SMT – Final thoughts

David Kauchak
CS159 – Fall 2014

Admin

Assignment 4b graded (except for 2 of them)
Assignment 6
MT lab on Thursday in Edmunds 105

Language translation

If we had the alignments

\[ p(f|e) \]

If we estimate this from a corpus, what does this represent?

Probability that \( f \) is aligned to \( e \) in the corpus.
If we had the alignments

\[ p(f \mid e) = \mathbb{I} \]

If we have the alignments, how do we estimate this?

Number of times \( f \) is aligned to \( e \) in the corpus

\[
p(f \mid e) = \frac{\text{count}(f \rightarrow e)}{\text{count}(e)}
\]

Input: corpus of English/Foreign sentence pairs along with alignment

for \((E, F)\) in corpus:
    for \(e\) in \(E\):
        for \(f\) in \(F\):
            if \(f\) aligned-to \(e\) in \((E,F)\):
                \(\text{count}(e,f)++\)
                \(\text{count}(e)++\)
        for all \((e,f)\) in \(\text{count}\):
            \(p(f\mid e) = \text{count}(e,f) / \text{count}(e)\)
Without the alignments

With alignments:

Without alignments:

Instead of actual counts, use "expected counts"

Probability of alignment

Without the alignments

With alignments:

Without alignments:

Probability of alignment

Without the alignments

With alignments:

Without alignments:

No. \( p(f_2 | e_2) \) is over the whole corpus!
Probability of alignment

\[ p(f \mid e) = \frac{\sum_{(f \rightarrow e) \in (E,F)} p(f \rightarrow e)}{\sum_{(f \rightarrow e) \in (E,F)}} \]

In this example, there are three options.

\[ p(f_2 \rightarrow e_2) : \text{Over all options, how likely does} \]
\[ \text{the model think it is to align } f_2 \text{ to } e_2. \]

How do we calculate this value?

\[ p(f_2 \rightarrow e_2) = \frac{p(f_2 \mid e_2)}{p(f_2 \mid e_1) + p(f_2 \mid e_2) + p(f_2 \mid e_3)} \]

Without the alignments

Input: corpus of English/Foreign sentence pairs along with alignment

for (E, F) in corpus:
  for e in E:
    for f in F:
      \[ p(f \rightarrow e) = \frac{\sum p(f \mid e)}{\sum p(f \mid e)} \]
      count(e,f) += p(f \rightarrow e)
      count(e) += p(f \rightarrow e)

for all (e,f) in count:
  \[ p(f \mid e) = \frac{\text{count}(e,f)}{\text{count}(e)} \]

Benefits of word-level model

Rarely used in practice for modern MT system

Mary did not slap the green witch

Maria no dió una botefada a la bruja verde

Two key side effects of training a word-level model:
• Word-level alignment
• \( p(f \mid e) \): translation dictionary

How do I get this?
Word alignment

100 iterations:

<table>
<thead>
<tr>
<th></th>
<th>0.005</th>
<th>0.995</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(casa</td>
<td>green)</td>
<td></td>
</tr>
<tr>
<td>p(VERDE</td>
<td>house)</td>
<td></td>
</tr>
<tr>
<td>p(VERDE</td>
<td>house)</td>
<td></td>
</tr>
<tr>
<td>p(la</td>
<td>casa</td>
<td>0.005</td>
</tr>
</tbody>
</table>

How should these be aligned?

Why?

Word-level alignment

alignment(E,F) = \arg_{i,j} \max p(A,F \mid E)

Which for IBM model 1 is:

alignment(E,F) = \arg_{i,j} \max \prod_{n=1}^{m} p(f_i \mid e_n)

Given a trained model (i.e. p(f|e) values), how do we find this?

Align each foreign word (f in F) to the English word (e in E) with highest p(f|e)

a_i = \arg_{f_i \in F} \max p(f_i \mid e_j)

Word-alignment Evaluation

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

How good of an alignment is this?

