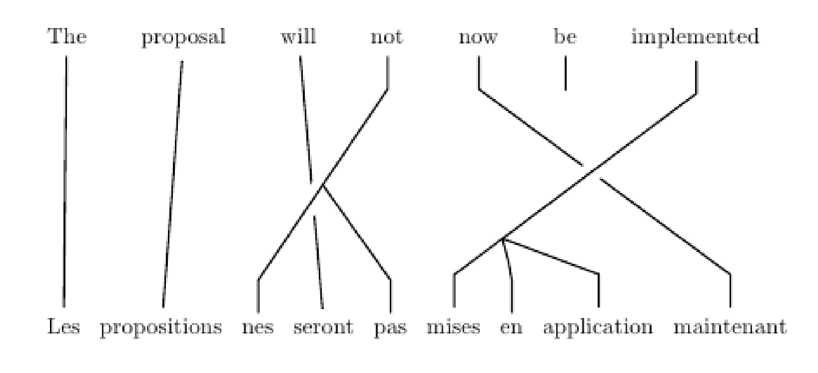
WORD ALIGNMENT MODELS

David Talbot

Autumn 2020

Yandex School of Data Analysis

WORD ALIGNMENT MODELS



Brown et al. (1990).

'The Mathematics of Machine Translation', Brown et al. (1993).

$$e^* = \operatorname{argmax} \Pr(e)\Pr(f|e)$$

‘The Mathematics of Machine Translation‘, Brown et al. (1993).

$$e^* = \text{argmax}\, \Pr(e)\Pr(f|e)$$

Why is modelling $\Pr(f|e)$ easier than modelling $\Pr(e|f)$ if we want to translate from f to e?

What's a simple generative model for sentences $\Pr(f|e)$?

What's a simple generative model for sentences $\Pr(\mathrm{f}|\mathrm{e})$?

How can word alignments simplify our model of sentences?

$$\Pr(\mathrm{f}, \mathrm{a}|\mathrm{e})$$

What's a simple generative model for sentences $\Pr(\mathrm{f}|\mathrm{e})$?

How can word alignments simplify our model of sentences?

$$\Pr(\mathrm{f}, \mathrm{a}|\mathrm{e})$$

Allow f to depend only on those e aligned to it

What's a simple generative model for sentences $\Pr(\mathrm{f}|\mathrm{e})$?

How can word alignments simplify our model of sentences?

$$\Pr(\mathrm{f}, \mathrm{a}|\mathrm{e})$$

Allow f to depend only on those e aligned to it

How do hidden variables complicate the parameter estimation?

What's a simple generative model for sentences $\Pr(\mathrm{f}|\mathrm{e})$?

How can word alignments simplify our model of sentences?

$$\Pr(\mathrm{f}, \mathrm{a}|\mathrm{e})$$

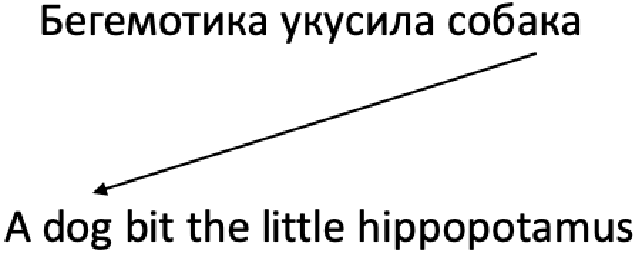
Allow f to depend only on those e aligned to it

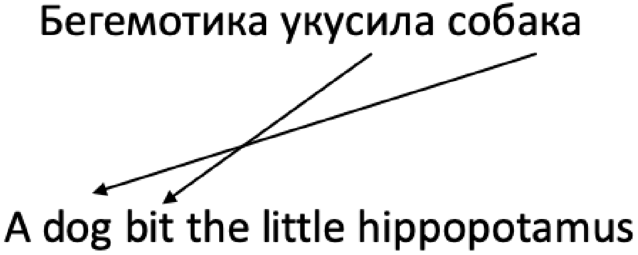
How do hidden variables complicate the parameter estimation?

$$\Pr(\mathrm{f}|\mathrm{e}) = \sum_{\mathrm{a} \in \mathcal{A}} \Pr(\mathrm{f}, \mathrm{a}|\mathrm{e})$$

