Learning Translation Models: Word Alignment
However, the sky remained clear under the strong north wind.
However, the sky remained clear under the strong north wind.
\[ p(\text{heads}) \quad \text{and} \quad 1 - p(\text{heads}) \]
\[ p(\text{heads}) ? \]

\[ p(\text{data}) = p(\text{heads})^7 \times [1 - p(\text{heads})]^3 \]
can be derived analytically using Lagrange multipliers
Optimization
Maximum Likelihood Estimation

The *maximum likelihood estimate* of a set of parameters is the one that solves:

$$\arg\max_{\text{parameters}} p(\text{data}|\text{parameters})$$
The *maximum likelihood estimate* of a set of parameters is the one that solves:

$$\text{argmax}_{\text{parameters}} p(\text{data}|\text{parameters})$$

maximum likelihood
Language Models Again

Maximum likelihood estimate of a probability in an n-gram language model:

\[ p(word_n|word_1...word_{n-1}) = \frac{\text{count}(word_1...word_n)}{\text{count}(word_1...word_{n-1})} \]

What happens if we have not seen an n-gram?
Maximum Likelihood Estimation

- If an event is not observed, its maximum likelihood estimate will be zero. This is an example of *overfitting*, and is often not what we want.

- In the case of language models, we do not want any non-zero probabilities!

- Can be solved by various techniques to move probability mass onto the unobserved events.
Maximum Likelihood

his associates are not strong.
the clients and the associates are enemies.
the modern groups sell strong pharmaceuticals.
his associates are not strong.
the clients and the associates are enemies.
the modern groups sell strong pharmaceuticals.
Maximum Likelihood

his associates are not strong.
the clients and the associates are enemies.
the modern groups sell strong pharmaceuticals.
Maximum Likelihood

\[ p(w) \quad p(w | \text{strong}) \]

\[ W: \text{and, are, associates, clients, enemies, groups, his, not, pharmaceuticals, sell, strong, the} \]
Maximum Likelihood

\[ \lambda \cdot p(w) + (1 - \lambda) \cdot p(w|\text{strong}) \]
Backed-off Maximum Likelihood

\[ \lambda \cdot p(w) + (1 - \lambda) \cdot p(w|\text{strong}) \]

\( w \)
Smoothing & Backoff

Intuitively, we can think of these techniques as optimizing likelihood under a definition that accounts for unseen data.

Further reading:

An empirical study of smoothing techniques for language modeling.

Translation Models

Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。

However, the sky remained clear under the strong north wind.
Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。

However, the sky remained clear under the strong north wind.

\[ p(English|Chinese)? \]
Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。
IBM Model 1

Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。
Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。
Although north wind howls, but sky still very clear.

$p(\text{English length} | \text{Chinese length})$
IBM Model 1

Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。ε

\[ p(\text{English length}|\text{Chinese length}) \]
Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。
Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。
Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。
IBM Model 1

Although north wind howls, but sky still very clear.

However
Although north wind howls, but sky still very clear.

However

$p(\text{English word} | \text{Chinese word})$
IBM Model 1

Although north wind howls, but sky still very clear.

However
Although north wind howls, but sky still very clear.

However
Although north wind howls, but sky still very clear.

However,
Although north wind howls, but sky still very clear.

However,
IBM Model 1

Although north wind howls, but sky still very clear.

However, the
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.
IBM Model 1
IBM Model 1

- Word translation probabilities.
IBM Model 1

- Word translation probabilities.
- No real ordering model.
- This is left to the LM.
IBM Model 1

- Word translation probabilities.
- No real ordering model.
- This is left to the LM.
IBM Model 1

- Word translation probabilities.
- No real ordering model.
- This is left to the LM.
$p(\text{despite} \mid \text{虽然})$

