Markus Kängsepp
Ian Mackerracher
Andre Tättar

Milestone 4: Intermediate Report (Team B)

1. Models trained

1.1 Hyperparameter tuning based on “Massive Exploration of Neural Machine Translation Architectures”

In this milestone we trained model with hyperparameters based on “Massive Exploration of Neural Machine Translation Architectures”. We changed embedding dimension from 500 to 512, also increased the encoder and decoder depth to 4 and changed attention type to Bahdanau, and changed the encoder to bidirectional. We also changed the optimization type from SGD to Adam and changed learning rate used in the article.

The model training looked well for first five epochs, however after that the accuracy started to degrade and we did not get any useful results.

List of trained models:

1. *tuned-model_acc_58.73_ppl_11.84_e1.pt*
2. *tuned-model_acc_60.92_ppl_9.96_e2.pt*
3. *tuned-model_acc_61.98_ppl_9.19_e3.pt*
4. *tuned-model_acc_62.68_ppl_8.81_e4.pt*
5. *tuned-model_acc_63.23_ppl_8.51_e5.pt*
6. *tuned-model_acc_40.00_ppl_60.34_e6.pt*
7. *tuned-model_acc_24.59_ppl_1112.12_e7.pt*
8. *tuned-model_acc_28.11_ppl_240.63_e8.pt*
9. *tuned-model_acc_27.94_ppl_258.12_e9.pt*
10. *tuned-model_acc_27.62_ppl_243.02_e10.pt*

From here we can see that after the 5th epoch, the accuracies and perplexities started to go down. Also tried to translate with the best model and it gave about 5 points lower score.

1.2 Beam size 10

Next we tried the translation half of hyperparameter tuning based on the article. For that we used beam size 10. We expect better translation quality with higher beam size, so let’s check some sentences with bad translations.

Markus Kängsepp
Ian Mackerracher
Andre Tättar

Examples:

1. ID:73
2. Source:
    a. Läti riigivõlg on Euroopa Liidus üks väiksemaid ning eelduste kohaselt jääb see keskmise tähtaja jooksul märgatavalt alla riigi koguvõla kriteeriumi (60% SKT-st), nagu määratletud Maastrichti lepingus.
3. Human:
    a. The level of the general government debt in Latvia is among the lowest in the European Union and is expected to remain considerably below the gross government debt volume criterion in the medium-term (60% of GDP) as defined in the Maastricht Treaty .
4. Base-line:
    a. Latvia ' s public debt is one of the smallest , and is expected to remain significantly below the State gross debt criterion in the medium term (60% of GDP) , as defined in the Maastricht Treaty .
5. New translation:
    a. Latvia ' s sovereign debt is one of the smallest in the European Union and is expected to remain significantly below the State gross debt criterion in the medium term (60% of GDP) , as defined in the Maastricht Treaty .

1. ID:108
2. Source: Riigi haldusasutuste napp personal vähendab nende suutlikkust tagada ELi direktiivide ülevõtmisprotsessi kvalitatiivne ja õigeaegne vastavus Läti õigussüsteemile.
3. Human:
    a. Limited human resources of public administration institutions reduce their ability for a qualitatively and timely controlled compliance of EU directives' transposition process with the Latvian legal system .
4. Base-line:
    a. The shortage of public administrations in the country reduces their capacity to ensure a qualitative and timely compliance with the transposition process of EU directives .
5. New Translation:
    a. The scarce staff of public administrations will reduce their capacity to ensure a qualitative and timely alignment of the EU directives to the Latvian legal system .

1. ID:121
2. Source:
    a. Kulutamisotsuste fookuses on tänapäeva maailma poolt meie ühiskonnale esitatud väljakutsetele vastamine ELi kodanike parema elu huvides.
3. Human:
    a. The focus of spending decisions is on meeting the challenges of the modern world to our society in the interests of a better life for the citizens of the EU .
4. Machine:

    a. The focus of the Cultural Decisions on the challenges of today's world is to meet the challenges of the EU's citizens in the interests of better life .
5. Machine:
    a. The focus of spending decisions on the challenges of today's world is to meet the challenges of the EU's citizens in the interests of better life .

Mostly the translations were almost the same, however for some cases the translations with beam 10 gave better results and for some worse. But overall the results were little better.

