Machine Translation: Training

Introduction to Natural Language Processing
Computer Science 585—Fall 2009
University of Massachusetts Amherst

David Smith

Training

Which features of data predict good translations?

Training: Generative/Discriminative

- Generative
  - –Maximum likelihood training: max p(data)
  - –“Count and normalize”
  - –Maximum likelihood with hidden structure
    - Expectation Maximization (EM)
- Discriminative training
  - –Maximum conditional likelihood
  - –Minimum error/risk training
  - –Other criteria: perceptron and max. margin

“Count and Normalize”

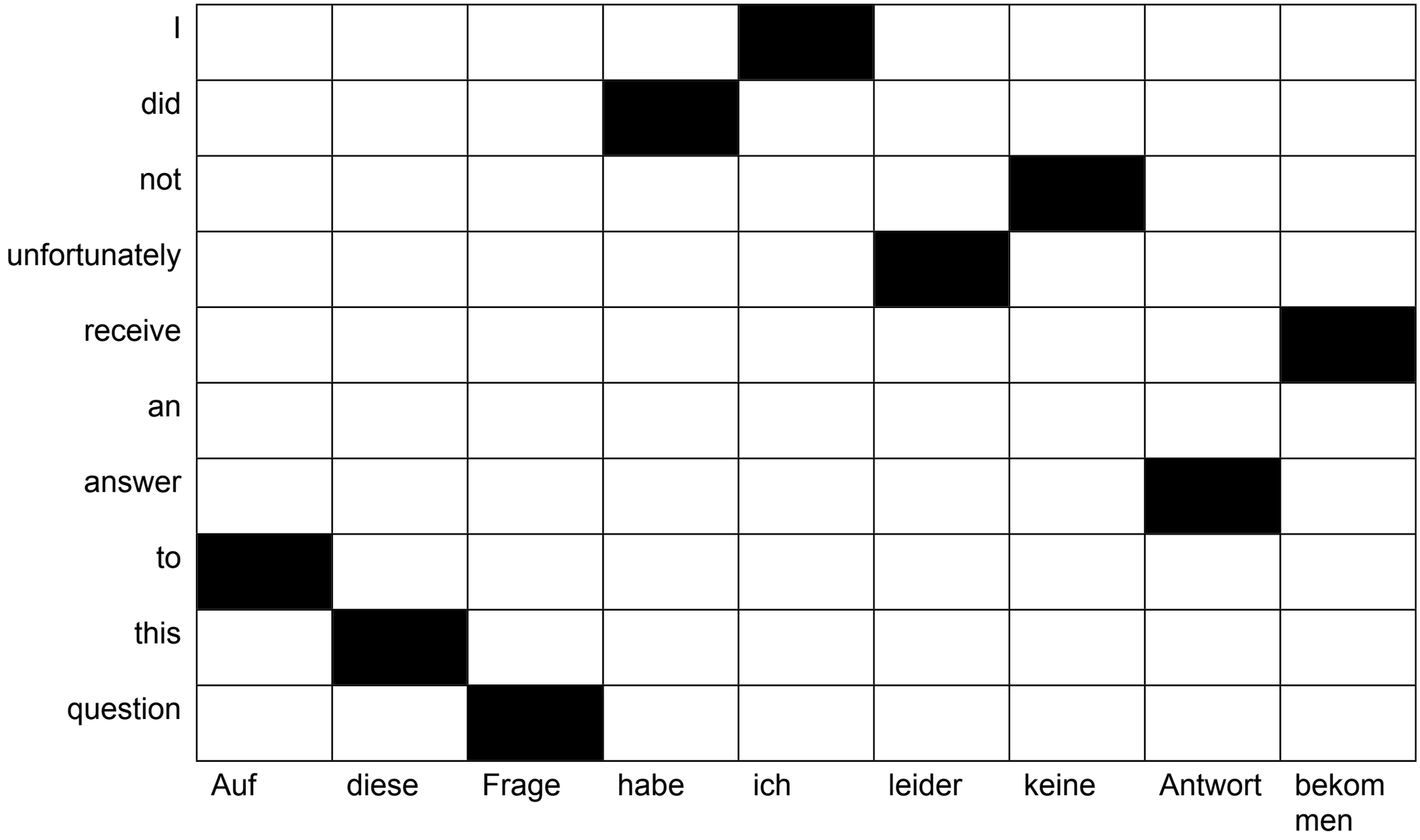
- Language modeling example: assume the probability of a word depends only on the previous 2 words.

$$p(\text{disease} \mid \text{into the}) = \frac{p(\text{into the disease})}{p(\text{into the})}$$

- p(disease|into the) = 3/20 = 0.15
- “Smoothing” reflects a prior belief that p(breech|into the) > 0 despite these 20 examples.

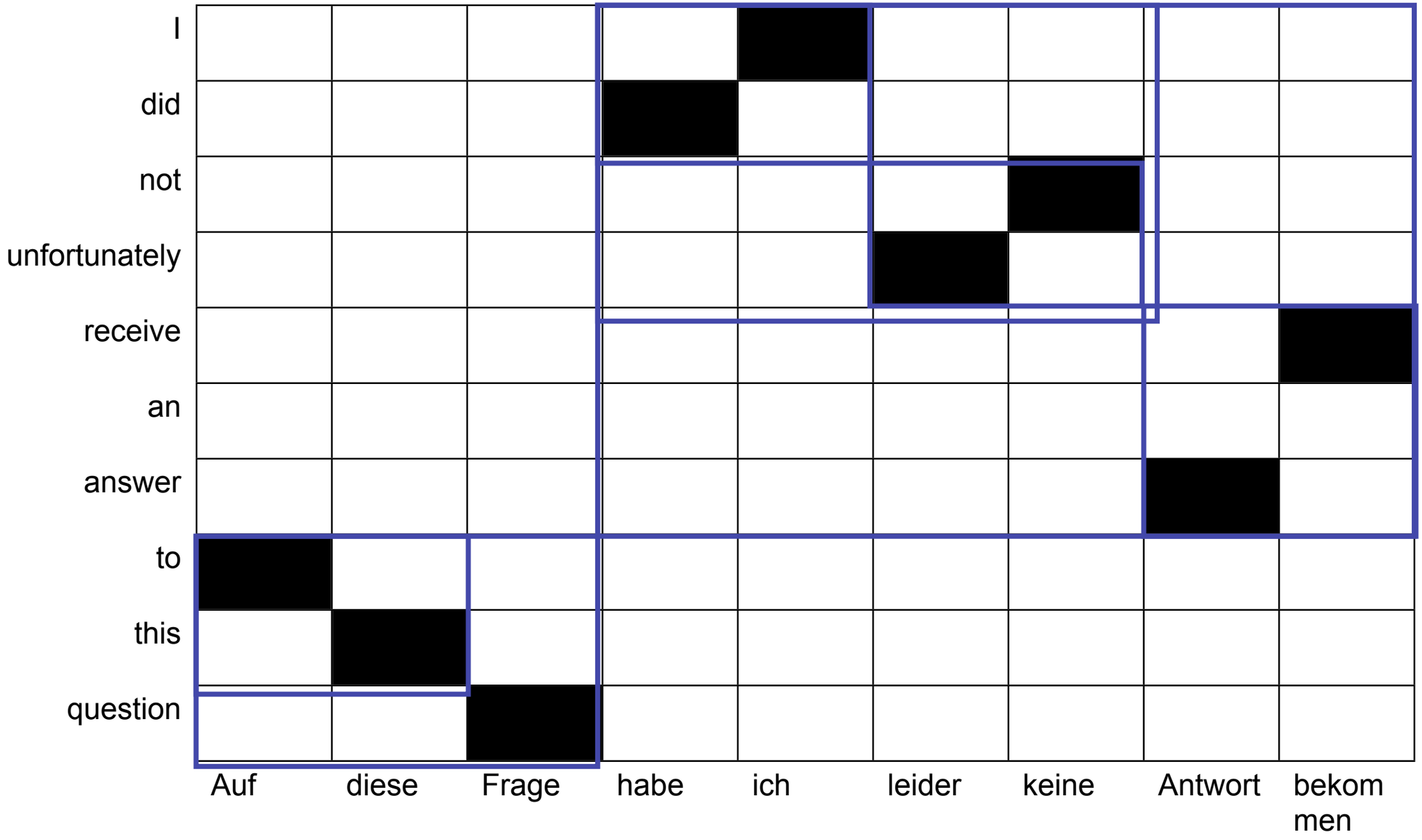
... into the programme ...
... into the **disease** ...
... into the **disease** ...
... into the correct ...
... into the next ...
... into the national ...
... into the integration ...
... into the Union ...
... into the Union ...
... into the Union ...
... into the sort ...
... into the internal ...
... into the general ...
... into the budget ...
... into the **disease** ...
... into the legal …
... into the various ...
... into the nuclear ...
... into the bargain ...
... into the situation ...

