MT, SS 21: Problem Set 1

Sebastian Padó

April 20, 2021

Please submit this problem set as a PDF file until Tuesday April 27 morning (10am).

As instructed in the lecture, please form groups of (up to) three students and submit one solution per group. I strongly recommend to work on the problems individually first and then work together to create a joint solution.

This problem set will be discussed in the tutorial session on Thu **Apr 29**. Feel also free to bring any other questions that you might have. Ideally, submit them to the Wiki beforehand.

Problem 1) Parallel and Comparable Corpora

a) The first lecture has explained that the translation model is learned from parallel data. Why are there relatively few parallel corpora?

b) For which domains do you think large parallel corpora exist?

c) What problem arises thus for the translation of "typical" texts (e.g. on the web)?

d) Most multilingual newswire corpora are *comparable*, not *parallel*: Matching documents cover the same events, but use essentially independent sentences and discourse structure. What are the problems for applying the alignment methods sketched in the first lecture?

e) Consider the following corpus. Does this look more like a parallel or a comparable corpus? Identify evidence for either position.

Wie eine bizarre Dünenlandschaft wirkt der Tagebau Welzow-Süd in Brandenburg – ganz weit und leer. Anders sieht es im Freizeitpark Tropical Island aus, ebenfalls in Brandenburg, rund eine Million Besucher im Jahr: Bei Wasserrauschen, Vogelgezwitscher und lautem Planschen vor Palmenkulisse soll sich Südsee-Feeling einstellen.

What looks like a sandy desert is the Welzow-Süd opencast mine in Brandenburg. Tropical Island, also located in Brandenburg, is a water park that attracts roughly one million visitors a year. Rushing water, twittering birds and children splashing happily beneath palm trees evokes a South Sea atmosphere.

Problem 2) Word Translation Probability

You have seen a definition of $t(y|x)$ as the lexical translation probability of a source language word x being translated into a target language word y. Are $t(x|y)$ and $t(y|x)$ (i.e., keeping the words identical, and switching the direction of translation) generally the same, or not? If yes, please argue why. If not, please give two examples where this is not the case (choose the language pair yourself).

Problem 3) Google Translate

One of the best-known publicly available translation services is Google Translate (`translate.google.com`). The first lecture should already have given you some background to better understand Google Translate's behavior.

a) Experiment with Google Translate German → English translation. Construct a couple of sentences (or paste some from German websites) that are translated badly. Try to find sentences where the problem is arguably not because you chose an arcane domain or genre, but to specific linguistic phenomena. Submit at least three sentences and point out the problems.

b) Now experiment with English → German translation. It is presumably worse then German → English. Why is translating into German more difficult?

(If you don't speak German, take another target language and ask the more general question of whether translating into English or out of English is more difficult – and why.)
