Sequence transduction

CMSC 723 / LING 723 / INST 725

Hal Daumé III [he/him]

17 Oct 2019

(many slides c/o Marine Carpuat or Graham Neubig or Kevin Knight)

Announcements, logistics

- Don’t forget P1 pitches!
  - If you need resources, please ask (eg LDC-$walled datasets)
- We’re working on HW4
  - will be about stuff covered up to and including today
  - hopefully posted by next class
- Grading
  - HW2 by Tuesday
  - Midterm by Oct 29

Last time

- Sequence labeling as independent predictions
- Structured perceptron for sequence labeling
- Do we really need structured features?
- Recurrent neural network taggers

Today

- Sequence generation when no 1-1 correspondence between input and output
- Encoder-decoder
- Attention
- Evaluation
- Later directions

DETECT LANGUAGE | ARABIC | JAPANESE | ENGLISH

بالحضارة الإسلامية.. موسيقي يرعى العلماء ومسيحي
طب و150 مليون دولار للمستنصرية

almanh altaelimiat bialhadarat al'iislamia.. musiqiin yareaa aleulam
tullab altibi w150 mlywn dular lilmustansaria

Arabic text	Swahili text	
Latin name	Japanese sp	
Long document	Short docu	
Utterance	Response	
Image	Text	
Speech	Transcript	Speech r
Spanish text	SQL Query	Code gene
Python code	Aramaic text	Comment
Question	Answer	Question a
NWS data	English text	Natural la
…	…	…

Sentence: *What is the capital of Germany?*

(Semantic parsing)

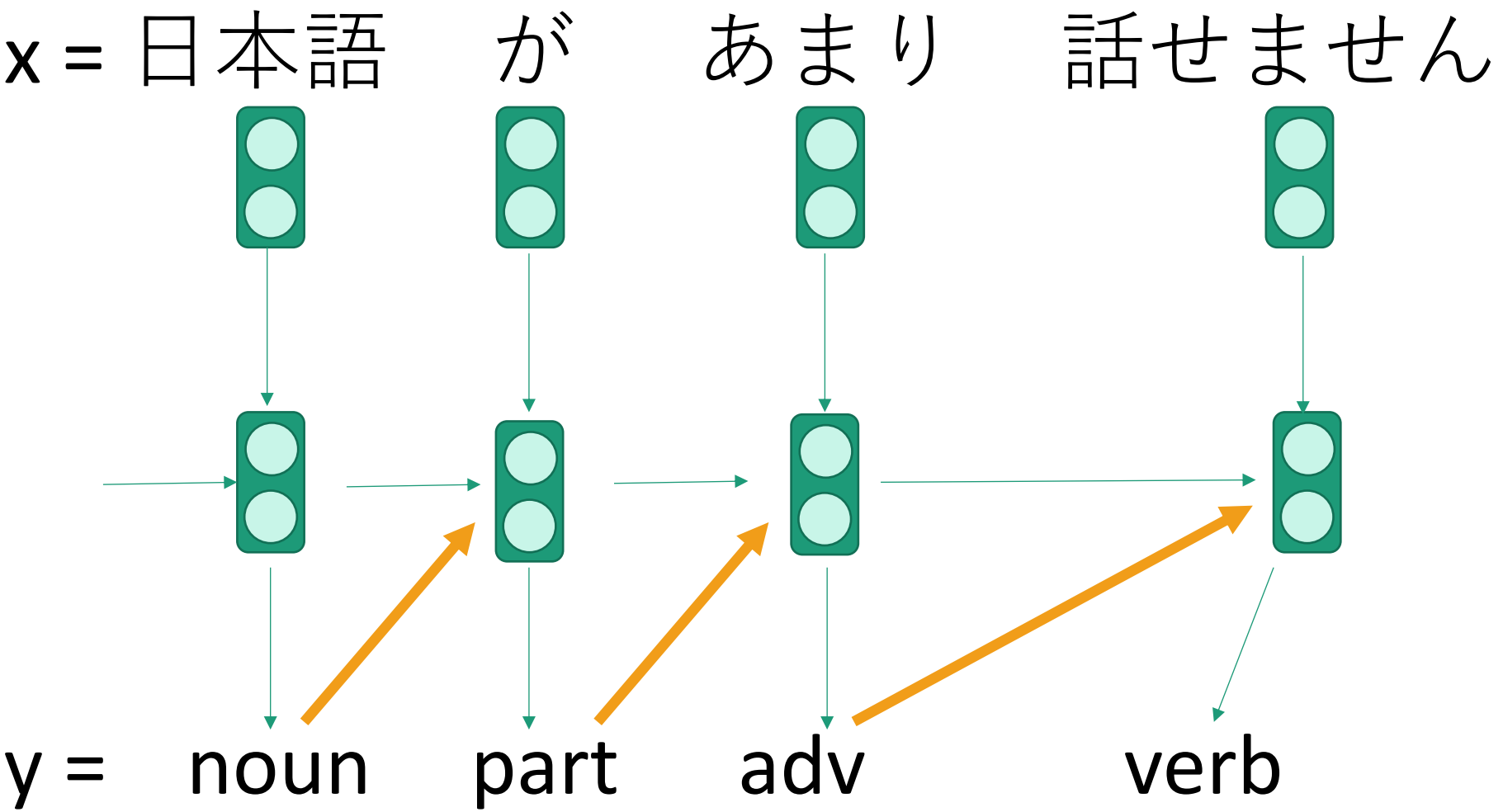
Logical form:

