

CA446: Statistical Machine Translation

Exercises

Thursday, 20th March 2014

1. Given a sentence aligned corpus {das Haus - the house, das Buch - the book, ein Buch - a book},

calculate the lexical probabilities of

- $t(\textit{the}|\textit{das})$
- $t(\textit{house}|\textit{das})$
- $t(\textit{book}|\textit{das})$
- $t(\textit{a}|\textit{das})$

after the first iteration of EM.

2. Given the sentence pair *the house* - *la maison*, calculate $p(a, f|e)$ for all alignments and then perform the factoring-out trick.
3. Provide examples where unigram, bigram, trigram and 4-gram language models would fail to capture a grammatical constraint of the English language.