CSCI 5582 Artificial Intelligence

Lecture 25

Jim Martin

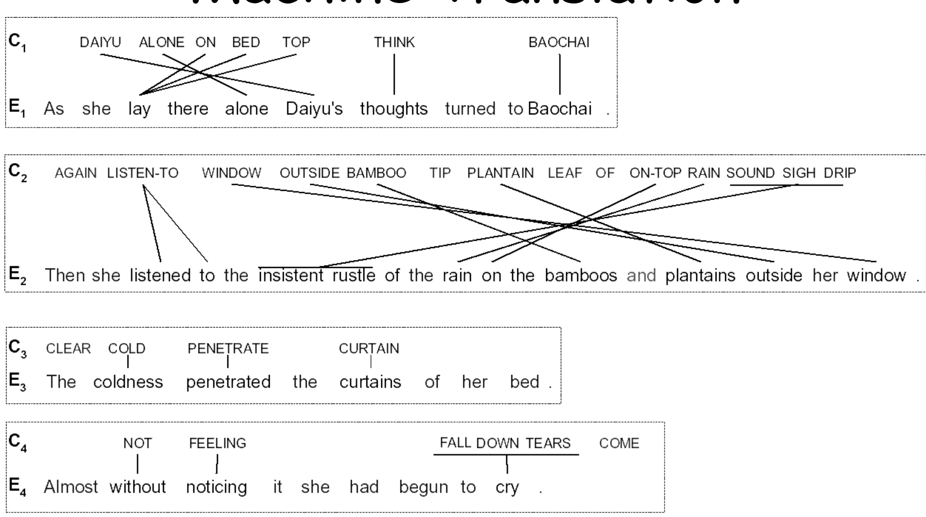
Today 12/5

- Machine Translation
  - Review MT
    - Models
    - Training
    - Decoding
  - Phrase-based models
  - Evaluation

