Andre Tättar
Markus Kängsepp
Ian Mackerracher

Milestone 3 - NMT Improvements (TEAM B)

Issues of the NMT baseline model

1. Large chunks of translation are missing, mostly because couple of sentences are together in input and model thinks of “dot” as end of sentence.

Examples:

1. ID: 123
    a. Source: 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.
    b. Human: 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 .
    c. Machine: This part of our website will find information on how Parliament will organise its work through the various committees .

Solution:

2. Replace all dots before last one with special symbol.
    a. Might have problems with keeping a lot of context from previous sentence parts. So might create (noisy) bad output as long sentences are problematic in NMT.
3. Alternate solution: split sentences.
    a. One problem: Correct sentence splitting (don’t want to split on every dot, only every sentence). Might lead to wrong outputs, if sentences wrongly tokenized.
4. Translate sentence by chunks (using Sockeye)
    a. Might be best. Also implemented into the Sockeye software.

2. Adequacy and Fluency problems in many sentences

Examples:

1. ID: 490
    a. Source: 1. info saamine kaugetest kohtadest.

   b. Human: 1 . getting information from distant places .
   c. Machine: Getting out of the remote places .
2. ID: 199
   a. Source: Keskkonna hea tervis on eurooplaste ja nende riikide valitsuste jaoks oluline teema.
   b. Human: A healthy environment is a big issue for Europeans and their governments .
   c. Machine: Good health is an important issue for Europeans and governments of these countries .

Solution:

1. Coverage and Context gates for improving translations.
   a. Promising results from the authors of the research.
      i. Coverage - indicates whether or not the source word has been translated or not.
      ii. Context gate - dynamically control the ratios at which source and target contexts contribute to the generation of target words.
      iii. Research available here: https://github.com/tuzhaopeng/NMT
2. Hyperparameter tuning based on “Massive Exploration of Neural Machine Translation Architectures”
   a. Paper showed that just by changing some key parameters or RNN models, a significant performance can be achieved (BLEU score) compared to other sophisticated models.
   b. Use LSTMs instead of GRU/vanilla. (Probably done by default)
   c. Bidirectional LSTMs
   d. Other parameters that can be changed (....)
   e. We could get a new better baseline.
   f. Compare current number to their proposed best numbers. How different are them?

Hyperparameter	Value
embedding dim	512
rnn cell variant	LSTMCell
encoder depth	4
decoder depth	4
attention dim	512
attention type	Bahdanau
encoder	bidirectional
beam size	10
length penalty	1.0

Table 7: Hyperparameter settings for our final combined model, consisting of all of the individually optimized values.

Model	newstest14	newstest15
Ours (experimental)	22.03	24.75
Ours (combined)	22.19	25.23
OpenNMT	19.34	-
Luong	20.9	-
BPE-Char	21.5	23.9
BPE	-	20.5
RNNSearch-LV	19.4	-
RNNSearch	-	16.5
Deep-Att*	20.6	-
GNMT*	24.61	-
Deep-Conv*	-	24.3

Table 8: Comparison to RNNSearch (Jean et al., 2015), RNNSearch-LV (Jean et al., 2015), BPE (Sennrich et al., 2016b), BPE-Char (Chung et al., 2016), Deep-Att (Zhou et al., 2016), Luong (Luong et al., 2015a), Deep-Conv (Gehring et al., 2016), GNMT (Wu et al., 2016), and OpenNMT (Klein et al., 2017). Systems with an * do not have a public implementation.

3. Words with ambiguous meaning are translated incorrectly.

Examples:

1. ID: 967

a. Source: Kommentaarides kasutatud nime lisamine või muutmine
b. Reference: Add or change the name used in comments
c. Machine: Inclusion or modification of the name used in the comments

Solution:

1. Include POS tags as features in the encoding layer

4. Named entities are translated.

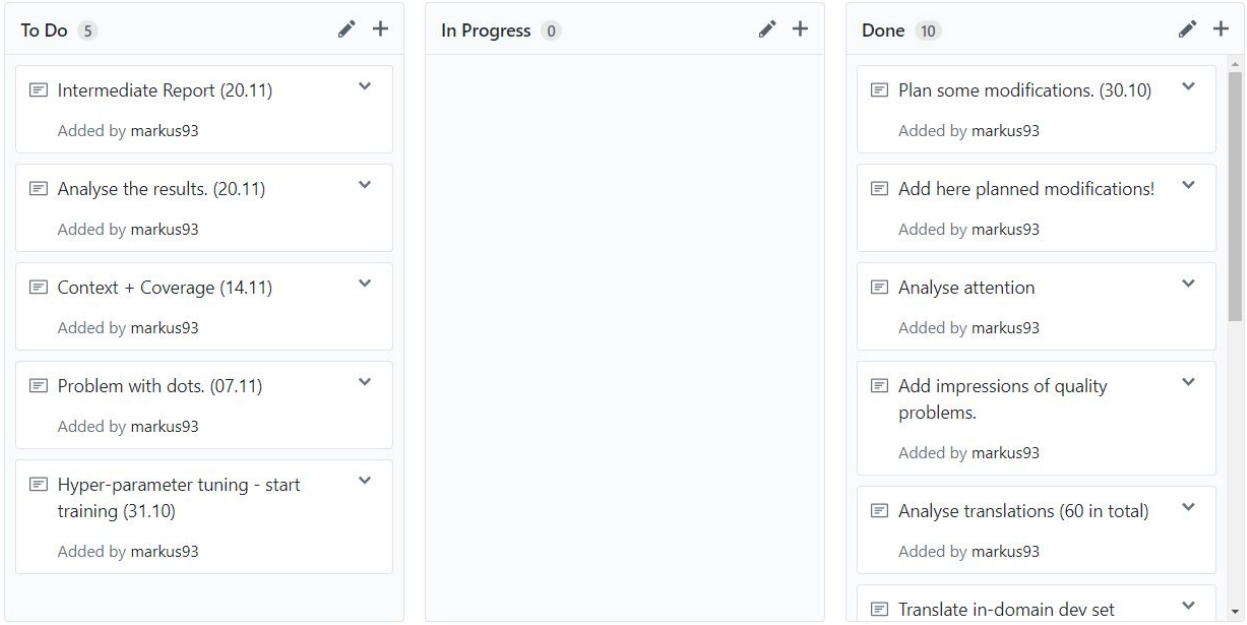
Examples:

1. ID: 333
    a. Source : Aeg / ELECTROLUXi masinad vastavad üldistele ohutusstandarditele ning masinate ohutust käsitlevatele õigusaktidele.
    b. Reference : The safety of AEG / ELECTROLUX appliances complies with the industry standards and with legal requirements on the safety of appliances .
    c. Machine : Time / Electrolux machines meet general safety standards and safety legislation on machinery .

Solution:

1. Not sure yet.
    b. Idea: replace all named entities with token, e.g. <NE>
    c. REMARK: Look at the OpenNMT solution - keeping named entities when tag is Named entity for example
    d. REMARK: Also for numerics, consider replacing numerics with single token.

REMARK: BPE with semantical information - HOW ?? We don't want to do bpe splits on things like dog_Noun_dog, which is word_POS-TAG_LEMMA. So what happens there ? Should research this quickly.



Andre Tättar
Markus Kängsepp
Ian Mackerracher

Could not fit all the solutions into 3 weeks timeframe as all of them would need most likely retraining. So we skipped third and fourth issues for this milestone.