Result:

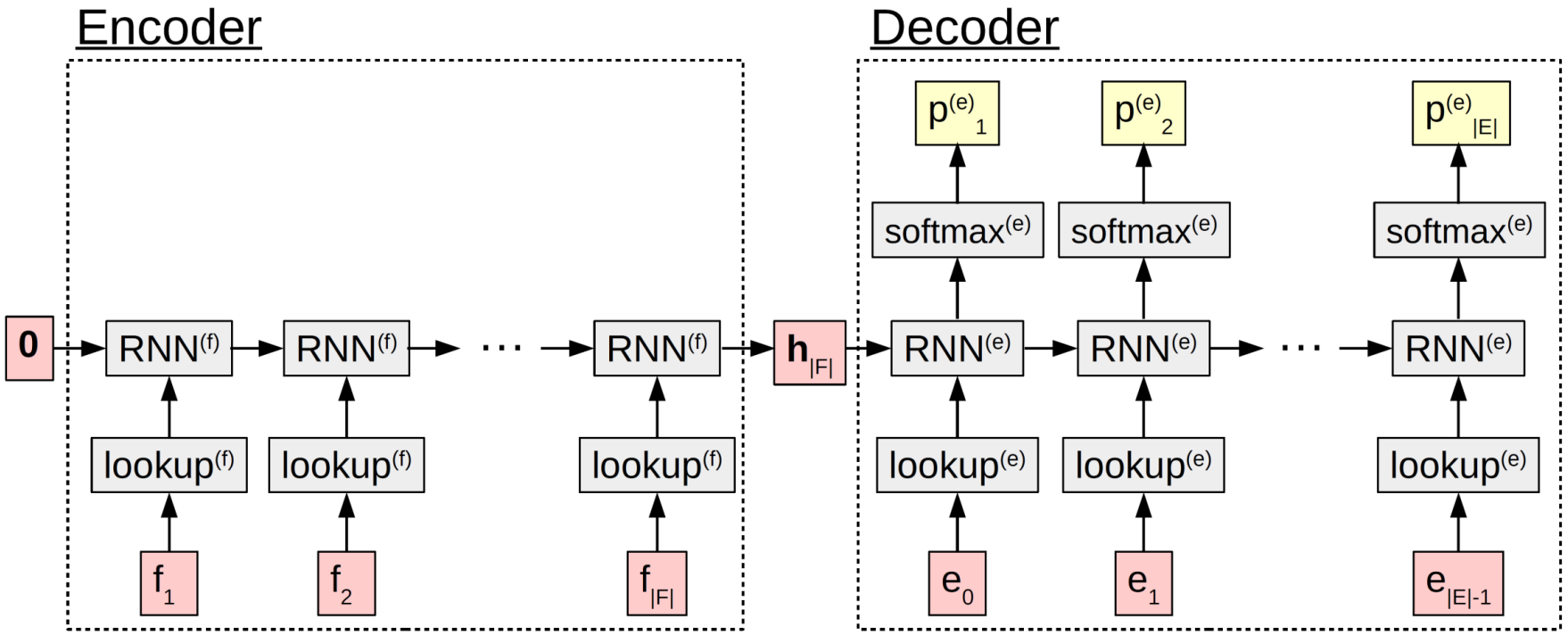
```
((type dailyweatherrecord)
 (date ((day 31)
        (month 5)
        (year 1994)))
 (temperature ((minimum ((unit degrees-c)
                         (number 12)))
               (maximum ((unit degrees-c)
                         (number 19)))))
 (rainfall ((unit millimetres)
            (number 3))))
```

The month was cooler and drier than average, with the average number of rain days, but ...

