Statistical Machine Translation

Homework 4

To be sent (pdf, Zip) to `waszczuk@phil.hhu.de` by 18.12.2018.

Exercise 1 - Practice

This week we implement the training for word alignment. Download the code (Eclipse project as a zip file) from the website. Start Eclipse and import the zip file as existing project. At the end of the session, you can export your project as zip archive and keep a copy of it (email, USB).

Make yourself familiar with the code. The main class is `de.hhu.phil.smt.ibm.Uebung4`. Your work is located in the `IMBModel1` class.

The goal is to implement the EM algorithm for IBM Model 1, see method `run()`. To do this, first think about which data structures to use for lexical translation probabilities $P(e|d)$ and for the frequencies $C(e|d)$ and $C(d)$ (in the code marked with `TODO (1)`, `TODO (2)` and `TODO (3)`). Note that the respective „tables “will be sparsely populated.

Implement the methods `findMostProbableAlignments`, `writeTransProbTable2File`, `writeMostProbableAlignments2File`. Further specifications can be found in the comments.

First use the example data in `toy.de/toy.en` to program and test. If your program calculates the same probability as found on paper (see exercise 2.1 below), use the data in `europarl-v7.de-en.*`.

`http://nlg.isi.edu/demos/picaro/` is a web tool for visualizing alignments. Enter the 3rd pair of the Europarl data (*ich bitte Sie ...*) with your best alignments (after 15 iterations). Make a screenshot of the resulting graph and add it to your solution.

General note: You do not have to stick exactly to the proposed structure of the code. If you are not sure, feel free to ask, but in any case, make sure that anyone can understand your code, for example by using comments. If the original entry point of the program no

longer works, please add a readme file that specifies how the code should be compiled and executed.

Exercise 2 - Theory

1

We have a corpus with two pairs of sentences:

1. (Hund bellte, dog barked)
2. (Hund, dog)

Calculate the lexical translation probabilities $P(e|d)$ for IBM Model 1 after two rounds of EM algorithm. Do not forget the zero alignments!

2

Given the tables with lexical translation probabilities below, calculate the translation probability of the following translations of the German sentence *das Haus ist klein*.

1. the house is small
2. the house is little
3. small house the is
4. the

Take in each case the most probable alignment (which you have to determine yourself).

(a) Is the IBM Model 1 as a translation model by itself a good model to find the best translation?
(b) Explain how the Noisy Channel Model, i.e., combining a translation model with a language model based on the Bayes' rule:

$$\arg\max_{\boldsymbol{e}} P(\boldsymbol{e} \mid \boldsymbol{f}) = \arg\max_{\boldsymbol{e}} \frac{P(\boldsymbol{f} \mid \boldsymbol{e}) \cdot P(\boldsymbol{e})}{P(\boldsymbol{e})} = \arg\max_{\boldsymbol{e}} P(\boldsymbol{f} \mid \boldsymbol{e}) \cdot P(\boldsymbol{e}) \tag{1}$$

fixes some problems of the IBM 1 translation model.

d =das		d =Haus		d =ist		d =klein	
e	$P(e\|d)$	e	$P(e\|d)$	e	$P(e\|d)$	e	$P(e\|d)$
the	0.7	house	0.8	is	0.8	small	0.4
that	0.15	building	0.16	's	0.16	little	0.4
which	0.075	home	0.02	exists	0.02	short	0.1
who	0.05	household	0.015	has	0.015	minor	0.06
this	0.025	shell	0.005	are	0.005	petty	0.04