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
USC/Information Sciences Institute
USC/Computer Science Department

Dan Klein
Computer Science Department
UC Berkeley

Admin

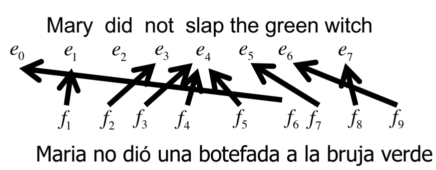
Assignment 6

Language translation

https://www.youtube.com/watch?v=Q6jzl_Oy2lQ
https://www.youtube.com/watch?v=vV1SkTdizZI

Benefits of word-level model

Rarely used in practice for modern MT system



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

p(casa | green)	0.005
p(verde | green)	0.995
p(la | green)	0

p(casa | house)	~1.0
p(verde | house)	~0.0
p(la | house)	~0.0

p(casa | the)	0.005
p(verde | the)	0
p(la | the)	0.995

green house

casa verde

How should these be aligned?

the house

la casa

Word alignment

100 iterations

p(casa | green)	0.005
p(verde | green)	0.995
p(la | green)	0

p(casa | house)	~1.0
p(verde | house)	~0.0
p(la | house)	~0.0

p(casa | the)	0.005
p(verde | the)	0
p(la | the)	0.995

green house

casa verde

Why?

the house

la casa

Word-level alignment

$$alignment(E,F) = \arg_A \max p(A,F \mid E)$$

Which for IBM model 1 is:

$$alignment(E,F) = \arg_A \max \prod_{i=1}^{|F|} p(f_i \mid e_{a_i})$$

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)

$$a_i = \arg_{j:1-|E|} \max p(f_i \mid e_j)$$

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?

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?

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 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

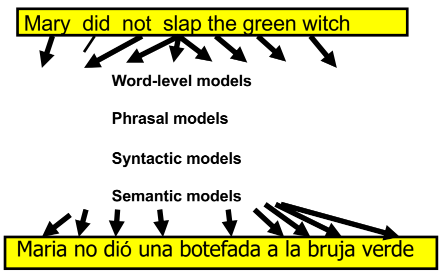
Translation modeling

Decoding

Parameter optimization

Evaluation

What kind of Translation Model?



Phrasal translation model

The models define probabilities over inputs

$$p(f \mid e)$$

Morgen | fliege | ich | nach Kanada | zur Konferenz

1. Sentence is divided into phrases

Phrasal translation model

The models define probabilities over inputs

$$p(f \mid e)$$

Morgen | fliege | ich | nach Kanada | zur Konferenz

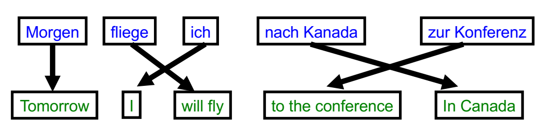
Tomorrow | will fly | I | In Canada | to the conference

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

Phrasal translation model

The models define probabilities over inputs

$$p(f \mid e)$$



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

Phrase table

natuerlich

Translation	Probability
of course	0.5
naturally	0.3
of course ,	0.15
, of course ,	0.05

Phrase table

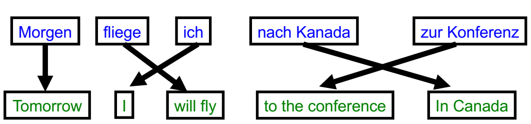
den Vorschlag

Translation	Probability
the proposal	0.6227
‘s proposal	0.1068
a proposal	0.0341
the idea	0.0250
this proposal	0.0227
proposal	0.0205
of the proposal	0.0159
the proposals	0.0159
the suggestions	0.0114
…	

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

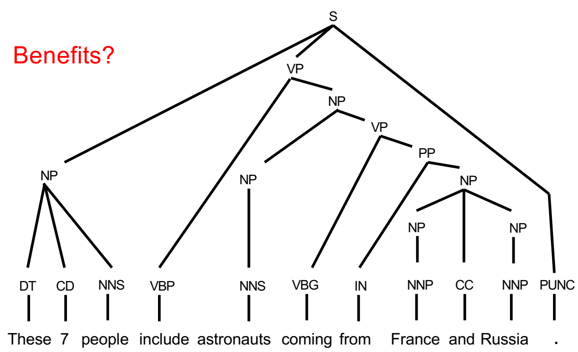
Easy to understand

Local context is very useful for disambiguating
- “Interest rate” → …
- “Interest in” → …

The more data, the longer the learned phrases
- Sometimes whole sentences!

