Introduction to Statistical Machine Translation

David Kauchak  
CS457 – Fall 2011

Some slides adapted from
Philipp Koehn  
Kevin Knight

Admin

• How are the projects going?
• Remaining classes applications
  3 MT  
  • General overview today  
  • Dive into one specific implementation next time  
  • MT lab
• Other applications
  • Information extraction
  • Information retrieval
  • Question answering/summarization

Language translation

Yo quiero  
Taco Bell

MT Systems

Where have you seen machine translation systems?
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

The classic acid test for natural language processing. Requires capabilities in both interpretation and generation.

People around the world stubbornly refuse to write everything in English.
Which is the human?
Pakistani President Musharraf Wins Senate Confidence Vote

Which is the human?
Pakistan President Pervez Musharraf Wins Senate Confidence Vote

Warren Weaver (1947)

Warren Weaver (1947)
Warren Weaver (1947)

decipherment is the analysis of documents written in ancient languages ...

The non-Turkish guy next to me is even deciphering Turkish! All he needs is a statistical table of letter-pair frequencies in Turkish ...

Can this be computerized?

Collected mechanically from a Turkish body of text, or corpus
“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.”
- Warren Weaver, March 1947

“... as to the problem of mechanical translation, I frankly am afraid that the [semantic] boundaries of words in different languages are too vague ... to make any quasi-mechanical translation scheme very hopeful.”
- Norbert Wiener, April 1947

Noisy channel

Some message is sent

along the way things get messed up

What was originally sent?

We have the mutated message, but would like to recover the original

If we know something about what goes on inside here, we might be to decode/recover the message.
“Hi bob”
I baab

“Hi bob”
‘H’s often get dropped
‘o’s go to ‘a’ sometimes
‘...’

“Hi bob”
I baab

Sometime, we don’t know what goes on or we only have a rough guess...
What then?
Noisy channel

A data driven approach: learn a model of the types of transformations that occur

Send a bunch of data through where we know what is being sent

How does a model $p(s \mid t)$ help us?

```
$"..." (s)  "..." (t)
```

Decode: $\arg\max p(s \mid t)$

Noisy channel

- source ($s$): what was originally sent
- target ($t$): what did we get through the noisy channel

We want a probabilistic model of the process:

$$p(s \mid t)$$

What is the probability of a given source sentence, given that we’ve seen target
Noisy channel model

\[ p(s \mid t) = \frac{p(t \mid s)p(s)}{p(t)} \]  
Bayes’ rule

- **p(t)**: how likely is it to receive the target message
- **p(s)**: What types of messages are likely to be sent? What do sent messages look like?
- **p(t \mid s)**: What types of transformation happen? How likely are we to generate \( t \) from \( s \)?

Applications: Speech recognition

- **source**
- **noise**
- **target**
- **words**
- **given the target audio, what were the original words?**

Applications: Speech recognition

- **channel model**
- **translation model**
- **language model**
- **source model**
- **expected words/word sequences**

\[ p(s \mid t) \propto p(t \mid s)p(s) \]
Applications: Spelling correction

Given the target text with misspellings, what were the original, correctly spelled words?

- basketball committee

Applications: Sentence compression

Given the target uncompressed text, what was the original compressed text?

- The dog runs slowly through the large park.
Applications: Machine translation

English text \rightarrow source

noise

given the target Chinese text, what was the original English text?

target

Applications: Machine translation

\[ p(s \mid t) \propto p(t \mid s)p(s) \]

channel model

translation model

language model

how do English words/phrases translate to Chinese?

what are likely English words/word sequences?

Data-Driven Machine Translation

Man, this is so boring...

Hmm, every time he sees "banco", he either types "bank" or "bench"... but if he sees "banco de..." he always types "bank", never 'bench'...

Welcome to the Chinese Room

You can teach yourself to translate Chinese using only bilingual data (without grammar books, dictionaries, any people to answer your questions...)

Translated documents
Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan:

1a. ok-voon ororok sprok .
1b. at-voon bichat dat .

7a. lalok farok ororok sprok izok enemok .
7b. wat jat bichat wat dat vat eneat .

2a. ok-drubel ok-voon anok plok sprok .
2b. at-drubel ok-voon peppat mat dat .

8a. lalok brok anok plok nok .
8b. iat lat peppat mat mut .

