

Natural Language Processing (CSE 447/547M): Bitext and Machine Translation

Noah Smith

© 2019

University of Washington
`nasmith@cs.washington.edu`

February 27, 2018

Evaluation

Intuition: good translations are **fluent** in the target language and **faithful** to the original meaning.

Bleu score (Papineni et al., 2002):

- ▶ Compare to a human-generated reference translation
- ▶ Or, better: multiple references
- ▶ Weighted average of n-gram precision (across different n)

There are some alternatives; most papers that use them report Bleu, too.

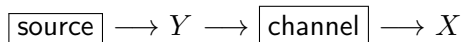
Warren Weaver to Norbert Wiener, 1947

One naturally wonders if the problem of translation could be conceivably treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'

Noisy Channel Models

Review

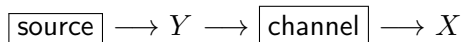
A pattern for modeling a pair of random variables, X and Y :



Noisy Channel Models

Review

A pattern for modeling a pair of random variables, X and Y :



- Y is the plaintext, the true message, the missing information, the output

Noisy Channel Models

Review

A pattern for modeling a pair of random variables, X and Y :

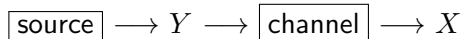
$$\boxed{\text{source}} \longrightarrow Y \longrightarrow \boxed{\text{channel}} \longrightarrow X$$

- ▶ Y is the plaintext, the true message, the missing information, the output
- ▶ X is the ciphertext, the garbled message, the observable evidence, the input

Noisy Channel Models

Review

A pattern for modeling a pair of random variables, X and Y :



- ▶ Y is the plaintext, the true message, the missing information, the output
- ▶ X is the ciphertext, the garbled message, the observable evidence, the input
- ▶ Decoding: select y given $X = x$.

$$\begin{aligned} y^* &= \operatorname{argmax}_y p(y \mid x) \\ &= \operatorname{argmax}_y \frac{p(x \mid y) \cdot p(y)}{p(x)} \\ &= \operatorname{argmax}_y \underbrace{p(x \mid y)}_{\text{channel model}} \cdot \underbrace{p(y)}_{\text{source model}} \end{aligned}$$

Bitext/Parallel Text

Let f and e be two sequences in \mathcal{V}^\dagger (French) and $\bar{\mathcal{V}}^\dagger$ (English), respectively.

In a noisy channel machine translation system, we could use this together with source/language model $p(e)$ to “decode” f into an English translation.

Where does the data to estimate this come from?

IBM Model 1

(Brown et al., 1993)

Let ℓ and m be the (known) lengths of \mathbf{e} and \mathbf{f} .

Latent variable $\mathbf{a} = \langle a_1, \dots, a_m \rangle$, each a_i ranging over $\{0, \dots, \ell\}$ (positions in \mathbf{e}).

- ▶ $a_4 = 3$ means that f_4 is “aligned” to e_3 .
- ▶ $a_6 = 0$ means that f_6 is “aligned” to a special NULL symbol, e_0 .

$$\begin{aligned} p(\mathbf{f} \mid \mathbf{e}, m) &= \sum_{a_1=0}^{\ell} \sum_{a_2=0}^{\ell} \cdots \sum_{a_m=0}^{\ell} p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, m) \\ &= \sum_{\mathbf{a} \in \{0, \dots, \ell\}^m} p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, m) \\ p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, m) &= \prod_{i=1}^m p(a_i \mid i, \ell, m) \cdot p(f_i \mid e_{a_i}) \\ &= \prod_{i=1}^m \frac{1}{\ell + 1} \cdot \theta_{f_i | e_{a_i}} = \left(\frac{1}{\ell + 1} \right)^m \prod_{i=1}^m \theta_{f_i | e_{a_i}} \end{aligned}$$

Example: f is German

Mr President , Noah's ark was filled not with production factors , but with living creatures .



Noahs Arche war nicht voller Produktionsfaktoren , sondern Geschöpfe .

$$\mathbf{a} = \langle 4, \dots \rangle$$

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, m) = \frac{1}{17 + 1} \cdot \theta_{\text{Noahs} \mid \text{Noah's}}$$

Example: f is German

Mr President , Noah's ark was filled not with production factors , but with living creatures .



Noahs Arche war nicht voller Produktionsfaktoren , sondern Geschöpfe .

$$\mathbf{a} = \langle 4, 5, \dots \rangle$$

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, m) = \frac{1}{17 + 1} \cdot \theta_{\text{Noahs}|\text{Noah's}} \cdot \frac{1}{17 + 1} \cdot \theta_{\text{Arche}|\text{ark}}$$

Example: f is German

Mr President , Noah's ark was filled not with production factors , but with living creatures .



