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?

   - Phrase–Based Translation attempts to translate multi–word units simultaneously; it often uses the Word–Based Translation as **alignments** to seed the process of finding sequences of tokens which are good translations.

   (a) What is the **decoding** problem in Machine Translation, and how might we solve it?

   - Decoding is the process of choosing the sentence in the target language which bests translates the source language (has the greatest probability).
   - A typical approach is by using the so–called "noisy channel" model, where we construct a **language model** to estimate $P(t)$ and a **translation model** (in the wrong direction) to estimate $P(s|t)$, and then choosing t such that $P(t)P(s|t)$ is maximised.
   - It is clearly impractical to calculate the probability of every sentence in the target language; however, it is often the case that words can only be plausibly translated in a few different ways, thereby making a **beam search** a sensibly approach.

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?

   - The core idea is that we are going to have a translation table which stores the probability of translating a word from the target language into every possible word in the source language (again, this is the wrong direction due to the noisy channel model).
   - The probability of a sentence can then be treated as a **uni-gram** probability, conditioned on how the tokens in the two sentences are aligned. Or, essentially, the product of all of the corresponding probabilities from the translation table.

   (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.

   - We need to establish the direction of translation before we begin (although it isn't important in this particular example): let's say, we're trying to translate language B into language A. Consequently, we want the alignments where we're translating A into B. (Again, the opposite direction, due to the "noisy channel" model.)

- We're going to initialise our translation table T with **uniform** values: every word from A is equally likely to be translated as every word from B:

T	casa	la	verde	Total
green	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1
house	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1
the	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1

- The other thing we'll want to establish is the set of possible alignments. This can be done exhaustively by hand, because the "sentences" under consideration are so short; however, even for longer sentences, this set of alignments is not so large so as to stress our computing resources.
- I'm going to follow the notebook in ignoring the possibility of alignment words from B with the null element of A. We can deal with the converse — where words in A align with the null element of B by simply not aligning them to anything. (Proper models deal with the former, at least.)
- Consequently, each of the two sentences (called I and II below) has four (2^2) possible alignments (where every token of B is accounted for), namely:
  - Ia: `green` aligns with `casa` and `house` aligns with `verde`
  - Ib: `green` aligns with `verde` and `house` aligns with `casa`
  - Ic: `green` aligns with `casa` and `verde` (`house` implicitly aligns to null)
  - Id: `house` aligns with `casa` and `verde`
  - IIa: `the` aligns with `la` and `house` aligns with `casa`
  - IIb: `the` aligns with `casa` and `house` aligns with `la`
  - IIc: `the` aligns with `la` and `casa`
  - IId: `house` aligns with `la` and `casa`
- Now, we're going to calculate the **expected** likelihood of each of these possible alignments, according to the following formula:

$$\hat{P}(F, A|E) \quad = \quad \frac{\epsilon}{(I+1)^J} \prod_{j=1}^{J} t(f_j|e_{a_j})$$

- A close inspection might lead us to say that the $(+1)$ should be excluded, because we're neglecting the null term from the A tokens, but it doesn't actually matter, as we'll see in a moment.
- For Ia, we observe the following:

$$\begin{aligned} \hat{P}(F, A|E) &= \frac{\epsilon}{(I+1)^J} t(\texttt{casa}|\texttt{green}) t(\texttt{verde}|\texttt{house}) \\ &= \frac{\epsilon}{(2+1)^2} (\frac{1}{3})(\frac{1}{3}) = \frac{\epsilon}{9}\frac{1}{9} \end{aligned}$$

- Because our translation table is uniform, every calculation will look the same.

