

11-731: Machine Translation

Spring 2007

Homework Assignment #2

Out: Wednesday, February 21, 2007

Due: Monday, March 19, 2007

Word Alignment and Bilingual Translation Lexicon Construction

Your task in this assignment is to develop and implement basic algorithms for word-aligning a given sentence-aligned parallel corpus between French and English, and for extracting a bilingual word translation lexicon from the word-aligned corpus. The quality of the bilingual lexicon that you develop will be evaluated by automatically comparing it to a given “gold-standard” high-quality lexicon.

Detailed Instructions:

1. Implement a basic word-alignment algorithm and an algorithm for extracting a word translation lexicon from the word-aligned corpus.

Training: Your developed module will be trained on a corpus of sentence-aligned parallel text. The corpus is a collection of pairs of files, where in each pair, one file contains sentences of language L1 and the other file contains the corresponding sentences for language L2. Each sentence starts on a new line. For each pair of corresponding files, the number of lines is the same.

Training Output: *A bilingual word translation lexicon.* The format of the lexicon file should be one “entry” per line, where each entry consists of a source-language word, a “tab”, followed by a target language word or multi-word translation. If a word has multiple possible translations, each should appear on a separate line.

2. **Retrieve your training data from the following URL:**

<http://www.isi.edu/natural-language/download/hansard/index.html>

3. **Download the following training and testing corpora:**

Training: The Senate Debates Training Set (182K sentence pairs)

Testing: The Senate Debates Testing Set (25K sentence pairs)

4. Train your algorithms on *just the training corpus*, and create the resulting French-to-English bilingual word translation lexicon. The testing corpus may be used during development to test how well your word alignment algorithm is working, but you *should not* use it as training data.
5. Write a program that can extract from your bilingual lexicon the subset of entries consisting of *only* the French words that appear in a given list of words. The words will be provided in a text file, one word per line.

6. In order to compute a quality score for your lexicon, write a program that given a “gold-standard” lexicon, calculates aggregate Precision, Recall and F1 measures for your lexicon. The measures are to be calculated as follows:
 - Calculate the total number of entries in your lexicon - CL
 - Calculate the total number of entries in the “gold-standard” lexicon – CG
 - Calculate the total number of entries that are identical in both lexicons – C
 - Precision = C/CL ; Recall = C/CG ; F1 = $2 * P * R / (P + R)$
7. The “gold-standard” lexicon will be provided to you in the last week before the homework is due. Extract a list of all French words that appear in the “gold-standard” lexicon. Using this list, extract the subset of entries from your bilingual translation lexicon that are needed for calculating the above defined scores, and evaluate your lexicon.

General Instructions and Comments:

- This assignment is to be performed individually.
- You may implement your algorithms in any programming language/environment of your choosing.
- Do not share your code with others or use code developed by others.
- Do not use any external *bilingual* resources other than the training corpus provided
- Consult with the instructor in case of any doubts about these instructions.
- **Note:** You are not expected to develop a very complex word-alignment algorithm for this assignment. You may wish to start with a simple approach based on counts (or relative counts) of French/English word pairs that co-occur in the parallel sentences. It is important, however, that you develop an implementation that can handle the large amounts of training data, and that you find effective solutions to the memory and runtime issues you encounter when handling such large amounts of data.
- Your grade on the assignment will be calculated by: 80% based on successfully completing the assignment, 20% based on the F1 score of your lexicon.

When complete, submit the following by email to: alavie@cs.cmu.edu

- Your code
- Your extracted bilingual word translation lexicons.
- Your table of Precision, Recall and F1 results.
- A brief write-up explaining how you did the work (and any intermediate results you found, e.g. on the test set), as well as design decisions you made, and how you dealt with the challenges of a large data set.