Phrase Models



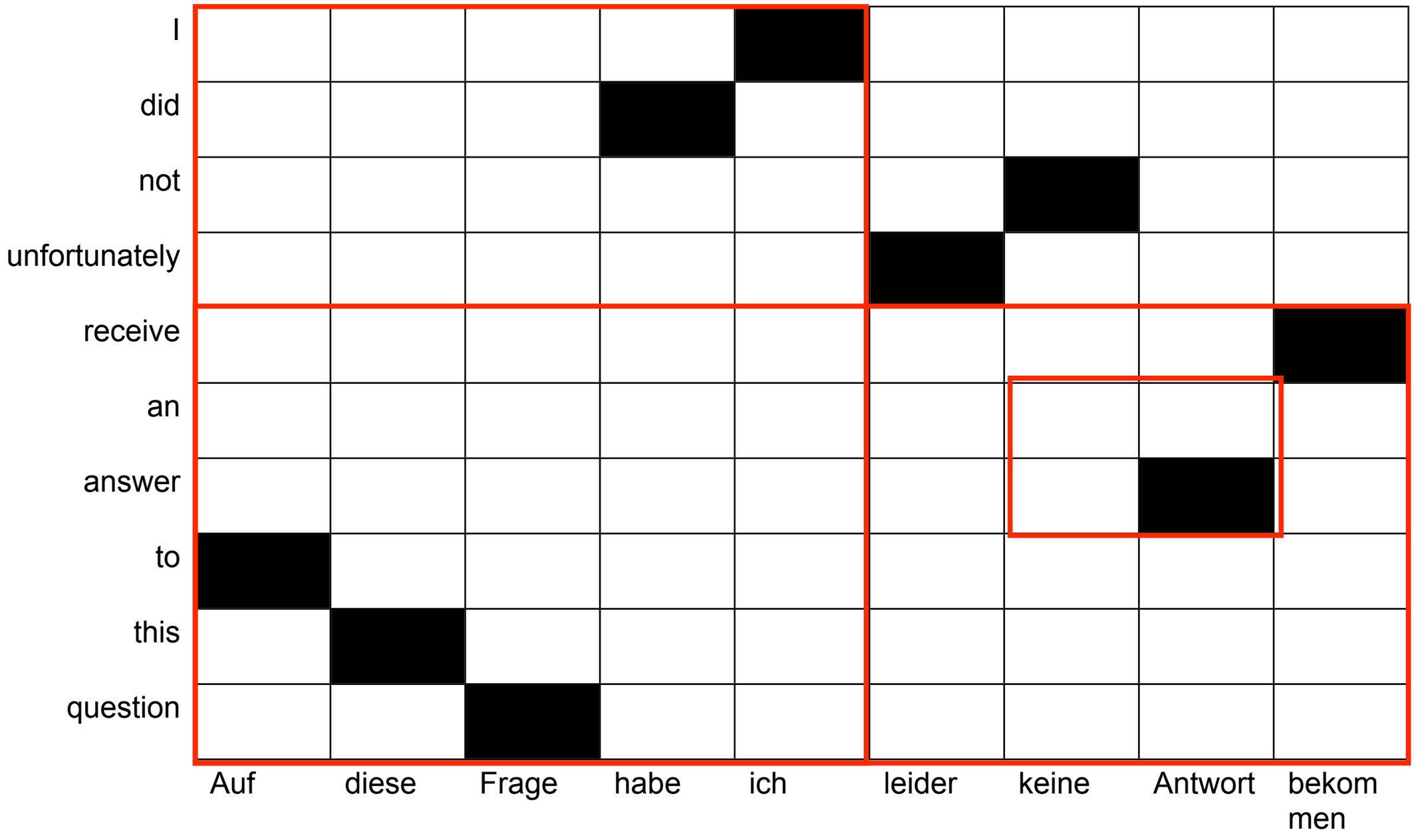
Assume word alignments are given.

Phrase Models



Some good phrase pairs.

Phrase Models



Some bad phrase pairs.

“Count and Normalize”

- Usual approach: treat relative frequencies of source phrase *s* and target phrase *t* as probabilities

$$p(s \mid t) \equiv \frac{count(s,t)}{count(t)} \qquad p(t \mid s) \equiv \frac{count(s,t)}{count(s)}$$

- This leads to overcounting when not all segmentations are legal due to unaligned words.

Hidden Structure

- But really, we don't observe word alignments.
- How are word alignment model parameters estimated?
- Find (all) structures consistent with observed data.
  - –Some links are incompatible with others.
  - –We need to score complete sets of links.

Hidden Structure and EM

- Expectation Maximization
  - Initialize model parameters (randomly, by some simpler model, or otherwise)
  - Calculate probabilities of hidden structures
  - Adjust parameters to maximize likelihood of observed data given hidden data
  - Iterate
- Summing over *all* hidden structures can be expensive
  - Sum over 1-best, *k*-best, other sampling methods

Discriminative Training

- Given a source sentence, give “good” translations a higher score than “bad” translations.
- We care about good translations, not a high probability of the training data.
- Spend less “energy” modeling bad translations.
- Disadvantages
  - We need to run the translation system at each training step.
  - System is tuned for one task (e.g. translation) and can’t be directly used for others (e.g. alignment)

“Good” Compared to What?

