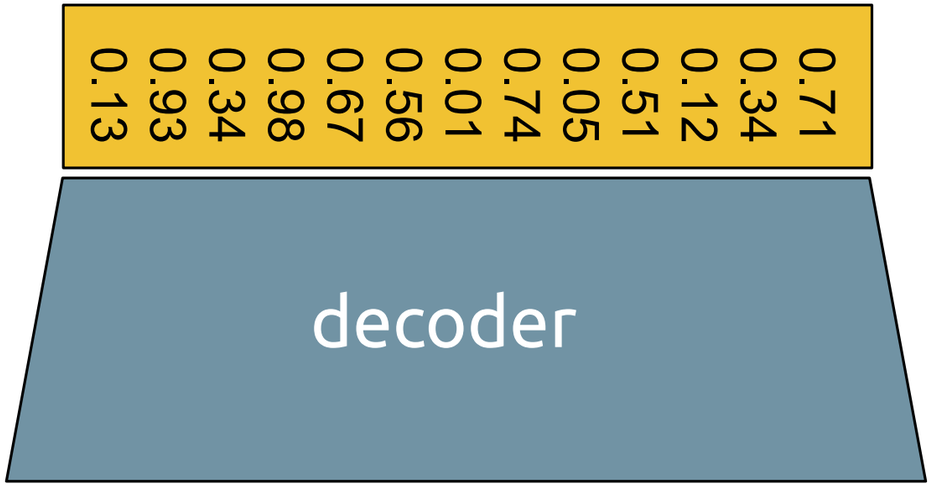
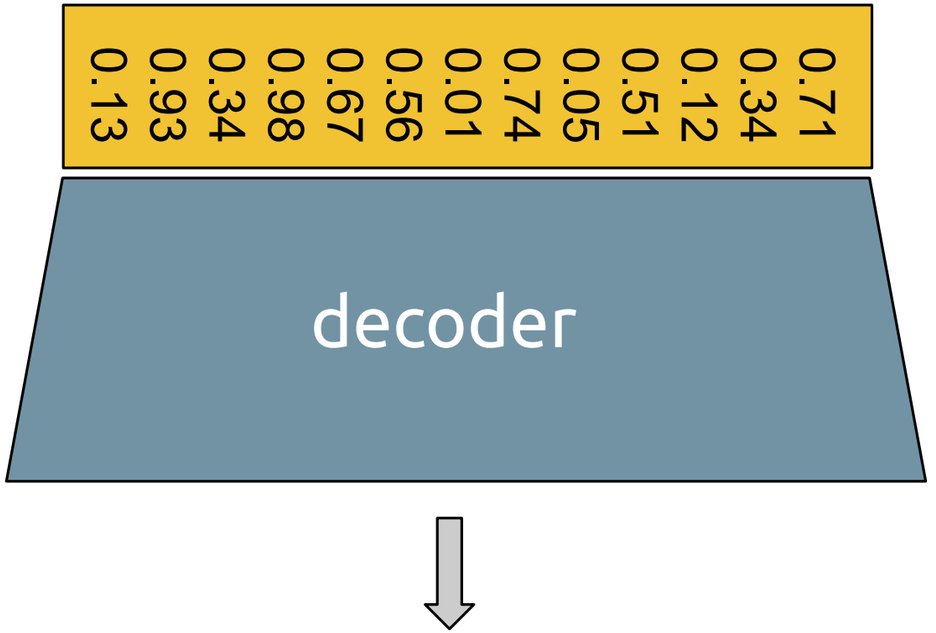
Sec 2.4

