SMT – Final thoughts

David Kauchak
CS159 – Spring 2019
What does being NP-complete imply?

Some slides adapted from
Philipp Koehn
School of Informatics
University of Edinburgh

Kevin Knight
UCL Institute of Computing Science

Dan Klein
Computer Science Department
UC Berkeley

Language translation
Yo quiero Taco Bell

https://www.youtube.com/watch?v=Q6jd_Oy29IQ
https://www.youtube.com/watch?v=mULShT9z2jI

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
• \( p(f | e) \): translation dictionary

How do I get this?
Word alignment

100 iterations

<table>
<thead>
<tr>
<th>Alignment</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>casa</td>
<td>green</td>
</tr>
<tr>
<td>verde</td>
<td>green</td>
</tr>
<tr>
<td>la</td>
<td>green</td>
</tr>
<tr>
<td>casa</td>
<td>house</td>
</tr>
<tr>
<td>verde</td>
<td>house</td>
</tr>
<tr>
<td>la</td>
<td>house</td>
</tr>
<tr>
<td>casa</td>
<td>the</td>
</tr>
<tr>
<td>verde</td>
<td>the</td>
</tr>
<tr>
<td>la</td>
<td>the</td>
</tr>
</tbody>
</table>

How should these be aligned?

casa verde

Verde house

la casa

Why?

green house
casa verde

the house
la casa

Word-level alignment

alignment(E, F) = argₙ maxₙ p(Aₙ, F | E)

Which for IBM model 1 is:

alignment(E, F) = argₙ maxₙ p(fᵢ | eᵣ)

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)

aᵣ = argₙ maxₙ p(fᵢ | eᵣ)
Word-alignment Evaluation

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.

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 and recall!

Word-alignment Evaluation

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 bofetada a la bruja verde

Phrasal translation model

The models define probabilities over inputs

\[
p(f \mid e)
\]

1. Sentence is divided into phrases

Phrasal translation model

The models define probabilities over inputs

\[
p(f \mid e)
\]

1. Sentence is divided into phrases
2. Phrases are translated (avoids a lot of weirdness from word-level model)
3. Phrases are reordered
Phrase table

<table>
<thead>
<tr>
<th>Translation</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>of course</td>
<td>0.5</td>
</tr>
<tr>
<td>naturally</td>
<td>0.3</td>
</tr>
<tr>
<td>of course,</td>
<td>0.15</td>
</tr>
<tr>
<td>, of course,</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Phrase table

<table>
<thead>
<tr>
<th>Translation</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>the proposal</td>
<td>0.6227</td>
</tr>
<tr>
<td>'s proposal</td>
<td>0.1068</td>
</tr>
<tr>
<td>a proposal</td>
<td>0.0341</td>
</tr>
<tr>
<td>the idea</td>
<td>0.0250</td>
</tr>
<tr>
<td>this proposal</td>
<td>0.0227</td>
</tr>
<tr>
<td>proposal</td>
<td>0.0205</td>
</tr>
<tr>
<td>of the proposal</td>
<td>0.0159</td>
</tr>
<tr>
<td>the proposals</td>
<td>0.0159</td>
</tr>
<tr>
<td>the suggestions</td>
<td>0.0114</td>
</tr>
</tbody>
</table>

Phrasal translation model

The models define probabilities over inputs 

\[ p(f \mid e) \]

Advantages?

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!
These 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:
- \( p(\text{foreign \textit{string}} \mid \text{English parse tree}) \)
- \( p(\text{foreign \textit{parse tree}} \mid \text{English parse tree}) \)

Why always English parse tree?

Tree to string rule

Tree to string rules examples

1. DT(these) \( \rightarrow \) 这
2. VBP(include) \( \rightarrow \) 中包括
3. VBP(includes) \( \rightarrow \) 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. NP(NNS(astronauts)) \( \rightarrow \) 宇航员，员
9. PUNC(.) \( \rightarrow \)
10. NP(x0:DT, CD(7), NNS(people)) \( \rightarrow \) x0, 7人
11. VP(VBG(coming), PP(IN(from), x0:NP)) \( \rightarrow \) 来自, x0
12. IN(from) \( \rightarrow \) 来自
13. NP(x0:NNP, x1:CC, x2:NNP) \( \rightarrow \) x0, x1, x2
14. VP(x0:VBP, x1:NP) \( \rightarrow \) x0, x1
15. S(x0:NP, x1:VP, x2:PUNC) \( \rightarrow \) x0, x1, x2
16. NP(x0:NP, x1:VP) \( \rightarrow \) x1, 的, x0
17. NP(DT("the"), x0:JJ, x1:NN) \( \rightarrow \) x0, x1

