WORD ALIGNMENT MODELS

David Talbot

Autumn 2018

Yandex School of Data Analysis

WORD ALIGNMENT MODELS

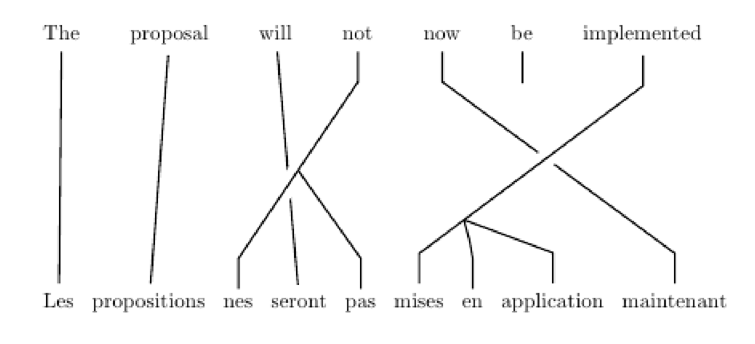
‘The Mathematics of Machine Translation‘, Brown et al. (1993).

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

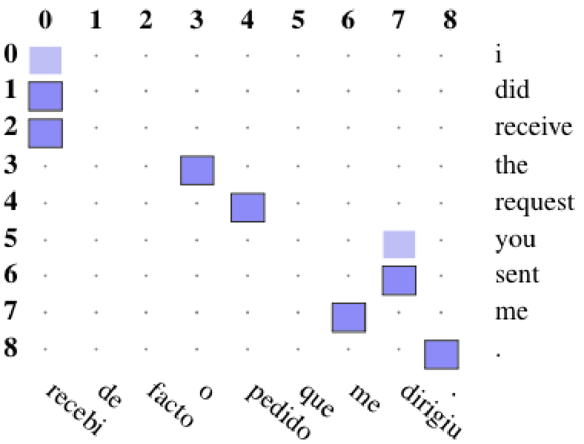
'The Mathematics of Machine Translation', Brown et al. (1993).

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

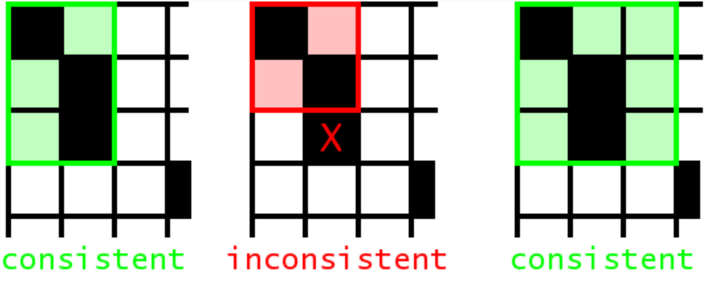
Why is modelling $\Pr(f|e)$ easier than modelling $\Pr(e|f)$ if we want to translate from f to e?



Brown et al. (1993).



Natural way to visualize an alignment.



Word alignments constrain the set of possible phrase pairs.

Formulated a generative model of parallel sentence pairs

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

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.

e = The dog bit the hippopotamus .

a =

f = Бегемотика укусила собака .

CONDITIONAL GENERATIVE MODEL

e = The dog bit the hippopotamus .

a = 5

f = Бегемотика укусила собака .

e = The dog bit the hippopotamus .

a = 5 3

f = Бегемотика укусила собака .

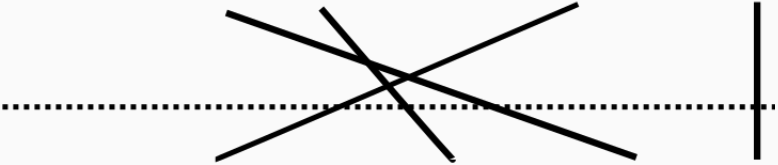
CONDITIONAL GENERATIVE MODEL

e = The dog bit the hippopotamus .