Output representations

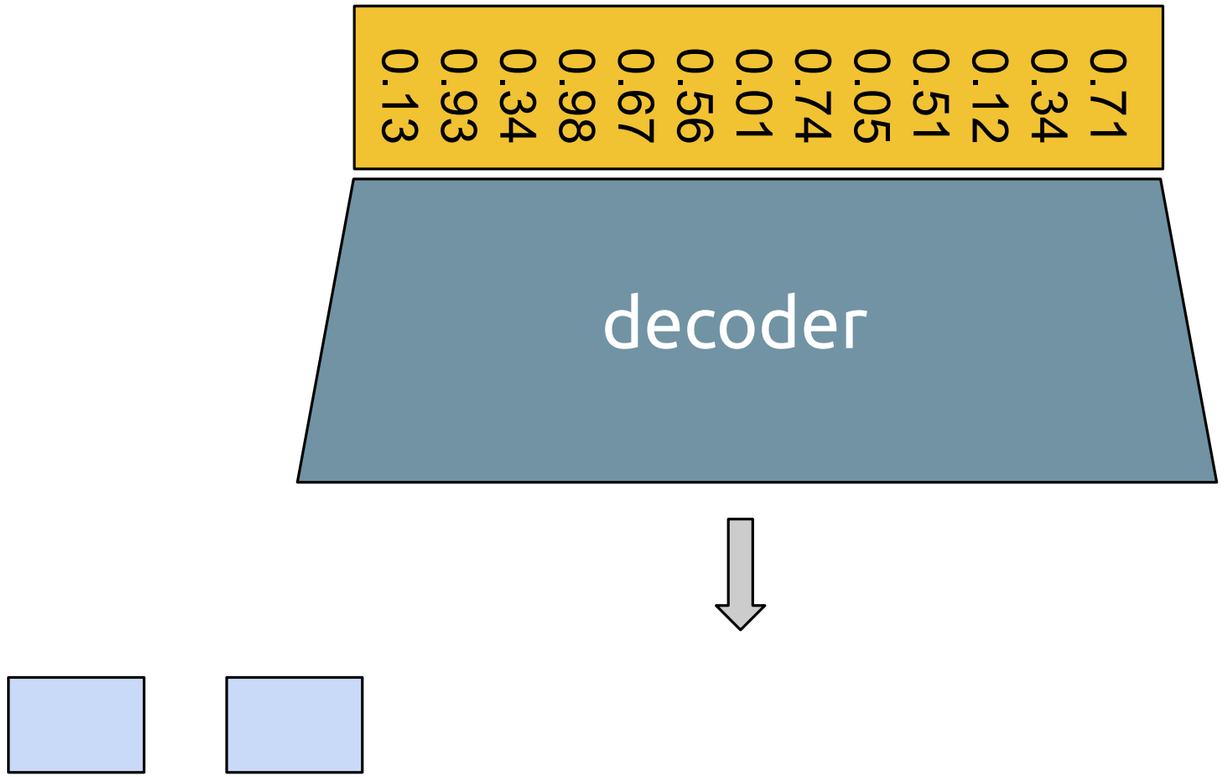
Output representation



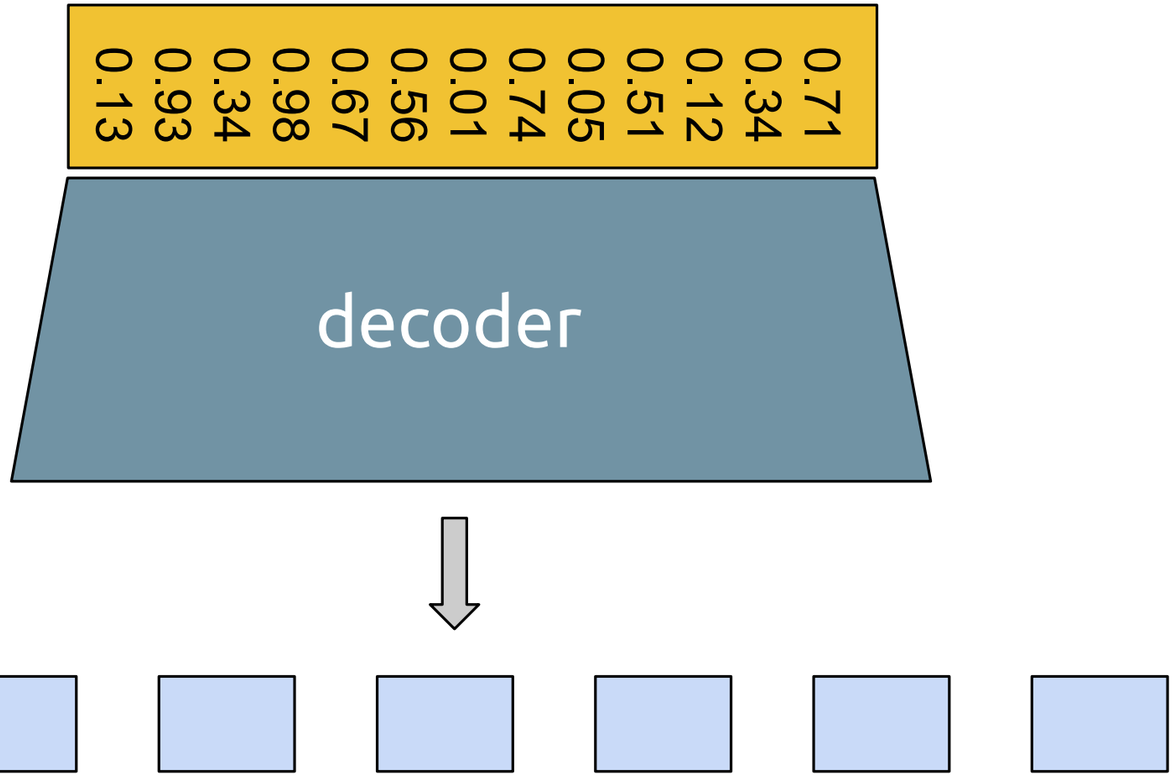
Output representation



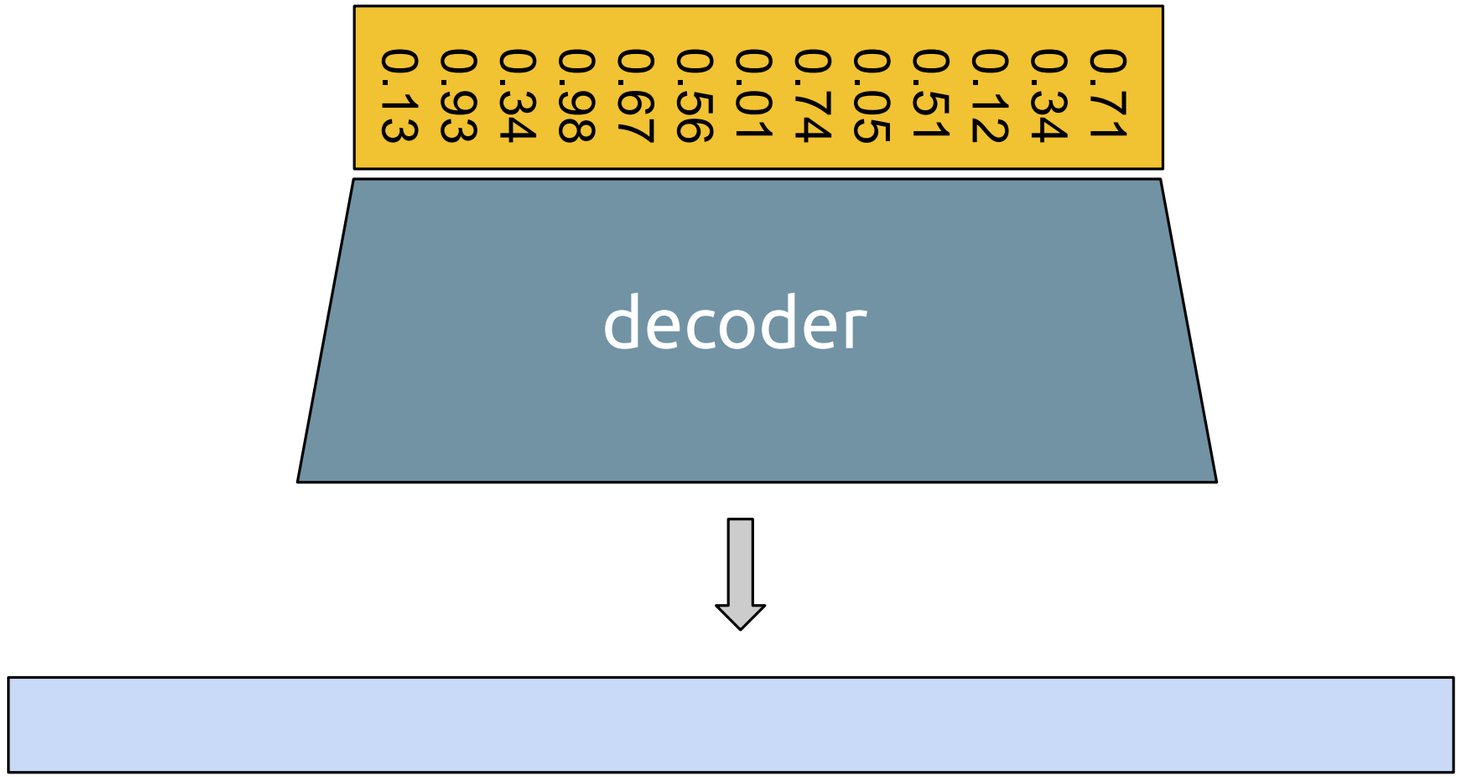
Output representation



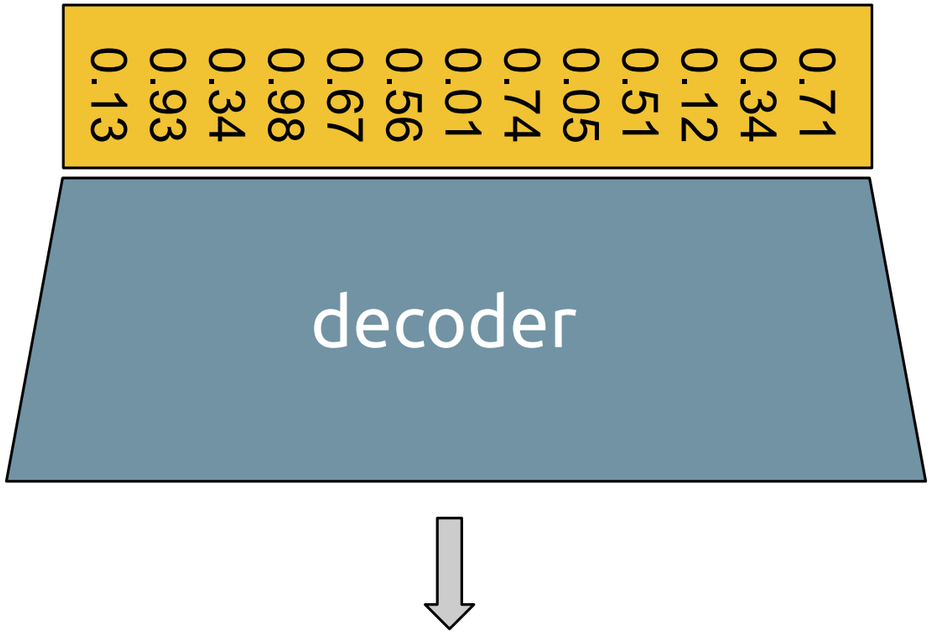
Output representation



Output representation



Output representation



What a wonderful tutorial!

Output representation

