

WORD ALIGNMENT MODELS

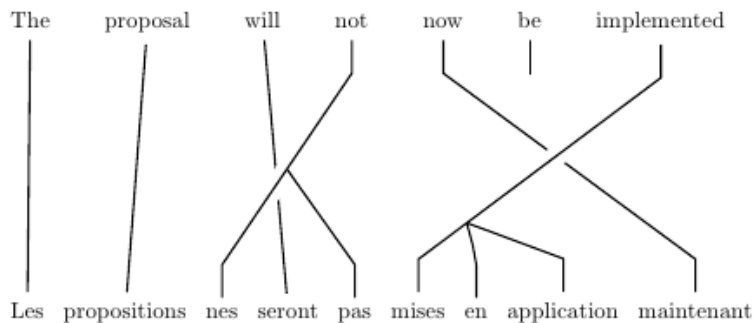
David Talbot

Autumn 2019

Yandex School of Data Analysis

WORD ALIGNMENT MODELS

THE FIRST WORD ALIGNMENT



Brown et al. (1990).

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

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

How can word alignments simplify our model of sentences?

$$\Pr(f, a|e)$$

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

How can word alignments simplify our model of sentences?

$$\Pr(f, a|e)$$

Allow f to depend only on those e aligned to it

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

How can word alignments simplify our model of sentences?

$$\Pr(f, a|e)$$

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 $\Pr(f|e)$?

How can word alignments simplify our model of sentences?

$$\Pr(f, a|e)$$

Allow f to depend only on those e aligned to it

How do hidden variables complicate the parameter estimation?

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

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

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

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

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

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

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

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

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

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

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

Formulated a generative model of parallel sentence pairs

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

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

$$D = \{(e_1, f_1), (e_2, f_2), (e_3, f_3), \dots\}.$$

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

We're given corpus of translated sentence pairs

$$D = \{(e_1, f_1), (e_2, f_2), (e_3, f_3), \dots\}.$$

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

$$\begin{aligned}\Pr(D|\theta) &\approx \prod_{k \in D} \Pr(f_k | e_k, \theta) \\ &= \prod_{k \in D} \sum_{a_k \in \mathcal{A}} \Pr(a_k, f_k | e_k, \theta) \\ &= \prod_{k \in D} \sum_{a_k \in \mathcal{A}} \underbrace{\Pr(a_k | e_k, \theta)}_{\text{Prior}} \underbrace{\Pr(f_k | e_k, a_k, \theta)}_{\text{Translation model}}\end{aligned}$$

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



Бегемотика укусила собака



A dog bit the little hippopotamus

Assumption 1

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

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

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 $t(f|dog)$ could be defined as

$$t(\text{собака}|dog) = 0.5$$

$$t(\text{собаку}|dog) = 0.3$$

$$t(\text{кошка}|dog) = 0.2.$$

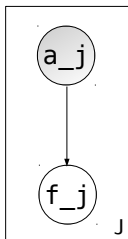
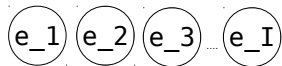
Assumption 3

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

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

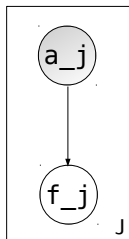
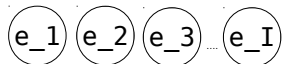
$$\begin{aligned}
\Pr(\mathbf{f}, \mathbf{a} | \mathbf{e}, \theta) &= \Pr(\mathbf{a} | \mathbf{e}, \theta) \Pr(\mathbf{f} | \mathbf{e}, \mathbf{a}, \theta) \\
&= \prod_{j=1}^J \Pr(a_j | \mathbf{a}_1^{j-1}, \mathbf{f}_1^{j-1}, \mathbf{e}, \theta) \Pr(f_j | \mathbf{a}_1^j, \mathbf{f}_1^{j-1}, \mathbf{e},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