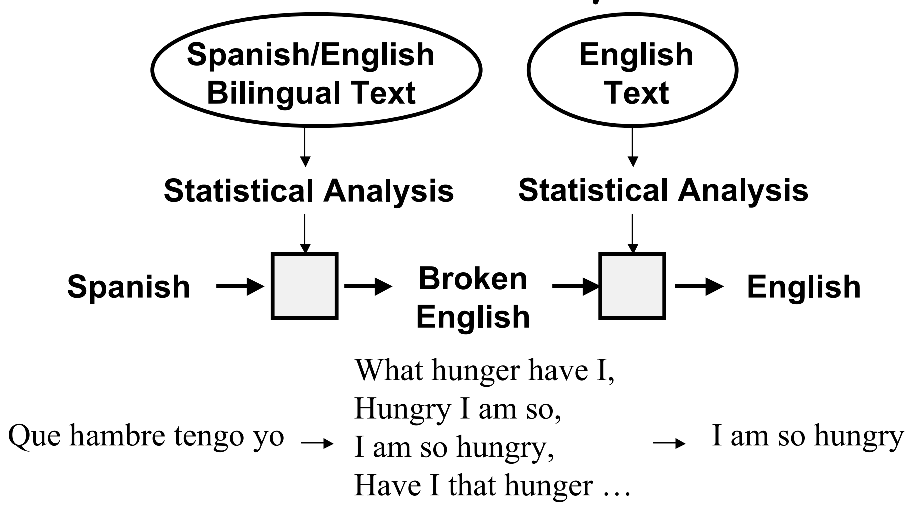
Readings

- Chapters 22 and 23 in Russell and Norvig
- Chapter 24 of Jurafsky and Martin

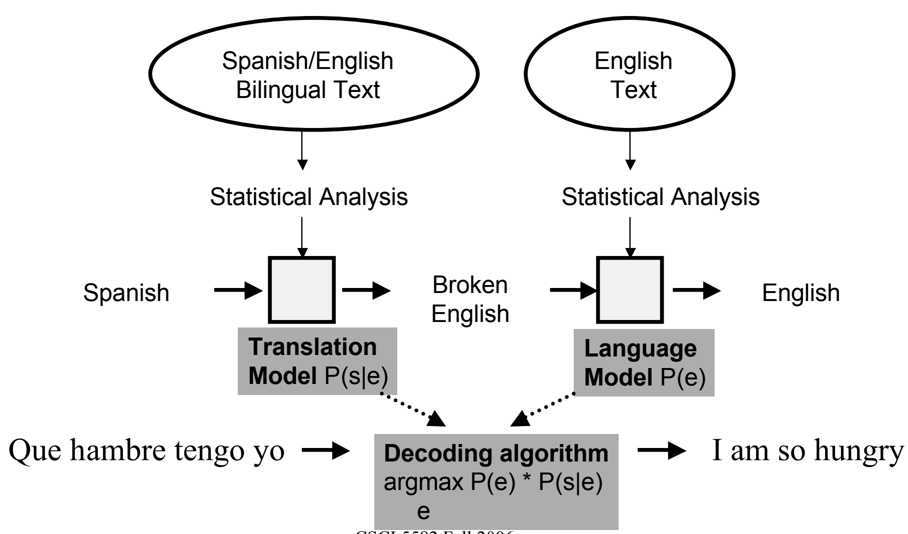
Machine Translation



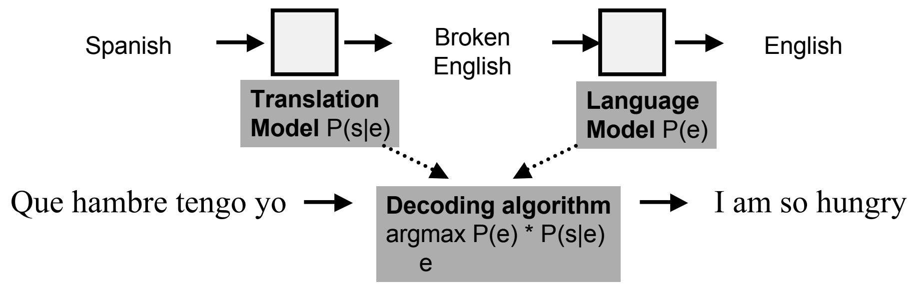
Statistical MT Systems



Statistical MT Systems



Bayes Rule



Given a source sentence s, the decoder should consider many possible translations … and return the target string e that maximizes

P(e | s)

By Bayes Rule, we can also write this as:

P(e) x P(s | e) / P(s)

and maximize that instead. P(s) never changes while we compare different e's, so we can equivalently maximize this: **P(e) x P(s | e)**

Four Problems for Statistical MT

- Language model
  - Given an English string e, assigns P(e) by the usual methods we've been using sequence modeling.
- Translation model
  - Given a pair of strings <f,e>, assigns P(f | e) again by making the usual markov assumptions
- Training
  - Getting the numbers needed for the models
- Decoding algorithm
  - Given a language model, a translation model, and a new sentence f … find translation e maximizing P(e) * P(f | e)

Language Model Trivia

- Google Ngrams data
  - Number of tokens:
  - 1,024,908,267,229
  - Number of sentences:
  - 95,119,665,584
  - Number of unigrams: 13,588,391
  - Number of bigrams: 314,843,401
  - Number of trigrams: 977,069,902
  - Number of fourgrams: 1,313,818,354
  - Number of fivegrams: 1,176,470,663

3 Models

- IBM Model 1
  - Dumb word to word
- IBM Model 3
  - Handles deletions, insertions and 1-to-N translations
- Phrase-Based Models (Google/ISI)
  - Basically Model 1 with phrases instead of words

Alignment Probabilities

- Recall what of all of the models are doing
  Argmax P(e|f) = P(f|e)P(e)

  In the simplest models P(f|e) is just direct word-to-word translation probs. So let's start with how to get those, since they're used directly or indirectly in all the models.

Training alignment probabilities

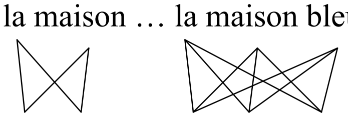
- Step 1: Get a parallel corpus
  - Hansards
    - Canadian parliamentary proceedings, in French and English
    - Hong Kong Hansards: English and Chinese
- Step 2: Align sentences
- Step 3: Use EM to train word alignments. Word alignments give us the counts we need for the word to word P(f|e) probs

Step 3: Word Alignments

- Of course, sentence alignments aren't what we need. We need word alignments to get the stats we need.
- It turns out we can bootstrap word alignments from raw sentence aligned data (no dictionaries)
- Using EM
  - Recall the basic idea of EM. A model predicts the way the world should look. We have raw data about how the world looks. Start somewhere and adjust the numbers so that the model is doing a better job of predicting how the world looks.

EM Training: Word Alignment Probs

… la maison … la maison bleue … la fleur …



… the house … the blue house … the flower …

All word alignments equally likely

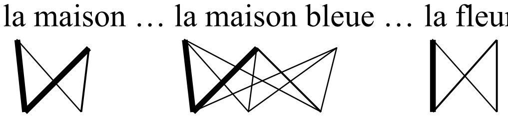
All P(french-word | english-word) equally likely.