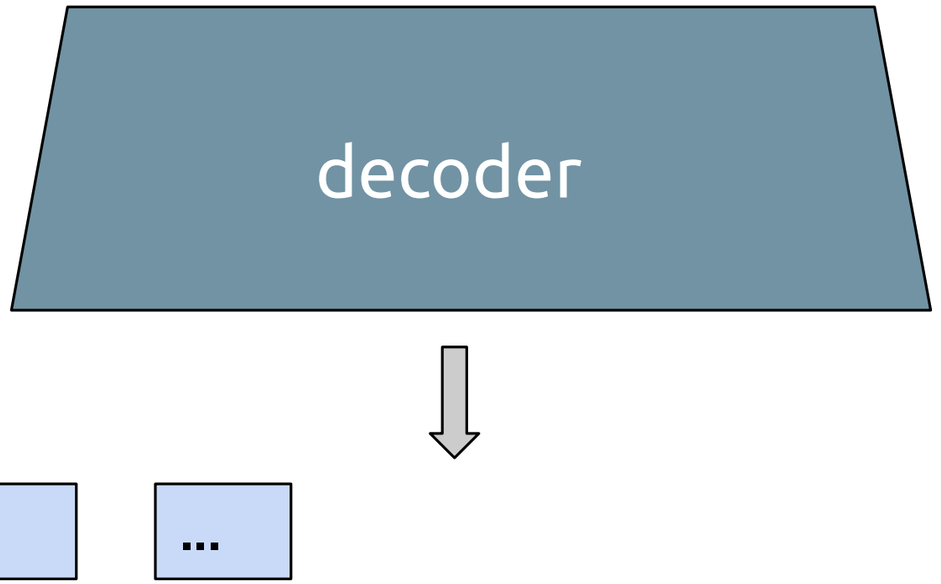
- Word (Bansal et al., 2018)
- Byte Pair Encoding (BPE) (Sperber et al., 2018)
- Character (Bérard et al., 2016; Weiss et al., 2017)

Output representation: Word

- Words as atomic unit
- Applicable only for small and high-repetitive datasets
- Tested in low-resource speech-to-text translation

Output representation: Word

- Words as atomic unit
- Applicable only for small and high-repetitive datasets
