Department of Computer Science
The University of Melbourne
COMP90042 WEB SEARCH AND TEXT ANALYSIS (Semester 1, 2017)

Workshop exercises: Week 12

Discussion

1. What is the difference between **Word–Based** and **Phrase–Based** Statistical Machine Translation?
   (a) What is the **decoding** problem in Machine Translation, and how might we solve it?
2. For the following "bi-text":

Language A	Language B
`green house`	`casa verde`
`the house`	`la casa`

   (a) What is the logic behind **IBM Model 1** for deriving word alignments?
   (b) Work through the first few iterations of using the **Expectation Maximisation** algorithm to build a translation table for this collection. Check your work by comparing to the `WSTA_N20_machine_translation.ipynb` output.

Programming

1. Using NLTK, find the Gale–Church sentence alignment of (the fragment of) the Europarl Corpus.
   (a) How many alignments are 1:1? 0:1? 1:2? 1:3?
   (b) What do you notice about sentences that participate in one–to-many alignments in the collection?

Catch-up

- What is **Machine Translation**?
- In a MT context, what is a **bitext**? What is the **sentence alignment** problem, and why is it important?
- What is a **word alignment** in MT? What is a **phrase table**?
- What is a **language model**? What is an n**-gram language model**?
- What is **Maximum Likelihood Estimation**?

Get ahead

- Read up on the some of the other IBM models. Explain why IBM Model 3 gives such a drastically different translation table to Model 1, on the given bi-text.