School of Computing and Information Systems
The University of Melbourne
COMP90042 WEB SEARCH AND TEXT ANALYSIS (Semester 1, 2019)

Workshop exercises: Week 12

Discussion

1. What aspects of human language make automatic translation difficult?
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 2 iterations of the **Expectation Maximisation** algorithm to build a translation table for this collection, based on IBM Model 1. Check your work by comparing to the `WSTA_N21_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 **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.
- (Harder) Consider the following aligned example, from Knight (1997)[1]:

CENTAURI	ARCTURAN
ok-voon ororok sprok.	at-voon bichat dat.
ok-drubel ok-voon anok plok sprok.	at-drubel at-voon pippat rrat dat.
erok sprok izok hihok ghirok.	totat dat arrat vat hilat.
ok-voon anok drok brok jok.	at-voon krat pippat sat lat.
wiwok farok izok stok.	totat jjat quat cat.
lalok sprok izok jok stok.	wat dat krat quat cat.
lalok farok ororok lalok sprok izok enemok.	wat jjat bichat wat dat vat eneat.
lalok brok anok plok nok.	iat lat pippat rrat nnat.
wiwok nok izok kantok ok-yurp.	totat nnat quat oloat at-yurp.
lalok mok nok yorok ghirok clok.	wat nnat gat mat bat hilat.
lalok nok crrrok hihok yorok zanzanok.	wat nnat arrat mat zanzanat.
lalok rarok nok izok hihok mok.	wat nnat forat arrat vat gat.

Translate the following Arcturan sentences into Centauri:

(a) iat lat pippat eneat hilat oloat at-yurp.

(b) totat nnat forat arrat mat bat.

(c) wat dat quat cat uskrat at-drubel.

[1] Knight, Kevin. (1997) *Automating knowledge acquisition for machine translation*. AI Magazine 18(4), *AAAI*. pp. 81–96.