1.3 Replace all dots before last one

In this step we preprocessed the data, so that all the dots before last dot we changed to "__DOT__". After we trained the model with same parameters as in base-line. After translation "__DOT__" signs were replaced with actual dots. The dot-model improved the sentences, where the dot was in the middle of the sentence. Also for sentences where there where number with dot in the beginning.

Examples:

1. **ID:123**
2. Source:
    a. Meie veebisaidi sellest osast leiate teavet, kuidas korraldab parlament erinevate komisjonide süsteemi abil oma tööd. Euroopa Parlamendi töö on oluline seepärast, et otsuseid uute Euroopa seaduste kohta teevad paljudes poliitikavaldkondades ühiselt parlament ja ministrite nõukogu, mis esindab liikmesriike.
3. Human translation:
    a. In this part of our website , you can find information on how the Parliament organises its work , through a system of specialised committees . the work of the European Parliament is important because in many policy areas , decisions on new European laws are made jointly by Parliament and the Council of Ministers , which represents Member States .
4. Base-line translation:
    a. This part of our website will find information on how Parliament will organise its work through the various committees .
5. **New translation:**
    a. This section of our website will find information on how Parliament operates its work through a system of various committees , and the work of the European Parliament is therefore important because decisions on new European laws are jointly made by the Parliament and the Council of Ministers .

1. **ID:420**
2. Source:
    a. Enne kui hakkame tulevikku avastama, vaatame, kuidas me jõudsime tänapäeva. alustame sellega, et läheme tagasi aastasse 1800, kus info leidmine toimus hoopis teisiti.
3. Human:

    a. Before we start to explore the future , let's see how we reached the present . we'll begin by going back to 1800 — when finding out information was very different .
4. Base-line translation
    a. Before we start exploring the future , we'll see how we made it to the day we go back to 1800 .
5. **New translation:**
    b. Before we start exploring the future , we'll see how we arrived today , so we go back to 1800 where finding the information was completely different .

Based on these examples we can see that given approach was also translating the other half of the sentence. However it did not put dot in between two sentences for some reason.

Also it helped for following type of sentences:
Example:

1. **ID:490**
2. Source:
    a. 2. selle saamine reaalajas.
3. Human:
    a. 2 . getting it live .
4. Base-line: It's getting in real time .
5. New Translation: 2 . getting it in real time .

1. **ID:494**
2. Source:
    a. 5. võimalus võtta osa ja kommenteerida.
3. Human:
    a. 5 . being able to take part and comment .
4. Base-line:
    a. 5 .
5. **New translations:**
    a. 5 . possibility to take part and comment .

However for some sentences it still did not translate the sentence after number and dot.

1.4 Context + coverage

Goal was to start training a model with context + coverage on third week, but it failed, because of some slurm error. We chose context gates + coverage (instead of attention) to hopefully increase translation fluency and adequacy in order to get better score.
So more in detail:

- Coverage - indicates whether or not source word has been translated or not
    - Proved to do less over/undertranslation
    - Coverage is attention type and is a upgrade to standard attention mechanism, study showed that this led to increase in BLEU by 1.8 points.

- Context gate - dynamically control the rations at which source and target contexts contribute to the generation of target words.
    - Improves adequacy
    - Studies show it increased BLEU by 1.6 points

Because studies show quite a nice increase with context-gates and coverage, hopefully it will work for us, since it had error the first try. Also this requires sockeye backend instead of OpenNMT, hopefully it will work and we get a better system.

2. Translating In-domain development and test set:

BLEU scores for de-preprocessed in-domain dev and test set. Where in-domain development set has 1000 sentences and test set has 50000 sentences, we also cleaned the datasets in order to clean these from empty and weird sentences. (What are weird sentences?)

2.1 Base-line model BLEU scores:

- In-domain dev set (accurate): 21.97
- Test set: 35.90

2.2 Dot-model BLEU scores:

- In-domain dev set (accurate): 22.35
- Test set: 35.89

2.3 Beam 10 model scores

- In-domain dev set (accurate): 22.11
- Test set: 35.77

3. Future work

- Get sockeye to work with context gates and coverage (using baseline data)
    - If sockeye was better, train final system with bpe on dots replaced data
    - We are hoping, this model will be better.
- Start playing with and tuning translation part, many methods here. A few:
    - Use ensemble methods
    - Find best beam size
    - Look into other translation parameters
    - Good, because doesn't require training, will be much quicker and possibly could use CPU-s for this task -> GPU queues are full.