Building a Phrase-based SMT System

Graham Neubig & Kevin Duh
Nara Institute of Science and Technology (NAIST)
5/10/2012
Phrase-based Statistical Machine Translation (SMT)

- Divide sentence into patterns, reorder, combine

Today I will give a lecture on machine translation.

今日は、機械翻訳の講義を行います。

- Statistical translation models, reordering models, language models learned from text
This Talk

1) What are the **steps** required to build a phrase-based machine translation system?

2) What **tools** implement these steps in Moses* (an open-source statistical MT system)?

3) What are some **research problems** related to each of these components?

* http://www.statmt.org/moses
Steps in Training a Phrase-based SMT System

- Collecting Data
- Tokenization
- Language Modeling
- Alignment
- Phrase Extraction/Scoring
- Reordering Models
- Decoding
- Evaluation
- Tuning
Collecting Data
Collecting Data

• **Sentence parallel data**
  - Used in: **Translation model/Reordering model**

<table>
<thead>
<tr>
<th>これはペンです。</th>
<th>This is a pen.</th>
</tr>
</thead>
<tbody>
<tr>
<td>昨日は友達と食べた。</td>
<td>I ate with my friend yesterday.</td>
</tr>
<tr>
<td>象は花が長い。</td>
<td>Elephants' trunks are long.</td>
</tr>
</tbody>
</table>

• **Monolingual data (in the target language)**
  - Used in: **Language model**

<table>
<thead>
<tr>
<th>This is a pen.</th>
<th>I ate with my friend yesterday.</th>
</tr>
</thead>
</table>
| Elephants' trunks are long. | }
Good Data is

- Big! →
- Clean
- In the same domain as test data
Collecting Data

- For academic workshops, data is prepared for us!

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>TED</td>
<td>Lectures</td>
<td>1.76M</td>
</tr>
<tr>
<td>News Commentary</td>
<td>News</td>
<td>2.52M</td>
</tr>
<tr>
<td>EuroParl</td>
<td>Political</td>
<td>45.7M</td>
</tr>
<tr>
<td>UN</td>
<td>Political</td>
<td>301M</td>
</tr>
<tr>
<td>Giga</td>
<td>Web</td>
<td>576M</td>
</tr>
</tbody>
</table>

e.g. IWSLT 2011 →

- In real systems
  - Data from government organizations, newspapers
  - Crawl the web
  - Merge several data sources
Research

- Finding bilingual pages [Resnik 03]
Research

- Finding bilingual pages [Resnik 03]
- Sentence alignment [Moore 02]

Editorial: Aging society does not necessarily spell doom

Longevity is something to be celebrated, but when it comes to the aging of Japanese society, it is often discussed in a pessimistic tone.

One reason for this is the continuing decline in people of working age. Learning that our society is shifting from one in which four working people financially support one senior citizen, to another in which each working person must support one senior citizen -- a so-called "piggyback" setup -- would make anyone anxious. And indeed, that is exactly what is happening.
Research

- Finding bilingual pages [Resnik 03]
- Sentence alignment [Moore 02]
- Crowd-sourcing data creation [Ambati 10]
- Mechanical Turk, duolingo, etc.
Tokenization
Tokenization

- **Example:** Divide Japanese into words
  
  太郎が花子を訪問した。
  
  太郎 が 花子 を 訪問 した 。

- **Example:** Make English lowercase, split punctuation
  
  Taro visited Hanako.
  
  taro visited hanako .
Tools for Tokenization

- **Most European languages**
  
  ```
  tokenize.perl en < input.en > output.en
  tokenize.perl fr < input.fr > output.fr
  ```

- **Japanese**
  
  ```
  MeCab: mecab -O wakati < input.ja > output.ja
  KyTea: kytea -notags < input.ja > output.ja
  JUMAN, etc.
  ```

- **Chinese**
  
  Stanford Segmenter, LDC, KyTea, etc...
Research

- What is good tokenization for machine translation?
  - Accuracy? Consistency? [Chang 08]
  - Matching target language words? [Sudoh 11]

  太郎 が 花子 を 訪問 した。

  Taro <ARG1> visited <ARG2> Hanako.

- Morphology (Korean, Arabic, Russian) [Niessen 01]

  단어란 도대체 무엇일까요?

  단어 란 도대체 무엇 일 까요?