- Compare current translation to
- Idea #1: a human translation. OK, but
  - Good translations can be very dissimilar
  - We’d need to find hidden features (e.g. alignments)
- Idea #2: other top *n* translations (the “n-best list”). Better in practice, but
  - Many entries in n-best list are the same apart from hidden links
- Compare with a **loss function** *L*
  - 0/1: wrong or right; equal to reference or not
  - Task-specific metrics (word error rate, BLEU, …)

MT Evaluation

* Intrinsic

  Human evaluation

  Automatic (machine) evaluation

* Extrinsic

  How useful is MT system output for…

  Deciding whether a foreign language blog is about politics?
  Cross-language information retrieval?
  Flagging news stories about terrorist attacks?
  …

Human Evaluation

Je suis fatigué.

	Adequacy	Fluency
Tired is I.	5	2
Cookies taste good!	1	5
I am exhausted.	5	5

Human Evaluation

PRO

High quality

CON

Expensive!

Person (preferably bilingual) must make a time-consuming judgment per system hypothesis.

Expense prohibits frequent evaluation of incremental system modifications.

Automatic Evaluation

PRO

Cheap. Given available reference translations, free thereafter.

CON

**We can only measure some proxy for translation quality.
(Such as N-Gram overlap or edit distance).**

Output of Chinese-English system

In the First Two Months Guangdong's Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong's export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province's total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong's export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province's total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion us dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Partially excellent translations

In the First Two Months Guangdong's Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong's export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province's total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong's export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province's total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion US dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Mangled grammar

In the First Two Months Guangdong's Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong's export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province's total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong's export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province's total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion us dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Evaluation of Machine Translation Systems

Bleu (Papineni, Roukos, Ward and Zhu, 2002):

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.

Unigram Precision

- **Unigram Precision** of a candidate translation:

$$\frac{C}{N}$$

  where N is number of words in the candidate, C is the number of words in the candidate which are in at least one reference translation.

e.g.,

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

$$Precision = \frac{17}{18}$$

(only *obeys* is missing from all reference translations)

Modified Unigram Precision

- Problem with unigram precision:

  Candidate: the the the the the the the

  Reference 1: the cat sat on the mat

  Reference 2: there is a cat on the mat

  precision = 7/7 = 1???

- **Modified unigram precision:** "Clipping"
  - Each word has a "cap". e.g., *cap(the) = 2*
  - A candidate word w can only be correct a maximum of $cap(w)$ times. e.g., in candidate above, $cap(the) = 2$, and *the* is correct twice in the candidate $\Rightarrow$

$$Precision = \frac{2}{7}$$

Modified N-gram Precision

- Can generalize modified unigram precision to other n-grams.
- For example, for candidates 1 and 2 above:

$$Precision_1(bigram) = \frac{10}{17}$$

$$Precision_2(bigram) = \frac{1}{13}$$

Precision Alone Isn't Enough

Candidate 1: of the

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.

$$Precision(unigram) = 1$$

$$Precision(bigram) = 1$$

But Recall isn't Useful in this Case

- Standard measure used in addition to precision is **recall**:

$$Recall = \frac{C}{N}$$

where C is number of n-grams in candidate that are correct, N is number of words in the references.

Candidate 1: I always invariably perpetually do.

Candidate 2: I always do

Reference 1: I always do

Reference 1: I invariably do

Reference 1: I perpetually do

Sentence Brevity Penalty

- Step 1: for each candidate, compute closest matching reference (in terms of length)
  e.g., our candidate is length 12, references are length $12, 15, 17$. Best match is of length 12.

- Step 2: Say l_i is the length of the i'th candidate, r_i is length of best match for the i'th candidate, then compute

$$brevity = \frac{\sum_i r_i}{\sum_i l_i}$$

  (I think! from the Papineni paper, although $brevity = \frac{\sum_i r_i}{\sum_i min(l_i, r_i)}$ might make more sense?)

- Step 3: compute brevity penalty

$$BP = \begin{cases} 1 & \text{If } brevity < 1 \\ e^{1-brevity} & \text{If } brevity \geq 1 \end{cases}$$

  e.g., if $r_i = 1.1 \times l_i$ for all i (candidates are always 10% too short) then $BP = e^{-0.1} = 0.905$

The Final Score

- Corpus precision for any n-gram is

$$p_n = \frac{\sum_{C \in \{Candidate\}} \sum_{ngram \in C} Count_{clip}(ngram)}{\sum_{C \in \{Candidate\}} \sum_{ngram \in C} Count(ngram)}$$

  i.e. number of correct ngrams in the candidates (after "clipping") divided by total number of ngrams in the candidates

- Final score is then

$$Bleu = BP \times (p_1 p_2 p_3 p_4)^{1/4}$$

  i.e., BP multiplied by the geometric mean of the unigram, bigram, trigram, and four-gram precisions

Automatic Evaluation: Bleu Score

hypothesis 1	I am exhausted
hypothesis 2	Tired is I
reference 1	I am tired
reference 2	I am ready to sleep now