Syntax-based models

Benefits?



Syntax-based models

Benefits

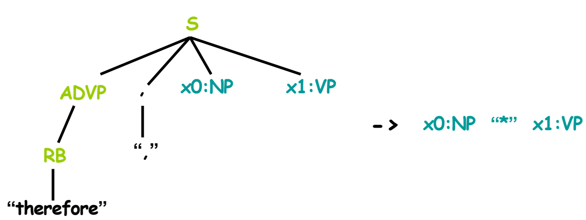
- Can use syntax to motivate word/phrase movement
- Could ensure grammaticality

Two main types:

- p(foreign *string* | English parse tree)
- p(foreign *parse tree* | English parse tree)

Why always English parse tree?

Tree to string rule



Tree to string rules examples

1. DT(these) → 这
2. VBP(include) → 中包括
3. VBP(includes) → 中包括
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) → x0 , 7人
11. VP(VBG(coming), PP(IN(from), x0:NP)) → 来自 ,x0
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC, x2:NNP) → x0 , x1 , x2
14. VP(x0:VBP, x1:NP) → x0 , x1
15. S(x0:NP, x1:VP, x2:PUNC) → x0 , x1, x2
16. NP(x0:NP, x1:VP) → x1 , 的 , x0
17. NP(DT(“the”), x0:JJ, x1:NN) → x0 , x1

Contiguous phrase pair substitution rules

Higher-level rules

Tree to string rules examples

1. DT(these) → 这
2. VBP(include) → 中包括
3. VBP(includes) → 中包括
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) → x0 , 7人
11. VP(VBG(coming), PP(IN(from), x0:NP)) → 来自 ,x0
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC, x2:NNP) → x0 , x1 , x2
14. VP(x0:VBP, x1:NP) → x0 , x1
15. S(x0:NP, x1:VP, x2:PUNC) → x0 , x1, x2
16. NP(x0:NP, x1:VP) → x1 , 的 , x0
17. NP(DT("the"), x0:JJ, x1:NN) → x0 , x1

Both VBP("include") and VBP("includes") will translate to "中包括" in Chinese.

Contiguous phrase pair substitution rules

Higher-level rules

Tree Transformations

1. DT(these) → 这
2. VBP(include) → 中包括
3. VBP(includes) → 中包括
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) → x0 , 7人
11. VP(VBG(coming), PP(IN(from), x0:NP)) → 来自 ,x0
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC, x2:NNP) → x0 , x1 , x2
14. VP(x0:VBP, x1:NP) → x0 , x1
15. S(x0:NP, x1:VP, x2:PUNC) → x0 , x1, x2
16. NP(x0:NP, x1:VP) → x1 , 的 , x0
17. NP(DT("the"), x0:JJ, x1:NN) → x0 , x1

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").

ase pair
es
plates)

Higher-level rules

Tree Transformations

1. DT(these) → 这
2. VBP(include) → 中包括
3. VBP(includes) → 中包括
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) → x0 , 7人
11. VP(VBG(coming), PP(IN(from), x0:NP)) → 来自 ,x0
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC, x2:NNP) → x0 , x1 , x2
14. VP(x0:VBP, x1:NP) → x0 , x1
15. S(x0:NP, x1:VP, x2:PUNC) → x0 , x1, x2
16. NP(x0:NP, x1:VP) → x1 , 的 , x0
17. NP(DT("the"), x0:JJ, x1:NN) → x0 , x1

Translate an English NP ("astronauts") modified by a gerund VP ("coming from France and Russia") as follows:
(1) translate the gerund VP,
(2) type the Chinese word "的",
(3) translate the NP.

hrase pair
Substitution rules
(alignment templates)