$p(\text{however} \mid \text{虽然})$

$p(\text{although} \mid \text{虽然})$

$p(\text{northern} \mid \text{北})$

$p(\text{north} \mid \text{北})$
IBM Model 1

\[ p(\text{despite}|\text{虽然}) \]

\[ p(\text{however}|\text{虽然}) \]

\[ p(\text{although}|\text{虽然}) \]

\[ p(\text{northern}|\text{北}) \]

\[ p(\text{north}|\text{北}) \]
IBM Model 1

\[ p(\textit{despite} | \text{虽然}) \quad ??? \]

\[ p(\textit{however} | \text{虽然}) \quad ??? \]

\[ p(\textit{although} | \text{虽然}) \quad ??? \]

\[ p(\textit{northern} | \text{北}) \quad ??? \]

\[ p(\textit{north} | \text{北}) \quad ??? \]

Side note: \( p(X | Y) \) is given here (but not always)
Translation Models

Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。

However, the sky remained clear under the strong north wind.

\[ p(\text{however} | \text{虽然}) = \frac{\text{# of times 虽然 aligns to However}}{\text{# of times 虽然 occurs}} \]
Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

\[
p(\text{however}\mid \text{虽然}) = \frac{\text{# of times 虽然 aligns to However}}{\text{# of times 虽然 occurs}}
\]
Likelihood Estimation for Model 1
Likelihood Estimation for Model 1

Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。
Likelihood Estimation for Model 1

Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。
Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。
Likelihood Estimation for Model 1

Although north wind howls, but sky still very clear.

雖然北風呼嘯，但天空依然十分清澈。
Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。

$p(\text{English length}|\text{Chinese length})$
Likelihood Estimation for Model 1

Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。
Likelihood Estimation for Model 1

Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

$p(\text{Chinese word position})$
Likelihood Estimation for Model 1

Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。
Likelihood Estimation for Model 1

Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。
Likelihood Estimation for Model 1

Although north wind howls, but sky still very clear.

However

\[
p(\text{English word}|\text{Chinese word})
\]
Although north wind howls, but sky still very clear.

However
Likelihood Estimation for Model 1

Although north wind howls, but sky still very clear.

However
Although north wind howls, but sky still very clear.

However,
Likelihood Estimation for Model 1

Although north wind howls, but sky still very clear.

However,
Likelihood Estimation for Model 1

Although north wind howls, but sky still very clear.

However, the
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

$p(\text{English length} | \text{Chinese length})$ observed
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

$p(\text{Chinese word position})$ uniform, no need to estimate.
Likelihood Estimation for Model 1

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

\[ p(\text{English word} | \text{Chinese word}) \quad \text{unobserved!} \]
Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。

the missing alignment is a \textit{latent variable}

However, the sky remained clear under the strong north wind.

\[ p(\text{English word}|\text{Chinese word}) \quad \text{unobserved!} \]
Likelihood Estimation for Model 1

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

\[ p(\text{English word}|\text{Chinese word}) \text{ unobserved!} \]
Likelihood Estimation for Model 1

Although north wind howls, but sky still very clear.

Parameters and alignments are both unknown.

If we knew the alignments, we could calculate the values of the parameters.

However, the sky remained clear under the strong north wind.

\[ p(\text{English word} | \text{Chinese word}) \text{ unobserved!} \]
Although north wind howls, but sky still very clear. Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。

Parameters and alignments are both unknown.

If we knew the alignments, we could calculate the values of the parameters.

If we knew the parameters, we could calculate the likelihood of the data.

However, the sky remained clear under the strong north wind.

$p(\text{English word}|\text{Chinese word})$ unobserved!
Parameters and alignments are both unknown.

If we knew the alignments, we could calculate the values of the parameters.

If we knew the parameters, we could calculate the likelihood of the data.

Although north wind howls, but sky still very clear.

$p(\text{English word}|\text{Chinese word})$ unobserved!
The Plan: Bootstrapping

- Arbitrarily select a set of parameters (say, uniform).
- Calculate *expected counts* of the unseen events.
- Choose new parameters to maximize likelihood, using expected counts as proxy for observed counts.
- Iterate.

