1.1 Mandatory Question：

How do the authors solve machine translation using language models?

Language Model, which is common to factorize the joint probabilities over symbols as the product of conditional probabilities, operates on a byte level, hence invertible detokenizers are used to standardized text and remove these tokenization / preprocessing artifacts.
In machine translation the authors sample from the model with greedy decoding and use the first generated sentence as the translation.

1.2 Mandatory Tutorial

What is “teacher forcing” and what is its benefit?

“Teacher forcing” is a training process in decoder, which use given previous characters of the target sequence (the internal state of encoder), instead of direct using in decoder previous predicted character.
If model doesn’t use Teacher forcing, a predicted error could lead to a series error in next prediction step, which means the weight and gradient at later steps are “useless”, consequently not good to convergence of loss function.

1.3 BiLSTM Baseline

Word accuracy: 0.837421
Word accuracy of bilstm: 0.8140964995269631

1.4 Data Analysis

EMPTY and *_MYJOIN_* are surrogate symbol of mapping.
a character maps to a character → relevant character in lemma
a character maps to null character → EMPTY in the lemma
a character maps to multi-characters → *_MYJOIN_* in the lemma

1.5 Data Preprocessing (Or Not)

Like sequence-to-sequence Tutorial (Populate the first character of target sequence with the start character, and use stop character)

1.6 Implementing the Sequence-to-Sequence Approach

run prediction
Word accuracy of seq2seq: 0.22442221921881336

1.7 Discussion

The performance of BiLSTM is better than sequence-to-sequence approach. Both of them have get into trouble in ireregular verbs (like stechen), but sequence-to-sequence approach has more problem on long verbs. (such as industrialisieren)