Noahs Arche war nicht voller Produktionsfaktoren , sondern Geschöpfe .

$$\mathbf{a} = \langle 4, 5, 6, \dots \rangle$$

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, m) = \frac{1}{17+1} \cdot \theta_{\text{Noahs}|\text{Noah's}} \cdot \frac{1}{17+1} \cdot \theta_{\text{Arche}|\text{ark}} \\ \cdot \frac{1}{17+1} \cdot \theta_{\text{war}|\text{was}}$$

Example: f is German

Mr President , Noah's ark was filled not with production factors , but with living creatures .



Noahs Arche war nicht voller Produktionsfaktoren , sondern Geschöpfe .

$$\mathbf{a} = \langle 4, 5, 6, 8, \dots \rangle$$

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, m) = \frac{1}{17+1} \cdot \theta_{\text{Noahs}|\text{Noah's}} \cdot \frac{1}{17+1} \cdot \theta_{\text{Arche}|\text{ark}} \\ \cdot \frac{1}{17+1} \cdot \theta_{\text{war}|\text{was}} \cdot \frac{1}{17+1} \cdot \theta_{\text{nicht}|\text{not}}$$

Example: f is German

Mr President , Noah's ark was filled not with production factors , but with living creatures .



Noahs Arche war nicht voller Produktionsfaktoren , sondern Geschöpfe .

$$\mathbf{a} = \langle 4, 5, 6, 8, 7, \dots \rangle$$

$$\begin{aligned} p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, m) &= \frac{1}{17+1} \cdot \theta_{\text{Noahs}|\text{Noah's}} \cdot \frac{1}{17+1} \cdot \theta_{\text{Arche}|\text{ark}} \\ &\cdot \frac{1}{17+1} \cdot \theta_{\text{war}|\text{was}} \cdot \frac{1}{17+1} \cdot \theta_{\text{nicht}|\text{not}} \\ &\cdot \frac{1}{17+1} \cdot \theta_{\text{voller}|\text{filled}} \end{aligned}$$

Example: f is German

Mr President , Noah's ark was filled not with production factors , but with living creatures .



Noahs Arche war nicht voller Produktionsfaktoren , sondern Geschöpfe .

$$\mathbf{a} = \langle 4, 5, 6, 8, 7, ?, \dots \rangle$$

$$\begin{aligned} p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, m) &= \frac{1}{17+1} \cdot \theta_{\text{Noahs}|\text{Noah's}} \cdot \frac{1}{17+1} \cdot \theta_{\text{Arche}|\text{ark}} \\ &\quad \cdot \frac{1}{17+1} \cdot \theta_{\text{war}|\text{was}} \cdot \frac{1}{17+1} \cdot \theta_{\text{nicht}|\text{not}} \\ &\quad \cdot \frac{1}{17+1} \cdot \theta_{\text{voller}|\text{filled}} \cdot \frac{1}{17+1} \cdot \theta_{\text{Produktionsfaktoren}|\text{?}} \end{aligned}$$

Example: f is German

Mr President , Noah's ark was filled not with production factors , but with living creatures .



Noahs Arche war nicht voller Produktionsfaktoren , sondern Geschöpfe .

$$\mathbf{a} = \langle 4, 5, 6, 8, 7, ?, \dots \rangle$$

$$\begin{aligned} p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, m) &= \frac{1}{17+1} \cdot \theta_{\text{Noahs}|\text{Noah's}} \cdot \frac{1}{17+1} \cdot \theta_{\text{Arche}|\text{ark}} \\ &\cdot \frac{1}{17+1} \cdot \theta_{\text{war}|\text{was}} \cdot \frac{1}{17+1} \cdot \theta_{\text{nicht}|\text{not}} \\ &\cdot \frac{1}{17+1} \cdot \theta_{\text{voller}|\text{filled}} \cdot \frac{1}{17+1} \cdot \theta_{\text{Produktionsfaktoren}|\text{?}} \end{aligned}$$

Problem: This alignment isn't possible with IBM Model 1! Each f_i is aligned to at most *one* e_{a_i} !

Example: \mathbf{f} is English

Mr President , Noah's ark was filled not with production factors , but with living creatures .



Noahs Arche war nicht voller Produktionsfaktoren , sondern Geschöpfe .

$$\mathbf{a} = \langle 0, \dots \rangle$$

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, m) = \frac{1}{10 + 1} \cdot \theta_{\text{Mr}|\text{NULL}}$$

Example: f is English

Mr President , Noah's ark was filled not with production factors , but with living creatures .



Noahs Arche war nicht voller Produktionsfaktoren , sondern Geschöpfe .

$$\mathbf{a} = \langle 0, 0, 0, \dots \rangle$$

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, m) = \frac{1}{10 + 1} \cdot \theta_{\text{Mr} \mid \text{NULL}} \cdot \frac{1}{10 + 1} \cdot \theta_{\text{President} \mid \text{NULL}} \\ \cdot \frac{1}{10 + 1} \cdot \theta_{, \mid \text{NULL}}$$

Example: f is English

Mr President , Noah's ark was filled not with production factors , but with living creatures .