- Guarantee: likelihood will be monotonically nondecreasing.
The Plan: Bootstrapping

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.
The Plan: Bootstrapping

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

if we had observed the alignment, this line would either be here (count 1) or it wouldn’t (count 0).
The Plan: Bootstrapping

Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。ε

if we had observed the alignment, this line would either be here (count 1) or it wouldn’t (count 0).

since we didn’t observe the alignment, we calculate the probability that it’s there.

However, the sky remained clear under the strong north wind.
Marginalize: sum all alignments containing the link

$p(\text{虽然北风呼啸，但天空依然十分清澈。}) +$

However, the sky remained clear under the strong north wind.

$p(\text{虽然北风呼啸，但天空依然十分清澈。}) +$

However, the sky remained clear under the strong north wind.

$p(\text{虽然北风呼啸，但天空依然十分清澈。})$

However, the sky remained clear under the strong north wind.
Divide by sum of all possible alignments

\[
p(\text{However , the sky remained clear under the strong north wind .}) + \\
\]

虽然 北 风 呼啸 , 但 天空 依然 十分 清澈 。

\[
p(\text{However , the sky remained clear under the strong north wind .}) + \\
\]

虽然 北 风 呼啸 , 但 天空 依然 十分 清澈 。

\[
p(\text{However , the sky remained clear under the strong north wind .}) \\
\]
Divide by sum of all possible alignments

\[ p(\text{However, the sky remained clear under the strong north wind.}) + \]

Is this hard? How many alignments are there?
Expectation Maximization

probability of an alignment.

\[ p(F, A|E) = p(I|J) \prod_{a_i} p(a_i = j)p(f_i|e_j) \]
Expectation Maximization

probability of an alignment.

\[ p(F, A|E) = p(I|J) \prod_{a_i} p(a_i = j)p(f_i|e_j) \]
Expectation Maximization

probability of an alignment.

$$p(F, A | E) = p(I | J) \prod_{a_i} p(a_i = j) p(f_i | e_j)$$

factors across words.

observed uniform
Expectation Maximization

marginal probability of alignments containing link

$$\sum_{a \in A: 北 \leftrightarrow north} p(north | 北) \cdot p(rest \ of \ a)$$
Expectation Maximization

marginal probability of alignments containing link

\[ p(\text{north} | \text{北}) \sum_{a \in A: \text{北} \leftrightarrow \text{north}} p(\text{rest of } a) \]
Expectation Maximization

marginal probability of alignments containing link

\[
p(\text{north} | \text{北}) \sum_{a \in A: \text{北} \leftrightarrow \text{north}} p(\text{rest of } a)
\]

\[
\sum_{c \in \text{Chinese words}} p(\text{north} | c) \sum_{a \in A: \leftrightarrow \text{north}} p(\text{rest of } a)
\]

marginal probability of all alignments
Expectation Maximization

marginal probability of alignments containing link

\[
p(\text{north} | \text{北}) \sum_{a \in A: \text{北} \leftrightarrow \text{north}} p(\text{rest of } a)
\]

\[
\sum_{c \in \text{Chinese words}} p(\text{north} | c) \sum_{a \in A: c \leftrightarrow \text{north}} p(\text{rest of } a)
\]

marginal probability of all alignments
Expectation Maximization

marginal probability of alignments containing link

\[ p(\text{north} | \text{北}) \sum_{a \in A: \text{北} \leftrightarrow \text{north}} p(\text{rest of } a) \]

\[ \sum_{c \in \text{Chinese words}} p(\text{north} | c) \sum_{a \in A: c \leftrightarrow \text{north}} p(\text{rest of } a) \]

identical!

marginal probability of all alignments
Expectation Maximization

\[
p(north|北) \quad \frac{p(north|北)}{\sum_{c \in \text{Chinese words}} p(north|c)}
\]
Expectation Maximization

marginal probability (expected count) of an alignment containing the link

\[
p(north | 北) \over \sum_{c \in Chinese \ words} p(north | c)
\]
Expectation Maximization

marginal probability (expected count) of an alignment containing the link

\[
p(\text{north} | \ 北) \over \sum_{c \in \text{Chinese words}} p(\text{north} | c)
\]

