Natural Language Processing

Anoop Sarkar
anoopsarkar.github.io/nlp-class

Simon Fraser University

October 16, 2014

Natural Language Processing

Anoop Sarkar
anoopsarkar.github.io/nlp-class

Simon Fraser University

Part 1: Generative Models for Word Alignment

Statistical Machine Translation

Noisy Channel Model

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \underbrace{\Pr(\mathbf{e})}_{\textbf{Language Model}} \cdot \underbrace{\Pr(\mathbf{f} \mid \mathbf{e})}_{\textbf{Alignment Model}}$$

Alignment Task

$$\begin{array}{l} \mathbf{e} \searrow \\ \quad \text{Program} \longrightarrow \Pr(\mathbf{e} \mid \mathbf{f}) \\ \mathbf{f} \nearrow \end{array}$$

Training Data

- **Alignment Model**: learn a mapping between $\mathbf{f}$and $\mathbf{e}$.
  Training data: lots of translation pairs between $\mathbf{f}$and $\mathbf{e}$.

Statistical Machine Translation

The IBM Models

- The first statistical machine translation models were developed at IBM Research (Yorktown Heights, NY) in the 1980s
- The models were published in 1993:
  Brown et. al. The Mathematics of Statistical Machine Translation.
  Computational Linguistics. 1993.
  `http://aclweb.org/anthology/J/J93/J93-2003.pdf`
- These models are the basic SMT models, called:
  IBM Model 1, IBM Model 2, IBM Model 3, IBM Model 4, IBM Model 5
  as they were called in the 1993 paper.
- We use $\mathbf{e}$and $\mathbf{f}$in the equations in honor of their system which translated from French to English.
  Trained on the Canadian Hansards (Parliament Proceedings)

Statistical Machine Translation

Generative Model of Word Alignment

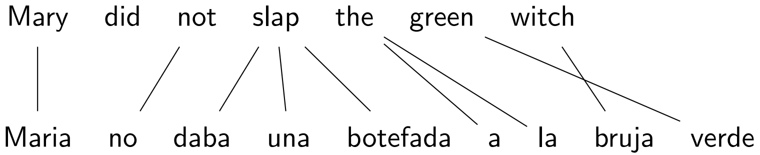
Word Alignments: IBM Model 3

Word Alignments: IBM Model 1

Generative Model of Word Alignment

- English **e**: Mary did not slap the green witch
- "French" **f**: Maria no daba una botefada a la bruja verde
- Alignment **a**: $\{1, 3, 4, 4, 4, 5, 5, 7, 6\}$
  e.g. $(f_8, e_{a_8}) = (f_8, e_7) = $ (bruja, witch)

Visualizing alignment **a**



Generative Model of Word Alignment

Data Set

- Data set $\mathcal{D}$ of N sentences:
$$\mathcal{D} = \{(\mathbf{f}^{(1)}, \mathbf{e}^{(1)}), \ldots, (\mathbf{f}^{(N)}, \mathbf{e}^{(N)})\}$$
- French $\mathbf{f}$: $(f_1, f_2, \ldots, f_I)$
- English $\mathbf{e}$: $(e_1, e_2, \ldots, e_J)$
- Alignment $\mathbf{a}$: $(a_1, a_2, \ldots, a_I)$

Generative Model of Word Alignment

Find the best alignment for each translation pair

$$\mathbf{a}^* = \arg\max_{\mathbf{a}} \Pr(\mathbf{a} \mid \mathbf{f}, \mathbf{e})$$

Alignment probability

$$\begin{aligned}
\Pr(\mathbf{a} \mid \mathbf{f}, \mathbf{e}) &= \frac{\Pr(\mathbf{a}, \mathbf{f}, \mathbf{e})}{\Pr(\mathbf{f}, \mathbf{e})} \\
&= \frac{\Pr(\mathbf{e}) \Pr(\mathbf{a}, \mathbf{f} \mid \mathbf{e})}{\Pr(\mathbf{e}) \Pr(\mathbf{f} \mid \mathbf{e})} \\
&= \frac{\Pr(\mathbf{a}, \mathbf{f} \mid \mathbf{e})}{\Pr(\mathbf{f} \mid \mathbf{e})} \\
&= \frac{\Pr(\mathbf{a}, \mathbf{f} \mid \mathbf{e})}{\sum_{\mathbf{a}} \Pr(\mathbf{a}, \mathbf{f} \mid \mathbf{e})}
\end{aligned}$$