Бегемотика укусила собака

A dog bit the little hippopotamus


Бегемотика укусила собака
A dog bit the little hippopotamus


Бегемотика укусила собака
A dog bit the little hippopotamus

EXAMPLE ALIGNMENT

MODELLING PARALLEL SENTENCE PAIRS

$$\Pr(\mathsf{f}|\mathsf{e}) = \sum_{\mathsf{a} \in \mathcal{A}} \Pr(\mathsf{f}, \mathsf{a}|\mathsf{e})$$

$$\Pr(\mathrm{f}|\mathrm{e}) = \sum_{\mathrm{a} \in \mathcal{A}} \Pr(\mathrm{f}, \mathrm{a}|\mathrm{e})$$

What is the cardinality of $\mathcal{A}$ for a single sentence pair $|\mathrm{e}| = I$ and $|\mathrm{f}| = J$?

$$\Pr(\mathrm{f}|\mathrm{e}) = \sum_{\mathrm{a}\in\mathcal{A}} \Pr(\mathrm{f}, \mathrm{a}|\mathrm{e})$$

What is the cardinality of $\mathcal{A}$ for a single sentence pair $|\mathrm{e}| = I$ and $|\mathrm{f}| = J$?

$$\Pr(f_1, ..., f_J|e_1, ..., e_I, \theta) = \sum_{a_1=1}^{I} ... \sum_{a_J=1}^{I} \Pr(f_1, ..., f_J, a_1, ..., a_J|e_1, ..., e_I, \theta)$$

$$\Pr(\mathrm{f}|\mathrm{e}) = \sum_{\mathrm{a}\in\mathcal{A}} \Pr(\mathrm{f}, \mathrm{a}|\mathrm{e})$$

What is the cardinality of $\mathcal{A}$ for a single sentence pair $|\mathrm{e}| = I$ and $|\mathrm{f}| = J$?

$$\Pr(f_1, ..., f_J|e_1, ..., e_I, \theta) = \sum_{a_1=1}^{I} ... \sum_{a_J=1}^{I} \Pr(f_1, ..., f_J, a_1, ..., a_J|e_1, ..., e_I, \theta)$$

Exact E-step is only tractable for a very limited set of models.

Formulated a generative model of parallel sentence pairs

$$\begin{aligned}\Pr(\mathsf{f}|\mathsf{e}) &= \sum_{\mathsf{a}\in\mathcal{A}} \Pr(\mathsf{a},\mathsf{f}|\mathsf{e}) \\ &= \sum_{\mathsf{a}\in\mathcal{A}} \underbrace{\Pr(\mathsf{a}|\mathsf{e})}_{\text{Prior}} \underbrace{\Pr(\mathsf{f}|\mathsf{e},\mathsf{a})}_{\text{Translation model}}\end{aligned}$$

where f is a French sentence, e is an English sentence and $\mathcal{A}$ is the set of all possible alignments for the sentence pair.

We're given corpus of translated sentence pairs $D = \{(\mathrm{e}_1, \mathrm{f}_1), (\mathrm{e}_2, \mathrm{f}_2), (\mathrm{e}_3, \mathrm{f}_3), ...\}$.

We assume these sentence pairs are distributed *i.i.d.*.

We're given corpus of translated sentence pairs $D = \{(\mathrm{e}_1, \mathrm{f}_1), (\mathrm{e}_2, \mathrm{f}_2), (\mathrm{e}_3, \mathrm{f}_3), \ldots\}$.

We assume these sentence pairs are distributed *i.i.d.*.

$$\begin{aligned}
\Pr(D|\theta) &\approx \prod_{k \in D} \Pr(\mathrm{f}_k | \mathrm{e}_k, \theta) \\
&= \prod_{k \in D} \sum_{\mathrm{a}_k \in \mathcal{A}} \Pr(\mathrm{a}_k, \mathrm{f}_k | \mathrm{e}_k, \theta) \\
&= \prod_{k \in D} \sum_{\mathrm{a}_k \in \mathcal{A}} \underbrace{\Pr(\mathrm{a}_k | \mathrm{e}_k, \theta)}_{\text{Prior}} \underbrace{\Pr(\mathrm{f}_k | \mathrm{e}_k, \mathrm{a}_k, \theta)}_{\text{Translation model}}
\end{aligned}$$

Assumption 1

Each French word f_j is generated independently given the English word to which it is aligned e_{a_j}, i.e.

$$\mathrm{Pr}(\mathrm{f}, \mathrm{a}|\mathrm{e}) \approx \prod_{j=1}^{J} \mathrm{Pr}(\mathrm{a}|\mathrm{e}, \theta)\mathrm{Pr}(f_j|e_{a_j}, \theta).$$

Assumption 2

We'll parameterize the translation model $\Pr(f_j|e_{a_j}, \theta)$ with a table of conditional probabilities $t(f|e)$.

