SFU NatLangLab

# Natural Language Processing


Anoop Sarkar

anoopsarkar.github.io/nlp-class

Simon Fraser University


September 28, 2017

# Natural Language Processing

Anoop Sarkar
anoopsarkar.github.io/nlp-class

Simon Fraser University

Part 1: Statistical Machine Translation

## Introduction to Statistical Machine Translation

# Basic Terminology

### Translation

We will consider translation of

- a source language string in French, called $\mathbf{f}$
- into a target language string in English, called $\mathbf{e}$.

### *A priori* probability: $\Pr(\mathbf{e})$

The chance that $\mathbf{e}$ is a valid English string.
What is better? Pr(*I like snakes*) or Pr(*snakes like I*)

### Conditional probability: $\Pr(\mathbf{f} \mid \mathbf{e})$

The chance of French string $\mathbf{f}$ given $\mathbf{e}$.
What is the chance of French string *maison bleue* given the English string *I like snakes*?

# Basic Terminology

### Joint probability: $\Pr(\mathbf{e}, \mathbf{f})$

The chance of both English string **e** and French string **f** occuring together.

- If **e** and **f** are independent (do not influence each other) then
$$\Pr(\mathbf{e}, \mathbf{f}) = \Pr(\mathbf{e}) \Pr(\mathbf{f})$$
- If **e** and **f** are not independent (they do influence each other) then
$$\Pr(\mathbf{e}, \mathbf{f}) = \Pr(\mathbf{e}) \Pr(\mathbf{f} \mid \mathbf{e})$$

Which one should we use for machine translation?

# Statistical Machine Translation

Given French string $\mathbf{f}$ find the English string $\mathbf{e}$ that maximizes $\Pr(\mathbf{e} \mid \mathbf{f})$

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

This finds the *most likely* translation $\mathbf{e}^*$

## Alignment Task

$$
\begin{tikzcd}
\mathbf{e} \arrow[dr] & & \\
& \text{Program} \arrow[r] & \Pr(\mathbf{e} \mid \mathbf{f}) \\
\mathbf{f} \arrow[ur] & &
\end{tikzcd}
$$

## Translation Task

$$
\begin{tikzcd}
& & \mathbf{e}_1 : \Pr(\mathbf{e}_1 \mid \mathbf{f}) \\
\mathbf{f} \arrow[r] & \text{Program} \arrow[ur] \arrow[dr] & \vdots \\
& & \mathbf{e}_n : \Pr(\mathbf{e}_n \mid \mathbf{f})
\end{tikzcd}
$$

# Bayes' Rule

### Bayes' Rule

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

### Exercise

Show the above equation using the definition of $P(\mathbf{e}, \mathbf{f})$ and the chain rule.

# Noisy Channel Model

## Use Bayes' Rule

$$
\begin{aligned}
\mathbf{e}^* &= \arg\max_{\mathbf{e}} \Pr(\mathbf{e} \mid \mathbf{f}) \\
&= \arg\max_{\mathbf{e}} \frac{\Pr(\mathbf{e}) \Pr(\mathbf{f} \mid \mathbf{e})}{\Pr(\mathbf{f})} \\
&= \arg\max_{\mathbf{e}} \Pr(\mathbf{e}) \Pr(\mathbf{f} \mid \mathbf{e})
\end{aligned}
$$

## Noisy Channel

- Imagine a French speaker has **e** in their head
- By the time we observe it, **e** has become “corrupted” into **f**
- To recover the most likely **e** we reason about
  1. What kinds of things are likely to be **e**
  2. How does **e** get converted into **f**

# Statistical Machine Translation

## Noisy Channel Model

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

## Training the components

- **Language Model**: *n*-gram language model with smoothing.
  Training data: lots of monolingual **e** text.
- **Alignment Model**: learn a mapping between **f** and **e**.
  Training data: lots of translation pairs between **f** and **e**.

# Word reordering in Translation

## Candidate translations

Every candidate translation $\mathbf{e}$ for a given $\mathbf{f}$ has two factors:

$$\Pr(\mathbf{e}) \Pr(\mathbf{f} \mid \mathbf{e})$$

What is the contribution of $\Pr(\mathbf{e})$?

## Exercise: Bag Generation

Put these words in order:
*have programming a seen never I language better*

## Exercise: Bag Generation

Put these words in order:
*actual the hashing is since not collision-free usually the is less perfectly the of somewhat capacity table*

# Word reordering in Translation

### Candidate translations

Every candidate translation **e** for a given **f** has two factors:

$$\Pr(\mathbf{e}) \Pr(\mathbf{f} \mid \mathbf{e})$$

What is the contribution of $\Pr(\mathbf{f} \mid \mathbf{e})$?

### Exercise: Bag Generation

Put these words in order:
*love John Mary*

### Exercise: Word Choice

Choose between two alternatives with similar scores $\Pr(\mathbf{f} \mid \mathbf{e})$:
*she is in the end zone*
*she is on the end zone*

# Statistical Machine Translation

## Noisy Channel Model

Every candidate translation **e** for a given **f** has two factors:

$$\Pr(\mathbf{e}) \Pr(\mathbf{f} \mid \mathbf{e})$$

## Translation Modeling

- $\Pr(\mathbf{f} \mid \mathbf{e})$ does not need to be perfect because of the $\Pr(\mathbf{e})$ factor.
- $\Pr(\mathbf{e})$ models **fluency**.
- $\Pr(\mathbf{f} \mid \mathbf{e})$ models the transfer of **content**.
- This a *generative model* of translation.

# $\Pr(\mathbf{f} \mid \mathbf{e})$: How does English become French?

## English $\Rightarrow$ Meaning $\Rightarrow$ French

- English to meaning representation:
  *John must not go* $\Rightarrow$ $\textsc{obligatory}(\textsc{not}(\textsc{go}(\textsc{john})))$
  *John may not go* $\Rightarrow$ $\textsc{not}(\textsc{permitted}(\textsc{go}(\textsc{john})))$
- Meaning representation to French

## English $\Rightarrow$ Syntax $\Rightarrow$ French

- Parsed English:
  *Mary loves soccer* $\Rightarrow$ *(S (NP Mary) (VP (V loves) (NP soccer)))*
- Parse tree to French parse tree:
  *(S (NP Mary) (VP (V loves) (NP soccer)))* $\Rightarrow$ *(S (NP Mary) (VP (V aime) (NP le football)))*

# $\Pr(\mathbf{f} \mid \mathbf{e})$: How does English become French?

**English words $\Rightarrow$ French words**

- Simplest model, map English words to French words
- Corresponds to an alignment between English and French:

$$\Pr(\mathbf{f} \mid \mathbf{e}) = \Pr(f_1, \ldots, f_I, a_1, \ldots, a_I \mid e_1, \ldots, e_J)$$

# 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 **e** and **f** in the equations in honor of their system which translated from French to English.
  Trained on the Canadian Hansards (Parliament Proceedings)

## Acknowledgements

Many slides borrowed or inspired from lecture notes by Michael Collins, Chris Dyer, Kevin Knight, Philipp Koehn, Adam Lopez, Graham Neubig and Luke Zettlemoyer from their NLP course materials.

All mistakes are my own.

A big thank you to all the students who read through these notes and helped me improve them.