How can we quantify this?
The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

How can we quantify this?

System:

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

Human:

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

Precision: $\frac{6}{7}$  
Recall: $\frac{6}{10}$

Problems for Statistical MT

Preprocessing
Language modeling
Translation modeling
Decoding
Parameter optimization
Evaluation
What kind of Translation Model?

- Word-level models
- Phrasal models
- Syntactic models
- Semantic models

Mary did not slap the green witch

Maria no dio una botefada a la bruja verde

Phrase-Based Statistical MT

Generative story has three steps:
1. Foreign input segmented into phrases
   - “phrase” is any sequence of words
2. Each phrase is probabilistically translated into English
   - P(to the conference | zur Konferenz)
   - P( into the meeting | zur Konferenz)
3. Phrases are probabilistically re-ordered

Morgen fliege ich nach Kanada zur Konferenz

Tomorrow I will fly to the conference in Canada

Advantages?

Advantages of Phrase-Based

- Many-to-many mappings can handle non-compositional phrases
- Easy to understand
- Local context is very useful for disambiguation
  - “Interest rate” → ...
  - “Interest in” → ...
- The more data, the longer the learned phrases
  - Sometimes whole sentences!
These 7 people include astronauts coming from France and Russia.

Syntax-based models

Benefits?

- Can use syntax to motivate word/phrase movement
- Could ensure grammaticality

Two main types:
- $p(\text{foreign string | English parse tree})$
- $p(\text{foreign parse tree | English parse tree})$

Tree to string rule examples

1. DT(thes) $\rightarrow$ 这
2. VBP(include) $\rightarrow$ 中包括
3. VBP(include) $\rightarrow$ 中包括
4. NNP(France) $\rightarrow$ 法国
5. CC(and) $\rightarrow$ 和
6. NNP(Russia) $\rightarrow$ 俄罗斯
7. IN(of) $\rightarrow$ 的
8. NNP(NNS(astronauts)) $\rightarrow$ 宇航员
9. PUNC(,) $\rightarrow$
10. NNP(x0:DT, CD(7), NNS(people)) $\rightarrow$ x0, 7人
11. VBP(coming), PP(In(from), x0:NPs) $\rightarrow$ 来自, x0
12. IN(from) $\rightarrow$ 来自
13. NNP(x0:NPs, x1:CC, x2:NPs) $\rightarrow$ x0, x1, x2
14. VBP(x0:VBP, x1:NPs) $\rightarrow$ x0, x1
15. S(x0:NP, x1:VP, x2:PUNC) $\rightarrow$ x0, x1, x2
16. NNP(x0:NPs, x1:VP) $\rightarrow$ x1, 的, x0
17. NP(DT("the"), x0:JJ, x1:NN) $\rightarrow$ x0, x1

Contiguous phrase pair substitution rules (alignment templates)

Higher-level rules
Decoding
Of all conceivable English word strings, find the one maximizing $P(e) \times P(f \mid e)$

Decoding is an NP-complete problem! (for many translation models)

Several decoding strategies are often available
These 7 people include astronauts coming from France and Russia.
The Problem: Learn Lambdas

\[
p(c|f) = \frac{p(f|c)p(c)}{p(f)}
\]

where

\[
p(f|c)^p p(c)^p p(e| f)^p \text{ length}(e)^p ...
\]

\[
\sum p(f|c)^p p(c)^p p(e| f)^p \text{ length}(e)^p ...
\]

\[
\exp(\lambda \log p(f|c) + \lambda_2 \log p(e) + \lambda_3 \log p(c|f) + \lambda_4 \text{ length}(c) ...)
\]

\[
\sum \exp(\lambda \log p(f|c) + \lambda_2 \log p(e) + \lambda_3 \log p(c|f) + \lambda_4 \text{ length}(c) ...)
\]

\[
- \frac{\exp \left( \sum \lambda_h(f, c) \right)}{\sum \exp \left( \sum \lambda_h(f, c') \right)}
\]

How should we optimize these?