E.g. for Russian to English translation the table $t(\text{f}|dog)$ could be defined as

$$t(\text{собака}|\text{dog}) = 0.5$$

$$t(\text{собаку}|\text{dog}) = 0.3$$

$$t(\text{кошка}|\text{dog}) = 0.2.$$

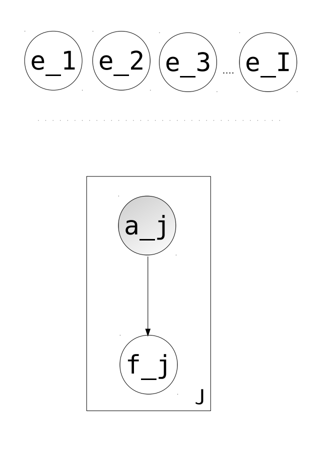
Assumption 3

We'll simplify the 'prior' $\mathrm{Pr}(a|\mathrm{e},\theta)$ by assuming that a_j depends only on a subset of the other alignments, i.e.

$$\mathrm{Pr}(\mathrm{f},\mathrm{a}|\mathrm{e}) \approx \prod_{j=1}^{J} \mathrm{Pr}(a_j|\mathrm{a}_{\mathrm{subset}},\mathrm{e},\theta)\mathrm{Pr}(f_j|e_{a_j},\theta).$$

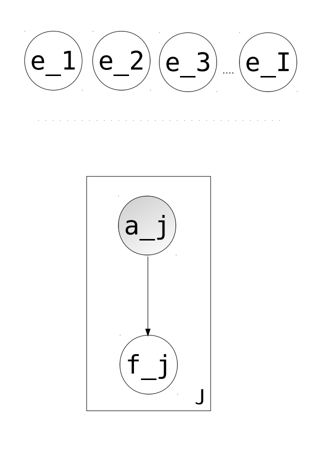
$$
\begin{aligned}
\Pr(\mathsf{f}, \mathsf{a} | \mathsf{e}, \theta) &= \Pr(\mathsf{a} | \mathsf{e}, \theta) \Pr(\mathsf{f} | \mathsf{e}, \mathsf{a}, \theta) \\
&= \prod_{j=1}^{J} \Pr(a_j | \mathsf{a}_1^{j-1}, \mathsf{f}_1^{j-1}, \mathsf{e}, \theta) \Pr(f_j | \mathsf{a}_1^{j}, \mathsf{f}_1^{j-1}, \mathsf{e}, \theta) \\
&\approx \prod_{j=1}^{J} \Pr(a_j | \mathsf{a}_1^{j-1}, \mathsf{f}_1^{j-1}, \mathsf{e}, \theta) \Pr(f_j | e_{a_j}, \theta) \\
&\approx \prod_{j=1}^{J} \underbrace{\Pr(a_j | \mathsf{a}_{subset}, \mathsf{e}, \theta)}_{\text{prior model}} \underbrace{\Pr(f_j | e_{a_j}, \theta)}_{\text{translation model}}
\end{aligned}
$$

IBM MODEL 1: UNIFORM PRIOR



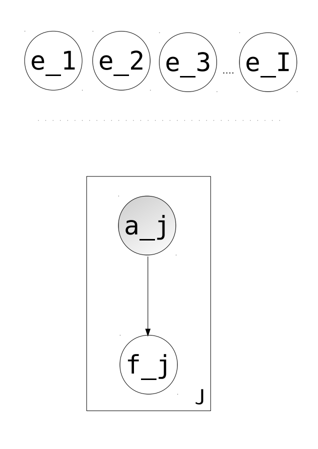
$$\Pr(\mathrm{f}, \mathrm{a} | \mathrm{e}, \theta) \approx \prod_{j=1}^{J} \Pr(f_j, a_j | \mathrm{e}, \theta)$$