RNN sequence labeling



Encoder-decoder model



Encoder-decoder model

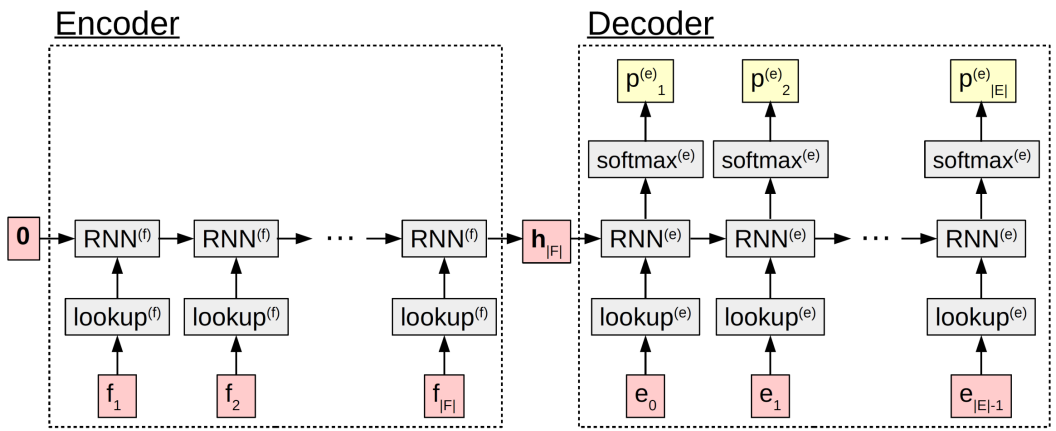
$$\boldsymbol{m}_t^{(f)} = M_{\cdot,f_t}^{(f)}$$

$$\boldsymbol{h}_t^{(f)} = \begin{cases} \mathrm{RNN}^{(f)}(\boldsymbol{m}_t^{(f)}, \boldsymbol{h}_{t-1}^{(f)}) & t \geq 1, \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

$$\boldsymbol{m}_t^{(e)} = M_{\cdot,e_{t-1}}^{(e)}$$

$$\boldsymbol{h}_t^{(e)} = \begin{cases} \mathrm{RNN}^{(e)}(\boldsymbol{m}_t^{(e)}, \boldsymbol{h}_{t-1}^{(e)}) & t \geq 1, \\ \boldsymbol{h}_{|F|}^{(f)} & \text{otherwise.} \end{cases}$$

$$\boldsymbol{p}_t^{(e)} = \mathrm{softmax}(W_{hs}\boldsymbol{h}_t^{(e)} + b_s)$$



Problem with encoder-decoder model

- Long-distance dependencies remain a problem

- A single vector represents the entire source sentence
  - No matter its length

Issues with Distributional Semantics

- How to compose meanings of larger phrases and sentences from lexical representations? (many recent proposals…)
- None of the proposals for compositionality capture the full representational or inferential power of FOPC (Grefenstette, 2013).

"You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector!"

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok .
1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .
2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghirok .
3b. totat dat arrat vat hilat .

4a. ok-voon anok drok brok jok .
4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .
5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .
6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .
7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .
8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .
9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghirok clok .
10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .
11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .
12b. wat nnat forat arrat vat gat .

What was your theory of language?