Higher-level rules

Tree Transformations

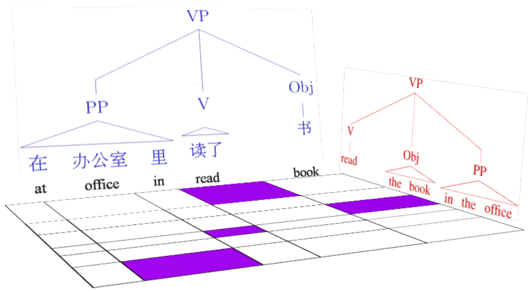
1. DT(these) → 这
2. VBP(include) → 中包括
3. VBP(includes) → 中包括
4. NNP(France) → 法国
5. CC(and) → 和
6. NNP(Russia) → 俄罗斯
7. IN(of) → 的
8. NP(NNS(astronauts)) → 宇航，员
9. PUNC(.) → .
10. NP(x0:DT, CD(7), NNS(people) → x0 , 7
11. VP(VBG(coming), PP(IN(from), x0:NP)) →
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC,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

Contiguous phrase pair
Substitution rules
(alignment templates)

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

Higher-level rules

Tree to tree example



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) * P(f | 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) * P(f | e)

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

– Not guaranteed to find the max

Many different approaches to decoding

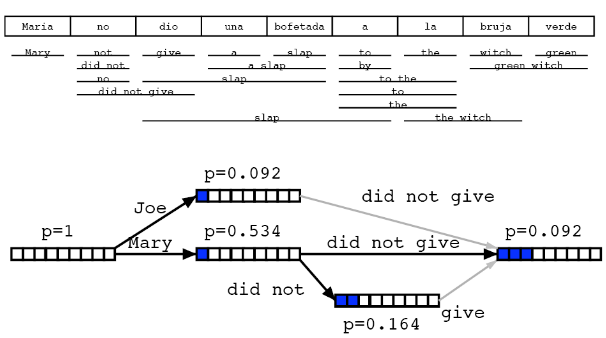
Phrase-Based Decoding

What is the best translation?

Phrase-Based Decoding

These 7 people include astronauts coming from France and Russia.

Hypotheis Lattices



Problems for Statistical MT

Preprocessing

Language modeling

Translation modeling

Decoding

Parameter optimization

Evaluation

The Problem: Learn Lambdas

$$p(e \mid f) = \frac{p(f \mid e)p(e)}{p(f)}$$

$$= \frac{p(f \mid e)^{\lambda_1} p(e)^{\lambda_2}}{\sum_{e'} p(f \mid e')^{\lambda_1} \lambda_2 p(e')^{\lambda_2}}$$

$$= \frac{p(f \mid e)^{\lambda_1} p(e)^{\lambda_2} p(e \mid f)^{\lambda_3} length(e)^{\lambda_4} \ldots}{\sum_{e'} p(f \mid e')^{\lambda_1} p(e')^{\lambda_2} p(e' \mid f)^{\lambda_3} length(e')^{\lambda_4} \ldots}$$

$$= \frac{\exp(\lambda_1 \log p(f \mid e) + \lambda_2 \log p(e) + \lambda_3 \log p(e \mid f) + \lambda_4 length(e) \ldots)}{\sum_{e'} \exp(\lambda_1 \log p(f \mid e') + \lambda_2 \log p(e') + \lambda_3 \log p(e' \mid f) + \lambda_4 length(e') \ldots)}$$

$$= \frac{\exp\left(\sum_i \lambda_i h_i(f,e)\right)}{\sum_{e'} \exp\left(\sum_i \lambda_i h_i(f,e')\right)}$$

How should we optimize these?

The Problem: Learn Lambdas

$$p(e \mid f) = \frac{p(f \mid e)p(e)}{p(f)}$$

$$= \frac{p(f \mid e)^{\lambda_1} p(e)^{\lambda_2}}{\sum_{e'} p(f \mid e')^{\lambda_1} \lambda_2 p(e')^{\lambda_2}}$$

$$= \frac{p(f \mid e)^{\lambda_1} p(e)^{\lambda_2} p(e \mid f)^{\lambda_3} length(e)^{\lambda_4} \ldots}{\sum_{e'} p(f \mid e')^{\lambda_1} p(e')^{\lambda_2} p(e' \mid f)^{\lambda_3} length(e')^{\lambda_4} \ldots}$$

$$= \frac{\exp(\lambda_1 \log p(f \mid e) + \lambda_2 \log p(e) + \lambda_3 \log p(e \mid f) + \lambda_4 length(e) \ldots)}{\sum_{e'} \exp(\lambda_1 \log p(f \mid e') + \lambda_2 \log p(e') + \lambda_3 \log p(e' \mid f) + \lambda_4 length(e') \ldots)}$$

$$= \frac{\exp\left(\sum_i \lambda_i h_i(f,e)\right)}{\sum_{e'} \exp\left(\sum_i \lambda_i h_i(f,e')\right)}$$

