6.881: Natural Language Processing
Machine Translation I

Philipp Koehn
CS and AI Lab
MIT

November 18, 2004
Outline

- Lecture I
  - Introduction to Machine Translation
  - Principles of Statistical MT
  - Word-Based Models
  - Phrase-Based Models

- Lecture II
  - Beam Search Decoding
  - Discriminative Training
  - The Challenge of Syntax
Machine Translation

- Task: Make sense of foreign text like

- One of the oldest problems in Artificial Intelligence

- AI-hard: reasoning and world knowledge required
A Brief History

- 1950s: success in code breaking $\rightarrow$ investment in MT
- 1964: ALPAC report killed MT research funding
- Late 1970s: Transformer MT
- 1980s: commercial companies like Systran founded
- 1980s: Example-Based MT, Linguistic Knowledge MT
- 1990s: Statistical MT
- 2000s: Commercial companies sell statistical MT systems (Language Weaver, IBM)
The Machine Translation Pyramid

- Foreign words
- Foreign syntax
- Foreign semantics
- Interlingua
- English syntax
- English semantics
- English words
The Machine Translation Pyramid

However, the currently best performing statistical machine translation systems are still crawling at the bottom.

Philipp Koehn, CSAIL, MIT
Data for Statistical MT: Parallel Corpora

- French-English: 180 million words
  - UN, Canadian Hansards, European Parliament
- Chinese-English: 130 million words
  - UN, Xinhua news corpora
- Arabic-English: 100 million words
  - UN, news corpora
- German-English: 30 million words
  - European Parliament
Statistical MT Systems

- Spanish/English Bilingual Text
  - Statistical Analysis
  - Spanish
  - Que hambre tengo yo
  - I am so hungry

- English Text
  - Statistical Analysis
  - Broken English
  - What hunger have I
  - I am so hungry

- English
  - I am so hungry

...
Statistical MT Systems (2)

Spanish/English Bilingual Text → Statistical Analysis → Spanish → Translation Model → Decoding Algorithm \[ \text{argmax} \ P(e) \times p(s \vert e) \] → Broken English → English Text → Statistical Analysis → English Text → Language Model
Three Problems in Statistical MT

- **Translation Model**
  - given a pair of strings \(< f, e >\), assigns \(P(f|e)\) by formula
  - \(< f, e >\) look like translations \(\Rightarrow\) high \(P(f|e)\)
  - \(< f, e >\) don’t look like translations \(\Rightarrow\) low \(P(f|e)\)

- **Language Model**
  - given an English string \(e\), assigns \(P(e)\) by formula
  - good English string \(\Rightarrow\) high \(P(e)\)
  - bad English string \(\Rightarrow\) low \(P(e)\)

- **Decoding Algorithm**
  - given a language model, a translation model and a new sentence \(f\), find translation \(e\) maximizing \(P(e) \times P(f|e)\)
Statistical Modeling

Mary did not slap the green witch

Maria no daba una bofetada a la bruja verde

- Learn $P(f|e)$ from a parallel corpus
- Not sufficient data to estimate $P(f|e)$ directly
Mary did not slap the green witch

Maria no daba una bofetada a la bruja verde

- Break the process into smaller steps
Statistical Modeling (3)

- Probabilities for smaller steps can be learned
Statistical Modeling (4)

- Generate a story how an English string $e$ gets to be a foreign string $f$
  - Choices in story are decided by reference to parameters
  - e.g., $p(\text{bruja}|\text{witch})$

- Formula for $P(f|e)$ in terms of parameters
  - usually long and hairy, but mechanical to extract from the story

- Training to obtain parameter estimates from possibly incomplete data
  - off-the-shelf EM
IBM Model 4

- **Fertility**
  - $n(3|\text{slap})$ number of words generated
  - $n(0|\text{do})$ dropped words

- **NULL Word**
  - parameter $p_0$
  - allows generation of additional words

- **Translation**
  - $t(\text{la}|\text{the})$ lexical translation

- **Distortion**
  - $d(-1|\ldots)$ reordering
Learning the Parameters

... la maison ... la maison blue ... la fleur ...

... the house ... the blue house ... the flower ...