Attention model intuition

- Encode each word in source sentence into a vector
- When decoding, perform a linear combination of these vectors, weighted by "attention weights"
- Use this combination when predicting next word

[Bahdanau et al. 2015]

Attention model
Source word representations

- We can use representations from bidirectional RNN encoder

$$\overrightarrow{\boldsymbol{h}}_j^{(f)} = \text{RNN}(\text{embed}(f_j), \overrightarrow{\boldsymbol{h}}_{j-1}^{(f)})$$
$$\overleftarrow{\boldsymbol{h}}_j^{(f)} = \text{RNN}(\text{embed}(f_j), \overleftarrow{\boldsymbol{h}}_{j+1}^{(f)}).$$

$$\boldsymbol{h}_j^{(f)} = [\overleftarrow{\boldsymbol{h}}_j^{(f)}; \overrightarrow{\boldsymbol{h}}_j^{(f)}].$$

- And concatenate them in a matrix

$$H^{(f)} = \text{concat\_col}(\boldsymbol{h}_1^{(f)}, \ldots, \boldsymbol{h}_{|F|}^{(f)}).$$

Attention model
Create a source context vector

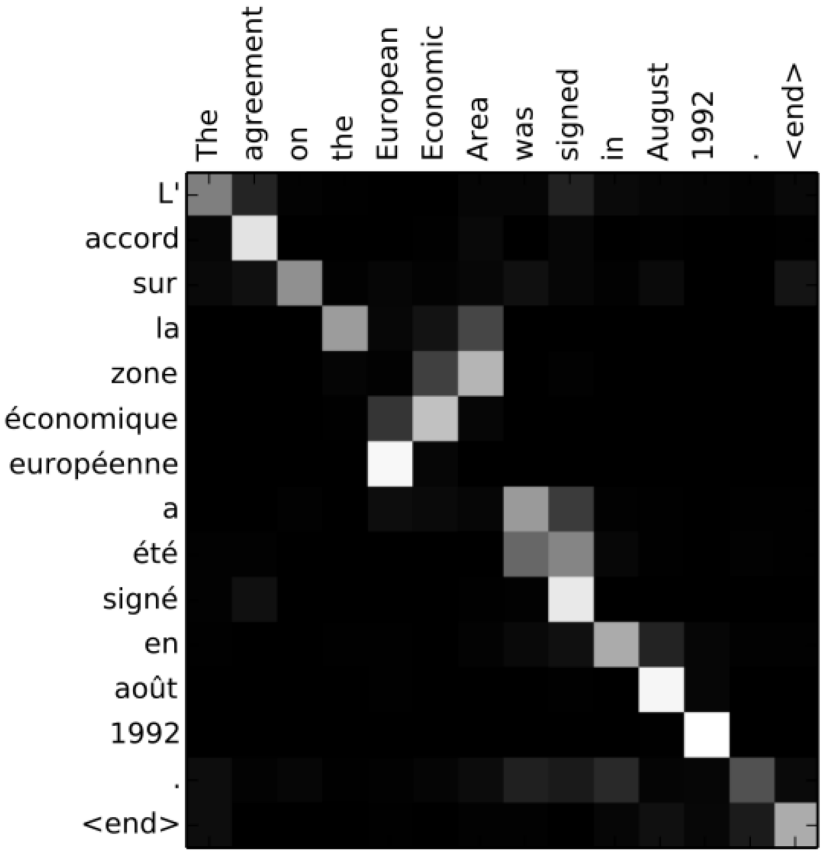
$$\boldsymbol{c}_t = H^{(f)}\boldsymbol{\alpha}_t.$$

Context vector

Attention vector

- Attention vector:
  - Entries between 0 and 1
  - Interpreted as weight given to each source word when generating output at time step t

Attention model
Illustrating attention weights



Attention model
How to calculate attention scores

$$\boldsymbol{h}_t^{(e)} = \text{enc}([\text{embed}(e_{t-1}); \boldsymbol{c}_{t-1}], \boldsymbol{h}_{t-1}^{(e)}).$$

$$a_{t,j} = \text{attn\_score}(\boldsymbol{h}_j^{(f)}, \boldsymbol{h}_t^{(e)}).$$

$$\boldsymbol{\alpha}_t = \text{softmax}(\boldsymbol{a}_t).$$

$$\boldsymbol{p}_t^{(e)} = \text{softmax}(W_{hs}[\boldsymbol{h}_t^{(e)}; \boldsymbol{c}_t] + b_s).$$

