1 Statistical Machine Translation

1.1 Background: Noisy Model

The general aim is to maximise faithfulness and fluency. We use the 'Noisy Channel' model, which takes the (reverse) perspective that a message has been corrupted in some way and the task is to find the original message.

If we were translating from a foreign sequence $F = f_1, f_2, .., f_n$ into an English sentence $E = e_1, e_2, .., e_n$, then:

$$\hat{E} = argmax_E P(E|F) \tag{1}$$

$$= argmax_E P(F|E)P(E) \tag{2}$$

$P(F|E)$ **is the translation model (corresponding to faithfulness) and requires a parallel corpus to learn probabilities.** $P(E)$ **is our language model (corresponding to fluency) and needs only a single language corpus (i.e. an n-gram language model).**

1.2 Naive Translation Model (Phrase Based)

Idea: Use a phrase (sequence of words) and/or single words as the fundamental units. Assuming that the parallel corpus is phrase aligned and may feature one-to-many alignments, the steps are:

1. Group English words into phrases $e_1, e_2, .., e_I$ and $f_1, f_2, .., f_i$
2. Translate each e_i into f_i
3. (optional) Reorder

The overall translation probability becomes:

$$P(F|E) = \prod^{I} \phi(f_j, e_i) d(a_i - b_{i-1}) \tag{3}$$

$\phi(f_i, e_i)$ is the translation probability - learnt from the corpus counts

$d(a_i - b_i) = \alpha^{|a_i - b_i|}$ is the distortion probability, a weighting to penalise larger offsets. a_i is foreign word start position for phrase i, b_i is the foreign word finish position e.g. if the foreign phrase starts at 3 and finishes at 5, then $d(a_i - b_i) = \alpha^{|3-5|}$

1.3 Word Alignment Translation Model: IBM Model 1

1.3.1 Definition

As direct phrase alignments are difficult, the problem might get broken down into one of word alignments.

$$P(F|E) = \sum^{A} P(F, A|E) \tag{4}$$

Let us assume the following:

- One to many translations possible (e.g. I^J possible alignments)
- Words may be dropped from the source sequence
- Words may be generated from NULL in the source sequence: $((I+1)^J$ alignments now

Then the generative steps of producing F from E are:

1. Choose a Foreign sentence length $F = f_1, f_2, .., f_J$
2. Choose an English sentence length $E = e_1, e_2, .., e_I$
3. Choose word alignments for the foreign sentence $A = a_1, a_2, .., a_J$
4. For each position in F, generate a word f_j from the aligned word in E (e_{aj}) with probability $t(f_j|e_{aj})$

1.3.2 Decoding: Computing $P(F|E)$ and the most probable alignment

Using the word alignment model, the probability of generating F from E is the probability of F from E with some alignment.

$$P(F, A|E) = P(F|E, A)P(A|E) \tag{5}$$

$$P(F, A|E) = \big(\prod^{J} t(f_j|e_{aj})\big) \cdot \big(\frac{\epsilon}{(I+1)^J}\big) \tag{6}$$

The alignment term (second term above) is assumed uniform for all alignments and comes about from knowing there are $(I+1)^J$ possible alignments to choose from where J (the length of F) is chosen with some small (uniform) probability represented by ϵ

A **is our decoding (and is the latent variable). We find it by:**

$$\hat{A} = argmax_A \prod^{J} t(f_j|e_{aj}) \tag{7}$$

1.3.3 Training the model with EM: An example

The parameters to be learned are $t(f_i|e_{aj})$

Two sentence pairs:
("green house", "casa verde") and ("the house", "la casa")

The vocabularies are:
V_E:("the", "green", "house") and V_F:("la", "casa", "verde")

- initialising t with uniform probabilities:

$$t(casa|green) = \frac{1}{3}$$
$$t(casa|house) = \frac{1}{3}$$
$$t(casa|the) = \frac{1}{3}$$
$$t(verde|green) = \frac{1}{3}$$
$$...$$

- E: Compute $P(A, F|E) = \prod^{J} t(f_i|e_{aj})$ for all possible sentences and their alignments (e.g. 2 sentences consisting of 2 words each equates to 4 alignments total). Normalise the results (sum over A)
- E: Get totals by adding up any duplicate pairs (e.g. "casa" and "house" are in both sentences so the counts add)
- M: For each English word e_i normalise $t(f_j|e_i)$