- **Incomplete Data**
  - English and foreign words, but no connections between them

- **Chicken and egg problem**
  - if we had the connections, we could estimate the parameters of our generative story
  - if we had the parameters, we could estimate the connections
EM Algorithm

- Incomplete data
  - if we had complete data, we could estimate model
  - if we had model, we could fill in the gaps in the data

- EM in a nutshell
  - initialize model parameters (e.g. uniform)
  - assign probabilities to the missing data
  - estimate model parameters from completed data
  - iterate
EM Algorithm (2)

... la maison ... la maison blue ... la fleur ...

... the house ... the blue house ... the flower ...

- Initial step: all connections equally likely
- Model learns that, e.g., la is often connected with the
EM Algorithm (3)

... la maison ... la maison blue ... la fleur ...

... the house ... the blue house ... the flower ...

- After one iteration
- Connections, e.g., between la and the are more likely
EM Algorithm (4)

... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...

- After another iteration
- It becomes apparent that connections, e.g., between fleur and flower are more likely (pigeon hole principle)
EM Algorithm (5)

... la maison ... la maison bleu ... la fleur ...  
/ \    / \  / \  / \  / \  / \  
/    \  /   \ /   \ /   \ /   \ /   \  
... the house ... the blue house ... the flower ...

- Convergence
- Inherent hidden structure revealed by EM
EM Algorithm (6)

\[
\begin{align*}
\text{\ldots la maison \ldots la maison bleu \ldots la fleur \ldots} \\
\text{\ldots the house \ldots the blue house \ldots the flower \ldots} \\
\end{align*}
\]

\[
p(la | \text{the}) = 0.453 \\
p(le | \text{the}) = 0.334 \\
p(\text{maison} | \text{house}) = 0.876 \\
p(\text{bleu} | \text{blue}) = 0.563 \\
\ldots
\]

- Parameter estimation from the connected corpus
More detail on the IBM Models

- “A Statistical MT Tutorial Workbook” (Knight, 1999)
- “The Mathematics of Statistical Machine Translation”
  (Brown et al., 1993)
- Downloadable software: Giza++, ReWrite decoder
Word Alignment

- Notion of word alignments valuable
- Trained humans can achieve high agreement
- Shared task at data-driven MT workshop at NAACL/HLT
Improved Word Alignments

- Improving IBM Model word alignments with heuristics
  [Och and Ney, 2000, Koehn et al., 2003]
  - one-to-many problem of IBM Models
  - bidirectionally aligned corpora $e \rightarrow f, f \rightarrow e$
  - take intersection of alignment points
    (high precision, low recall)
  - grow additional alignment points
    (increase recall while preserving precision)
Improved Word Alignments (2)

- Intersection of bidirectional alignments
Improved Word Alignments (3)

- Grow additional alignment points
Other Methods for Word Alignment

- Comparison of various methods in journal “Computation Linguistics”, March 2003 [Och and Ney, 2003]
- Shared task at data-driven MT workshop HLT/NAACL 2003
Flaws of Word-Based MT

- **Multiple English words for one German word**
  
  German: Zeitmangel erschwert das Problem.
  
  Gloss: LACK OF TIME MAKES MORE DIFFICULT THE PROBLEM.
  
  Correct translation: Lack of time makes the problem more difficult.
  
  MT output: Time makes the problem.

- **Phrasal translation**
  
  German: Eine Diskussion erübrigt sich demnach.
  
  Gloss: A DISCUSSION IS MADE UNNECESSARY ITSELF THEREFORE.
  
  Correct translation: Therefore, there is no point in a discussion.
  
  MT output: A debate turned therefore.
Flaws of Word-Based MT (2)

- Syntactic transformations

  German: Das ist der Sache nicht angemessen.
  Gloss: THAT IS THE MATTER NOT APPROPRIATE.
  Correct translation: That is not appropriate for this matter.
  MT output: That is the thing is not appropriate.