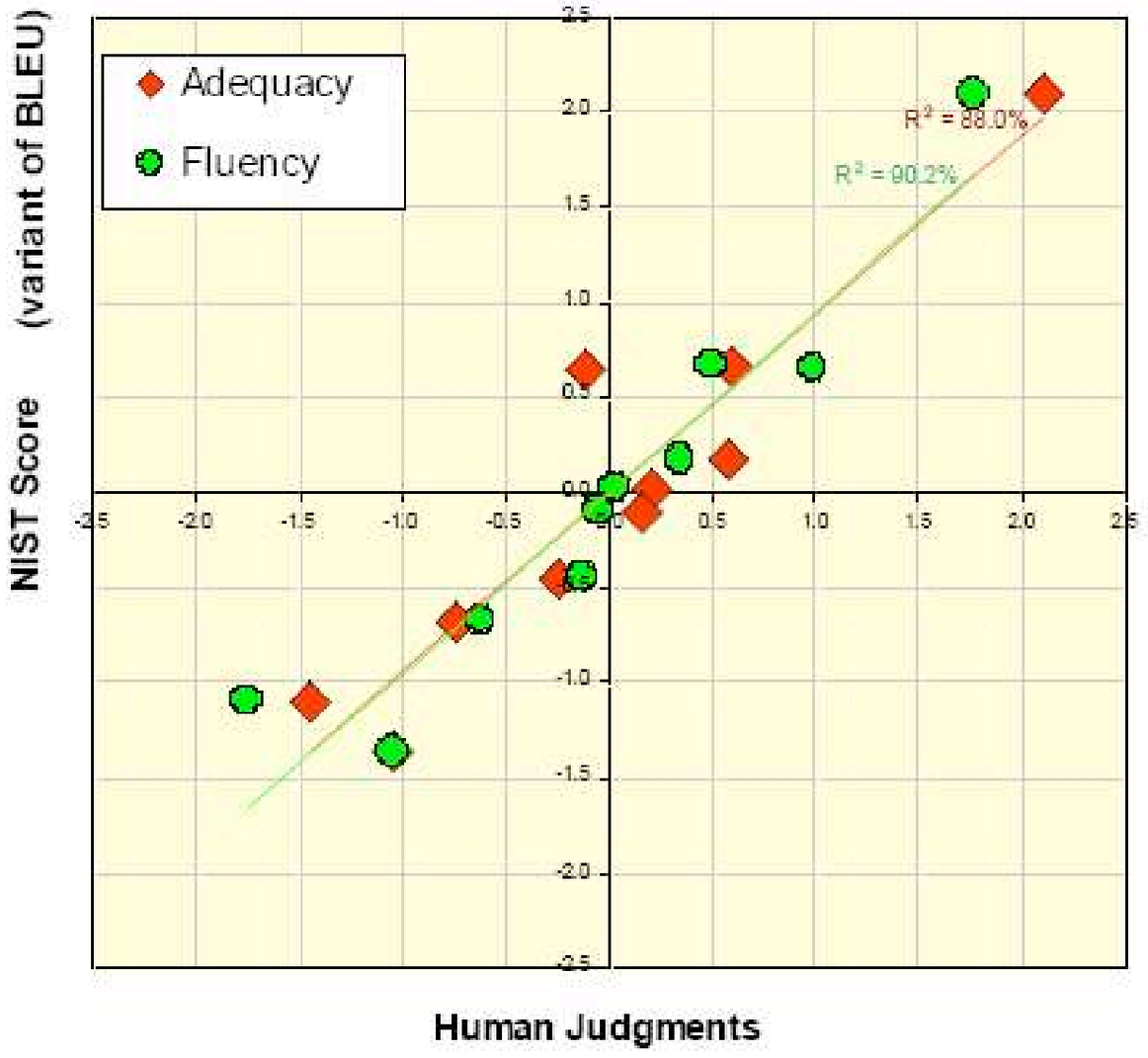
Automatic Evaluation: Bleu Score

		1-gram	2-gram	3-gram
hypothesis 1	I am exhausted	3/3	1/2	0/1
hypothesis 2	Tired is I	1/3	0/2	0/1
hypothesis 3	I I I	1/3	0/2	0/1

reference 1: I am tired

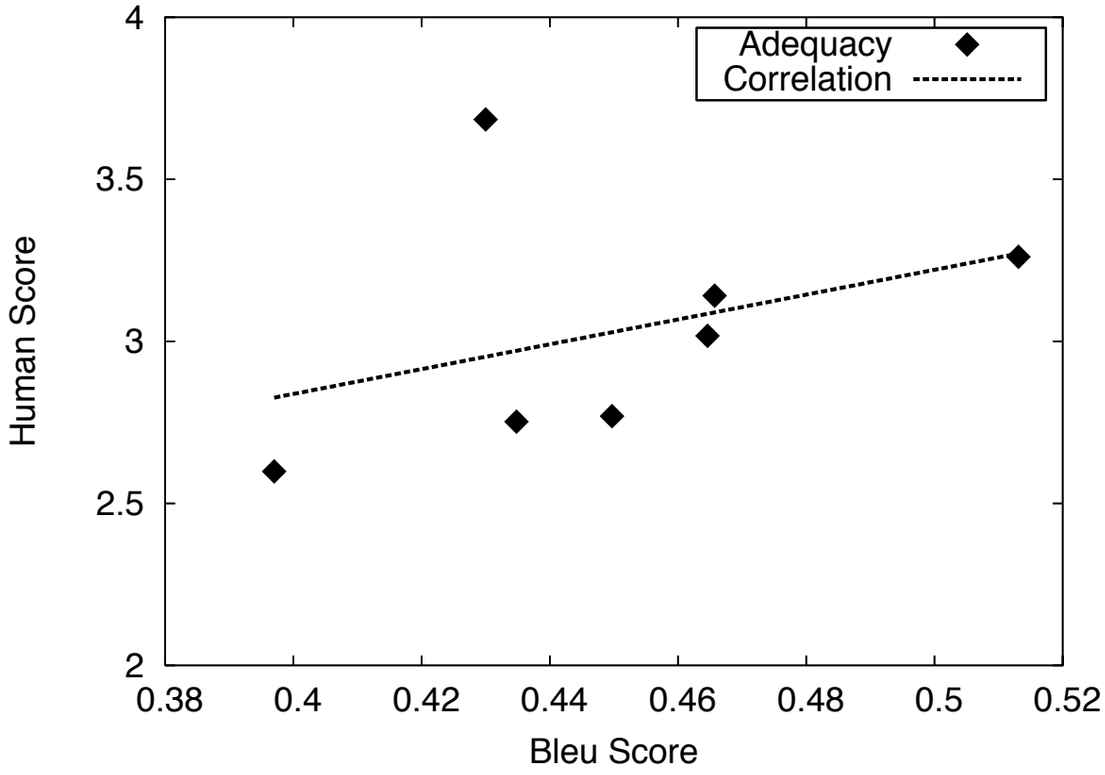
reference 2: I am ready to sleep now and so exhausted

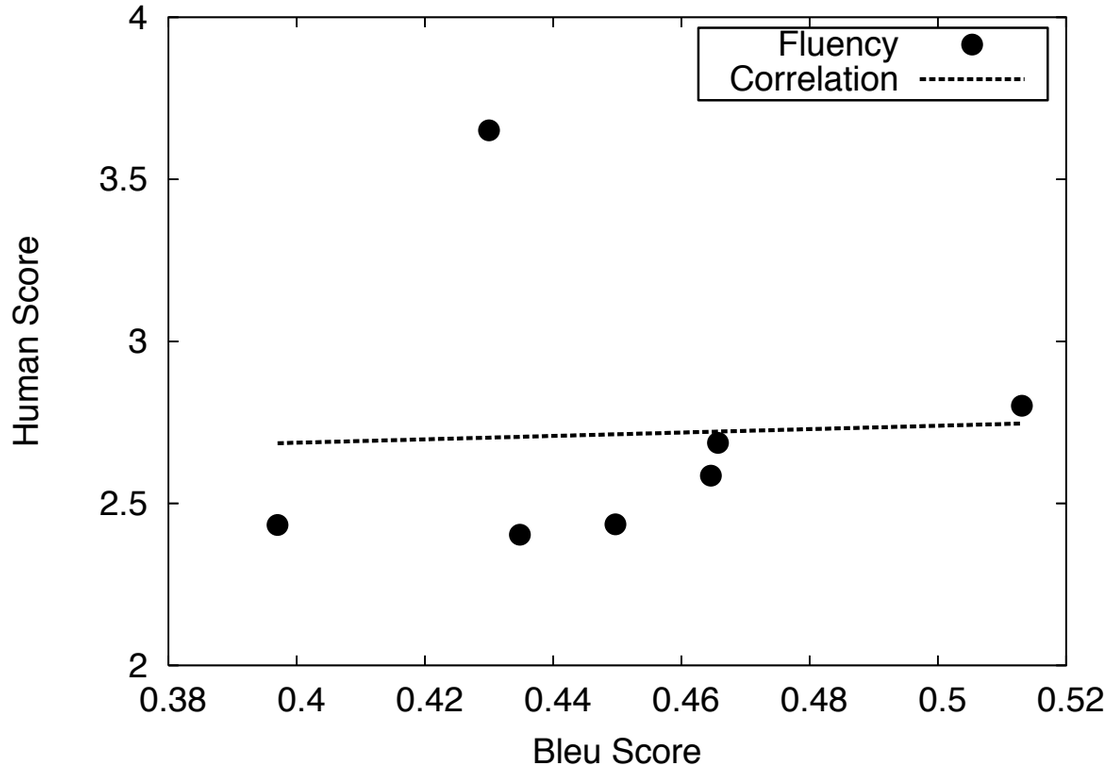
How Good are Automatic Metrics?