- Unsupervised learning [Chung 09, Neubig 12]
Language Modeling
Language Modeling

- Assign a probability to each sentence

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>E1: Taro visited Hanako</td>
<td>P(E1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E2: the Taro visited the Hanako</td>
<td>LM</td>
<td>P(E2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E3: Taro visited the bibliography</td>
<td>P(E3)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- More fluent sentences get higher probability

\[
P(E1) > P(E2) \quad \text{and} \quad P(E1) > P(E3)\]
n-gram Models

• We want the probability of

\[ P(W = \text{“Taro visited Hanako”}) \]

• n-gram model calculates one word at a time

  • Condition on n-1 previous words
    e.g. 2-gram model

\[ P(w_1 = \text{“Taro”}) \ast P(w_2 = \text{“visited”} \mid w_1 = \text{“Taro”}) \]

  \[ \ast P(w_3 = \text{“Hanako”} \mid w_2 = \text{“visited”}) \]

  \[ \ast P(w_4 = \text{”</s>”} \mid w_3 = \text{“Hanako”}) \]

NOTE:
sentence ending symbol </s>
Tools

- SRILM Toolkit:
  
  Train:
  
  \text{ngram-count} \text{ -order 5 -interpolate -kndiscount -unk}
  \text{-text input.txt -lm lm.arpa}

  Test:
  
  \text{ngram -lm lm.arpa -ppl test.txt}

- Others: KenLM, RandLM, IRSTLM
Research Problems

- Is there anything that can beat n-grams? [Goodman 01]
  - Fast to compute
  - Easy to integrate into decoding
  - Surprisingly strong
- Other methods
  - Syntactic LMs [Charniak 03]
  - Neural networks [Bengio 06]
  - Model M [Chen 09]
  - etc...
Alignment
Alignment

- **Find which words correspond to each-other**

太郎 が 花子 を 訪問 した。

| taro visited hanako |

- **Done automatically with probabilistic methods**

P( 花子 |hanako) = 0.99
P( 太郎 |taro) = 0.97
P(visited| 訪問 ) = 0.46
P(visited| した ) = 0.04
P( 花子 |taro) = 0.0001
IBM/HMM Models

- One-to-many alignment model

- IBM Model 1: No structure ("bag of words")
- IBM Models 2-5, HMM: Add more structure
Combining One-to-Many Alignments

• Several different heuristics
Tools

- **mkcls**: Find bilingual classes

  ホテル の 受付 → 35 49 12
  the hotel front desk → 23 35 12 19

- **GIZA++**: Find alignments using IBM models (uses classes from mkcls for smoothing)

  ホテル の 受付 + 35 49 12 → ホテル の 受付
  the hotel front desk + 23 35 12 19 → the hotel front desk

- **symal**: Combine alignments in both directions

- (Included in *train-model.perl* of Moses)
Research Problems

- Does alignment actually matter? [Aryan 06]
- Supervised alignment models [Fraser 06, Haghhighi 09]
- Alignment using syntactic structure [DeNero 07]
- Phrase-based alignment models [Marcu 02, DeNero 08]
Phrase Extraction
Phrase Extraction

- Use alignments to find phrase pairs

ホテルの受付 → hotel front desk
ホテルの受付 → the hotel front desk
ホテルの → hotel
ホテルの → the hotel
受付 → front desk
Phrase Scoring

• Calculate 5 standard features
  • **Phrase Translation Probabilities:**
    \[ P(f|e) = \frac{c(f,e)}{c(e)} \quad P(e|f) = \frac{c(f,e)}{c(f)} \]
    e.g. \( \frac{c(\text{ホテル の, the hotel})}{c(\text{the hotel})} \)
  • **Lexical Translation Probabilities**
    – Use word-based translation probabilities (IBM Model 1)
    – Helps with sparsity
    \[ P(f|e) = \prod_f \frac{1}{|e|} \sum_e P(f|e) \]
    e.g.
    \[ (P(\text{ホテル |the})+P(\text{ホテル |hotel}))/2 \times (P(\text{の |the})+P(\text{の |hotel}))/2 \]
  • **Phrase penalty:** 1 for each phrase
Tools

- **extract**: Extract all the phrases
- **phrase-extract/score**: Score the phrases
- (Included in train-model.perl)
Research

- Domain adaptation of translation models [Koehn 07, Matsoukas 09]
- Reducing phrase table size [Johnson 07]
- Generalized phrase extraction (Geppetto toolkit) [Ling 10]
- Phrase sense disambiguation [Carpuat 07]
Reordering Models
Lexicalized Reordering

- Probability of monotone, swap, discontinuous

```
the thin man visited Taro
```

細い → the thin

高 monotone probability

太郎 を → Taro

高 swap probability

- Conditioning on input/output, left/right, or both
Tools

- **extract**: Same as phrase extraction
- **lexical-reordering/score**: Scores lexical reordering
- (included in train-model.perl)
Research