a = 5 3 2

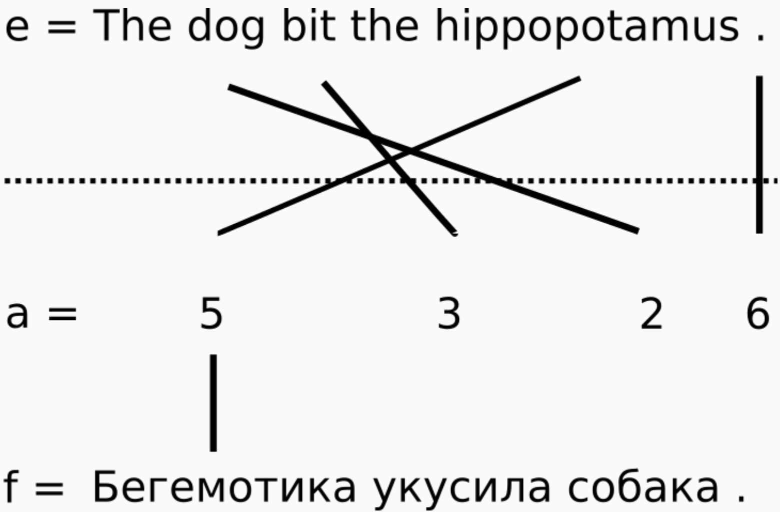
f = Бегемотика укусила собака .

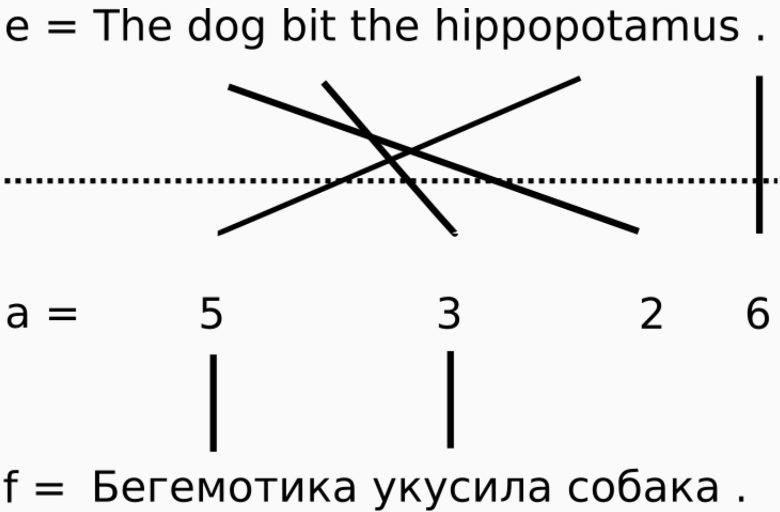
e = The dog bit the hippopotamus .



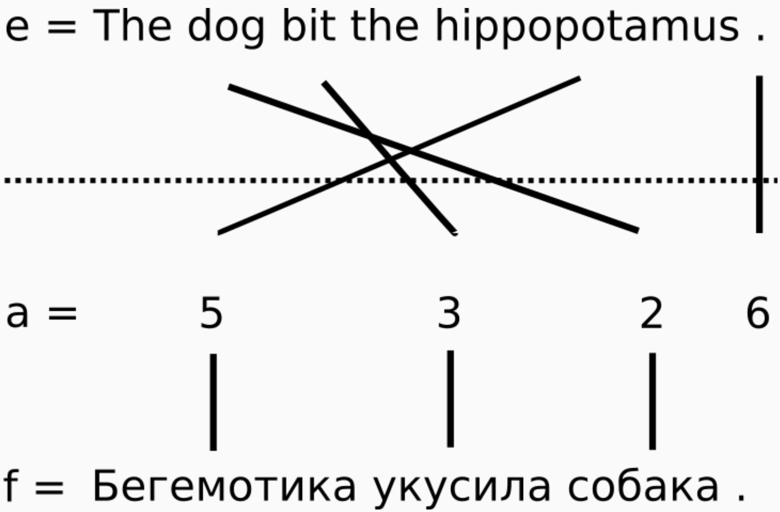
a = 5 3 2 6

f = Бегемотика укусила собака .