slide from G. Doddington (NIST)

Correlation? [Callison-Burch et al., 2006]



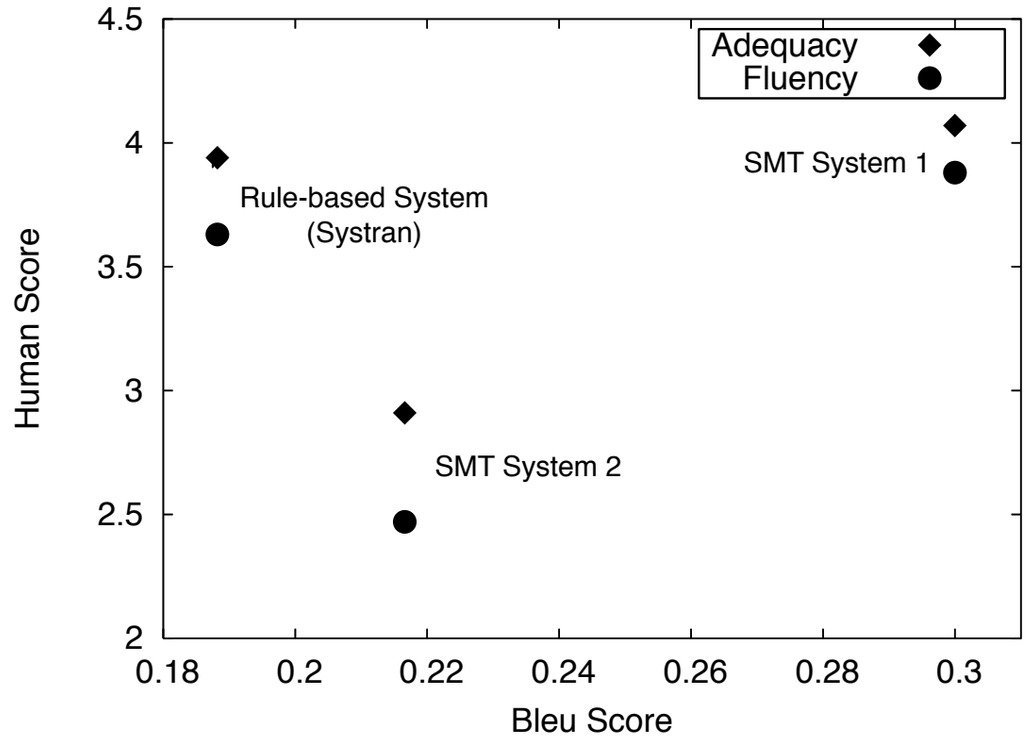


[from Callison-Burch et al., 2006, EACL]

- DARPA/NIST MT Eval 2005
  - Mostly statistical systems (all but one in graphs)
  - One submission **manual post-edit** of statistical system's output

→ Good adequacy/fluency scores *not reflected* by BLEU

Correlation? [Callison-Burch et al., 2006]



- Comparison of

[from Callison-Burch et al., 2006, EACL]

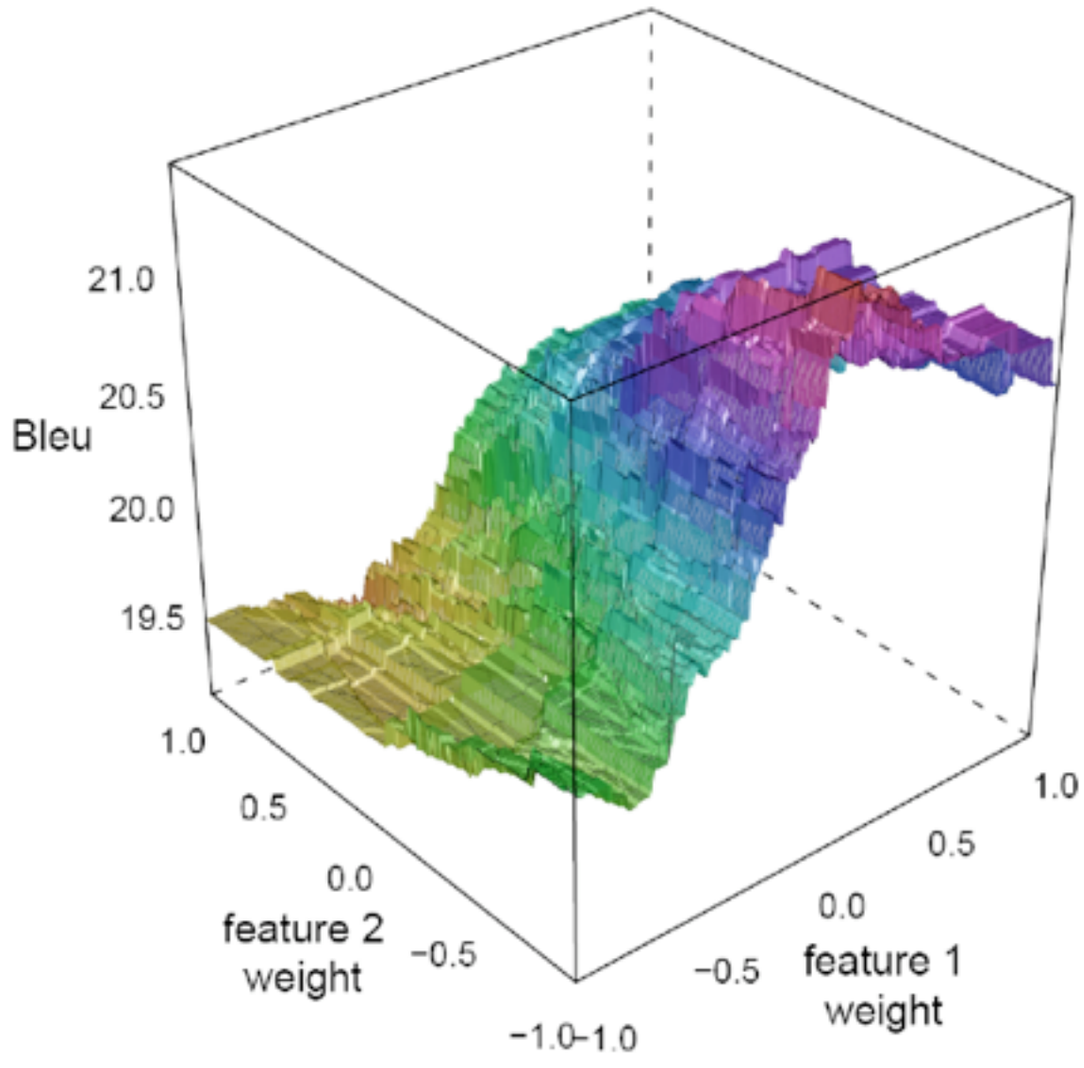
  - *good statistical* system: high BLEU, high adequacy/fluency
  - *bad statistical* sys. (trained on less data): low BLEU, low adequacy/fluency
  - *Systran*: lowest BLEU score, but high adequacy/fluency