For each sentence, use this quantity instead of 0 or 1
Translation Models

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

\[ p(\text{however} \mid \text{虽然}) = \frac{\text{# of times 虽然 aligns to However}}{\text{# of times 虽然 occurs}} \]
Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

\[ p(\text{however}|\text{虽然}) = \frac{\text{Expected} \ # \ of \ times \ 虽然 \ \text{aligns to} \ \text{However}}{\# \ of \ times \ 虽然 \ occurs} \]
Expectation Maximization

Why does this even work?

\[
p(north | 北) \over \sum_{c \in \text{Chinese words}} p(north | c)
\]
Expectation Maximization

Observation 1: We are still solving a maximum likelihood estimation problem.
Observation 1: We are still solving a maximum likelihood estimation problem.

\[ p(\text{Chinese}|\text{English}) = \sum_{\text{alignments}} p(\text{Chinese, alignment}|\text{English}) \]
Expectation Maximization

Observation 1: We are still solving a maximum likelihood estimation problem.

\[ p(\text{Chinese} | \text{English}) = \sum_{\text{alignments}} p(\text{Chinese}, \text{alignment} | \text{English}) \]

MLE: choose parameters that maximize this expression.
Observation 1: We are still solving a maximum likelihood estimation problem.

\[
p(\text{Chinese}|\text{English}) = \sum_{\text{alignments}} p(\text{Chinese}, \text{alignment}|\text{English})
\]

MLE: choose parameters that maximize this expression.

Minor problem: there is no analytic solution.
**Expectation Maximization**

Observation 1: We are still solving a maximum likelihood estimation problem.

\[ p(\text{Chinese}|\text{English}) = \sum_{\text{alignments}} p(\text{Chinese}, \text{alignment}|\text{English}) \]

**MLE:** choose parameters that maximize this expression.

Minor problem: there is no analytic solution.

Remember: likelihood is monotonically non-decreasing!
... and, likelihood is *convex* for this model:
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

What are some things this model doesn’t account for?
Summary

• We can formulate learning as an optimization problem: choose parameters that optimize some function, such as likelihood.

• Supervised: maximum likelihood.

  • Beware of overfitting.

• Unsupervised: expectation maximization.

• Many, many, many other algorithms.

• Thursday: learning better translation models.

  • Also: first “Language in 10 minutes”
Alignment: Challenge Problem 1

Aligning words is a key task in data-driven machine translation. We start with a large parallel corpus where the translation of each sentence is known. So the input consists of sentence pairs like this one:

\[ \text{le droit de permis passe donc de } $25 \text{ à } $500. \quad \text{— we see the licence fee going up from } $25 \text{ to } $500. \]

It is relatively easy to obtain data in this form, though not completely trivial. This particular example is from the Canadian Hansards, proceedings of government meetings that are required to be published in both English and French. To align the sentences we can exploit document boundaries and structural cues arising from the fact that the documents are translations of each other, such as the fact that translated sentences will be in the same order, of similar length, and so forth. The problem is that we don't know the word-to-word correspondences in each sentence. That's where you come in: your challenge is to write a program that aligns the words automatically. For the sentence above, you might output the following correspondences:

\[ \text{le — the, droit — fee, permis — license, passe — going, passe — up, donc — from, $ — $, 25 — 25, à — to, $ — $, 50 — 50} \]

Some words might be unaligned (e.g. we and see), while some words might be aligned to multiple words in the corresponding sentence (e.g. passe is aligned to going up). Sometimes words aren't exact translations of each other. That's ok: even experienced bilinguals will sometimes disagree on the best alignment of a sentence. But while word alignment doesn't capture every nuance, it is very useful for learning translation models or bilingual dictionaries.

Getting Started

Run this command:

```bash
git clone https://github.com/alopez/dreamt.git
```