Statistical Machine Translation

Generative Model of Word Alignment

Word Alignments: IBM Model 3

Word Alignments: IBM Model 1

Word Alignments: IBM Model 3

Generative "story" for $P(\mathbf{a}, \mathbf{f} \mid \mathbf{e})$

Mary did not slap the green witch (fertility)

Mary not slap slap slap the the green witch (translate)

Maria no daba una botefada a la verde bruja (reorder)

Maria no daba una botefada a la bruja verde

Word Alignments: IBM Model 3

Fertility parameter

$$n(\phi_j \mid e_j) : n(3 \mid slap); n(0 \mid did)$$

Translation parameter

$$t(f_i \mid e_{a_i}) : t(bruja \mid witch)$$

Distortion parameter

$$d(f_{pos} = i \mid e_{pos} = j, I, J) : d(8 \mid 7, 8, 6)$$

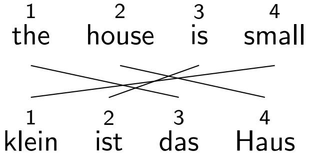
Word Alignments: IBM Model 3

Generative model for $P(\mathbf{a}, \mathbf{f} \mid \mathbf{e})$

$$\begin{aligned} P(\mathbf{a}, \mathbf{f} \mid \mathbf{e}) &= \prod_{j=1}^{J} n(\phi_i \mid e_i) \\ &\times \prod_{i=1}^{I} t(f_i \mid e_{a_j}) \\ &\times \prod_{i=1}^{I} d(i \mid j, I, J) \end{aligned}$$

Word Alignments: IBM Model 3

Sentence pair with alignment $\mathbf{a} = (4, 3, 1, 2)$

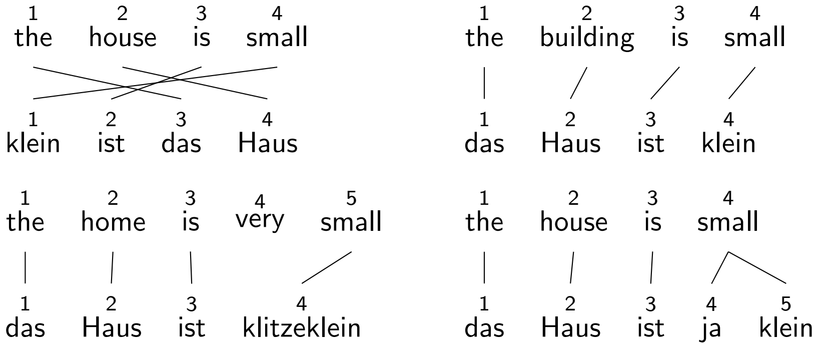


If we know the parameter values we can easily compute the probability of this aligned sentence pair.

$\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) =$

$n(1 \mid \text{the})$	$\times$	$t(\text{das} \mid \text{the})$	$\times$	$d(3 \mid 1, 4, 4) \times$
$n(1 \mid \text{house})$	$\times$	$t(\text{Haus} \mid \text{house})$	$\times$	$d(4 \mid 2, 4, 4) \times$
$n(1 \mid \text{is})$	$\times$	$t(\text{ist} \mid \text{is})$	$\times$	$d(2 \mid 3, 4, 4) \times$
$n(1 \mid \text{small})$	$\times$	$t(\text{klein} \mid \text{small})$	$\times$	$d(1 \mid 4, 4, 4)$

Word Alignments: IBM Model 3



Parameter Estimation

- What is $n(1 \mid \text{very}) = ?$ and $n(0 \mid \text{very}) = ?$
- What is $t(\text{Haus} \mid \text{house}) = ?$ and $t(\text{klein} \mid \text{small}) = ?$
- What is $d(1 \mid 4, 4, 4) = ?$ and $d(1 \mid 1, 4, 4) = ?$

Word Alignments: IBM Model 3

1	2	3	4
the	house	is	small
klein	ist	das	Haus
1	2	3	4

1	2	3	4
the	building	is	small
das	Haus	ist	klein
1	2	3	4

1	2	3	4	5
the	home	is	very	small
das	Haus	ist	klitzeklein	
1	2	3	4	

1	2	3	4	
the	house	is	small	
das	Haus	ist	ja	klein
1	2	3	4	5

Parameter Estimation: Sum over all alignments

$$\sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \sum_{\mathbf{a}} \prod_{i=1}^{I} n(\phi_{a_i} \mid e_{a_i}) \times t(f_i \mid e_{a_i}) \times d(i \mid a_i, \mathbf{f}_{\mathrm{len}}, \mathbf{e}_{\mathrm{len}})$$

Word Alignments: IBM Model 3

Summary

- If we know the parameter values we can easily compute the probability $\Pr(\mathbf{a} \mid \mathbf{f}, \mathbf{e})$ given an aligned sentence pair
- If we are given a corpus of sentence pairs with alignments we can easily learn the parameter values by using relative frequencies.
- If we do not know the alignments then perhaps we can produce all possible alignments each with a certain probability?