- Now, we're going to make a **maximum** likelihood estimate of each entry in our translation table. We do this by summing the expected probability of the alignment for each possible translation.
- For `green`:
  - It aligns with `casa` in Ia ($\frac{\epsilon}{9}\frac{1}{9}$) and Ic (same), to give a total of $\frac{\epsilon}{9}\frac{2}{9}$.
  - It aligns with `verde` in Ib ($\frac{\epsilon}{9}\frac{1}{9}$) and Ic (same), to give a total of $\frac{\epsilon}{9}\frac{2}{9}$.
  - It never aligns with `la`, because they don't appear in a sentence together.
- Let's summarise our likelihoods in the (un-normalised) translation table.

T	casa	la	verde	Total
green	$\frac{\epsilon}{9}\frac{2}{9}$	0	$\frac{\epsilon}{9}\frac{2}{9}$	$\frac{\epsilon}{9}\frac{4}{9}$
house	$\frac{\epsilon}{9}\frac{4}{9}$	$\frac{\epsilon}{9}\frac{2}{9}$	$\frac{\epsilon}{9}\frac{2}{9}$	$\frac{\epsilon}{9}\frac{8}{9}$
the	$\frac{\epsilon}{9}\frac{2}{9}$	$\frac{\epsilon}{9}\frac{2}{9}$	0	$\frac{\epsilon}{9}\frac{4}{9}$

- We will now normalise the rows so that they look like probabilities. Doing this causes all of the $\frac{\epsilon}{9}$ terms to vanish; consequently, we will just ignore them for the rest of the steps below.
- After simplifying, here is the new translation table:

T	casa	la	verde	Total
green	$\frac{1}{2}$	0	$\frac{1}{2}$	1
house	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	1
the	$\frac{1}{2}$	$\frac{1}{2}$	0	1

- At this point, it perhaps isn't obvious that this table will give us better alignment estimates, but it does:
- For Ia, we observe the following (ignoring the ϵ term):

$$\begin{aligned}\hat{P}(F,A|E) &= t(\texttt{casa}|\texttt{green})t(\texttt{verde}|\texttt{house}) \\ &= (\frac{1}{2})(\frac{1}{4}) = \frac{1}{8}\end{aligned}$$

- For Ib:

$$\begin{aligned}\hat{P}(F,A|E) &= t(\texttt{verde}|\texttt{green})t(\texttt{casa}|\texttt{house}) \\ &= (\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}\end{aligned}$$

- For Ic:

$$\begin{aligned}\hat{P}(F,A|E) &= t(\texttt{casa}|\texttt{green})t(\texttt{verde}|\texttt{green}) \\ &= (\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}\end{aligned}$$

- For Id:

$$\begin{aligned}\hat{P}(F,A|E) &= t(\texttt{casa}|\texttt{house})t(\texttt{verde}|\texttt{house}) \\ &= (\frac{1}{2})(\frac{1}{4}) = \frac{1}{8}\end{aligned}$$

- The calculations for II are similar.

- Updating the alignment counts for `green` gives us:
  - It aligns with `casa` in Ia ($\frac{1}{8}$) and Ic ($\frac{1}{4}$), to give a total of $\frac{3}{8}$.
  - It aligns with `verde` in Ib ($\frac{1}{4}$) and Ic (same), to give a total of $\frac{1}{2}$.
  - It never aligns with `la`.
- Our un-normalised counts (neglecting the ϵ terms) are now:

T	`casa`	`la`	`verde`	Total
`green`	$\frac{3}{8}$	0	$\frac{1}{2}$	$\frac{7}{8}$
`house`	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{5}{4}$
`the`	$\frac{3}{8}$	$\frac{1}{2}$	0	$\frac{7}{8}$

- We can see that we have correctly observed that `green` is most likely to be `verde`, `house` to be `casa`, and `the` to be `la`. Summarising the normalised probabilities:

T	`casa`	`la`	`verde`	Total
`green`	$\frac{3}{7}$	0	$\frac{4}{7}$	1
`house`	$\frac{3}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	1
`the`	$\frac{3}{7}$	$\frac{4}{7}$	0	1

- Further iterations will continue to improve these counts, and to observe that Ib and IIa are the most likely alignments.