EM Training Constraint

- Recall what we're doing here… Each English word has to translate to some french word.
- But its still true that

EM for training alignment probs

… la maison … la maison bleue … la fleur …



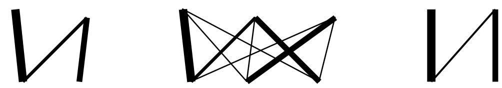
… the house … the blue house … the flower …

"la" and "the" observed to co-occur frequently,
so P(la | the) is increased.

Slide from Kevin Knight

EM for training alignment probs

… la maison … la maison bleue … la fleur …



… the house … the blue house … the flower …

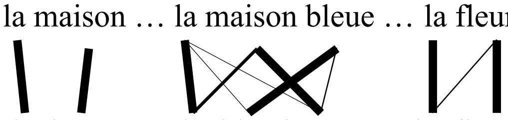
“house” co-occurs with both “la” and “maison”, but P(maison | house) can be raised without limit, to 1.0, while P(la | house) is limited because of “the”

(pigeonhole principle)

Slide from Kevin Knight

EM for training alignment probs

… la maison … la maison bleue … la fleur …



… the house … the blue house … the flower …

settling down after another iteration

Slide from Kevin Knight

EM for training alignment probs

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

Inherent hidden structure revealed by EM training!
For details, see:

- •Section 24.6.1 in the chapter
- “A Statistical MT Tutorial Workbook” (Knight, 1999).
- “The Mathematics of Statistical Machine Translation” (Brown et al, 1993)
- Free Alignment Software: GIZA++

Slide from Kevin Knight

Direct Translation

… la maison … la maison bleue … la fleur …

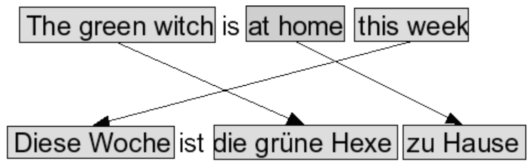
… the house … the blue house … the flower …

P(juste | fair) = 0.411
P(juste | correct) = 0.027
P(juste | right) = 0.020
…

New French sentence

Possible English translations, rescored by language model

Phrase-Based Translation



- Generative story here has three steps
  1) Discover and align phrases during training
  2) Align and translate phrases during decoding
  3) Finally move the phrases around

Phrase-based MT

- Language model P(E)
- Translation model P(F|E)
  - Model
  - How to train the model
- Decoder: finding the sentence E that is most probable

Generative story again

1) Group English source words into phrases $e_1, e_2, \ldots, e_n$
2) Translate each English phrase e_i into a Spanish phrase f_j.
   - The probability of doing this is $\phi(f_j|e_i)$
3) Then (optionally) reorder each Spanish phrase
   - We do this with a **distortion** probability
   - A measure of distance between positions of a corresponding phrase in the 2 languages
   - "What is the probability that a phrase in position X in the English sentences moves to position Y in the Spanish sentence?"

Distortion probability

- The distortion probability is parameterized by
  - The start position of the foreign (Spanish) phrase generated by the *i*th English phrase e_i.
  - The end position of the foreign (Spanish) phrase generated by the I-1th English phrase e_{i-1}.
- We'll call the distortion probability d(.)

Final translation model for phrase-based MT

Position	1	2	3	4	5
English	Mary	did not	slap	the	green witch
Spanish	Maria	no	dió una bofetada	a la	bruja verde

$$\begin{aligned} P(F|E) &= P(\text{Maria}, \text{Mary}) \times d(1) \times P(\text{no}|\text{did not}) \times d(1) \times \\ &\quad P(\text{dió una bofetada}|\text{slap}) \times d(1) \times P(\text{a la}|\text{the}) \times d(1) \times \\ &\quad P(\text{bruja verde}|\text{green witch}) \times d(1) \end{aligned}$$

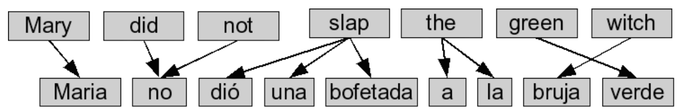
Training P(F|E)

- What we mainly need to train is $\phi(f_j|e_i)$
- Assume as before we have a large bilingual training corpus
- And suppose we knew exactly which phrase in Spanish was the translation of which phrase in the English
- We call this a **phrase alignment**
- If we had this, we could just count-and-divide:

$$\phi(\bar{f}, \bar{e}) = \frac{\text{count}(\bar{f}, \bar{e})}{\sum_{\bar{f}} \text{count}(\bar{f}, \bar{e})}$$