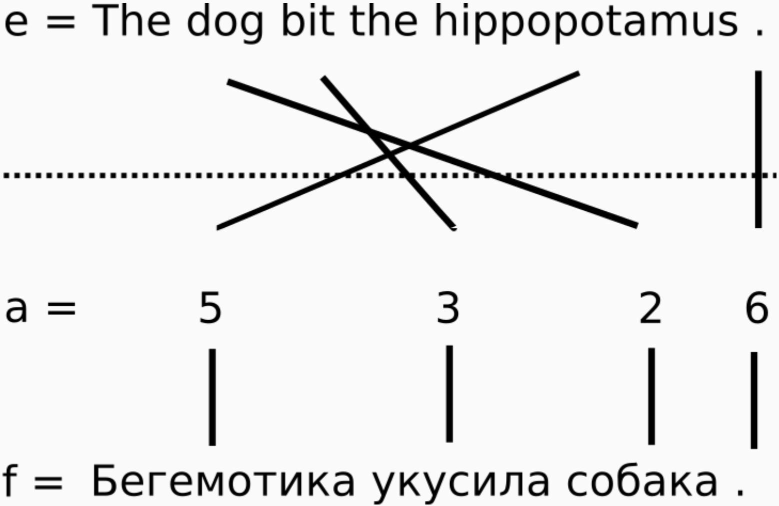

e = The dog bit the hippopotamus .
a = 5 3 2 6
f = Бегемотика укусила собака .


e = The dog bit the hippopotamus .
a = 5 3 2 6
f = Бегемотика укусила собака .

CONDITIONAL GENERATIVE MODEL



CONDITIONAL GENERATIVE MODEL



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

We assume these sentence pairs are distributed *i.i.d.* given the parameters θ,

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

We assume these sentence pairs are distributed *i.i.d.* given the parameters θ,

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

Bias-variance trade-off

Simple models (few parameters) generalize better to new data, but may not capture the structure of the data (e.g. unigram *n*-gram model).

Complex models (many parameters) capture the structure of the training data, but generalize less well to new data (e.g. unsmoothed 5-gram model).

How can word alignments simplify our model of $\Pr(f|e)$?

How can word alignments simplify our model of $\Pr(f|e)$?

How do hidden variables complicate the choice of model structure?

How can word alignments simplify our model of $\Pr(f|e)$?

How do hidden variables complicate the choice of model structure?

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

How can word alignments simplify our model of $\Pr(f|e)$?

How do hidden variables complicate the choice of model structure?

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

What if we limit ourselves to a single alignment per target word?

How can word alignments simplify our model of $\Pr(f|e)$?

How do hidden variables complicate the choice of model structure?

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

What if we limit ourselves to a single alignment per target word?

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

How can word alignments simplify our model of $\Pr(f|e)$?

How do hidden variables complicate the choice of model structure?

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

What if we limit ourselves to a single alignment per target word?

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

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

Assumption 1

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

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

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 $t(\mathrm{f}|dog)$ could be defined as

$$t(\text{собака}|\text{dog}) = 0.5$$

$$t(\text{собаку}|\text{dog}) = 0.3$$

$$t(\text{кошка}|\text{dog}) = 0.2.$$

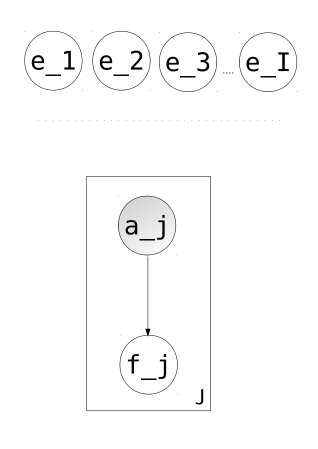
Assumption 3

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

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

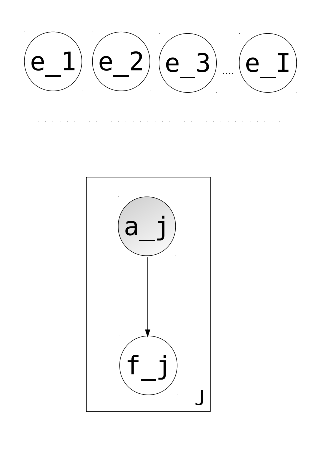
$$
\begin{aligned}
\Pr(\mathsf{f}, \mathsf{a} | \mathsf{e}, \theta) &= \Pr(\mathsf{a}|\mathsf{e}, \theta)\Pr(\mathsf{f}|\mathsf{e}, \mathsf{a}, \theta) \\
&= \prod_{j=1}^{J} \Pr(a_j | \mathsf{a}_1^{j-1}, \mathsf{f}_1^{j-1}, \mathsf{e}, \theta)\Pr(f_j | \mathsf{a}_1^{j}, \mathsf{f}_1^{j-1}, \mathsf{e},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

