BM1 Advanced Natural Language Processing
University of Potsdam, Winter 2016/17

Assignment 5: Word Alignments

Tatjana Scheffler, Ph.D.

Due: Friday, January 20, 10:00 a.m.

Aligning sentences and words are central tasks in statistical machine translation. In this assignment, you get to implement a word aligner. Given pairs of aligned sentences in two languages, source and target, the goal is to align source words to their target translations. The resulting alignments might contain unaligned or multiply aligned words, i.e., the word alignments are generally m:n, which makes the task challenging.

This assignment is almost identical to the first assignment of a well-known online course on SMT, which is available here: `http://mt-class.org/jhu/hw1.html`. Read this website, and note in particular the link to a tutorial by Adam Lopez on how to implement IBM Model 1 (this contains helpful pseudocode). Your tasks are as follows:

1. Clone the repository from `https://github.com/alopez/en600.468.git`, using Git. Observe that the repository contains some code and a dataset of 100,000 English-French sentence pairs (`hansards.e` and `hansards.f`). The first 37 sentence pairs are manually aligned, and these manual alignments are encoded in file `hansards.a`.

2. Get to know the code of the aligner (folder `aligner`). It provides a very simple baseline system; test it through the provided command-line interface. Submit the Alignment Error Rate (AER) of the baseline system. Observe that it is terrible.

3. Improve over the baseline by implementing an aligner based on IBM Model 1 (see "The Challenge" section of the JHU course for a more detailed description). IBM Model 1 is a probabilistic model that generates each word of the English sentence independently, conditioned on some word in the Foreign sentence. Thus, its key features are the word translation parameters $P(e|f)$, which are in turn learned from the data using expectation maximization (EM). These are the two focus points for your implementation. Feel free to use NLTK if you find it helpful, but note that anything in the `nltk.align` package is disallowed in this assignment.

4. In addition to what is required in the JHU assignment, experiment also with GIZA++ (`https://github.com/moses-smt/giza-pp`), a de facto

standard state-of-the-art word alignment toolkit for NLP. Namely, compare your IBM Model 1 to GIZA++ Model 1 and another Model i of your choice.

Submission Submit your code and document all your evaluation results. Choose at least one alignment visualization from your system in comparison to the baseline system and GIZA++ for discussion in your write-up.

Extra credit. Implement any of the suggestions for improving over your IBM Model 1 as suggested in the task description from the JHU course, i.e., implement an aligner that improves over your IBM Model 1 AER.

Submit your solutions and code via email to `johannsmeier@uni-potsdam.de`