- Tested in low-resource speech-to-text translation



Output representation: BPE

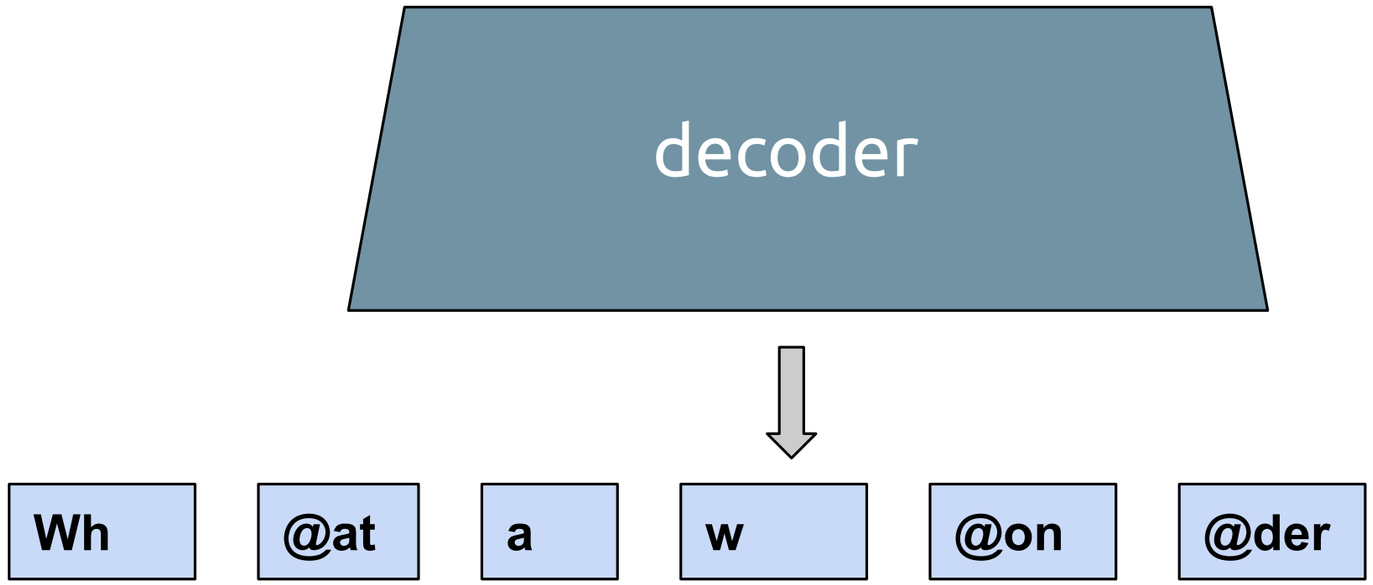
- Introduced in Neural Machine Translation to fit a large vocabulary in memory
- Each target sentence splits in sub-word units
- Iterative approach merging the most frequently co-occurring characters or character sequences
- Widely used in several NLP tasks