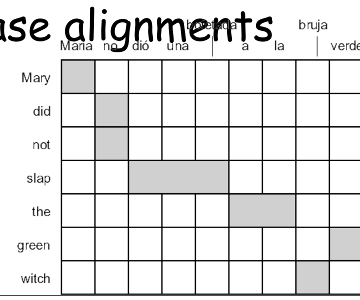
But we don't have phrase alignments

- What we have instead are word alignments:



Getting phrase alignments

- To get phrase alignments:
  1) We first get word alignments
     How? EM as before…
  2) Then we "symmetrize" the word alignments into phrase alignments



Final Problem

- Decoding…
  - Given a trained model and a foreign sentence produce
    - Argmax P(e|f)
    - Can't use Viterbi it's too restrictive
    - Need a reasonable efficient search technique that explores the sequence space based on how good the options look…
      - A*

A*

- Recall for A* we need
  - Goal State
  - Operators
  - Heuristic

A*

- Recall for A* we need

– Goal State	Good coverage of source
– Operators	Translation of phrases/words distortions deletions/insertions
– Heuristic	Probabilities (tweaked)

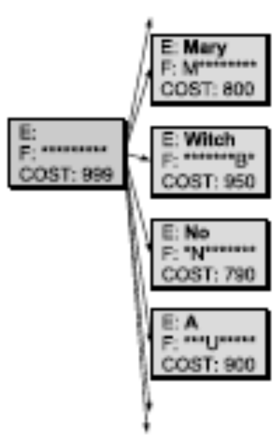
A* Decoding

- Why not just use the probability as we go along?
  - Turns it into Uniform-cost not A*
  - That favors shorter sequences over longer ones.
  - Need to counter-balance the probability of the translation so far with its "progress towards the goal".

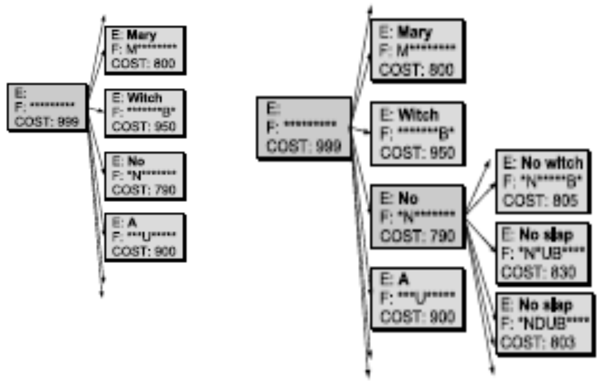
A*/Beam

- Sorry…
  - Even that doesn't work because the space is too large
  - So as we go we'll prune the space as paths fall below some threshold

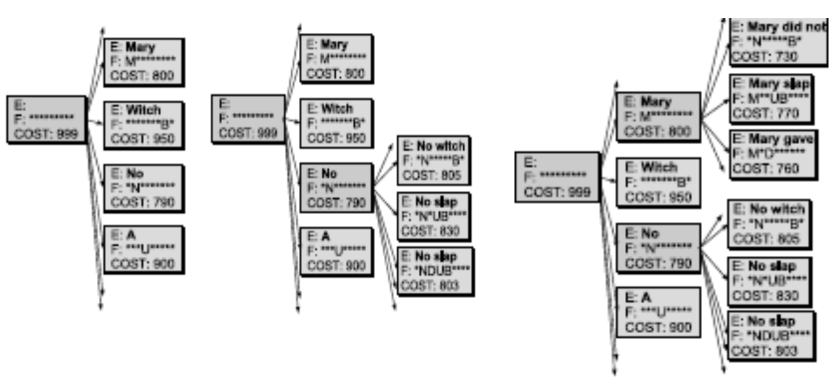
A* Decoding



A* Decoding



A* Decoding



Break

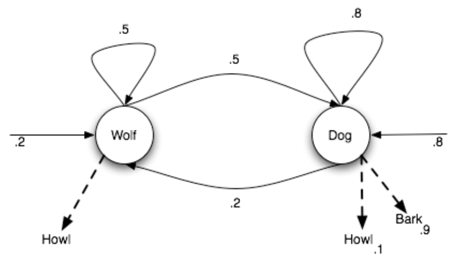
- Homework
  - I'm going to send out an additional test set
- Last quiz…
  - Next
- Average over the quizzes
  - 81% with a sd of 11…
    - That's (q1/55 + q2/50 + q3/50)/3

Break

- Quiz
  - True
  - Forward
  - EM
  - W,W,D
  - Yes
  - Anything

WWD

Hard way
WWD
WDD
DWD
DDD



YES

- Red, Square
  - YES
    - P(Red|Yes)P(Square|Yes)P(Yes)
      = .5*.5*.6 = .15
  - NO
    - P(Red|No)P(Square|No) =
      = .5*.5*.4=.1