  German: Den Vorschlag lehnt die Kommission ab.
  Gloss: THE PROPOSAL REJECTS THE COMMISSION OFF.
  Correct translation: The commission rejects the proposal.
  MT output: The proposal rejects the commission.
Phrase-Based Translation

- Foreign input is segmented in phrases
  - any sequence of words, not necessarily linguistically motivated
- Each phrase is translated into English
- Phrases are reordered
Advantages of Phrase-Based Translation

- Many-to-many Translation
- Use of local context in translation
- Allows translation of non-compositional phrases
- The more data, the longer phrases can be learned
Three Phrase-Based Translation Models

- Word alignment induced phrase model [Koehn et al., 2003]
- Alignment templates [Och et al., 1999]
- Joint phrase model [Marcu and Wong, 2002]
Word Alignment Induced Phrases

- Collect all phrase pairs that are consistent with the word alignment
  - a phrase alignment has to contain all alignment points for all words it covers
Word Alignment Induced Phrases (2)

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green)
Word Alignment Induced Phrases (3)

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green), (Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch)
Word Alignment Induced Phrases (4)

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch),
(verde, green), (Maria no, Mary did not), (no daba una bofetada, did not slap),
(daba una bofetada a la, slap the), (bruja verde, green witch),
(Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch)
Word Alignment Induced Phrases (5)

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch),
(verde, green), (Maria no, Mary did not), (no daba una bofetada, did not slap),
(daba una bofetada a la, slap the), (bruja verde, green witch),
(Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la, Mary did not slap the),
(daba una bofetada a la bruja verde, slap the green witch)
Word Alignment Induced Phrases (6)

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch),
(verde, green), (Maria no, Mary did not), (no daba una bofetada, did not slap),
(daba una bofetada a la, slap the), (bruja verde, green witch),
(Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la, Mary did not slap the),
(daba una bofetada a la bruja verde, slap the green witch),
(no daba una bofetada a la bruja verde, did not slap the green witch),
(Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)
Word Alignment Induced Phrases (7)

- Given the collected phrase pairs, estimate the phrase translation probability distribution by relative frequency:

\[ \phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{f},\bar{e})}{\sum_{\bar{f}} \text{count}(\bar{f},\bar{e})} \]

- No smoothing is performed
Word Alignment Induced Phrases (8)

- Lexical weighting:

\[ p_w(f|\bar{e}, a) = \prod_{i=1}^{n} \frac{1}{|\{j|(i,j)\in a\}|} \sum_{\forall(i,j)\in a} w(f_i|e_j) \]

\[ p_w(\text{a la bruja verde}|\text{the green witch}) = \]

\[ w(\text{a}|\text{the}) \times w(\text{la}|\text{the}) \times \]

\[ w(\text{verde}|\text{green}) \times \]

\[ w(\text{bruja}|\text{witch}) \]
Joint Phrase Model

- Direct phrase alignment of parallel corpus
  [Marcu and Wong, 2002]

- Generative story
  - a number of concepts are created
  - each concept generates a foreign and English phrase
  - the English phrases are reordered
Limits of Phrase Models

- Non-contiguous phrases
  - German: *Ich habe das Auto gekauft*
  - English: *I bought the car*
  - good phrase pair: *habe ... gekauft == bought*

- Syntactic transformations
  - German: *Den Antrag verabschiedet das Parlament*
  - English gloss: *The draft approves the Parliament*
  - case marking that indicates that “the draft” is object is lost during translation
Phrase-Based MT: Do it yourself

- Phrase-based MT has currently best performance
- Corpora available at LDC, ISI, other places
  - e.g., Europarl http://www.isi.edu/~koehn/europarl/
- Giza++ toolkit available at RWTH Aachen
  http://www-i6.informatik.rwth-aachen.de/web/Software/GIZA++.html
- Language model available at SRI
- Pharaoh decoder available at ISI
  - http://www.isi.edu/licensed-sw/pharaoh/
Output of Chinese-English System

In the First Two Months Guangdong’s Export of High-Tech Products 3.76 Billion US Dollars
Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong’s export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province’s total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong’s export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province’s total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion US dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem
Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui fbwer nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

MIT statistical MT system [Koehn, 2004], NIST Eval 2002 test set, about 100 million words training data
Partially Excellent Translations

In the First Two Months Guangdong’s Export of High-Tech Products 3.76 Billion US Dollars
Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong’s export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province’s total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong’s export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province’s total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion US dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem
Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui fbwer nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Partially excellent translations
In the First Two Months Guangdong’s Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong’s export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province’s total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong’s export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province’s total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion US dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui fbwer nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Mangled Grammar