IBM MODEL 1: UNIFORM PRIOR



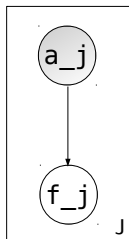
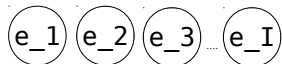
$$\Pr(f, a|e, \theta) \approx \prod_{j=1}^J \Pr(f_j, a_j|e, \theta)$$

IBM MODEL 1: UNIFORM PRIOR



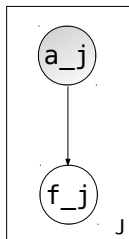
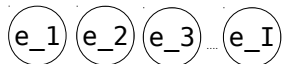
$$\begin{aligned}\Pr(\mathbf{f}, \mathbf{a} | \mathbf{e}, \theta) &\approx \prod_{j=1}^J \Pr(f_j, a_j | \mathbf{e}, \theta) \\ &= \prod_{j=1}^J \Pr(a_j | \mathbf{e}) \Pr(f_j | \mathbf{e}, a_j, \theta)\end{aligned}$$

IBM MODEL 1: UNIFORM PRIOR

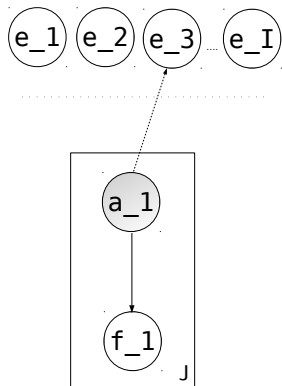


$$\begin{aligned}\Pr(\mathbf{f}, \mathbf{a} | \mathbf{e}, \theta) &\approx \prod_{j=1}^J \Pr(f_j, a_j | \mathbf{e}, \theta) \\ &= \prod_{j=1}^J \Pr(a_j | \mathbf{e}) \Pr(f_j | \mathbf{e}, a_j, \theta) \\ &\approx \prod_{j=1}^J \epsilon \Pr(f_j | e_{a_j}, \theta)\end{aligned}$$

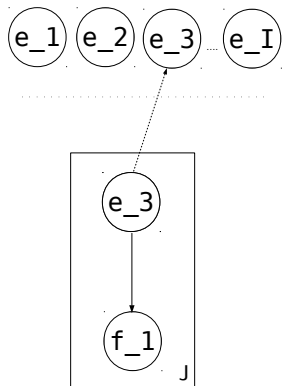
IBM MODEL 1: UNIFORM PRIOR



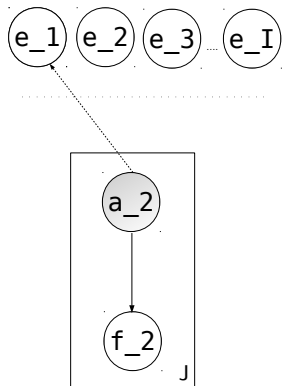
$$\begin{aligned}\Pr(f, a|e, \theta) &\approx \prod_{j=1}^J \Pr(f_j, a_j|e, \theta) \\ &= \prod_{j=1}^J \Pr(a_j|e) \Pr(f_j|e, a_j, \theta) \\ &\approx \prod_{j=1}^J \epsilon \Pr(f_j|e_{a_j}, \theta) \\ &\propto \prod_{j=1}^J t(f_j|e_{a_j})\end{aligned}$$



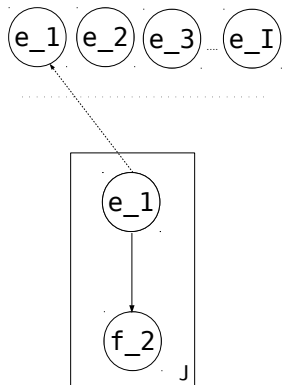
$$\begin{aligned} \Pr(f, a|e, \theta) &\approx \prod_{j=1}^J \Pr(f_j, a_j | e_{a_j}, \theta) \\ &= \Pr(f_1, a_1 = 3 | e_3, \theta) \dots \end{aligned}$$



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



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



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

COINS?