3a. awk sprok izok hihok ghirok .
3b. totat dat arrat vat hilat .

9a. totat dat arrat vat hilat .
9b. totat dat arrat vat hilat .

4a. ok-voon anok druk brok jot .
4b. at-voon kret peppat sat lat .

10a. lalok nok nok yorok ghirok clok .
10b. wat mut mat bat hilat .

5a. awk sprok izok stok .
5b. totat jat quit cat .

11a. totat jat quit cat .
11b. totat jat quit cat .

6a. lalok sprok izok jot stok .
6b. wat lat jat quit cat .

12a. lalok rerek nok izok hihok nok .
12b. wat mut mut arrat vat gat .

Your assignment, translate this to Arcturan:

farok crrrok hihok yorok clok kantok ok-yurp
Your assignment, translate this to Arcturan:

1a. ok-voon ororok sprok.
1b. at-voon bichat dat.
2a. ok-drubel ok-voon anok pliek sprok.
2b. at-drubel at-voon peppat mat dat.
3a. ok-voon izok bitrok ghirok.
3b. totat dart arrat vat hilat.
4a. ok-voon anok druk brok jok.
4b. at-voon krat pippat sat lat.
5a. wiwok sprok izok.
5b. totat jat quat cat.
6a. lalok sprok izok.
6b. wat dart jat quat cat.
7a. lalok farok ororok lalok sprok izok enemok.
7b. wat jat bichat wat dat vat enemat.
8a. lalok druk anok pliek nok.
8b. iat lat peppat mat mat.
9a. totat dart arrat vat hilat.
9b. totat dart arrat vat hilat.
10a. lalok mok nok yorok ghirok clok.
10b. wat jat jat cat mat bat hilat.
11a. lalok druk anok pliek nok.
11b. iat lat peppat mat mok.
12a. lalok mok nok yorok ghirok.
12b. wat jat jat cat mat bat hilat.
### Centauri/Arcturan [Knight, 1997]

#### Your assignment, translate this to Arcturan:

<table>
<thead>
<tr>
<th>English</th>
<th>Arcturan</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. ok-voon ororok sprok .</td>
<td>farok kemok iblok yerkok (cik) kantok ok-yarp</td>
</tr>
<tr>
<td>1b. at-voon bichat dat .</td>
<td>7a. lubok kemok ororok lubok sprok izok emenom .</td>
</tr>
<tr>
<td>1b. at-voon bichat dat .</td>
<td>7b. wat jat bichat wat dat vat eneat .</td>
</tr>
<tr>
<td>2a. ok-drubel ok-voon anok plok sprok .</td>
<td>8a. lubok break anok plok nok .</td>
</tr>
<tr>
<td>2b. at-drubel at-voon peppat rat dat .</td>
<td>8b. iat lat peppat rat mat .</td>
</tr>
<tr>
<td>3a. erok sprok izok lubok ghirok .</td>
<td>9a. wewok nok izok kantok ok-yarp .</td>
</tr>
<tr>
<td>3b. totat dat arrat vat hilat .</td>
<td>9b. totat mat quiet okolat at-yarp .</td>
</tr>
<tr>
<td>4a. ok-voon anok druk brok jok .</td>
<td>10a. iat lat nok izok yerkok clok .</td>
</tr>
<tr>
<td>4b. at-voon keat peppat sat lat .</td>
<td>10b. wat jat lat rat bat hilat .</td>
</tr>
<tr>
<td>5a. wewok fopok izok nok .</td>
<td>11a. lubok muk izok kemok kemok noromok .</td>
</tr>
<tr>
<td>5b. totat jar quiet cat .</td>
<td>11b. wat mat quiet rat otomat .</td>
</tr>
<tr>
<td>6a. lalok sprok izok jok steok .</td>
<td>12a. lalok marok nok izok lubok nok .</td>
</tr>
<tr>
<td>6b. wat lat dat quiet cat .</td>
<td>12b. wat mat quiet rat vat jat .</td>
</tr>
</tbody>
</table>

#### Your assignment, put these words in order:

- jat, arrat, mat, bat, oloat, at-yarp

---

#### Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan:

- farok kemok iblok yerkok (cik) kantok ok-yarp

#### Centauri/Arcturan [Knight, 1997]

Your assignment, put these words in order:

- jat, arrat, mat, bat, oloat, at-yarp
### Clients do not sell pharmaceuticals in Europe => Clientes no venden medicinas en Europa

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. Garcia and associates .</td>
<td>7a. the clients and the associates are enemies .</td>
</tr>
<tr>
<td>1b. Garcia y asociados .</td>
<td>7b. los clientes y los asociados son enemigos .</td>
</tr>
<tr>
<td>2a. Carlos Garcia has three associates .</td>
<td>8a. the company has three groups .</td>
</tr>
<tr>
<td>2b. Carlos Garcia tiene tres asociados .</td>
<td>8b. la empresa tiene tres grupos .</td>
</tr>
<tr>
<td>3a. his associates are not strong .</td>
<td>9a. its groups are in Europe .</td>
</tr>
<tr>
<td>3b. sus asociados no son fuertes .</td>
<td>9b. sus grupos estan en Europa .</td>
</tr>
<tr>
<td>4a. Garcia has a company also .</td>
<td>10a. the modern groups sell strong pharmaceuticals .</td>
</tr>
<tr>
<td>4b. Garcia tambien tiene una empresa .</td>
<td>10b. los grupos modernos venden medicinas fuertes .</td>
</tr>
<tr>
<td>5a. its clients are angry .</td>
<td>11a. the groups do not sell zenzanine .</td>
</tr>
<tr>
<td>5b. sus clientes estan enfadados .</td>
<td>11b. los grupos no venden zanzanina .</td>
</tr>
<tr>
<td>6a. the associates are also angry .</td>
<td>12a. the small groups are not modern .</td>
</tr>
<tr>
<td>6b. los asociados tambien estan enfadados</td>
<td>12b. los grupos pequenos no son modernos .</td>
</tr>
</tbody>
</table>

#### Data available

- **Many languages**
  - European corpus has all European languages
    - [http://www.statmt.org/europarl/](http://www.statmt.org/europarl/)
  - From a few hundred thousand sentences to a few million
  - French/English from French parliamentary proceedings
  - Lots of Chinese/English and Arabic/English from government projects/interests
    - Chinese-English: 440 million words (15-20 million sentence pairs)
    - Arabic-English: 790 million words (30-40 million sentence pairs)
  - Smaller corpora in many, many other languages
- **Lots of monolingual data available in many languages**
- **Even less data with multiple translations available**
- **Available in limited domains**
  - most data is either news or government proceedings
  - some other domains recently, like blogs

#### Statistical MT Overview

- **Bilingual data**
- **Monolingual data**
- **Find the best translation given the foreign sentence and the model**
- **English sentence**
Statistical MT

- We will model the translation process probabilistically
- Given a foreign sentence to translate, for any possible English sentence, we want to know the probability that sentence is a translation of the foreign sentence
- If we can find the most probable English sentence, we’re done

\[ p(\text{english sentence} \mid \text{foreign sentence}) \]

Noisy channel model

\[ p(e \mid f) \propto p(f \mid e)p(e) \]

- Translation model
- Language model
- How do foreign sentences get translated to English sentences?
- What do English sentences look like?

Translation model

- The models define probabilities over inputs
  \[ p(f \mid e) \]

Morgen fliege ich nach Kanada zur Konferenz

Tomorrow I will fly to the conference in Canada

What is the probability that the English sentence is a translation of the foreign sentence?
Translation model

• The models define probabilities over inputs $p(f \mid e)$

Translation model

• The models define probabilities over inputs $p(f \mid e)$

Language model

• The models define probabilities over inputs $p(e)$

What is a probability distribution?

• A probability distribution defines the probability over a space of possible inputs
• For the language model, what is the space of possible inputs?
  – A language model describes the probability over ALL possible combinations of English words
• For the translation model, what is the space of possible inputs?
  – ALL possible combinations of foreign words with ALL possible combinations of English words
One way to think about it…

Spanish (foreign) → Translation model → Broken English → language model → English

Que hambre tengo yo → What hunger have I,
Hungry I am so, I am so hungry,
Have I that hunger …

Translation

\[ p(e \mid f) \propto p(f \mid e)p(e) \]

- Let’s assume we have a translation model and a language model
- Given a foreign sentence, what question do we want to ask to translate that sentence into English?

