Karlsruhe Institute of Technology
Elizabeth Salesky
Institute for Anthropomatics and Robotics, Interactive Systems Labs
elizabeth.salesky@kit.edu

MT Praktikum - Alignment

11. Juli 2018

Environment Setup

First we need to have access to our cluster enviroment.
If you can use the available machines in the pool room, please log in with username: smt[30-45] (any number between 30 to 45, for example smt35), password=123456.

If you use your own laptop, you can directly connect to the cluster using ssh, using this command:

```
ssh smt[30-45]@i13hpc1.ira.uka.de
```

Next, please log into i13hpc28 or i13hpc29, using the following commands:

`ssh i13hpc1` *(this is our login server; if you use your laptop, you are already here)*
`ssh i13hpc28` **or** `ssh i13hpc29`

From there, go to your working directory:

```
cd /project/smtstud/ss18/systems/{username}/
```

Now enter the pre-installed virtual environment using these 2 commands:

```
bash
. /project/smtstud/ss18/commands/setup.sh
```

If you see the (praktikum) at the beginning of your terminal line, the setup was successful.

This directory contains today's data and scripts. Copy this directory into your working directory:

```
/project/smtstud/ss18/data/alignments
```

NMT: Alignments

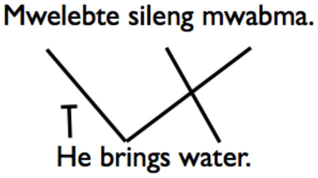
Log into rg3hpc1, and then to i13hpc28 or i13hpc29. Once you are in either i13hpc28 or i13hpc29, move into your working directory.

First we are going to look at alignments. You have been given two data files and also a file with word alignments (en.mi.udhr.aligned). The data is the Universal Declaration of Human Rights in both English and Maori, aligned by line. In this parallel data, we assume that each line is a sentence.

The alignment file has the same number of lines as your two UDHR file, and each line gives a list of word alignments. E.g. a line in the '.aligned' file that says '0-0 0-2 1-3' means that the 0th word in the Maori line is aligned to the 0th word in the English line, the 0th word in the Maori UDHR is also aligned to the 2nd word in the English UDHR, and the 1st word in Maori UDHR is aligned to the 3rd word in the English UDHR. The format is maoriIndex-englishIndex.

`alignments.py` reads in the text files and alignments and computes all values you need to answer these questions; you just need to read and understand this script to figure out which values to print to the answers to the questions below.

Reminder, fertility is the number of tokens a word is aligned to. E.g. 'brings' below has a fertility of 2: it is aligned to two words.
The alignments shown for these example sentences would be 0-1 1-2 2-1.



1. What is the average fertility of word tokens in Maori and in English?

2. What is the average fertility of word types in Maori and in English?

3. What's the difference between word type and word token average fertility in each language? What does that mean?

4. Does Maori or English have a higher ratio of word type/token average fertility? What does that mean?

5. How many tokens are null-aligned in Maori? In English?

6. How many types are null-aligned in Maori? In English?

7. What is the highest fertility for a type in Maori? In English? What does this mean?

8. How many types have fertilities greater than 1 in Maori? In English? Why might this be?

Expectation Maximization (EM)