Given a data set with foreign/English sentences, find the λ's that:

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

Problems for Statistical MT

Preprocessing

Language modeling

Translation modeling

Decoding

Parameter optimization

Evaluation

MT Evaluation

How do we do it?

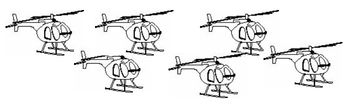
What data might be useful?

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

All sentence pairs → Training sentence pairs

All 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.

Ideas?

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

Reference (human) translation:
The U.S. island of Guam is maintaining a high state of alert <u>after the</u> Guam <u>airport and its</u> 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 <u>the airport</u> .

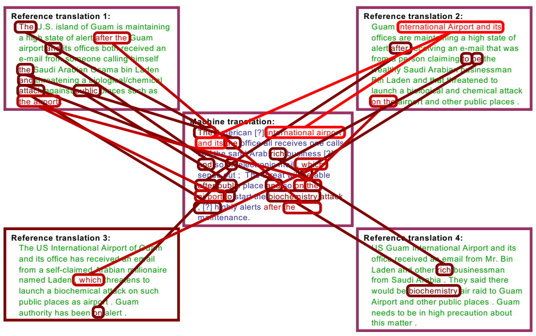
Machine translation:
The American [?] international <u>airport and its</u> the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on <u>the airport</u> to start the biochemistry attack , [?]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

Basic idea:

Combination of n-gram precisions of varying size

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

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

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.*

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

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: 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)

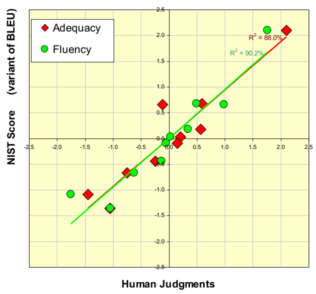
Reference (human) translation:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

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.

- 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



slide from G. Doddington (NIST)

BLEU in Action

枪手被警方击毙。	**(Foreign Original)**
the gunman was shot to death by the police .	**(Reference Translation)**
the gunman was police kill .	**#1**
wounded police jaya of	**#2**
the gunman was shot dead by the police .	**#3**
the gunman arrested by police kill .	**#4**
the gunmen were killed .	**#5**
the gunman was shot to death by the police .	**#6**
gunmen were killed by police ?SUB>0 ?SUB>0	**#7**
al by the police .	**#8**
the ringer is killed by the police .	**#9**
police killed the gunman .	**#10**

BLEU in Action

枪手被警方击毙。	**(Foreign Original)**
the gunman was shot to death by the police .	**(Reference Translation)**
the gunman was police kill .	**#1**
wounded police jaya of	**#2**
the gunman was shot dead by the police .	**#3**
the gunman arrested by police kill .	**#4**
the gunmen were killed .	**#5**
the gunman was shot to death by the police .	**#6**
gunmen were killed by police ?SUB>0 ?SUB>0	**#7**
al by the police .	**#8**
the ringer is killed by the police .	**#9**
police killed the gunman .	**#10**

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

BLEU in Action

枪手被警方击毙。	**(Foreign Original)**	
the gunman was shot to death by the police .	**(Reference Translation)**	
the gunman was police kill .	**#1**	Machine
wounded police jaya of	**#2**	Machine
the gunman was shot dead by the police .	**#3**	**Human**
the gunman arrested by police kill .	**#4**	Machine
the gunmen were killed .	**#5**	Machine
the gunman was shot to death by the police .	**#6**	**Human**
gunmen were killed by police ?SUB>0 ?SUB>0	**#7**	Machine
al by the police .	**#8**	Machine
the ringer is killed by the police .	**#9**	Machine
police killed the gunman .	**#10**	**Human**

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 …