Alex Flückiger
S12-452-223
21. Mai 2018

VL Machine Translation: Assignment 5

Report

Für das Training des MT-Systems von Deutsch nach Englisch wurde der bereitgestellte Europarl-Korpus verwendet. Als Preprocessing-Schritte wurden die Moses-Scripts zur Interpunktions-Normalisierung, Tokenisierung und Truecasing verwendet. Ausserdem wurde ein Joint Byte Pair Encoding (BPE) Modell mit 100'000 Iterationen von subword-nmt angewendet, das auf den aufbereiteten Trainings-Dateien beider Sprachen trainiert wurde. Gemäss den Empfehlungen wurden nur Wörter mit BPE encodiert, die mindestens 50 Mal in einer Sprache vorkommen, um beim Testen keine unbekannten Tokens zu produzieren, die nur in der jeweils anderen Sprache vorkommen.

Insgesamt wurden drei eigene Modelle getestet, die allesamt auf GRU-Zellen basieren statt auf LSTM-Zellen, wie in der Baseline. Dadurch wird ein schnelleres Training erreicht, was aber vermutlich zu Lasten der Performance geht. GRU können teilweise schlechter mit langen Dependenzen in Sequenzen umgehen. Für Hyperparameter, die hier nicht explizit erwähnt sind, wurden die Default-Werte aus der Baseline übernommen.
Neben dem umgestellten Preprocessing und dem erweiterten Training (mehr Epochen) liegt dem ersten eigenen Modell nur die Änderung von LSTM- zu GRU-Zellen zugrunde. Im zweiten Modell wurde Bidirektionalität für das Encoding implementiert, damit beim Decoding sowohl Informationen über den Anfang wie auch den Schluss des Input-Satzes zur Verfügung stehen. Zusätzlich zur Bidirektionalität wurde auch ein Dropout von 0.3 für Encoder und Decoder definiert (nicht für die Embeddings). Da sich der Dropout auf die aktiven Neuronen bezieht (keep probability) statt auf deren Inaktivierung, wie fälschlicherweise gedacht, hat dies zu einem starken Underfitting des RNN geführt. Im dritten Modell wurde dies korrigiert, indem stattdessen eine Dropout keep probability von 0.7 definiert wurde. Vermutlich ist diese Massnahme gegen Overfitting aber wahrscheinlich kontraproduktiv bei der "geringen" Anzahl an Epochen und führt zu einer langsameren Konvergenz.

Die Ergebnisse finden sich in der Tabelle unten. Mit diesen Erweiterungen konnte der BLEU-Score gegenüber der stärkeren Baseline um ca. 3.5 Punkte auf 30.33 verbessert werden.

Ausserdem wurde noch eine Unterdrückung von UNK-Tokens in Übersetzungen implementiert. Dies macht allerdings keinen Unterschied, da unbekannte Tokens zumindest im Dev-Set nie als wahrscheinlichste Übersetzung gewählt wurden. Dies dürfte der Regelfall sein für fast alle Übersetzungen, sofern das Vokabular nicht zu klein gewählt und BPE angewendet wurde.

Eigentlich wollte ich auch noch Beamsearch für ein bereits trainiertes Modell implementieren, was aber zu viel Basteln erforderlich gemacht hätte. Beamsearch sollte bereits von Anfang an im Computational Graph definiert werden.

Alex Flückiger
S12-452-223
21. Mai 2018

Bzgl. Übungsorganisation: Nächstes Mal für diese Übung mindestens 3 Wochen einplanen und evtl. Pflicht zur Gruppenbildung sowie Restriktion für GRU-Zellen. Ansonsten reichen die vorhandenen Ressourcen bzw. Zeit nicht aus verschiedene Dinge auszutesten. Dies ist Schade, da dann nicht Wissen und Motivation die limitierenden Faktoren sind. :)

Model	**Description**	**Epochs**	**BLEU (Dev-Set)**
1_baseline	Vanilla LSTM	6	19.59
2_baseline	Vanilla LSTM (larger vocabulary)	10?	26.8
1_custom	GRU	9	29.1
2_custom	Bi-GRU Dropout_keep_prob 0.3	8	26.8
3_custom	Bi-GRU Dropout_keep_prob 0.7	10	30.33