\[ \arg \max_e p(e \mid f) \propto p(f \mid e)p(e) \]

Statistical MT Overview

Translation

Foreign sentence → Decoder (what English sentence is most probable given foreign sentence with learned models)

Bilingual data

preprocessing

monolingual data

“nice” fragment aligned data

Translation model

Language model

learned parameters

Problems for Statistical MT

- Preprocessing
  - How do we get aligned bilingual text?
  - Tokenization
  - Segmentation (document, sentence, word)
- Language modeling
  - Given an English string e, assigns \( P(e) \) by formula
- Translation modeling
  - Given a pair of strings \(<f,e>\), assigns \( P(f \mid e) \) by formula
- Decoding
  - Given a language model, a translation model, and a new sentence f … find translation e maximizing \( P(e) \cdot P(f \mid e) \)
- Parameter optimization
  - Given a model with multiple feature functions, how are they related? What are the optimal parameters?
- Evaluation
  - How well is a system doing? How can we compare two systems?
Problems for Statistical MT
• Preprocessing
• Language modeling
• Translation modeling
• Decoding
• Parameter optimization
• Evaluation

Data
We want pairs of aligned sentences/text fragments. How do we get them?

From No Data to Sentence Pairs
• Easy way 1: Linguistic Data Consortium (LDC)
• Easy way 2: pay $$$
  – Suppose one billion words of parallel data were sufficient
  – At 20 cents/word, that’s $200 million
• Hard way: Find it, and then earn it!
  – De-formatting
  – Remove strange characters
  – Character code conversion
  – Document alignment
  – Sentence alignment
  – Tokenization (also called Segmentation)
If you don't get the characters right...

ISO-8859-2 (Latin2)

ISO-8859-6 (Arabic)

Chinese?

- GB Code
- GBK Code
- Big 5 Code
- CNS-11643-1992
- ...


Document Alignment

- Input:
  - Big bag of files obtained from somewhere, believed to contain pairs of files that are translations of each other.

- Output:
  - List of pairs of files that are actually translations.
• Input:
  – Big bag of files obtained from somewhere, believed to contain pairs of files that are translations of each other.

• Output:
  – List of pairs of files that are actually translations.

Sentence Alignment

1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.

1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.

Sentence Alignment

1. The old man is happy.  El viejo está feliz.
2. He has fished many times.  Porque ha pescado muchos veces.
3. His wife talks to him.  Su mujer habla con él.
4. The fish are jumping.  Los tiburones esperan.
5. The sharks await.
**Sentence Alignment**

1. The old man is happy. He has fished many times.
2. His wife talks to him.
3. The sharks await.

---

**Tokenization (or Segmentation)**

- **English**
  - Input (some byte stream):
    "There," said Bob.
  - Output (7 "tokens" or "words"):
    "There," said Bob.

- **Chinese**
  - Input (byte stream):
    美国夏威夷州机场及其办公室将接收一份来自鲜压机和重量标签等的电子邮件。
  - Output:
    美国 夏威夷州 机场 及 其 办公室 将 接收 一 份 来自 鲜压机 和 重量 标签 等 的 电子邮件。

---

**Problems for Statistical MT**

- Preprocessing
- Language modeling
- Translation modeling
- Decoding
- Parameter optimization
- Evaluation

---

**Language Modeling**

- Most common: n-gram language models
- More data the better (Google n-grams)
- Domain is important
Problems for Statistical MT

- Preprocessing
- Language modeling
- Translation modeling
- Decoding
- Parameter optimization
- Evaluation

MT Pyramid

Translation Model

Learn How to Translate from Data

Direct Estimation:

Mary did not slap the green witch

not enough data for this (most input sentences unseen)

Maria no dio una bofetada a la bruja verde
Generative Model

Break up process into smaller steps:

Mary did not slap the green witch

sufficient statistics for smaller steps

Maria no dio una botefada a la bruja verde

What kind of Translation Model?

May use syntactic and semantic representations:

Mary did not slap the green witch

Word-level models

Phrasal models

Syntactic models

Semantic models

Maria no dio una botefada a la bruja verde

The Classic Translation Model

Word Substitution/Permutation [IBM Model 3, Brown et al., 1993]

Generative story:

Mary did not slap the green witch

n(3|slap)

Mary not slap slap slap the green witch

P-Null

t(la|the)

d(ji)

Maria no dio una botefada a la verde bruja

Maria no dio una botefada a la bruja verde

Probabilities can be learned from raw bilingual text.