How Good are Automatic Metrics?

- Do n-gram methods like BLEU overly favor certain types of systems?
- Automatic metrics still useful
  - During development of one system, a better BLEU indicates a better system
  - Evaluating different systems has to depend on human judgement
  - What are some other evaluation ideas?

Minimizing Error/Maximizing Bleu

- Adjust parameters to minimize error (L) when translating a training set
- Error as a function of parameters is
  - *nonconvex*: not guaranteed to find optimum
  - *piecewise constant*: slight changes in parameters might not change the output.
- Usual method: optimize one parameter at a time with linear programming



Generative/Discriminative Reunion

- Generative models can be cheap to train: “count and normalize” when nothing’s hidden.
- Discriminative models focus on problem: “get better translations”.
- Popular combination
  - Estimate several generative translation and language models using relative frequencies.
  - Find their optimal (log-linear) combination using discriminative techniques.

Generative/Discriminative Reunion

Score each hypothesis with several generative models:

$$\theta_1 p_{phrase}(\bar{s} \mid \bar{t}) + \theta_2 p_{phrase}(\bar{t} \mid \bar{s}) + \theta_3 p_{lexical}(s \mid t) + \mathbf{L} \; + \theta_7 p_{LM}(\bar{t}) + \theta_8 \#\text{words} + \mathbf{L}$$

If necessary, renormalize into a probability distribution:

$$Z = \sum_k \exp(\grave{\mathbf{e}} \bullet \mathbf{f}_k)$$

Unnecessary if thetas sum to 1 and p's are all probabilities.

where *k* ranges over all hypotheses. We then have

$$p(t_i \mid s) = \frac{1}{Z} \exp(\grave{\mathbf{e}} \bullet \mathbf{f})$$

Exponentiation makes it positive.

for any given hypothesis *i*.

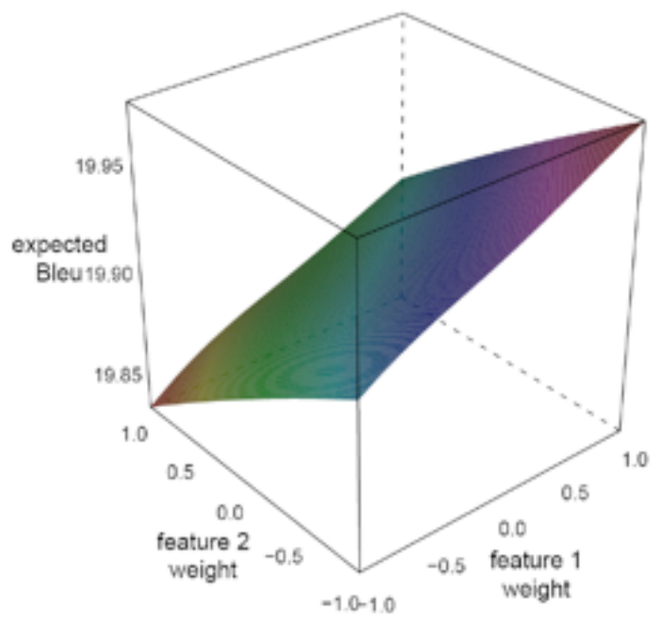
Minimizing Risk

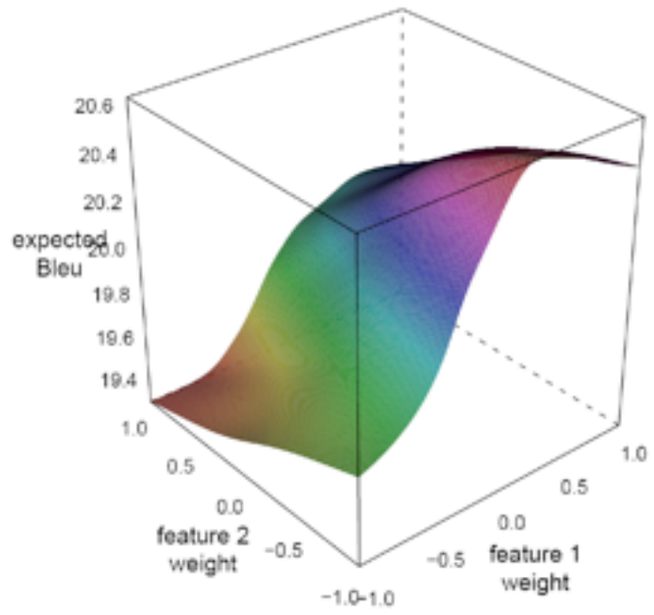
Instead of the error of the 1-best translation, compute **expected error** (risk) using *k*-best translations; this makes the function differentiable.

Smooth probability estimates using gamma to even out local bumpiness. Gradually increase gamma to approach the 1-best error.