IBM MODEL 1: UNIFORM PRIOR



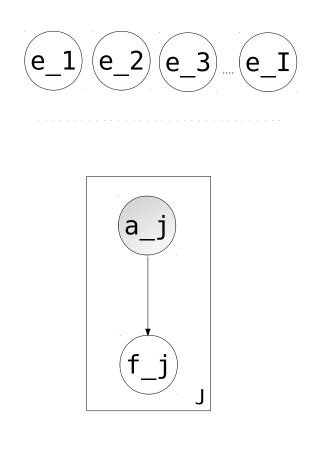
$$\Pr(\mathrm{f}, \mathrm{a} | \mathrm{e}, \theta) \approx \prod_{j=1}^{J} \Pr(f_j, a_j | \mathrm{e}, \theta)$$

IBM MODEL 1: UNIFORM PRIOR



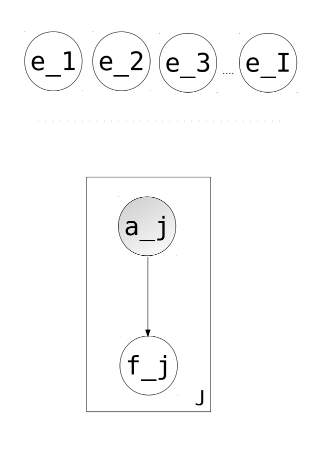
$$\begin{aligned} \Pr(\mathrm{f}, \mathrm{a} | \mathrm{e}, \theta) &\approx \prod_{j=1}^{J} \Pr(f_j, a_j | \mathrm{e}, \theta) \\ &= \prod_{j=1}^{J} \Pr(a_j | \mathrm{e}) \Pr(f_j | \mathrm{e}, a_j, \theta) \end{aligned}$$

IBM MODEL 1: UNIFORM PRIOR

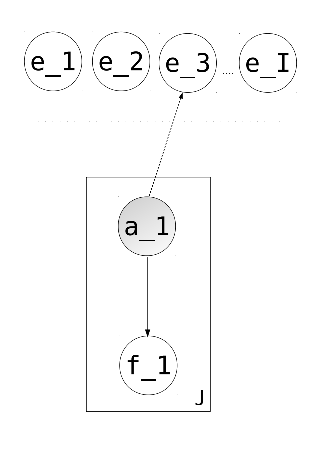


$$\begin{aligned}
\Pr(\mathrm{f}, \mathrm{a} | \mathrm{e}, \theta) &\approx \prod_{j=1}^{J} \Pr(f_j, a_j | \mathrm{e}, \theta) \\
&= \prod_{j=1}^{J} \Pr(a_j | \mathrm{e}) \Pr(f_j | \mathrm{e}, a_j, \theta) \\
&\approx \prod_{j=1}^{J} \epsilon \Pr(f_j | e_{a_j}, \theta)
\end{aligned}$$

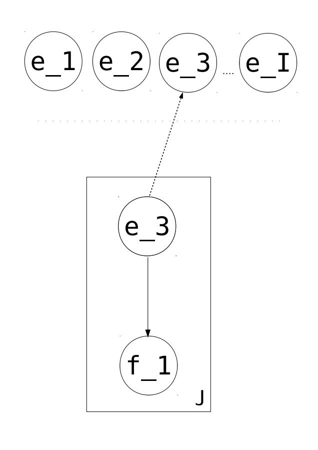
IBM MODEL 1: UNIFORM PRIOR



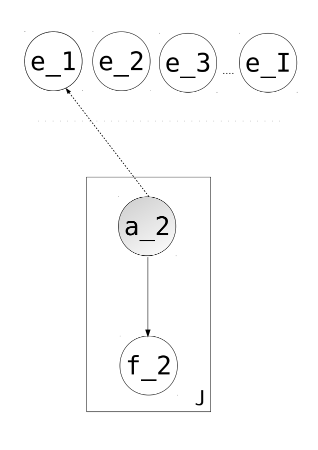
$$
\begin{aligned}
\Pr(\mathrm{f}, \mathrm{a} | \mathrm{e}, \theta) &\approx \prod_{j=1}^{J} \Pr(f_j, a_j | \mathrm{e}, \theta) \\
&= \prod_{j=1}^{J} \Pr(a_j | \mathrm{e}) \Pr(f_j | \mathrm{e}, a_j, \theta) \\
&\approx \prod_{j=1}^{J} \epsilon \Pr(f_j | e_{a_j}, \theta) \\
&\propto \prod_{j=1}^{J} t(f_j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



