

Report for Homework 2: Translation

1 Weakness of IBM alignment model

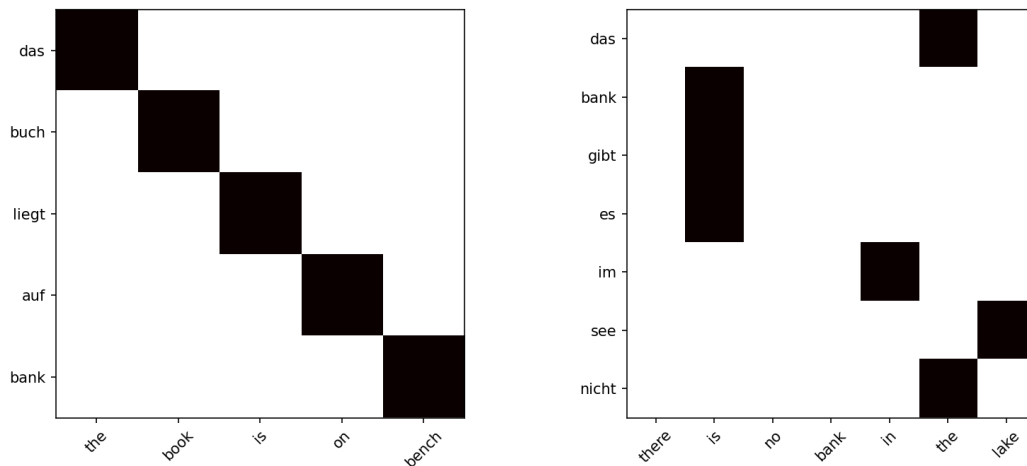


Figure 1: Alignment examples of IBM model.

1.1 Well Performed Example

In the first example, the foreign sentence is *Das buch liegt auf bank*, which corresponds to the English sentence *The book is on bench*. The model performs really well in this example, in which all common words exist with their common meanings.

1.2 Failed Example

In the second example, the foreign sentence is *Das bank gibt es im see nicht*, which corresponds to the English sentence *There is no bank in the lake*. The model fails catastrophically in this example, in which the meaning of bank should be inferred through contextual information.

1.3 Weakness of IBM Model

- Ignore contextual information.** IBM Model is a word-bag-based statistical translation model. It has an assumption that all words can be translated independently without considering contextual information. In examples, we can observe that when the word has multiple meanings and needs to be inferred with contextual information, it fails. For example, bank in German has a more common meaning bench in the training set and the model fails to align bank to bank in the other sentence. In addition, the word-bag-based translation model tends to fail to generate other language sentences with the right grammar.
- Limited Vocabulary.** IBM Model modeled the sentence pairs with a limited vocabulary set, which means this model is unable to process rare words that do not exist in the training set.

2 Findings of Attention Mechanism in Translation Model

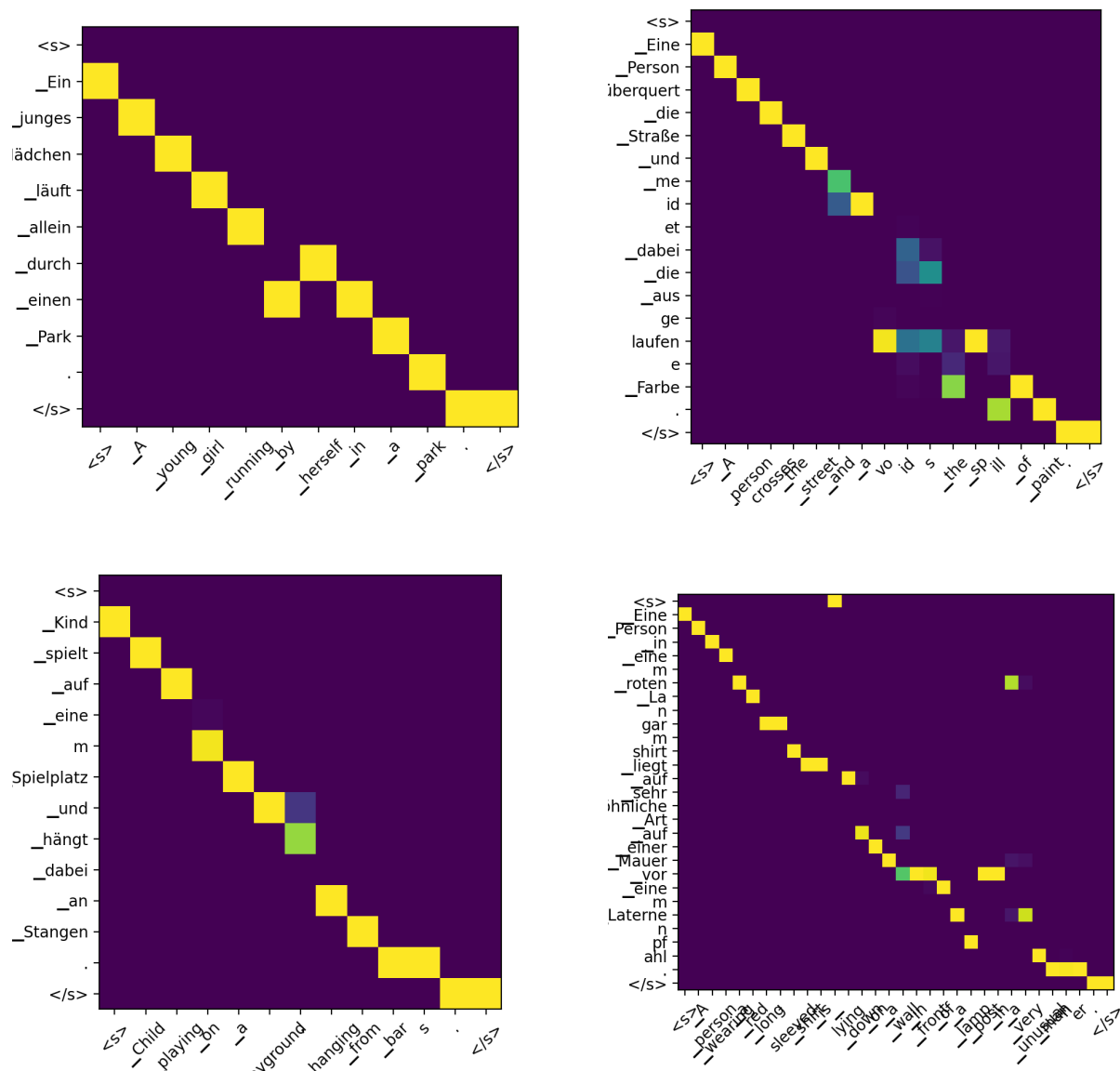


Figure 2: Examples of attention weights distribution in translation task.

Most of the time, the attention is really concentrated, in this case, is similar to the IBM alignment examples.

There are differences, which enable the attention-enhanced rnn translation model to perform well. Figures show that the attention is always one token behind the corresponding token, considering the decoder received the former information, this mechanism helps the decoder unit to output contextual information. Also, we can observe that when contextual-information-needed tokens exist, the attention will be dispersed. These tokens are always related to number, time, gender, etc. To correctly predict these tokens, it really makes sense to integrate more contextual information.