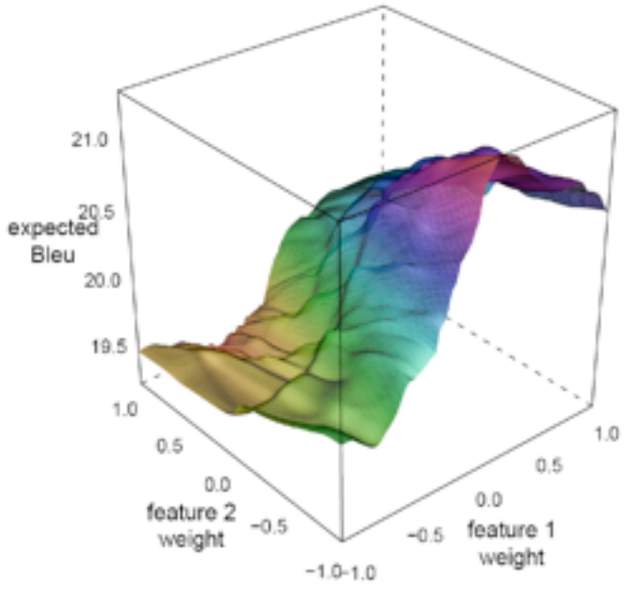
$$\mathrm{E}_{p_{\gamma,\grave{\mathbf{e}}}}[L(s,t)]$$

$$p_{\gamma,\theta}(t_i \mid s_i) = \frac{[\exp \grave{\mathbf{e}} \bullet \mathbf{f}_i]^{\gamma}}{\sum_{k'} [\exp \grave{\mathbf{e}} \bullet \mathbf{f}_{k'}]^{\gamma}}$$

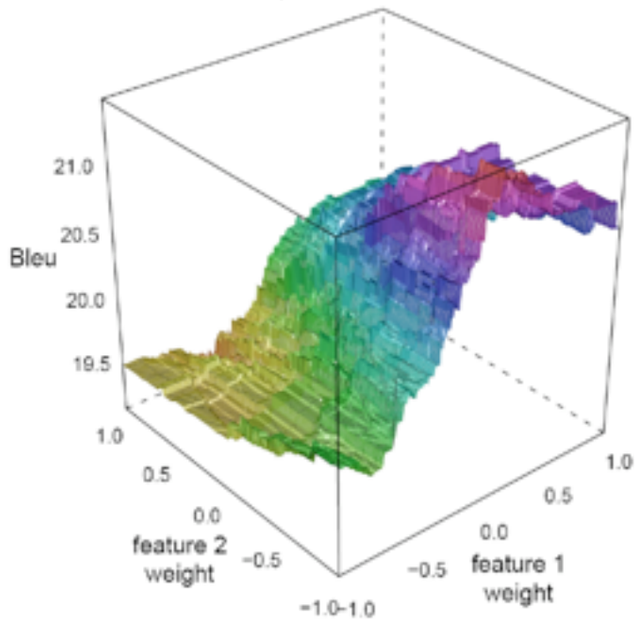




$\gamma = 0.1$



$\gamma = 1$



$\gamma = 10$

$\gamma = \infty$

Case Study: Inversion Transduction Grammar

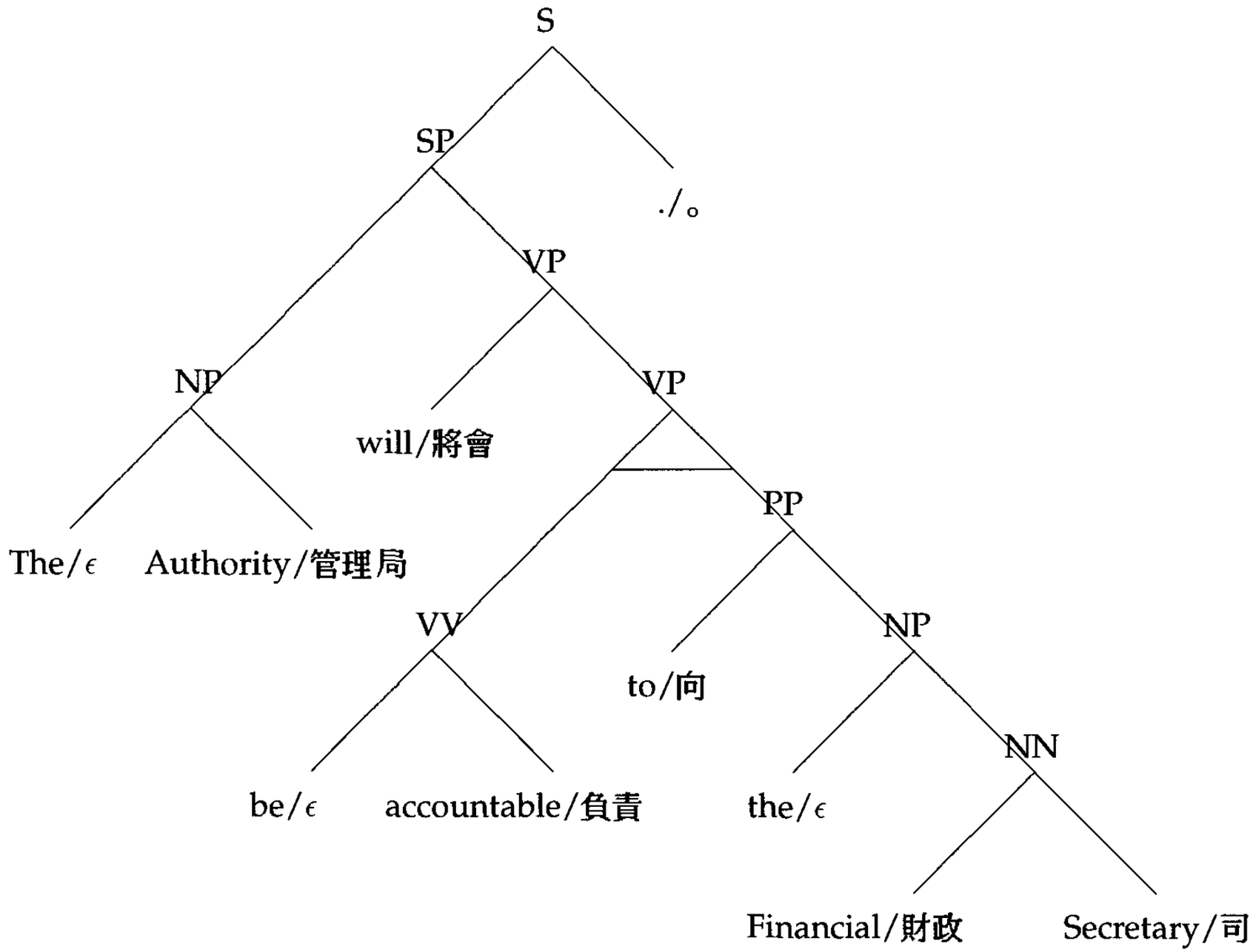
Syntactically-Motivated Distortion

The Authority will be accountable to the Financial Secretary.