IBM MODEL 1: UNIFORM PRIOR



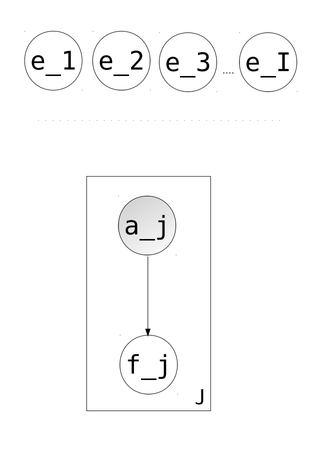
$$
\begin{aligned}
\Pr(\mathrm{f}, \mathrm{a}|\mathrm{e}, \theta) &\approx \prod_{j=1}^{J} \Pr(f_j, a_j|\mathrm{e}, \theta) \\
&= \prod_{j=1}^{J} \Pr(a_j|\mathrm{e}) \Pr(f_j|\mathrm{e}, a_j, \theta)
\end{aligned}
$$

IBM MODEL 1: UNIFORM PRIOR



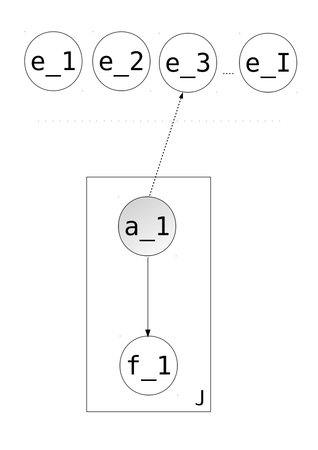
$$
\begin{aligned}
\Pr(\mathrm{f}, \mathrm{a} | \mathrm{e}, \theta) &\approx \prod_{j=1}^{J} \Pr(f_j, a_j | \mathrm{e}, \theta) \\
&= \prod_{j=1}^{J} \Pr(a_j | \mathrm{e}) \Pr(f_j | \mathrm{e}, a_j, \theta) \\
&\approx \prod_{j=1}^{J} \epsilon \Pr(f_j | e_{a_j}, \theta)
\end{aligned}
$$

IBM MODEL 1: UNIFORM PRIOR



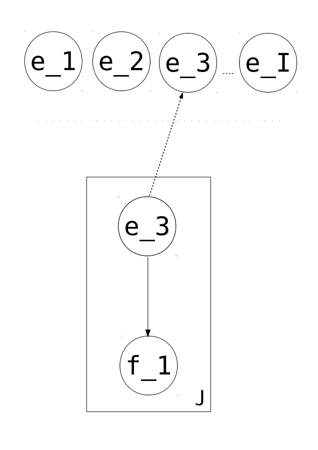
$$
\begin{aligned}
\Pr(\mathrm{f}, \mathrm{a} | \mathrm{e}, \theta) &\approx \prod_{j=1}^{J} \Pr(f_j, a_j | \mathrm{e}, \theta) \\
&= \prod_{j=1}^{J} \Pr(a_j | \mathrm{e}) \Pr(f_j | \mathrm{e}, a_j, \theta) \\
&\approx \prod_{j=1}^{J} \epsilon \Pr(f_j | e_{a_j}, \theta) \\
&\propto \prod_{j=1}^{J} t(f_j | e_{a_j})
\end{aligned}
$$

IBM MODEL 1

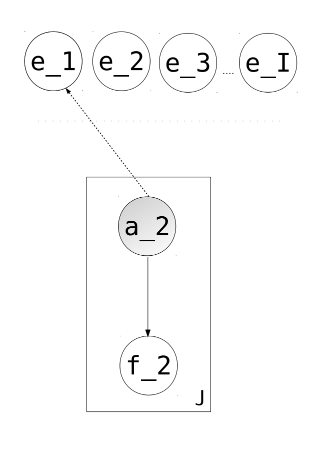


$$
\begin{aligned}
\Pr(\mathrm{f}, \mathrm{a} | \mathrm{e}, \theta) &\approx \prod_{j=1}^{J} \Pr(f_j, a_j | e_{a_j}, \theta) \\
&= \Pr(f_1, a_1 = 3 | e_3, \theta) \ldots
\end{aligned}
$$

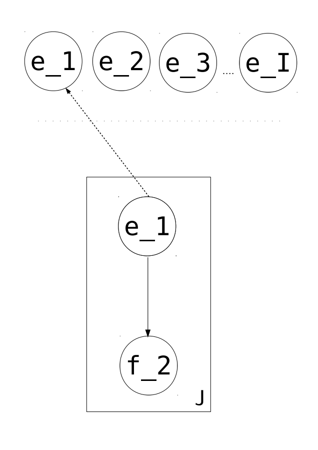
IBM MODEL 1



$$
\begin{aligned}
\Pr(\mathrm{f}, \mathrm{a}|\mathrm{e}, \theta) &\approx \prod_{j=1}^{J} \Pr(f_j, a_j | e_{a_j}, \theta) \\
&= \Pr(f_1, a_1 = 3 | e_3, \theta) \ldots \\
&\approx t(f_1, | e_3) \ldots
\end{aligned}
$$