IBM Model 3 is too hard: Let us try learning only $t(f_i \mid e_{a_i})$

$$\sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \sum_{\mathbf{a}} \prod_{i=1}^{I} n(\phi_{a_i} \mid e_{a_i}) \times t(f_i \mid e_{a_i}) \times d(i \mid a_i, \mathbf{f}_{\mathrm{len}}, \mathbf{e}_{\mathrm{len}})$$

Statistical Machine Translation

Generative Model of Word Alignment

Word Alignments: IBM Model 3

Word Alignments: IBM Model 1

Word Alignments: IBM Model 1

Alignment probability

$$\Pr(\mathbf{a} \mid \mathbf{f}, \mathbf{e}) = \frac{\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})}{\sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})}$$

Example alignment

1	2	3	4
the	house	is	small
|	|	|	|
1	2	3	4
das	Haus	ist	klein

$$\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \prod_{i=1}^{I} t(f_i \mid e_{a_i})$$

$$\begin{aligned}\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \\ & t(\text{das} \mid \text{the}) \times \\ & t(\text{Haus} \mid \text{house}) \times \\ & t(\text{ist} \mid \text{is}) \times \\ & t(\text{klein} \mid \text{small})\end{aligned}$$

Word Alignments: IBM Model 1

Generative "story" for Model 1

the	house	is	small	(translate)
↓	↓	↓	↓	
das	Haus	ist	klein	

$$\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \prod_{i=1}^{I} t(f_i \mid e_{a_i})$$

Finding the best word alignment: IBM Model 1

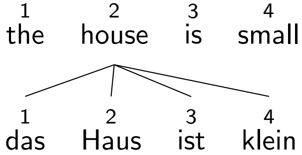
Compute the arg max word alignment

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} \Pr(\mathbf{a} \mid \mathbf{e}, \mathbf{f})$$

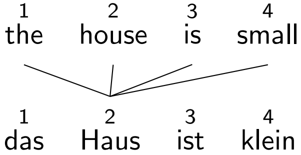
- For each f_i in $(f_1, \ldots, f_I)$ build $\mathbf{a} = (\hat{a}_1, \ldots, \hat{a}_I)$

$$\hat{a}_i = \arg\max_{a_i} t(f_i \mid e_{a_i})$$

Many to one alignment ✓



One to many alignment ✗



Word Alignments: IBM Model 1

Alignment probability

$$
\begin{aligned}
\Pr(\mathbf{a} \mid \mathbf{f}, \mathbf{e}) &= \frac{\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})}{\Pr(\mathbf{f} \mid \mathbf{e})} \\
&= \frac{\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})}{\sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})} \\
&= \frac{\prod_{i=1}^{I} t(f_i \mid e_{a_i})}{\sum_{\mathbf{a}} \prod_{i=1}^{I} t(f_i \mid e_{a_i})}
\end{aligned}
$$

Computing the denominator

- The denominator above is summing over J^I alignments
- An interlude on how compute the denominator faster ...

Word Alignments: IBM Model 1

Sum over all alignments

$$\sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \sum_{a_1=1}^{J} \sum_{a_2=1}^{J} \cdots \sum_{a_I=1}^{J} \prod_{i=1}^{I} t(f_i \mid e_{a_i})$$

Assume (f_1, f_2, f_3) and (e_1, e_2)

$$\sum_{a_1=1}^{2} \sum_{a_2=1}^{2} \sum_{a_3=1}^{2} t(f_1 \mid e_{a_1}) \times t(f_2 \mid e_{a_2}) \times t(f_3 \mid e_{a_3})$$

Word Alignments: IBM Model 1

Assume (f_1, f_2, f_3) and (e_1, e_2): $I = 3$ and $J = 2$

$$\sum_{a_1=1}^{2} \sum_{a_2=1}^{2} \sum_{a_3=1}^{2} t(f_1 \mid e_{a_1}) \times t(f_2 \mid e_{a_2}) \times t(f_3 \mid e_{a_3})$$