Preprocessing Steps

TRAIN-DE
perl ../mosesdecoder/scripts/tokenizer/normalize-punctuation.perl < corpus.train.de > corpus.train.de.norm

perl ../mosesdecoder/scripts/tokenizer/tokenizer.perl -l de -q < corpus.train.de.norm > corpus.train.de.tok

perl ../mosesdecoder/scripts/recaser/train-truecaser.perl --corpus corpus.train.de.tok --model truecase-model.de

perl ../mosesdecoder/scripts/recaser/truecase.perl --model truecase-model.de < corpus.train.de.tok > corpus.train.de.tc

TRAIN-EN
perl ../mosesdecoder/scripts/tokenizer/normalize-punctuation.perl < corpus.train.en > corpus.train.en.norm

perl ../mosesdecoder/scripts/tokenizer/tokenizer.perl -l en -q < corpus.train.en.norm > corpus.train.en.tok

perl ../mosesdecoder/scripts/recaser/train-truecaser.perl --corpus corpus.train.en.tok --model truecase-model.en

perl ../mosesdecoder/scripts/recaser/truecase.perl --model truecase-model.en < corpus.train.en.tok > corpus.train.en.tc

BPE
python ../subword-nmt/learn_joint_bpe_and_vocab.py -i corpus.train.de.tc corpus.train.en.tc -s 10000 -o bpe.codes.de_en --write-vocabulary vocab.de vocab.en

python ../subword-nmt/apply_bpe.py -c bpe.codes.de_en --vocabulary vocab.de --vocabulary-threshold 50 < corpus.train.de.tc > corpus.train.de.bpe
python ../subword-nmt/apply_bpe.py -c bpe.codes.de_en --vocabulary vocab.en --vocabulary-threshold 50 < corpus.train.en.tc > corpus.train.en.bpe

DEV-DE
perl ../mosesdecoder/scripts/tokenizer/normalize-punctuation.perl < corpus.dev.de > corpus.dev.de.norm

perl ../mosesdecoder/scripts/tokenizer/tokenizer.perl -l de -q < corpus.dev.de.norm > corpus.dev.de.tok

perl ../mosesdecoder/scripts/recaser/truecase.perl --model truecase-model.de < corpus.dev.de.tok > corpus.dev.de.tc

python ../subword-nmt/apply_bpe.py -c bpe.codes.de_en --vocabulary vocab.de --vocabulary-threshold 50 < corpus.dev.de.tc > corpus.dev.de.bpe

TEST-DE
perl ../mosesdecoder/scripts/tokenizer/normalize-punctuation.perl < corpus.test.de > corpus.test.de.norm

perl ../mosesdecoder/scripts/tokenizer/tokenizer.perl -l de -q < corpus.test.de.norm > corpus.test.de.tok

perl ../mosesdecoder/scripts/recaser/truecase.perl --model truecase-model.de < corpus.test.de.tok > corpus.test.de.tc

python ../subword-nmt/apply_bpe.py -c bpe.codes.de_en --vocabulary vocab.de --vocabulary-threshold 50 < corpus.test.de.tc > corpus.test.de.bpe

TF Training

daikon train --source corpus/corpus.train.de.bpe --target corpus/corpus.train.en.bpe --sample_after_epoch

Alex Flückiger
S12-452-223
21. Mai 2018

TF Translation

daikon translate -i corpus/corpus.dev.de.bpe -o corpus/translation.en

Postprocessing

sed "s/\@\@ //g" < translation.en > translation.nobpe.en

perl ../mosesdecoder/scripts/recaser/detruecase.perl \
< translation.nobpe.en > translation.ntc.en

perl ../mosesdecoder/scripts/tokenizer/detokenizer.perl -l en -q \
< translation.ntc.en > translation.detok.en

perl ../mosesdecoder/scripts/tokenizer/normalize-punctuation.perl \
< translation.detok.en > translation.norm.en

BLEU
python ../compute_bleu.py -trg corpus.dev.en -trans translation.norm.en

Other Commands

ssh fla@35.185.16.67

ssh -L 16006:127.0.0.1:6007 fla@35.185.16.67

tensorboard --logdir logs --port=6007
CUDA_VISIBLE_DEVICES=2

scp -r daikon fla@35.185.16.67:daikon/.