管理局將會向財政司負責。

(*Authority will to Financial Secretary accountable.*)

Syntactically-Motivated Distortion



ITG Overview

- Special case of synchronous CFG
- One, joint nonterminal per bilingual node
- Children are translated monotonically, or reversed
- Binarized normal form
- Mostly used for exact, polytime alignment

ITG Rules

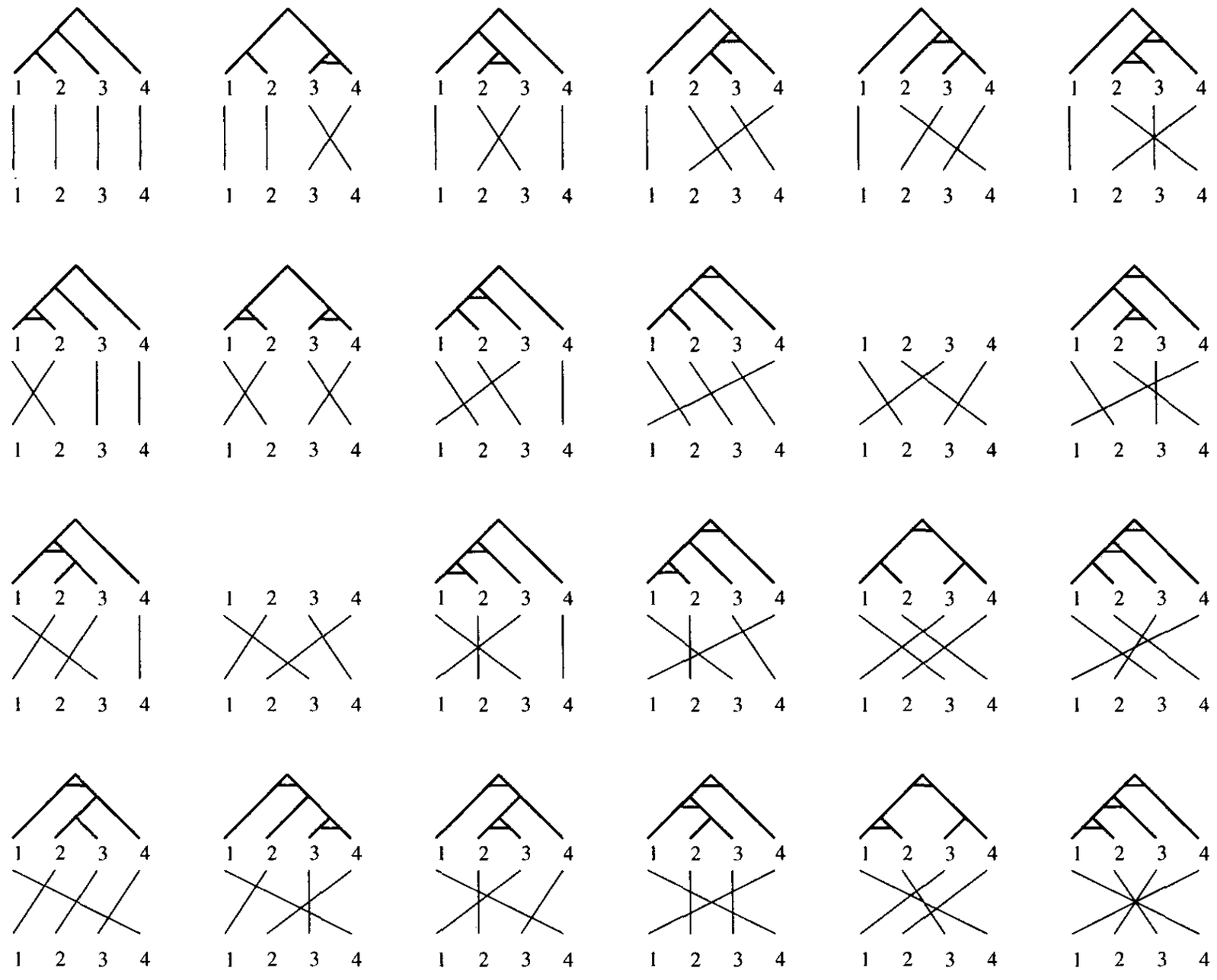
S	→	[SP Stop]
SP	→	[NP VP] | [NP VV] | [NP V]
PP	→	[Prep NP]
NP	→	[Det NN] | [Det N] | [Pro] | [NP Conj NP]
NN	→	[A N] | [NN PP]
VP	→	[Aux VP] | [Aux VV] | [VV PP]
VV	→	[V NP] | [Cop A]
Det	→	the/ϵ
Prep	→	to/向
Pro	→	I/我 | you/你
N	→	authority/管理局 | secretary/司
A	→	accountable/負責 | financial/財政
Conj	→	and/和
Aux	→	will/將會
Cop	→	be/ϵ
Stop	→	./。
VP	→	⟨VV PP⟩

ITG Alignment

Where is the Secretary of Finance when needed ?

財政 司 有需要 時 在 那裡 ?

Legal ITG Alignments



Bracketing ITG

$$
\begin{array}{lcll}
\mathrm{A} & \xrightarrow{a} & [\mathrm{A}\ \mathrm{A}] & \\
\mathrm{A} & \xrightarrow{a} & \langle \mathrm{A}\ \mathrm{A} \rangle & \\
\mathrm{A} & \xrightarrow{b_{ij}} & u_i / v_j & \text{for all } i,j \text{ English-Chinese lexical translations} \\
\mathrm{A} & \xrightarrow{b_{i\epsilon}} & u_i / \epsilon & \text{for all } i \text{ English vocabulary} \\
\mathrm{A} & \xrightarrow{b_{\epsilon j}} & \epsilon / v_j & \text{for all } j \text{ Chinese vocabulary}
\end{array}
$$

Removing Spurious Ambiguity

$$
\begin{array}{lcll}
\mathrm{A} & \xrightarrow{a} & [\mathrm{A}\ \mathrm{B}] & \\
\mathrm{A} & \xrightarrow{a} & [\mathrm{B}\ \mathrm{B}] & \\
\mathrm{A} & \xrightarrow{a} & [\mathrm{C}\ \mathrm{B}] & \\
\mathrm{A} & \xrightarrow{a} & [\mathrm{A}\ \mathrm{C}] & \\
\mathrm{A} & \xrightarrow{a} & [\mathrm{B}\ \mathrm{C}] & \\
\mathrm{B} & \xrightarrow{a} & \langle \mathrm{A}\ \mathrm{A} \rangle & \\
\mathrm{B} & \xrightarrow{a} & \langle \mathrm{B}\ \mathrm{A} \rangle & \\
\mathrm{B} & \xrightarrow{a} & \langle \mathrm{C}\ \mathrm{A} \rangle & \\
\mathrm{B} & \xrightarrow{a} & \langle \mathrm{A}\ \mathrm{C} \rangle & \\
\mathrm{B} & \xrightarrow{a} & \langle \mathrm{B}\ \mathrm{C} \rangle & \\
\mathrm{C} & \xrightarrow{b_{ij}} & u_i/v_j & \text{for all } i,j \text{ English-Chinese lexical translations} \\
\mathrm{C} & \xrightarrow{b_{i\epsilon}} & u_i/\epsilon & \text{for all } i \text{ English vocabulary} \\
\mathrm{C} & \xrightarrow{b_{\epsilon j}} & \epsilon/v_j & \text{for all } j \text{ Chinese vocabulary}
\end{array}
$$