Output representation: BPE

- Training and test data are split based on a learned vocabulary
- After translation, BPEs converted into words

Output representation: BPE

- Training and test data are split based on a learned vocabulary
- After translation, BPEs converted into words

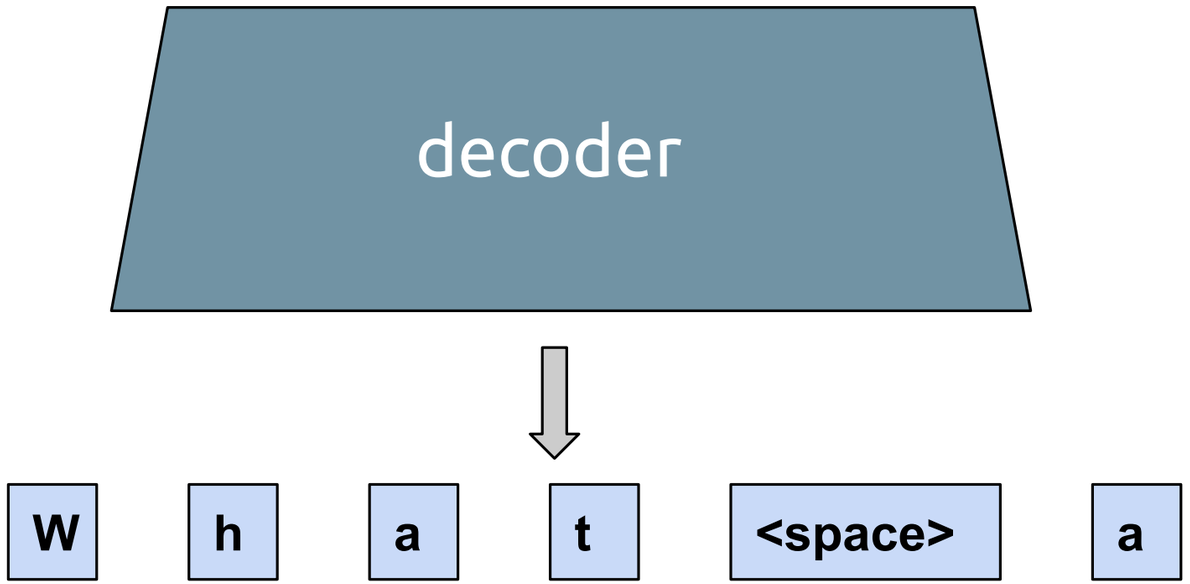


Output representation: Characters

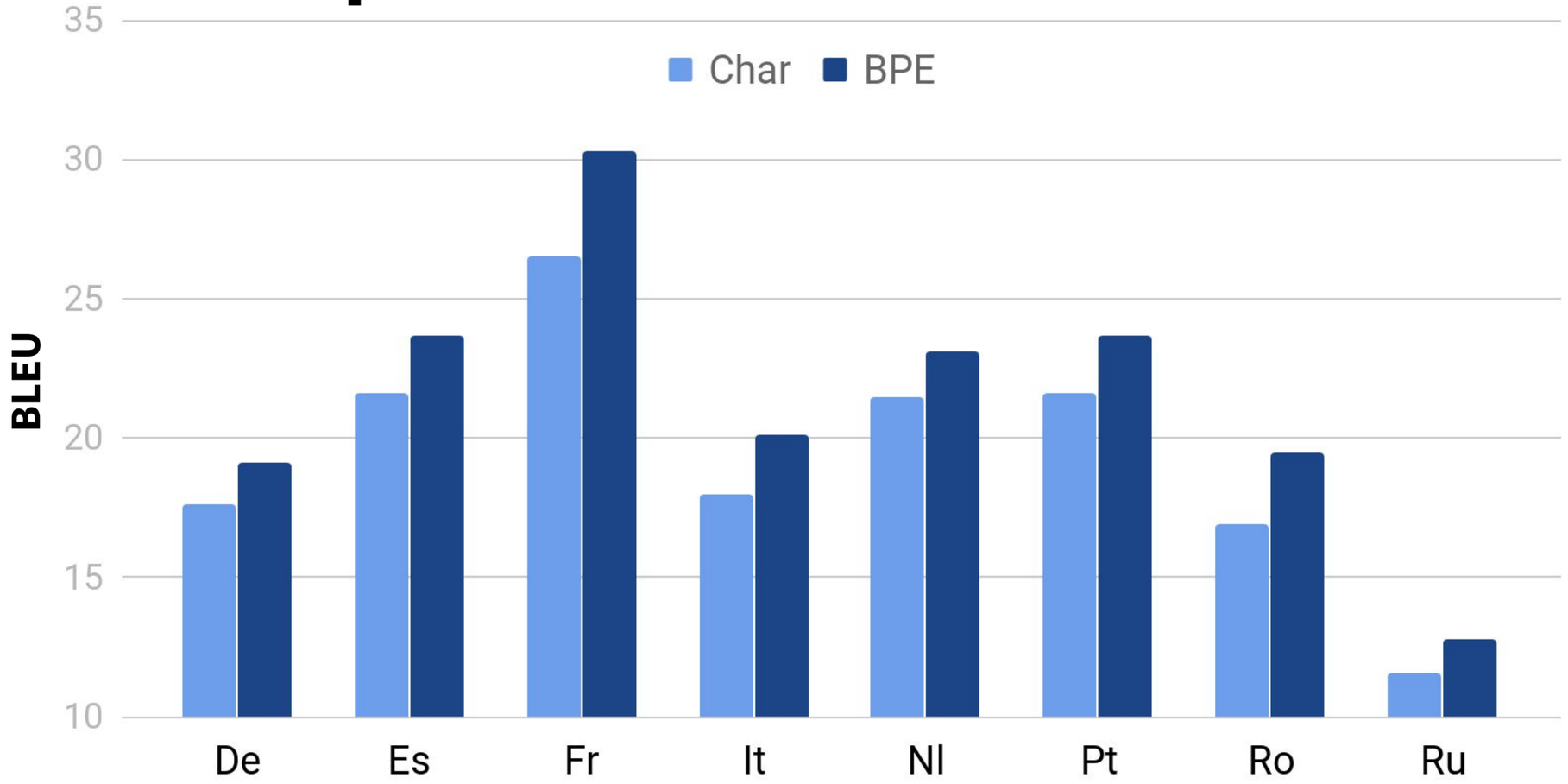
- Each sentence splits in characters with a special symbol for the empty space
- Training and test data are split
- After translation, characters converted into words

Output representation: Characters

- Each sentence splits in characters with a special symbol for the empty space
- Training and test data are split
- After translation, characters converted into words