Anything

- All three features give .6 accuracy. Doesn't matter which is chosen it's arbitrary

F1

R: 1,2,3,7,10: 3Y,2N	3 Right
G: 5,6,8: 2Y, 1N	2 Right
B: 4,9: 1Y, 1N	1 Right

Evaluation

- There are 2 dimensions along which MT systems can be evaluated
  - Fluency
    - How good is the output text as an example of the target language
  - Fidelity
    - How well does the output text convey the source text
      - Information content and style

Evaluating MT: Human tests for **fluency**

- Rating tests: Give the raters a scale (1 to 5) and ask them to rate
  - Or distinct scales for
    - Clarity, Naturalness, Style
  - Or check for specific problems
    - Cohesion (Lexical chains, anaphora, ellipsis)
      - Hand-checking for cohesion.
    - Well-formedness
      - 5-point scale of syntactic correctness

Evaluating MT: Human tests for **fidelity**

- Adequacy
  - Does it convey the information in the original?
  - Ask raters to rate on a scale
    - Bilingual raters: give them source and target sentence, ask how much information is preserved
    - Monolingual raters: give them target + a good human translation

Evaluating MT: Human tests for **fidelity**

- Informativeness
  - Task based: is there enough info to do some task?

Evaluating MT: Problems

- Asking humans to judge sentences on a 5-point scale for 10 factors takes time and $$$ (weeks or months!)
- Can't build language engineering systems if we can only evaluate them once every quarter!!!!
- Need a metric that we can run every time we change our algorithm.
- It's OK if it isn't perfect, just needs to correlate with the human metrics, which we could still run in periodically.

Automatic evaluation

- Assume we have one or more human translations of the source passage
- Compare the automatic translation to these human translations using some simple metric
  - Bleu
  - NIST
  - Meteor
  - Precision/Recall

BiLingual Evaluation Understudy (BLEU)

- Automatic Technique
- Requires the pre-existence of Human (Reference) Translations
- Approach:
  - Produce corpus of high-quality human translations
  - Judge "closeness" numerically (word-error rate)
  - Compare n-gram matches between candidate translation and 1 or more reference translations

Slide from Bonnie Dorr

BLEU Evaluation Metric

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?
    - An n-gram is an sequence of n words
  - 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)

Slide from Bonnie Dorr

BLEU Evaluation Metric

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.

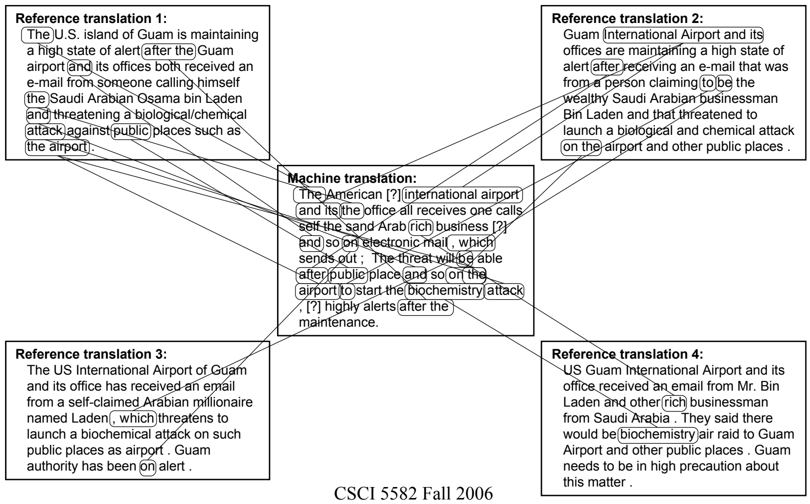
- BLEU4 formula
  (counts n-grams up to length 4)

exp (1.0 * log p1 +
0.5 * log p2 +
0.25 * log p3 +
0.125 * log p4 –
max(words-in-reference / words-in-machine – 1, 0)

p1 = 1-gram precision
P2 = 2-gram precision
P3 = 3-gram precision
P4 = 4-gram precision

Slide from Bonnie Dorr

Multiple Reference Translations



CSCI 5582 Fall 2006

Slide from Bonnie Dorr

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**

CSCI 5582 Fall 2006

Slide from Bonnie Dorr

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!)

Slide from Bonnie Dorr

Bleu Comparison

Chinese-English Translation Example:

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

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

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 the directions of the party.

Slide from Bonnie Dorr

BLEU Tends to Predict Human Judgments