$$\begin{aligned}\Pr(\mathrm{f},\mathrm{a}|\mathrm{e},\theta) &\approx \prod_{j=1}^{J}\Pr(f_j,a_j|e_{a_j},\theta)\\ &= \Pr(f_1,a_1=3|e_3,\theta)\ldots\\ &\approx t(f_1,|e_3)\ldots\end{aligned}$$



$$
\begin{aligned}
\Pr(\mathrm{f}, \mathrm{a}|\mathrm{e}, \theta) &\approx \prod_{j=1}^{J} \Pr(f_j, a_j|e_{a_j}, \theta) \\
&= \Pr(f_1, a_1 = 3|e_3, \theta) \ldots \\
&\approx t(f_1, |e_3) \ldots \\
&\approx t(f_1, |e_3) t(f_2|e_1) \ldots
\end{aligned}
$$

- Co-occurrence: if an input word occurs commonly with an output word, it acquires more weight.

- Co-occurrence: if an input word occurs commonly with an output word, it acquires more weight.

- Explaining away: once an output word is explained by one input, then the other inputs become less important.

- Co-occurrence: if an input word occurs commonly with an output word, it acquires more weight.

- Explaining away: once an output word is explained by one input, then the other inputs become less important.

- (Assuming a uniform prior)

$$\Pr(a_j = i | \mathrm{e}, \mathrm{f}) = \frac{\Pr(f_j | e_i)}{\sum_k \Pr(f_j | e_k)}$$

LIMITATIONS

- No guarantee that all input words will be aligned (i.e. explain some output).

LIMITATIONS

- No guarantee that all input words will be aligned (i.e. explain some output).

- HMM can force some dependency between decisions but no guarantee that it will use all the input.

LIMITATIONS

- No guarantee that all input words will be aligned (i.e. explain some output).

- HMM can force some dependency between decisions but no guarantee that it will use all the input.

- More complex models can guarantee all input words are aligned but ....

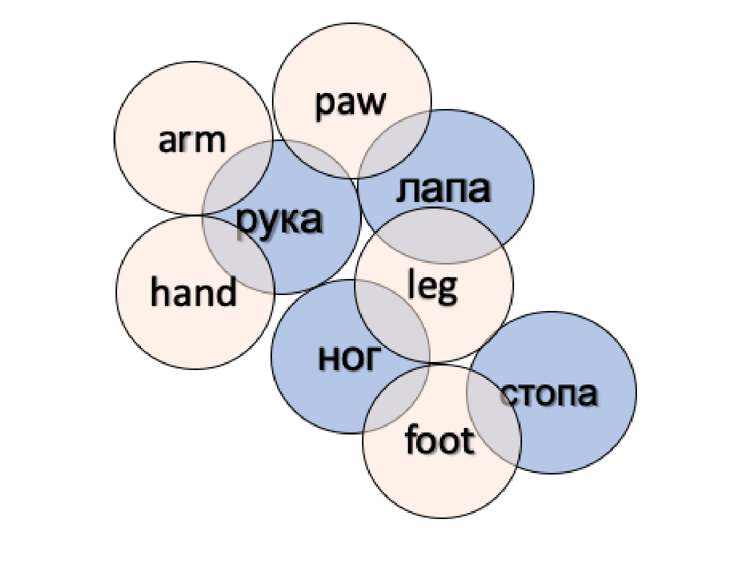

Бегемотика укусила собака
A dog bit the little hippopotamus


Бегемотика укусила собака
A dog bit the little hippopotamus

EXAMPLE ALIGNMENT


null
Бегемотика укусила собака
A dog bit the little hippopotamus

EXAMPLE ALIGNMENT


paw
arm
лапа
рука
hand
leg
ног
стопа
foot

The expected log-likelihood for f given e under IBM Model 1 is

$$\begin{aligned}\mathbb{E}[log(\mathrm{f}|\mathrm{e},\theta)] &= \sum_{j=1}^{J}\sum_{i=1}^{I} \Pr(a_j = i|\mathrm{f},\mathrm{e},\theta)\log \Pr(f_j, a_j = i|e_i, \theta) \\ &= \sum_{j=1}^{J}\sum_{i=1}^{I} \Pr(a_j = i|\mathrm{f},\mathrm{e},\theta)\log t(f_j|e_i) + C.\end{aligned}$$

To apply EM we need to compute $\Pr(a_j = i|\mathrm{f},\mathrm{e},\theta)$ for each source and target pair and then maximize this term w.r.t. our parameters $\theta = t(f|e)$.