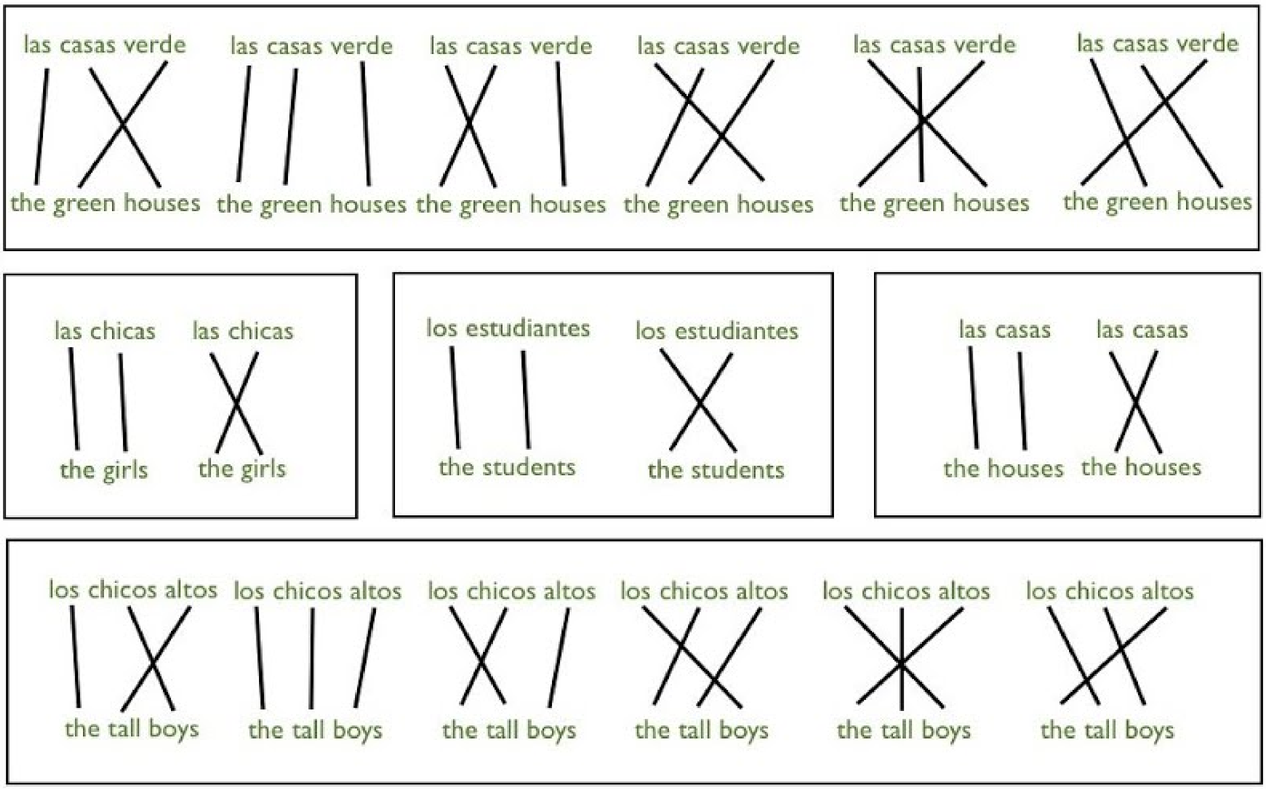
Expectation Maximization: recall this is how IBM Model 1 is trained.

For word alignments, we want to probabilities of alignments (a) given translation probabilities between words (e, f). With this algorithm, we first pretend we have the translation probabilities. Then, we use these probabilities to estimate the probability of all alignments. Specifically, compute $p(a|e, f)$ for all possible alignments. Then, we compute $p(a, f|e)$ for all alignments. Then, we use these probabilistic alignments to re-estimate the translation probabilities.We iterate this process until our probabilities converge.

Here, we will illustrate EM with a simple example. You do not need to code here: the equations are already in Excel, you just need to modify the initial probabilities, and think about the effects.

To look at this, first `scp` EMAlignments.xlsx to your local machine or laptop.

Below are the five sentences to be aligned, along with all of the possible alignments for each sentence pair. There are 8 Spanish word types and 7 English word types. We estimate p(Spanish Word — English Word) for each English word type, and use these probabilities only (not p(E—S)) to estimate alignment probabilities.



The first EM run is initialized to uniform translation probabilities (p(Spanish Word — English Word) =1/8 for all Spanish words and all English words).

- How many iterations does EM run before the translation probabilities are no longer changing with additional iterations?
- What translation probabilities has it learned? What has it learned correctly and what hasn't it been able to learn correctly?
- Why couldn't it learn everything that we may have wanted it to learn?
- Play with the initial translation probabilities to see what happens in terms of number of iterations and the final learned alignments.