

11-731: Machine Translation

Homework Assignment #5:

Out: Wednesday, February 18, 2009

Due: 2:00 pm Wednesday, February 25, 2009

In this homework you will generate a phrase table, where the phrases are annotated with syntactic categories. Phrase table generation steps are similar to homework 4. Use the setup provided at [/afs/cs.cmu.edu/project/cmt-55/lti/Courses/731/homework/HW5](http://afs/cs.cmu.edu/project/cmt-55/lti/Courses/731/homework/HW5).

Notice that the phrase extraction is already done up to step-5 of the training script. (i.e. extracting phrases for both directions, producing the files `extract` and `extract.inv`). Also notice that the phrases in these two files are already annotated with the sentence ID in the corpus from which the phrases were extracted.

You are also given parse-trees (in `/parse-trees/europarl-v3b.es-en.top10k.en.pp.parsed`), generated by Stanford parser¹ for the English side of the training corpus.

Your task is to label the extracted phrases with syntactic categories from the English parse-trees and generate a phrase table.

The steps involved are:

- i) Extract syntactic phrases from the parse-trees

For each training sentence, extract all syntactic phrases from the parse tree of the English sentence, along with the syntactic label of the phrase.

For example, for the first sentence in the corpus:

```
(ROOT (NP (NP (NN Resumption)) (PP (IN of) (NP (DT the) (NN session))))))
```

You could extract phrases such as:

```
(NP "the session")
```

```
(PP "of the session")
```

....

Notice that the sentences in the parse tree are in mixed case, and therefore will have to be converted into lowercase when matching with the phrases extracted by the `train-phrase-model.perl` script. The parser has also converted some punctuations into special tags (e.g. right parenthesis "(" in to `-LRB-`). Those will have to be converted back in to the original form.

¹ <http://nlp.stanford.edu/software/lex-parser.shtml>

ii) Label the extracted phrases in `extract.inv` and `extract` with the syntactic categories

Match between the syntactic phrases found for each training sentence in (i) and all the extracted phrases for that sentence as they appear in `extract.inv`. If the phrase is found, the phrase pair should be labeled with the syntactic label, by adding it to the English side of the phrase. If not found, the phrase should be labeled as "NONE" (indicating that it is not a syntactic phrase). Here is an example of how you should label:

Syntactic phrase (from the parse-tree):

(PP "of the session")

Extracted phrases (in `extract.inv`):

Sentence_ID=1: of the ||| del ||| 0-0 1-0

Sentence_ID=1: of the session ||| del período de sesiones ||| 0-0 1-0 2-1 2-3

....

After labeling:

NONE of the ||| del ||| 0-0 1-0

PP of the session ||| del período de sesiones ||| 0-0 1-0 2-1 2-3

....

You should also remove the sentence ID tag from the phrase pair.

Repeat the same tagging for `extract`. Note that, in `extract`, the English phrase is in the second block. For each phrase pair in `extract` and `extract.inv`, the Spanish and English phrases should be the same, except for the alignments, which are in opposite directions.

iii) Generate the labeled phrase table

Once all the phrases in `extract.inv` and `extract` are properly labeled run step-6 in the training script to generate the aggregate phrase table and calculate feature scores.

Note that, because we introduced a new token to the English side without properly altering the alignment information, the feature scores generated by this step will no longer be valid. As we are only interested in the distribution of the phrases, for the homework it will not be an issue.

iv) Analysis

Analyze the generated phrase table and provide the following details. You can either modify the scripts (`AnalyzePhraseTable1.sh` and `AnalyzePhraseTable2.sh`) from homework 4 to support the following analysis or write your own scripts.

- a. For source phrases up to length 7, give the following statistics:
 - i. the number of different source phrases
 - ii. average number of translations
 - iii. average number of labels per source phrase

- iv. average number of labels per phrase pair
- b. What fraction of source phrases have **only** the label "NONE" (i.e. are non-syntactic)
- c. What fraction of phrase pairs have **only** the label "NONE" (i.e. are non-syntactic)
- d. Overall distribution over all the labels (i.e. fraction of total of phrase pairs that are labeled by each label)
- v) Compare the resulting phrase table with the phrase table you generated for homework 4. How do the overall sizes of the phrase tables compare?

Please submit your write-up by email (as an attachment) to the class TA by the due date. In the write-up, also provide links (don't attach) to the generated files.