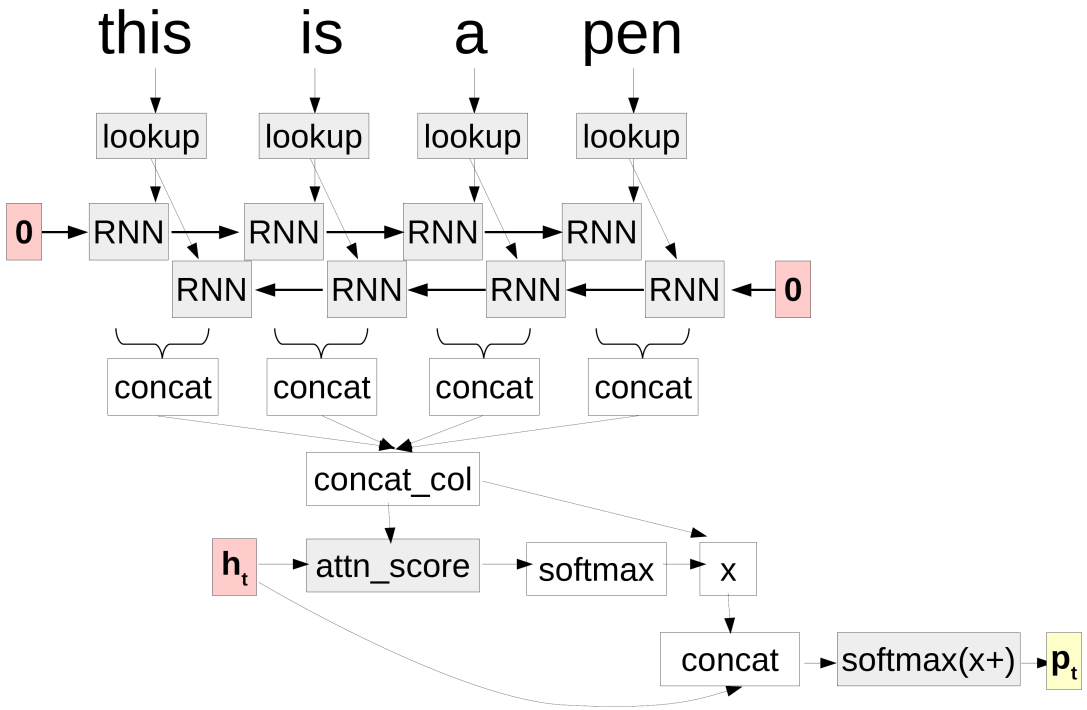


Figure 28: A computation graph for attention.

Attention model
Various ways of calculating attention score

- Dot product

$$\text{attn\_score}(\boldsymbol{h}_j^{(f)}, \boldsymbol{h}_t^{(e)}) := \boldsymbol{h}_j^{(f)\intercal} \boldsymbol{h}_t^{(e)}$$

- Bilinear function

$$\text{attn\_score}(\boldsymbol{h}_j^{(f)}, \boldsymbol{h}_t^{(e)}) := \boldsymbol{h}_j^{(f)\intercal} W_a \boldsymbol{h}_t^{(e)}$$