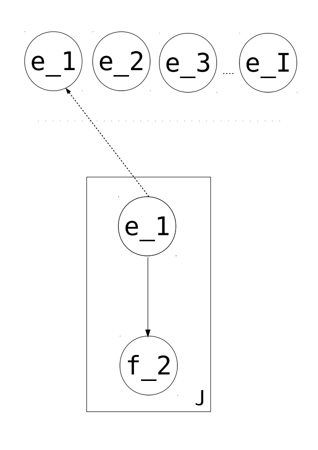
$$
\begin{aligned}
\Pr(\mathrm{f}, \mathrm{a} | \mathrm{e}, \theta) &\approx \prod_{j=1}^{J} \Pr(f_j, a_j | e_{a_j}, \theta) \\
&= \Pr(f_1, a_1 = 3 | e_3, \theta) \dots
\end{aligned}
$$



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



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



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

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

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

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, $\Pr(a_j = i|\mathrm{f}, \mathrm{e}, \theta)$ can be computed as follows

$$\begin{align}
\Pr(\mathrm{a}|\mathrm{f}, \mathrm{e}, \theta) &= \frac{\Pr(\mathrm{f}, \mathrm{a}|\mathrm{e}, \theta)}{\sum_k \Pr(\mathrm{f}, \mathrm{a}' = k|\mathrm{e}, \theta)} \tag{1}\\
&= \frac{\Pr(a_j = i|\mathrm{e}, \theta)\Pr(f_j|a_j = i, \mathrm{e}, \theta)}{\sum_{k=1}^{I} \Pr(a_j = k|\mathrm{e}, \theta)\Pr(f_j|a_j = k, \mathrm{e}, \theta)} \tag{2}\\
&= \frac{\epsilon t(f_j|e_i)}{\sum_{k=1}^{I} \epsilon t(f_j|e_k)} \tag{3}\\
&= \frac{t(f_j|e_i)}{\sum_{k=1}^{I} t(f_j|e_k)}. \tag{4}
\end{align}$$

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:

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

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

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

assignment: due 31st october, 2018 (23.59)

Improve the alignments of Model 1 as measured by AER.

ASSIGNMENT: DUE 31ST OCTOBER, 2018 (23.59)

Improve the alignments of Model 1 as measured by AER.

If this is your first attempt:

Achieve AER below 0.30 on 10k sentences or less.

assignment: due 31st october, 2018 (23.59)

Improve the alignments of Model 1 as measured by AER.

If this is your first attempt:

Achieve AER below 0.30 on 10k sentences or less.

If this is your second attempt:

Achieve AER below 0.20 on 10k sentences.

assignment: due 31st october, 2018 (23.59)

Improve the alignments of Model 1 as measured by AER.

If this is your first attempt:

Achieve AER below 0.30 on 10k sentences or less.

If this is your second attempt:

Achieve AER below 0.20 on 10k sentences.

Identify at least one problem based on error analysis.

ASSIGNMENT: DUE 31ST OCTOBER, 2018 (23.59)

Improve the alignments of Model 1 as measured by AER.

If this is your first attempt:
Achieve AER below 0.30 on 10k sentences or less.

If this is your second attempt:
Achieve AER below 0.20 on 10k sentences.

Identify at least one problem based on error analysis.

Improve the model by at least 10 percent (relative AER).

ASSIGNMENT: DUE 31ST OCTOBER, 2018 (23.59)

Improve the alignments of Model 1 as measured by AER.

If this is your first attempt:
  Achieve AER below 0.30 on 10k sentences or less.

If this is your second attempt:
  Achieve AER below 0.20 on 10k sentences.

Identify at least one problem based on error analysis.

Improve the model by at least 10 percent (relative AER).

Include code and a one page report (in English or Russian).

Suggestions:

- More complex prior (e.g. Model 2, HMM etc.)
- Better regularization (parameter tying, priors over parameters, smoothing etc.)
- Adding constraints (priors?) from a dictionary, character-level model, etc.
- Using linguistic annotations (see assignment data)
- Using a pivot language (see additional data provided)