik-Sequencer	Musik@@ -Sequ@@ en@@ cer
Live-Auftritte	Live-@@ Auftri@@ tte
Siftables	Si@@ ft@@ ables
Führungsstimme	Führungs@@ stimme
Bass	B@@ ass
Schlagzeug	Schlag@@ zeug
einspeisen	ein@@ speisen

4. Preprocess the English reference data the same way. Copy the files and preprocess it.

   `/project/smtstud/ss17/data/nmt/test.en`
   `/project/smtstud/ss17/bin/Preprocess.en.sh`

5. Measure the performance in BLEU. Copy the file, and run it as ./BLEU_evaluation.sh hypfile reffile.

   `/project/smtstud/ss17/bin/BLEU_evaluation.sh`

   - How high is the BLEU score?

   21.38

6. Copy an example file into your directory.

   `/project/smtstud/ss17/data/nmt/example.de`

   Feel free to change text in the text. Apply the same steps before (Process - BPE) and translate it. How is the translation?

Characteristics of NMT

1. Copy another example file into your directory.

   `/project/smtstud/ss17/data/nmt/test_shift.de`

   It contains partial test sentences, whose words are appended line by line. Apply preprocessing and BPE in order to translate it.

   - How is the translation for partial sentences?

   The performance of NMT suffers greatly for partial sentences.

2. Copy another example file into your directory.

   `/project/smtstud/ss17/data/nmt/lecturetest.de`

   It contains lecture excerpts. Apply preprocessing and BPE in order to translate it.

   - How is the translation? What is the most distinctive characteristics? Why?

   Often rare words are not processed well. Repetition happens around rare or unknown words.

Attention in NMT

1. We are going to examine attention matrix. Copy the following test set into your directory.

   `/project/smtstud/ss17/data/nmt/test16.head.de`

   Run `/project/smtstud/ss17/bin/Attention.sh` for the input file `test16.head.de`. This may take upto several minutes. Look into the file `test16.head.de.attention` for the attention values.

2. You can create a drawing based on attentional values. Note that attention file should contain the attentional values for one sentence only.

   `python /project/smtstud/ss17/bin/draw.py $attentionFile`

   For example, you can try out following files for attentionFile.

   `/project/smtstud/ss17/data/nmt/sent1.de.attention`
   `/project/smtstud/ss17/data/nmt/sent4.de.attention`
   `/project/smtstud/ss17/data/nmt/sent14.de.attention`

   - What information does the attentional matrix carry?

   It contains alignment information between source and target sentences.