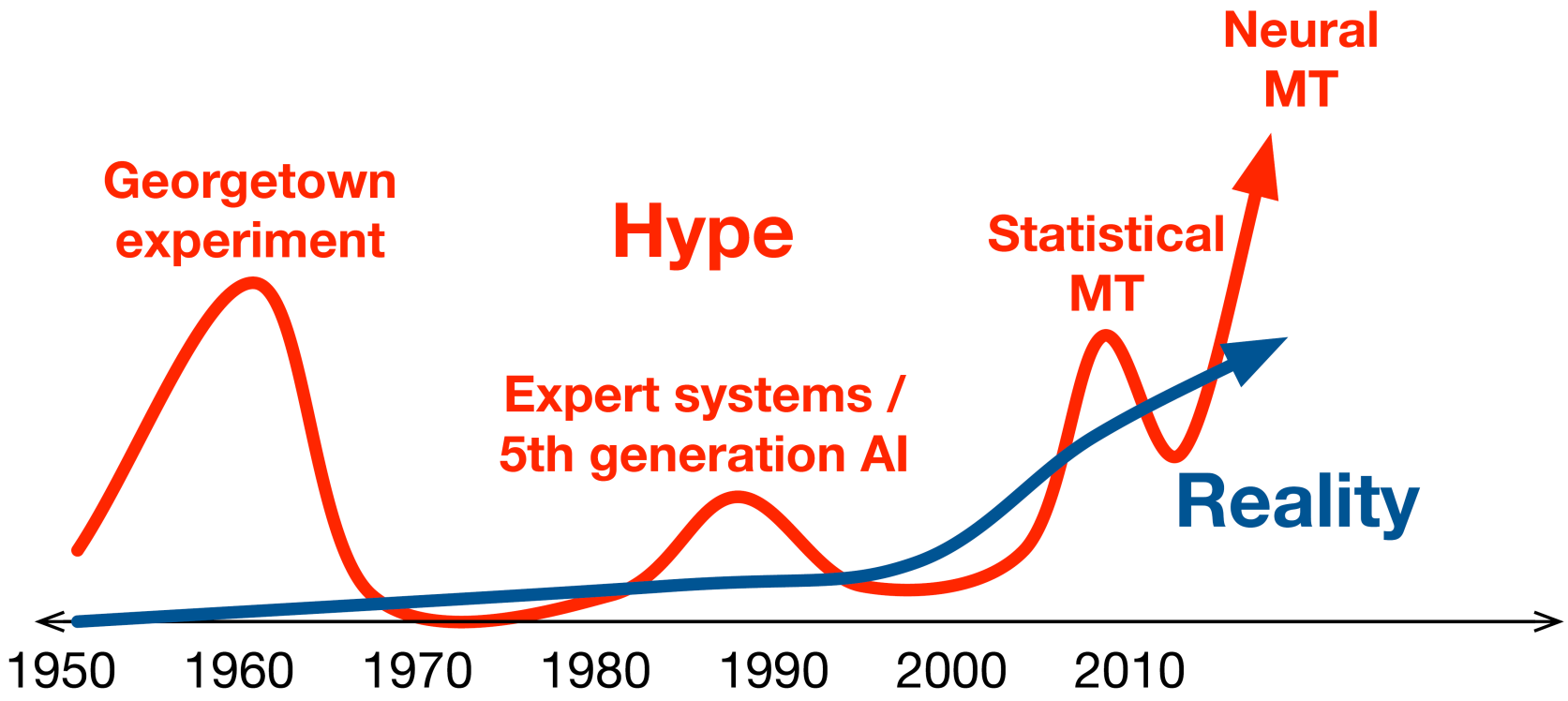
- Multi-layer perceptron (original formulation in Bahdanau et al.)

$$\text{attn\_score}(\boldsymbol{h}_t^{(e)}, \boldsymbol{h}_j^{(f)}) := \boldsymbol{w}_{a2}^{\intercal} \tanh(W_{a1}[\boldsymbol{h}_t^{(e)}; \boldsymbol{h}_j^{(f)}])$$

Advantages of attention

- Helps illustrate/interpret translation decisions
- Can help insert translations for OOV
  - By copying or look up in external dictionary
- Can incorporate linguistically motivated priors in model

MT History: Hype vs. Reality



1947

When I look at an article in Russian, I say to myself: This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.

Warren Weaver

1950s-1960s

- 1954 Georgetown-IBM experiment
  - 250 words, 6 grammar rules

- 1966 ALPAC report
  - Skeptical in research progress
  - Led to decreased US government funding for MT

Automatic Evaluation Metrics

- Goal: computer program that computes quality of translations
- Advantages: low cost, optimizable, consistent
- Basic strategy
  - Given: MT output
  - Given: human reference translation
  - Task: compute similarity between them