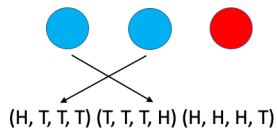
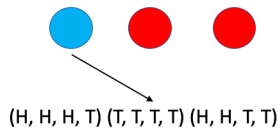


(H, H, H, T) (T, T, T, T) (H, H, T, T)

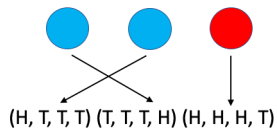
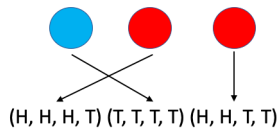


(H, T, T, T) (T, T, T, H) (H, H, H, T)

COINS?



COINS?



- 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)

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

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.

- 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

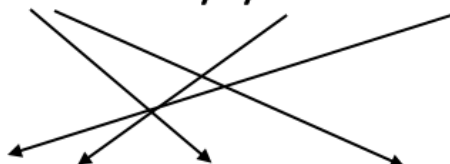


A dog bit the little hippopotamus

null

Бегемотика укусила собака

A dog bit the little hippopotamus

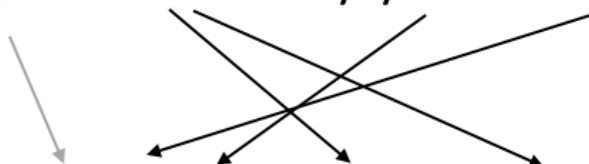


EXAMPLE ALIGNMENT

null

Бегемотика укусила собака

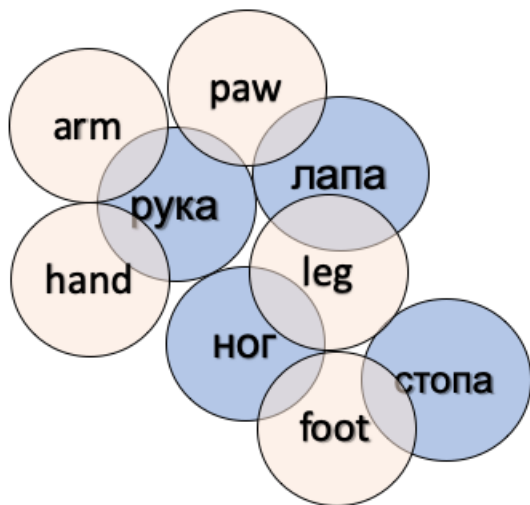
A dog bit the little hippopotamus



EXAMPLE ALIGNMENT



SEMANTIC DRIFT



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

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

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

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

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

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

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

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

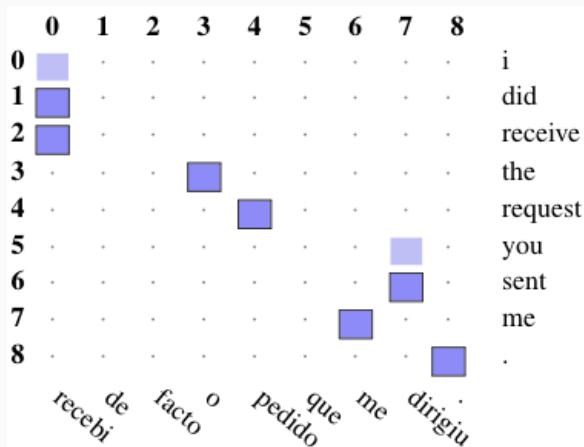
Given a golden set of manually created M consisting of probable P and sure S alignments. We can measure the error rate of an automatic alignment A :

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

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

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

WORD ALIGNMENT MATRIX



Natural way to visualize an alignment.

‘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).

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

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

IS ATTENTION AN ALIGNMENT?

- Attention scores

IS ATTENTION AN ALIGNMENT?

- 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

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

- Soft attention averages contexts

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

IS ATTENTION AN ALIGNMENT?

- Soft attention averages contexts

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

IS ATTENTION AN ALIGNMENT?

- Soft attention averages contexts

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

- So something like

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

IS ATTENTION AN ALIGNMENT?

- Soft attention averages contexts

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

- So something like

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

- Hard attention is closer to a probabilistic alignment

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