The posterior alignment probabilities, $\Pr(a_j = i|\mathsf{f}, \mathsf{e}, \theta)$ can be computed as follows

$$\Pr(\mathsf{a}|\mathsf{f}, \mathsf{e}, \theta) = \frac{\Pr(\mathsf{f}, \mathsf{a}|\mathsf{e}, \theta)}{\sum_k \Pr(\mathsf{f}, \mathsf{a}' = k|\mathsf{e}, \theta)} \tag{1}$$

$$= \frac{\Pr(a_j = i|\mathsf{e}, \theta)\Pr(f_j|a_j = i, \mathsf{e}, \theta)}{\sum_{k=1}^{I} \Pr(a_j = k|\mathsf{e}, \theta)\Pr(f_j|a_j = k, \mathsf{e}, \theta)} \tag{2}$$

$$= \frac{\epsilon t(f_j|e_i)}{\sum_{k=1}^{I} \epsilon t(f_j|e_k)} \tag{3}$$

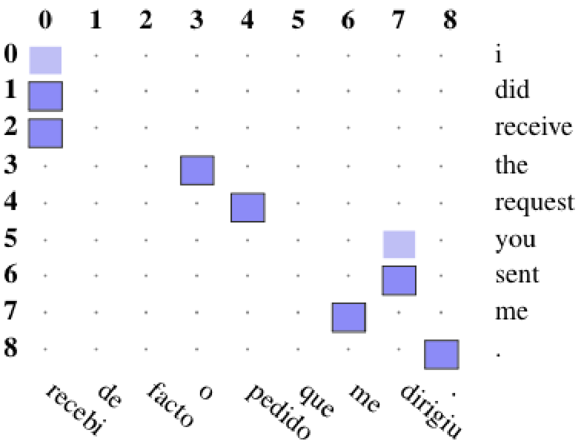
$$= \frac{t(f_j|e_i)}{\sum_{k=1}^{I} t(f_j|e_k)}. \tag{4}$$

Given a golden set of manually created M consisting of probable P and sure S alignments. We can measure the error rate of an automatic alignment A:

$$Precision(A; P) = \frac{|P \cap A|}{|A|}$$

$$Recall(A; S) = \frac{|S \cap A|}{|S|}$$

$$AlignmentErrorRate(A; S, P) = 1 - \frac{|P \cap A| + |S \cap A|}{|S| + |A|}.$$



Natural way to visualize an alignment.

IS ATTENTION AN ALIGNMENT?

- Attention scores

- Attention scores

$$\Pr(a_j = i | e_1, \dots, e_m, s_j) = \frac{e^{A_\theta(e_i, s_j)}}{\sum_k e^{A_\theta(e_k, s_j)}}$$

IS ATTENTION AN ALIGNMENT?

- Attention scores

$$\Pr(a_j = i | e_1, ..., e_m, s_j) = \frac{e^{A_\theta(e_i, s_j)}}{\sum_k e^{A_\theta(e_k, s_j)}}$$

- Soft attention averages contexts

$$c_j = \sum_i \Pr(a_j = i | e_1, ..., e_m, s_j) e_i$$

- Soft attention averages contexts

$$c_j = \sum_i \Pr(a_j = i | e_1, \ldots, e_m, s_j) e_i$$

IS ATTENTION AN ALIGNMENT?

- Soft attention averages contexts

$$c_j = \sum_i \Pr(a_j = i | e_1, ..., e_m, s_j) e_i$$

- So something like

$$\Pr(f_j | s_j, c_j) = \Pr(f_j | s_j, \sum_i \Pr(a_j = i | e_1, ..., e_m, s_i) e_i)$$

- Soft attention averages contexts

$$c_j = \sum_i \Pr(a_j = i | e_1, ..., e_m, s_j) e_i$$

- So something like

$$\Pr(f_j | s_j, c_j) = \Pr(f_j | s_j, \sum_i \Pr(a_j = i | e_1, ..., e_m, s_i) e_i)$$

- Hard attention is closer to a probabilistic alignment

$$\Pr(f_j | e_1, ..., e_m, s_j) = \sum_i \Pr(a_j = i | e_1, ..., e_m, s_t) \Pr(f_j | e_i, s_j)$$