Phrase-Based Statistical MT

- Foreign input segmented into phrases
  - "phrase" is any sequence of words
- Each phrase is probabilistically translated into English
  - P(to the conference / zur Konferenz)
  - P(into the meeting / zur Konferenz)
- Phrases are probabilistically re-ordered

Morgen fliege ich nach Kanada zur Konferenz

Tomorrow I will fly to the conference In Canada
Advantages of Phrase-Based

- Many-to-many mappings can handle non-compositional phrases
- Easy to understand
- Local context is very useful for disambiguating
  - "Interest rate" $\rightarrow$ …
  - "Interest in" $\rightarrow$ …
- The more data, the longer the learned phrases
  - Sometimes whole sentences

Syntax

These 7 people include astronauts coming from France and Russia.

Problems for Statistical MT

- Preprocessing
- Language modeling
- Translation modeling
- Decoding
- Parameter optimization
- Evaluation
Decoding

• Of all conceivable English word strings, find the one maximizing \( P(e) \times P(f | e) \)

• Decoding is an NP-complete problem (for many translation models – (Knight, 1999))

• Several decoding strategies are often available

Search

\[
\arg\max_p(p(f | e))p(e)
\]

\[
\text{partial trans. } f_1, f_2, f_3, f_4, f_5
\]

\[
\text{partial trans. } f_1, f_2, f_3, f_4, f_5
\]

\[
\text{partial trans. } f_1, f_2, f_3, f_4, f_5
\]
Search

\[
\text{arg}_{\text{max}} \ P(f \mid e) = \text{e}
\]

\[
\text{arg}_{\text{max}} \ P(e) \times P(f \mid e) / P(f) = \text{e}
\]

\[
\text{arg}_{\text{max}} \ P(e) \times P(f \mid e) = \text{e}
\]

Basic Model, Revisited

Problems for Statistical MT

- Preprocessing
- Language modeling
- Translation modeling
- Decoding
- Parameter optimization
- Evaluation

Basic Model, Revisited

\[
\text{arg}_{\text{max}} \ P(e \mid f) = \text{e}
\]

\[
\text{arg}_{\text{max}} \ P(e) \times P(f \mid e) / P(f) = \text{e}
\]

\[
\text{arg}_{\text{max}} \ P(e)^{2.4} \times P(f \mid e) \quad \ldots \text{works better!} \]

\[
\text{arg}_{\text{max}} \ P(e) \times P(f \mid e) = \text{e}
\]
Basic Model, Revisited

argmax \ P(e \mid f) = \ e

argmax \ P(e) \times \frac{P(f \mid e)}{P(f)} e

argmax \ P(e)^{2.4} \times P(f \mid e) \times \text{length}(e)^{1.1} e

Rewards longer hypotheses, since these are unfairly punished by \ P(e).

The Problem: Learn Lambdas

\begin{align*}
    p(e \mid f) &= \frac{p(f \mid e)p(e)}{p(f)} \\
    &= \frac{\sum p(f \mid e') p(e')}{{\lambda} \sum p(f \mid e') p(e')} \\
    &= \exp(\lambda \log \frac{p(f \mid e)}{p(f)} + \log p(e) + \log p(e \mid f) + \text{length}(e)...) \\
    &= \exp(\lambda_1 \log p(f \mid e') + \lambda_2 \log p(e') + \lambda_3 \log p(e' \mid f) + \lambda_4 \text{length}(e')...) \\
    &= \exp(\sum \lambda_k(f, e)) \\
    &= \frac{\exp(\sum \lambda_k(f, e'))}{\sum \exp(\sum \lambda_k(f, e'))}
\end{align*}

Given a data set with foreign/English sentences, find the \(\lambda\)'s that:

- maximize the likelihood of the data
- maximize an evaluation criterion

Basic Model, Revisited

argmax \ P(e)^{2.4} \times P(f \mid e) \times \text{length}(e)^{1.1} \times \text{KS}^{3.7} ...

Lots of knowledge sources vote on any given hypothesis.

“Knowledge source” = “feature function” = “score component”.

A feature function simply scores a hypothesis with a real value.

(May be binary, as in “e has a verb”).