$J^I = 2^3$ terms to be added:

$t(f_1 \mid e_1)$	$\times$	$t(f_2 \mid e_1)$	$\times$	$t(f_3 \mid e_1)$	$+$
$t(f_1 \mid e_1)$	$\times$	$t(f_2 \mid e_1)$	$\times$	$t(f_3 \mid e_2)$	$+$
$t(f_1 \mid e_1)$	$\times$	$t(f_2 \mid e_2)$	$\times$	$t(f_3 \mid e_1)$	$+$
$t(f_1 \mid e_1)$	$\times$	$t(f_2 \mid e_2)$	$\times$	$t(f_3 \mid e_2)$	$+$
$t(f_1 \mid e_2)$	$\times$	$t(f_2 \mid e_1)$	$\times$	$t(f_3 \mid e_1)$	$+$
$t(f_1 \mid e_2)$	$\times$	$t(f_2 \mid e_1)$	$\times$	$t(f_3 \mid e_2)$	$+$
$t(f_1 \mid e_2)$	$\times$	$t(f_2 \mid e_2)$	$\times$	$t(f_3 \mid e_1)$	$+$
$t(f_1 \mid e_2)$	$\times$	$t(f_2 \mid e_2)$	$\times$	$t(f_3 \mid e_2)$	

Word Alignments: IBM Model 1

Factor the terms:

$$\begin{array}{lcll}
(t(f_1 \mid e_1) \times t(f_2 \mid e_1)) & \times & (t(f_3 \mid e_1) + t(f_3 \mid e_2)) & + \\
(t(f_1 \mid e_1) \times t(f_2 \mid e_2)) & \times & (t(f_3 \mid e_1) + t(f_3 \mid e_2)) & + \\
(t(f_1 \mid e_2) \times t(f_2 \mid e_1)) & \times & (t(f_3 \mid e_1) + t(f_3 \mid e_2)) & + \\
(t(f_1 \mid e_2) \times t(f_2 \mid e_2)) & \times & (t(f_3 \mid e_1) + t(f_3 \mid e_2)) &
\end{array}$$

$$(t(f_3 \mid e_1) + t(f_3 \mid e_2)) \left(\begin{array}{lcll}
t(f_1 \mid e_1) & \times & t(f_2 \mid e_1) & + \\
t(f_1 \mid e_1) & \times & t(f_2 \mid e_2) & + \\
t(f_1 \mid e_2) & \times & t(f_2 \mid e_1) & + \\
t(f_1 \mid e_2) & \times & t(f_2 \mid e_2) &
\end{array} \right)$$

$$(t(f_3 \mid e_1) + t(f_3 \mid e_2)) \left(\begin{array}{lcll}
t(f_1 \mid e_1) & \times & (t(f_2 \mid e_1) + t(f_2 \mid e_2)) & + \\
t(f_1 \mid e_2) & \times & (t(f_2 \mid e_1) + t(f_2 \mid e_2)) &
\end{array} \right)$$

Word Alignments: IBM Model 1

Assume (f_1, f_2, f_3) and (e_1, e_2): $I = 3$ and $J = 2$

$$\prod_{i=1}^{3} \sum_{a_i=1}^{2} t(f_i \mid e_{a_i})$$

$I \times J = 2 \times 3$ terms to be added:

$$\begin{array}{lcll}
(t(f_1 \mid e_1) & + & t(f_1 \mid e_2)) & \times \\
(t(f_2 \mid e_1) & + & t(f_2 \mid e_2)) & \times \\
(t(f_3 \mid e_1) & + & t(f_3 \mid e_2)) &
\end{array}$$

Parameter Estimation: IBM Model 1

We wish to learn the parameters $t(\cdot \mid \cdot)$ that maximize the log-likelihood of the training data:

$$\arg\max_t L(t) = \arg\max_t \sum_s \log \Pr(\mathbf{f}^{(s)} \mid \mathbf{e}^{(s)}, t)$$

- We start with an initial estimate t_0
- Modify it iteratively to get $t_1, t_2, \ldots$
- Create t from previous time step t_{-1}

Acknowledgements

Many slides borrowed or inspired from lecture notes by Michael Collins, Chris Dyer, Kevin Knight, Adam Lopez, and Luke Zettlemoyer from their NLP course materials. All mistakes are my own.