Noahs Arche war nicht voller Produktionsfaktoren , sondern Geschöpfe .

$$\mathbf{a} = \langle 0, 0, 0, 1, \dots \rangle$$

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, m) = \frac{1}{10+1} \cdot \theta_{\text{Mr}|\text{NULL}} \cdot \frac{1}{10+1} \cdot \theta_{\text{President}|\text{NULL}} \\ \cdot \frac{1}{10+1} \cdot \theta_{,|\text{NULL}} \cdot \frac{1}{10+1} \cdot \theta_{\text{Noah's}|\text{Noahs}}$$

Example: f is English

Mr President , Noah's ark was filled not with production factors , but with living creatures .



Noahs Arche war nicht voller Produktionsfaktoren , sondern Geschöpfe .

$$\mathbf{a} = \langle 0, 0, 0, 1, 2, \dots \rangle$$

$$\begin{aligned} p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, m) &= \frac{1}{10+1} \cdot \theta_{\text{Mr}|\text{NULL}} \cdot \frac{1}{10+1} \cdot \theta_{\text{President}|\text{NULL}} \\ &\cdot \frac{1}{10+1} \cdot \theta_{,|\text{NULL}} \cdot \frac{1}{10+1} \cdot \theta_{\text{Noah's}|\text{Noahs}} \\ &\cdot \frac{1}{10+1} \cdot \theta_{\text{ark}|\text{Arche}} \end{aligned}$$

Example: f is English

Mr President , Noah's ark was filled not with production factors , but with living creatures .



Noahs Arche war nicht voller Produktionsfaktoren , sondern Geschöpfe .

$$\mathbf{a} = \langle 0, 0, 0, 1, 2, 3, \dots \rangle$$

$$\begin{aligned} p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, m) &= \frac{1}{10+1} \cdot \theta_{\text{Mr}|\text{NULL}} \cdot \frac{1}{10+1} \cdot \theta_{\text{President}|\text{NULL}} \\ &\cdot \frac{1}{10+1} \cdot \theta_{\text{,}|\text{NULL}} \cdot \frac{1}{10+1} \cdot \theta_{\text{Noah's}|\text{Noahs}} \\ &\cdot \frac{1}{10+1} \cdot \theta_{\text{ark}|\text{Arche}} \cdot \frac{1}{10+1} \cdot \theta_{\text{was}|\text{war}} \end{aligned}$$

Example: f is English

Mr President , Noah's ark was filled not with production factors , but with living creatures .



Noahs Arche war nicht voller Produktionsfaktoren , sondern Geschöpfe .

$$\mathbf{a} = \langle 0, 0, 0, 1, 2, 3, 5, \dots \rangle$$

$$\begin{aligned} p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, m) = & \frac{1}{10+1} \cdot \theta_{\text{Mr}|\text{NULL}} \cdot \frac{1}{10+1} \cdot \theta_{\text{President}|\text{NULL}} \\ & \cdot \frac{1}{10+1} \cdot \theta_{\text{,}|\text{NULL}} \cdot \frac{1}{10+1} \cdot \theta_{\text{Noah's}|\text{Noahs}} \\ & \cdot \frac{1}{10+1} \cdot \theta_{\text{ark}|\text{Arche}} \cdot \frac{1}{10+1} \cdot \theta_{\text{was}|\text{war}} \\ & \cdot \frac{1}{10+1} \cdot \theta_{\text{filled}|\text{voller}} \end{aligned}$$

Example: f is English

Mr President , Noah's ark was filled not with production factors , but with living creatures .



Noahs Arche war nicht voller Produktionsfaktoren , sondern Geschöpfe .

$$\mathbf{a} = \langle 0, 0, 0, 1, 2, 3, 5, 4, \dots \rangle$$

$$\begin{aligned} p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, m) = & \frac{1}{10+1} \cdot \theta_{\text{Mr}|\text{NULL}} \cdot \frac{1}{10+1} \cdot \theta_{\text{President}|\text{NULL}} \\ & \cdot \frac{1}{10+1} \cdot \theta_{,|\text{NULL}} \cdot \frac{1}{10+1} \cdot \theta_{\text{Noah's}|\text{Noahs}} \\ & \cdot \frac{1}{10+1} \cdot \theta_{\text{ark}|\text{Arche}} \cdot \frac{1}{10+1} \cdot \theta_{\text{was}|\text{war}} \\ & \cdot \frac{1}{10+1} \cdot \theta_{\text{filled}|\text{voller}} \cdot \frac{1}{10+1} \cdot \theta_{\text{not}|\text{nicht}} \end{aligned}$$

References I

- Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*, 2002.