- Still a very open research area (especially en ↔ ja)
- Change the translation model
  - Hierarchical phrase-based [Chiang 07]
  - Syntax-based translation [Yamada 01, Galley 06]
- Pre-ordering [Xia 04, Isozaki 10]
Decoding
Decoding

- Given the models, find the best answer (or n-best)

- Exact search is NP-hard! [Knight 99]
- Decoding uses beam-search to find an approximate solution [Koehn 03]
**Tools**

- **Moses!**
  
  moses -f moses.ini < input.txt > output.txt

- **Also:** moses_chart, cdec (for Hiero, syntax-based models)
Research

- Decoding for lattice input [Dyer 08]
- Decoding for syntax models [Mi 08]
- Minimum Bayes risk decoding [Kumar 04]
- Exact decoding [Germann 01]
Evaluation
Human Evaluation

- **Adequacy**: Is the meaning correct?
- **Fluency**: Is the sentence natural?
- **Pairwise**: Is X a better translation than Y?

太郎が花子を訪問した
Taro visited Hanako
the Taro visited the Hanako
Hanako visited Taro

<table>
<thead>
<tr>
<th>Adequate</th>
<th>Fluent</th>
<th>Better</th>
</tr>
</thead>
<tbody>
<tr>
<td>○</td>
<td>○</td>
<td>B, C</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>C</td>
</tr>
<tr>
<td>X</td>
<td>○</td>
<td></td>
</tr>
</tbody>
</table>
Automatic Evaluation

- How well does the translation match a reference?
  - (or multiple references: more than one correct translation)
- **BLEU**: n-gram precision, brevity penalty [Papineni 03]

Reference: Taro visited Hanako
System: the Taro visited the Hanako

Brevity: min(1, |System|/|Reference|) = min(1, 5/3)  
brevity penalty = 1.0

\[
\text{BLEU-2} = (\frac{3}{5} \times \frac{1}{4})^{1/2} \times 1.0 \\
= 0.387
\]

- Also **METEOR** (normalizes synonyms), **TER** (# of changes), **RIBES** (reordering)
Research

- **Metrics with focus** on a particular thing
  - Reordering [Isozaki 10]
  - Accuracy of meaning [Lo 11]
- **Tunable** metrics [Cer 10]
- **Metric aggregation** [Albrecht 07]
- **Crowdsourcing** human evaluation [Callison-Burch 11]
Tuning
## Tuning

- **Scores** of translation, reordering, and language models

<table>
<thead>
<tr>
<th></th>
<th>LM</th>
<th>TM</th>
<th>RM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taro visited Hanako</td>
<td>-4</td>
<td>-3</td>
<td>-1</td>
</tr>
<tr>
<td>the Taro visited the Hanako</td>
<td>-5</td>
<td>-4</td>
<td>-1</td>
</tr>
<tr>
<td>Hanako visited Taro</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
</tr>
</tbody>
</table>

**Best Score**

- ○ Taro visited Hanako: -2.2
- × the Taro visited the Hanako: -2.7
- × Hanako visited Taro: -2.3

- If we **add weights**, we can get better answers:

<table>
<thead>
<tr>
<th></th>
<th>LM</th>
<th>TM</th>
<th>RM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taro visited Hanako</td>
<td>0.2*4</td>
<td>0.3*3</td>
<td>0.5*1</td>
</tr>
<tr>
<td>the Taro visited the Hanako</td>
<td>0.2*5</td>
<td>0.3*4</td>
<td>0.5*1</td>
</tr>
<tr>
<td>Hanako visited Taro</td>
<td>0.2*2</td>
<td>0.3*3</td>
<td>0.5*2</td>
</tr>
</tbody>
</table>

**Best Score**

- ○ Taro visited Hanako: -2.2

- Tuning finds these weights: $w_{LM} = 0.2$, $w_{TM} = 0.3$, $w_{RM} = 0.5$
Tuning Methods

- Minimum error rate training: MERT [Och 03]

- Others: MIRA [Watanabe 07] (online update), PRO (ranking) [Hopkins 11]
Research

- Tuning with millions of features (e.g. MIRA, PRO)
- Tuning with lattices [Macherey 08]
- Speeding up tuning [Suzuki 11]
- Tuning with multiple metrics [Duh 12]
Last Words
Last Words

- MT is fun! Join us.
- Improving very quickly, but still many problems.
- System is big, but you can focus on one problem.

Thank You
Bibliography