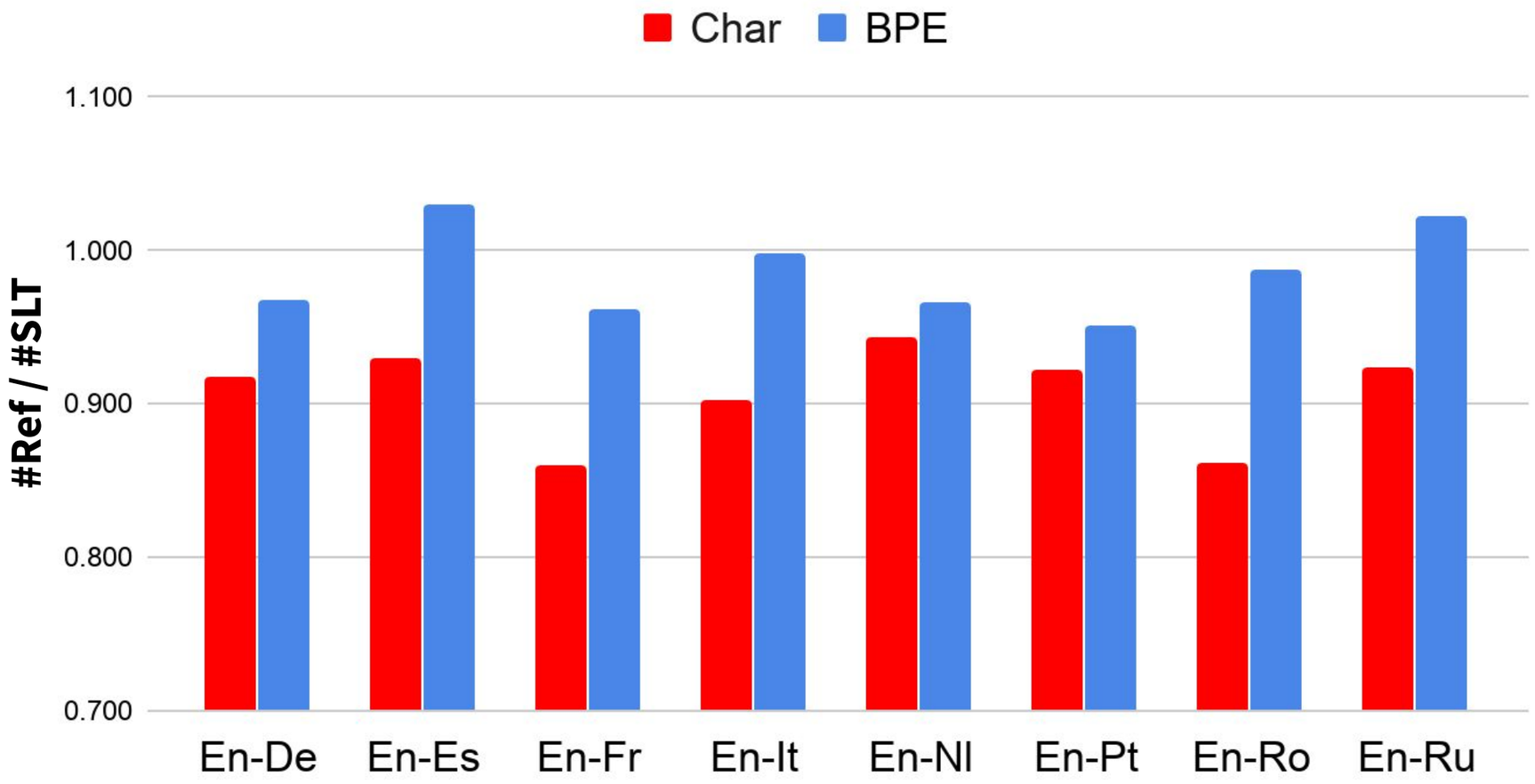


Translation performance (Di Gangi et al., 2020)