Contiguous phrase pair substitution rules

Higher-level rules
Tree to string rules examples

1. DT(these) → 这
2. VP(includes) → 由
3. VP(includes) → 由
4. NNP(France) → 法国
5. CC(and) → 和
6. NNP(Russia) → 俄罗斯
7. IN(of) → 的
8. NP(NNS(astronauts)) → 宇航员
9. PUNC(.,) → .
10. NP(x0:DT, CD(7), NNS(people) → 人, 7人
11. VP(VBG(coming), PPI(IN(from), x0:NP)) → 来自, x0
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC, x2:NNP) → 人, 7人
14. VP(VBP(von), x1:NP) → 人
15. S(x0:NP, x1:VP, x2:PUNC) → .
16. NP(x0:NP, x1:VP) → .
17. NP(DT("the"), x0:JJ, x1:NN) → .

Both VBP("include") and
VP("includes") will translate to "由" in Chinese.

Tree Transformations

1. DT(these) → 这
2. VP(includes) → 由
3. VP(includes) → 由
4. NNP(France) → 法国
5. CC(and) → 和
6. NNP(Russia) → 俄罗斯
7. IN(of) → 的
8. NP(NNS(astronauts)) → 宇航员
9. PUNC(.,) → .
10. VP(VBG(coming), PPI(IN(from), x0:NP)) → 来自, x0
11. IN(from) → 来自
12. NP(x0:DT, CD(7), NNS(people) → 人, 7人
13. VP(VBG(coming), PPI(IN(from), x0:NP)) → 来自, x0
14. NP(x0:DT("the"), x0:JJ, x1:NN) → .

The phrase "coming from" translates to "来自" only if followed by an NP (whose translation is then placed to the right of "来自").

Tree Transformations

1. DT(these) → 这
2. VP(includes) → 由
3. VP(includes) → 由
4. NNP(France) → 法国
5. CC(and) → 和
6. NNP(Russia) → 俄罗斯
7. IN(of) → 的
8. NP(NNS(astronauts)) → 宇航员
9. PUNC(.,) → .
10. VP(VBG(von), PPI(IN(from), x0:NP)) → 来自, x0
11. IN(from) → 来自
12. NP(x0:DT, CD(7), NNS(people) → 人, 7人
13. VP(VBG(von), PPI(IN(from), x0:NP)) → 来自, x0
14. NP(x0:DT("the"), x0:JJ, x1:NN) → .

To translate the JJ NN, just translate the JJ and then translate the NN (drop "the").
Decoding

Of all conceivable English word strings, find the one maximizing $P(e) \cdot P(f \mid e)$

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

What does this imply?

Decoding

Of all conceivable English word strings, find the one maximizing $P(e) \cdot P(f \mid e)$

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

- Not guaranteed to find the max

Many different approaches to decoding
These 7 people include astronauts coming from France and Russia.
The Problem: Learn Lambdas

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

\[ = \frac{\exp(\sum \lambda h(f, e))}{\sum \exp(\sum \lambda h(f, e'))} \]

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

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 against public places such as the airport.

Ideas?
Multiple Reference Translations

N-gram precision example

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.

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

N-gram precision example

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
N-gram precision example

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

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: 10/17
Trigrams: 7/16

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

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: 10/17
Trigrams: 7/16
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 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

$R^2 = 88.0\%$

$R^2 = 90.2\%$

BLEU in Action

Slide from G. Doddington (NIST)

<table>
<thead>
<tr>
<th>Machine</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>#1</td>
</tr>
<tr>
<td>#2</td>
<td>#2</td>
</tr>
<tr>
<td>#3</td>
<td>#3</td>
</tr>
<tr>
<td>#4</td>
<td>#4</td>
</tr>
<tr>
<td>#5</td>
<td>#5</td>
</tr>
<tr>
<td>#6</td>
<td>#6</td>
</tr>
<tr>
<td>#7</td>
<td>#7</td>
</tr>
<tr>
<td>#8</td>
<td>#8</td>
</tr>
<tr>
<td>#9</td>
<td>#9</td>
</tr>
<tr>
<td>#10</td>
<td>#10</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 to 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 …