Decoding of a Word

- The decoder computes a distribution over the vocabulary
  \[ p(y|h) = g(x, h, \theta) \]

- Vanilla attentive decoder: softmax over hidden state projection
  \[ e_{ij} = y_j^T W_o t_i \]
  \[ p(y_i|h) = \frac{\exp e_{ij}}{\sum_{k=0}^{|V|} \exp e_{kj}} \]

  ...\(y_i\) is an one-hot representation of the target word
  ...\(t_i\) is the output hidden state of the decoder
  ...\(W_o\) is a trainable parameter matrix

- The softmax computation is expensive (cannot be computed in parallel)
Limited Vocabulary

- Simple solution: limit the size of the vocabulary
- Usually around 50k
- Many words get thrown away

Ein brauner Hund rennt dem schwarzen Hund hinterher.

↓

Ein <unk> Hund rennt dem <unk> Hund <unk>.
Large vocabulary NMT


- During training, only a subset of the vocabulary by segmenting the training data
- During testing, compose the vocabulary from:
  - A list of $K$ most frequent words (15–50 thousand)
  - $K'$ candidate translations for each word (10–20 words)
    - Candidate translations are obtained from alignment
- Can use very large vocabulary (~500k)
Copying Source Words


- Learn alignment on the training data
- During truncating the vocabulary, keep track of relative positions of the `<unk>` tokens
- Postprocess the result by replacing the unknown tokens with dictionary translations of the corresponding source words
- If there is no dictionary translation, just copy the source word
Subword Units

Sennrich et al., 2016
(https://arxiv.org/abs/1508.07909)

- **Byte-pair encoding (BPE)**
- Build vocabulary from characters, merging to larger groups
- Stop when vocabulary limit is reached
- Rare words are modeled as groups of subwords
- There are no out-of-vocabulary tokens
Character-level Decoding

Chung et al., 2016 (https://arxiv.org/abs/1603.06147)

- Encoder works on byte-pair encoded source sentences
- Decoding is done one character at a time
https://arxiv.org/pdf/1703.04357

Question:
Compare the Nematus models with the models from Bahdanau et al., 2014. How do they differ? Think of at least three differences.