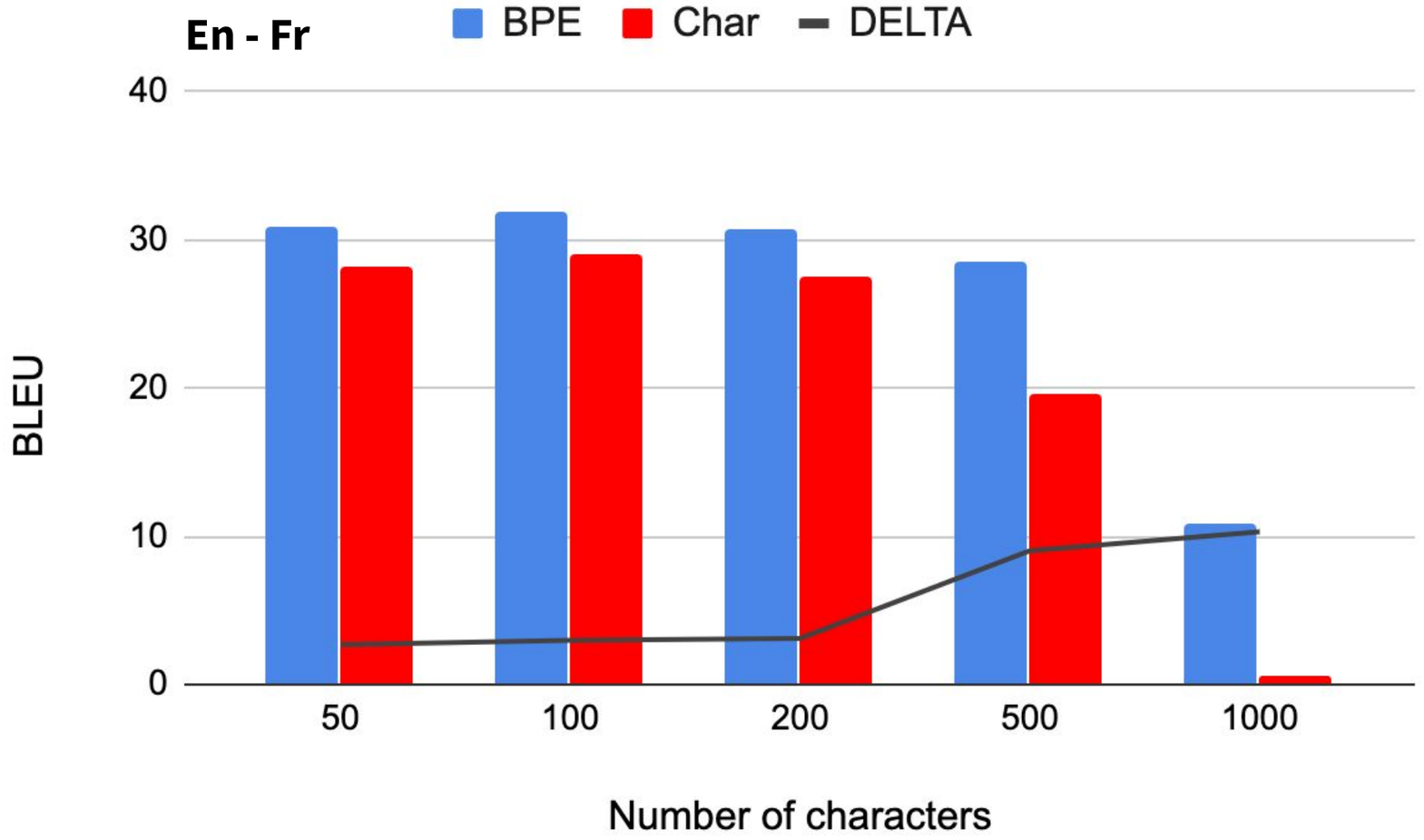
BPE outperforms Characters in all languages

