Université Paris Cité

Decoding in NMT

Guillaume Wisniewski
guillaume.wisniewski@u-paris.fr

November 2025

This lab aims at comparing different decoding methods for NMT : we will assume that the system has already been trained and focus on predicting new translation hypothesis. We will use `JoeyNMT` a pedagogical implementation of a modern NMT system (i.e. a transformer model).

A English → Portuguese `JoeyNMT` model can be downloaded from the lecture website. In all your experiments you should use the 1,000 longest sentences of the English-Portuguese corpus available on the Tatoeba website. [1]

`JoeyNMT`, as all other modern NMT systems, relies on a single configuration file to describe the architecture of the system (number of layers, size of hidden representation, ...), the training (scheduler, optimization method, ...) and even how new translation must be predicted. To use the model you have just downloaded you must edit the configuration file in the archive to update the paths to files and models on your environment.

1 Installing and running `JoeyNMT`

You can install `JoeyNMT` with the following command :

```
pip install joeynmt==2.1.0
```

(as you already know it is best to run the command in a virtual environment).

2 Impact of the beam size

1. Using the provided model, translate the test set using beams of size 1 (greedy), 5, 10, 15 and 20. For each decoding report :
   — the time needed to decode the test set ;
   — the `Bleu`, `chrF` and $\texttt{CometKiwi}_{22}^{\texttt{DA}}$ scores ;

1. https://tatoeba.org/fr/downloads

— the number of hypotheses that are exactly the same as the hypotheses of the greedy decoding ;
— the 5 hypotheses with the most differences from those generated by greedy decoding.

2. Comment
3. Are the variations in $\texttt{CometKiwi}_{22}^{\texttt{DA}}$ scores significant (you can look here to know how to interpret $\texttt{CometKiwi}_{22}^{\texttt{DA}}$ scores)

3 Evaluating Error Propagation

You will find on the lecture website, a script that allows to translate a sentence with a greedy decoder. We will now modify this script to show that NMT systems suffer from exposure bias.

Basically, the decoder is a model which successively estimates the probability $\mathbb{P}(w_t|\mathbf{s}, \mathbf{w}_{<t})$ of generating a word w_t in the target language, corresponding to the t-th word of the translation hypothesis, knowing the representation of the source sentence $\mathbf{s}$ constructed by the encoder and the target words already generated $\mathbf{w}_{<t}$. It is possible, at any time, to know whether the generated word is correct or not by comparing it with the t-th word of the reference. The decoder can therefore be evaluated simply by a 0/1 loss.

To demonstrate the exposure bias, it's enough (emphasis on enough) to compare the average 0/1 loss of a translation when the translation history $\mathbf{w}_{<t}$ contains no errors (i.e. it's made up of the first t words of the reference) and when it's predicted (i.e. it's made up of the t words predicted during normal greedy decoding).

4. Modify the provided script to implement the experiment I've just described. What can you conclude ?

4 ϵ-sampling Decoding

5. Drawing on your experience gained from implementing the previous experiment, implement a decoder using the ϵ-sampling strategy described in this article (section 4.3) with $\epsilon = 0.02$.
6. Using this decoding strategy, generate 200 translation hypotheses for each source sentence and determine the mean, max and min $\texttt{CometKiwi}_{22}^{\texttt{DA}}$ scores for each list. What can you conclude from this ?