

11-731: Machine Translation

Homework Assignment #3:

Out: Wednesday, February 4, 2009

Due: Wednesday, February 11, 2009

In this homework you will perform the second step of building a Statistical Machine Translation system: training word alignment models. You will use GIZA++¹ tool to generate different alignment models. Use the setup provided at [/afs/cs.cmu.edu/project/cmt-55/lti/Courses/731/homework/HW3](http://afs/cs.cmu.edu/project/cmt-55/lti/Courses/731/homework/HW3). This directory contains all the tools and data necessary for the homework. You don't have to copy the setup. Simply link to the data from your workspace.

The steps involved are:

i) Run GIZA++ to train models

We will use a subset of the training data used in the previous homework (training.en and training.es). The data is already preprocessed.

To train the models in Spanish->English direction:

```
trainGIZA++.sh training.en training.es GIZA.S2E
```

Go through each step in the script and understand what it is doing. In this run, we use the default parameters in GIZA++.

Look into the GIZA.S2E.log and see different training steps. At the end of the training you will see a set of output files. Alignment models are named as GIZA.S2E.A*.*.

GIZA.S2E.A1.5: IBM Model 1 alignment after 5 iterations;

GIZA.S2E.Ahmm.5: HMM alignment after 5 iterations;

GIZA.S2E.A3.5: IBM Model 3 alignment after 5 iterations;

GIZA.S2E.A3.final: IBM Model 4 alignment

Refer to GIZA++-v2/README for more details on each of the files GIZA++ produces.

ii) Evaluate the alignment accuracy for different models

We select a sample² from the alignment file and evaluate it with a manually generated alignment.

¹ <http://www.fjoch.com/GIZA++.html>

² Dev set from http://gps-tsc.upc.es/veu/LR/epps_ensp_alignref.php3

Select the last 100 sentences from the each of the 4 alignment files. (In the alignment files, this would correspond to last 300 lines).

You can use: `tail -n300 AlignmentFile > AlignmentFile.dev`

We will use a script from the `Lingua::AlignmentSet` toolkit³ to calculate the alignment error rate.

```
perl ./Lingua-AlignmentSet-1.1/bin/evaluate_alSet-1.1.pl -sub
AlignmentFile.dev -subf giza -ans hand-aligned/dev.engspa.naacl -ansf
naacl
```

(Make sure you set `PERL5LIB` path to `/afs/cs.cmu.edu/project/cmt-55/lti/Courses/731/homework/HW3/Lingua-AlignmentSet-1.1/blib/lib/`)

List the alignment error rates (AER) for each of the four models.

- iii) Repeat the training for the reverse direction (English->Spanish) by switching the source and target input files to GIZA++.

(i.e. `trainGIZA++.sh training.es training.en GIZA.E2S`).

Compute and report AER for each model similar to (ii). Use:

```
perl ./Lingua-AlignmentSet-1.1/bin/evaluate_alSet-1.1.pl -sub
AlignmentFile.dev -subf giza -ans hand-aligned/dev.spaeng.naacl -ansf
naacl
```

- iv) Write a simple script to compute the percentage of NULL alignments, and the percentage of words not-aligned. For both directions of training, report the two values for each alignment model you generated. State your observations. You don't need to submit the script.
- v) Write a simple script to compute the word fertility frequencies from the alignment model files. Plot the fertility value against the frequency. Do this for both directions of training. You don't need to submit the scripts.
- vi) Change alignment parameters (`p0` and `maxfertility`) and see how it affects the alignment accuracy. Give the best alignment accuracy you could achieve with the parameter settings you used.

³ <http://gps-tsc.upc.es/veu/personal/lambert/software/AlignmentSet.html>