Length comparison



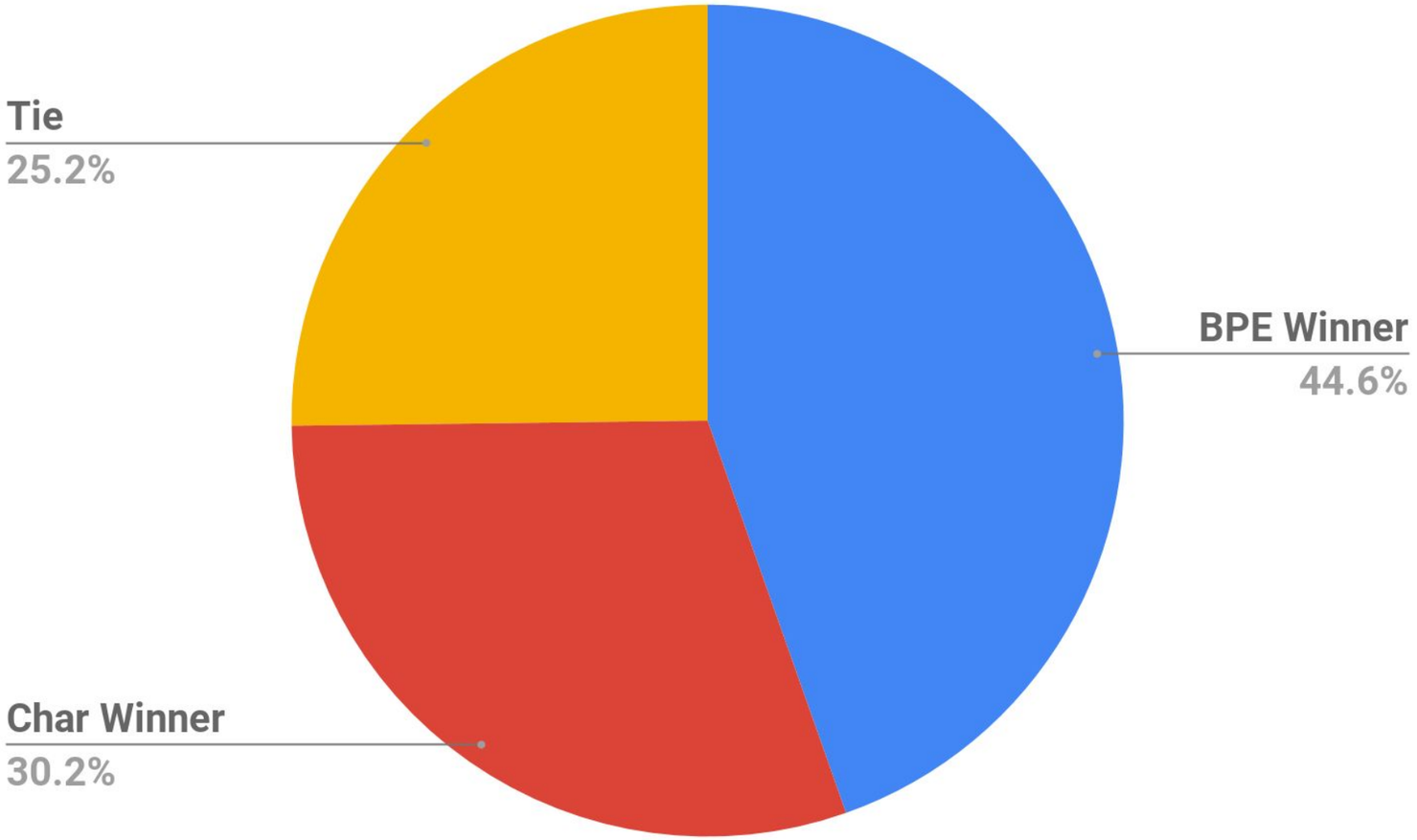
BPE produces longer sentences

Translation quality by sent. length



BPE better on longer sentences

Sentence Level Comparison



Chars better on lower quality translations