---

Problems for Statistical MT

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

---

MT Evaluation

How do we do it?

What data might be useful?

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

* maximize the likelihood of the data
* maximize an evaluation criterion
MT Evaluation

Source only

Manual:
- SSER (subjective sentence error rate)
- Correct/Incorrect
- Error categorization

Extrinsic:
Objective usage testing

Automatic:
- WER (word error rate)
- BLEU (Bilingual Evaluation Understudy)
- NIST

Automatic Evaluation

Common NLP/machine learning/AI approach

Training sentence pairs

Testing sentence pairs

Automatic Evaluation

Reference (human) translation: 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.

Machine translation: The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail . which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack . [?] highly alerts after the maintenance.

Machine translation 2: United States Office of the Guam International Airport and were received by a man claiming to be Saudi Arabian businessman Osama bin Laden, sent emails, threats to airports and other public places will launch a biological or chemical attack, remain on high alert in Guam.

BLEU Evaluation Metric (Papineni et al, ACL-2002)

Basic idea:
Combination of n-gram precisions of varying size

What percentage of machine n-grams can be found in the reference translation?

Reference (human) translation: 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.

Machine translation: The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail . which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack . [?] highly alerts after the maintenance.

Machine translation 2: United States Office of the Guam International Airport and were received by a man claiming to be Saudi Arabian businessman Osama bin Laden, sent emails, threats to airports and other public places will launch a biological or chemical attack, remain on high alert in Guam.
Candidate 1: It is a guide to action that ensures that the military always obey the commands of the party.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.
Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.
Reference 3: It is the practical guide for the army always to heed directions of the party.

What percentage of machine n-grams can be found in the reference translations? Do unigrams, bigrams and trigrams.
Candidate 1: It is a guide to action which ensures that the military always obey the commands of the party.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.
Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.
Reference 3: It is the practical guide for the army always to heed directions of the party.

Unigrams: 17/18
Bigrams: 10/17
Trigrams: 7/16

Candidate 2: It is to ensure the army forever hearing the directions guide that party commands.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.
Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.
Reference 3: It is the practical guide for the army always to heed directions of the party.

Unigrams: 12/14
Bigrams: 4/13
N-gram precision example 2

Candidate 2: It is to ensure the army forever hearing the directions guide that party commands.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.
Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.
Reference 3: It is the practical guide for the army always to heed directions of the party.

Unigrams: 12/14
Bigrams: 4/13
Trigrams: 1/12

N-gram precision

Candidate 1: It is a guide to action which ensures that the military always obey the commands of the party.

Unigrams: 17/18
Bigrams: 10/17
Trigrams: 7/16

Candidate 2: It is to ensure the army forever hearing the directions guide that party commands.

Unigrams: 12/14
Bigrams: 4/13
Trigrams: 1/12

Any problems/concerns?

N-gram precision example

Candidate 3: the
Candidate 4: It is a

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.
Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.
Reference 3: It is the practical guide for the army always to heed directions of the party.

What percentage of machine n-grams can be found in the reference translations? Do unigrams, bigrams and trigrams.

BLEU Evaluation Metric
(Papineni et al, ACL-2002)

N-gram precision (score is between 0 & 1)
- What percentage of machine n-grams can be found in the reference translation?

- Not allowed to use same portion of reference translation twice (can’t cheat by typing out "the the the the")

Brevity penalty
- Can’t just type out single word "the" (precision 1.0!)

*** Amazingly hard to "game" the system (i.e., find a way to change machine output so that BLEU goes up, but quality doesn’t)
BLEU Tends to Predict Human Judgments

\[ R^2 = 88.0\% \]

\[ R^2 = 90.2\% \]

<table>
<thead>
<tr>
<th>Human Judgments</th>
<th>NIST Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adequacy</td>
<td>Fluency</td>
</tr>
<tr>
<td>Linear (Adequacy)</td>
<td>Linear (Fluency)</td>
</tr>
</tbody>
</table>

slide from G. Doddington (NIST)

BLEU in Action

<table>
<thead>
<tr>
<th>(Foreign Original)</th>
<th>(Reference Translation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>the gunman was shot to death by the police.</td>
<td>#1 the gunman was police kill.</td>
</tr>
<tr>
<td>wounded police jaya of</td>
<td>#2 wounded police jaya of</td>
</tr>
<tr>
<td>the gunman was shot dead by the police.</td>
<td>#3 the gunman was shot dead by the police.</td>
</tr>
<tr>
<td>the gunman arrested by police kill.</td>
<td>#4 the gunman arrested by police kill.</td>
</tr>
<tr>
<td>the gunmen were killed.</td>
<td>#5 the gunmen were killed.</td>
</tr>
<tr>
<td>the gunman was shot to death by the police.</td>
<td>#6 the gunman was shot to death by the police.</td>
</tr>
<tr>
<td>al by the police.</td>
<td>#7 al by the police.</td>
</tr>
<tr>
<td>the ringer is killed by the police.</td>
<td>#8 the ringer is killed by the police.</td>
</tr>
<tr>
<td>police killed the gunman.</td>
<td>#9 police killed the gunman.</td>
</tr>
<tr>
<td>police killed the gunman.</td>
<td>#10 police killed the gunman.</td>
</tr>
</tbody>
</table>

green = 4-gram match (good!)
red = word not matched (bad!)

BLEU in Action

<table>
<thead>
<tr>
<th>(Foreign Original)</th>
<th>(Reference Translation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>the gunman was shot to death by the police.</td>
<td>#1 the gunman was police kill.</td>
</tr>
<tr>
<td>wounded police jaya of</td>
<td>#2 wounded police jaya of</td>
</tr>
<tr>
<td>the gunman was shot dead by the police.</td>
<td>#3 the gunman was shot dead by the police.</td>
</tr>
<tr>
<td>the gunman arrested by police kill.</td>
<td>#4 the gunman arrested by police kill.</td>
</tr>
<tr>
<td>the gunmen were killed.</td>
<td>#5 the gunmen were killed.</td>
</tr>
<tr>
<td>the gunman was shot to death by the police.</td>
<td>#6 the gunman was shot to death by the police.</td>
</tr>
<tr>
<td>al by the police.</td>
<td>#7 al by the police.</td>
</tr>
<tr>
<td>the ringer is killed by the police.</td>
<td>#8 the ringer is killed by the police.</td>
</tr>
<tr>
<td>police killed the gunman.</td>
<td>#9 police killed the gunman.</td>
</tr>
<tr>
<td>police killed the gunman.</td>
<td>#10 police killed the gunman.</td>
</tr>
</tbody>
</table>

green = 4-gram match (good!)
red = word not matched (bad!)
BLEU: Problems?

Doesn’t care if an incorrectly translated word is a name or a preposition
- gave it to Albright (reference)
- gave it at Albright (translation #1)
- gave it at altar (translation #2)

What happens when a program reaches human level performance in BLEU but the translations are still bad?
- maybe sooner than you think …

11 Human Translation Agencies Employed to Translate 100 Chinese News Articles

上个星期的战斗至少夺取了12个人的生命。
At least 12 people were killed in the battle last week.

Last week’s fight took at least 12 lives.
The fighting last week killed at least 12.
The battle of last week killed at least 12 persons.
At least 12 people lost their lives in last week’s fighting.
At least 12 persons died in the fighting last week.
At least 12 died in the battle last week.
At least 12 people were killed in the fighting last week.
During last week’s fighting, at least 12 people died.
Last week at least twelve people died in the fighting.
Last week’s fighting took the lives of twelve people.

Merging Translations
(Pang, Knight, and Marcu, NAACL-HLT 2003)

Create word graphs by merging paraphrases
⇒ from 10 sentences to over a thousand

11th human translation is often found in the graph!