Precision and Recall of Words

SYSTEM A: Israeli officials ~~responsibility~~ ~~of~~ airport ~~safety~~

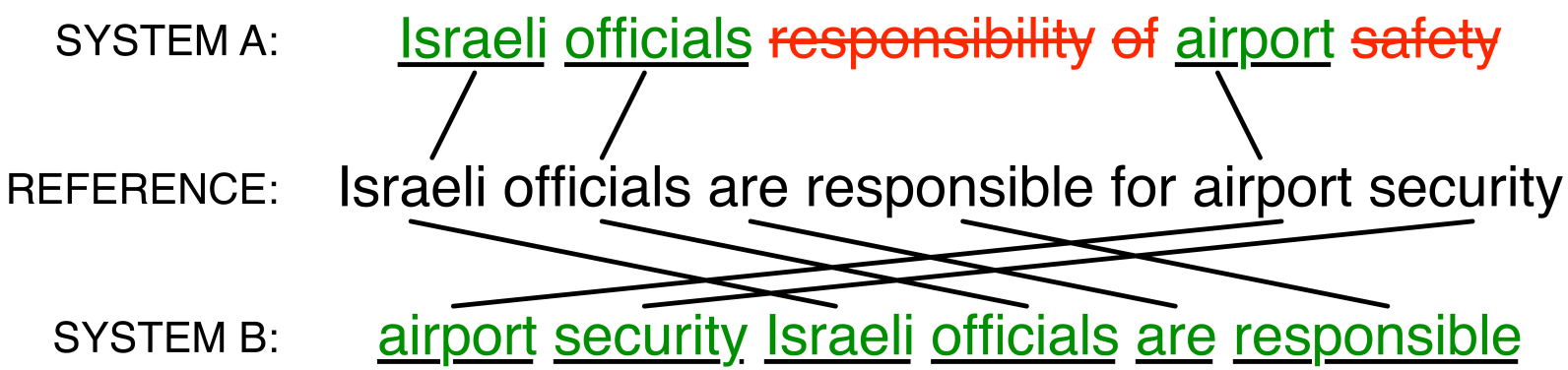
REFERENCE: Israeli officials are responsible for airport security

Precision $$\frac{\textit{correct}}{\textit{output-length}} = \frac{3}{6} = 50\%$$

Recall $$\frac{\textit{correct}}{\textit{reference-length}} = \frac{3}{7} = 43\%$$

F-measure $$\frac{\textit{precision} \times \textit{recall}}{(\textit{precision} + \textit{recall})/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%$$

Precision and Recall of Words



Metric	System A	System B
precision	50%	100%
recall	43%	100%
f-measure	46%	100%

flaw: no penalty for reordering

Word Error Rate

Minimum number of editing steps to transform output to reference

match: words match, no cost
substitution: replace one word with another
insertion: add word
deletion: drop word

Levenshtein distance

$$\text{WER} = \frac{\textit{substitutions} + \textit{insertions} + \textit{deletions}}{\textit{reference-length}}$$

WER example

		Israeli	officials	responsibility	of	airport	safety
	0	1	2	3	4	5	6
Israeli	1	0	1	2	3	4	5
officials	2	1	0	1	2	3	4
are	3	2	1	1	2	3	4
responsible	4	3	2	2	2	3	4
for	5	4	3	3	3	3	4
airport	6	5	4	4	4	3	4
security	7	6	5	5	5	4	4

		airport	security	Israeli	officials	are	responsible
	0	1	2	3	4	5	6
Israeli	1	1	2	2	3	4	5
officials	2	2	2	3	2	3	4
are	3	3	3	3	3	2	3
responsible	4	4	4	4	4	3	2
for	5	5	5	5	5	4	3
airport	6	5	6	6	6	5	4
security	7	6	5	6	7	6	5

Metric	System A	System B
word error rate (WER)	57%	71%

BLEU
Bilingual Evaluation Understudy

N-gram overlap between machine translation output and reference translation

Compute precision for n-grams of size 1 to 4

Add brevity penalty (for too short translations)

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \left(\prod_{i=1}^{4} \textit{precision}_i\right)^{\frac{1}{4}}$$

Typically computed over the entire corpus, not single sentences

Multiple Reference Translations

To account for variability, use multiple reference translations

- n-grams may match in any of the references
- closest reference length used

Example

SYSTEM: Israeli officials (2-GRAM MATCH) responsibility of (2-GRAM MATCH) airport (1-GRAM) safety

REFERENCES:
Israeli officials are responsible for airport security
Israel is in charge of the security at this airport
The security work for this airport is the responsibility of the Israel government
Israeli side was in charge of the security of this airport

BLEU examples

SYSTEM A: Israeli officials responsibility of airport safety
2-GRAM MATCH 1-GRAM MATCH

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible
2-GRAM MATCH 4-GRAM MATCH

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

Semantics-aware metrics: e.g., METEOR

Partial credit for matching stems

SYSTEM	Jim went home
REFERENCE	Joe goes home

Partial credit for matching synonyms

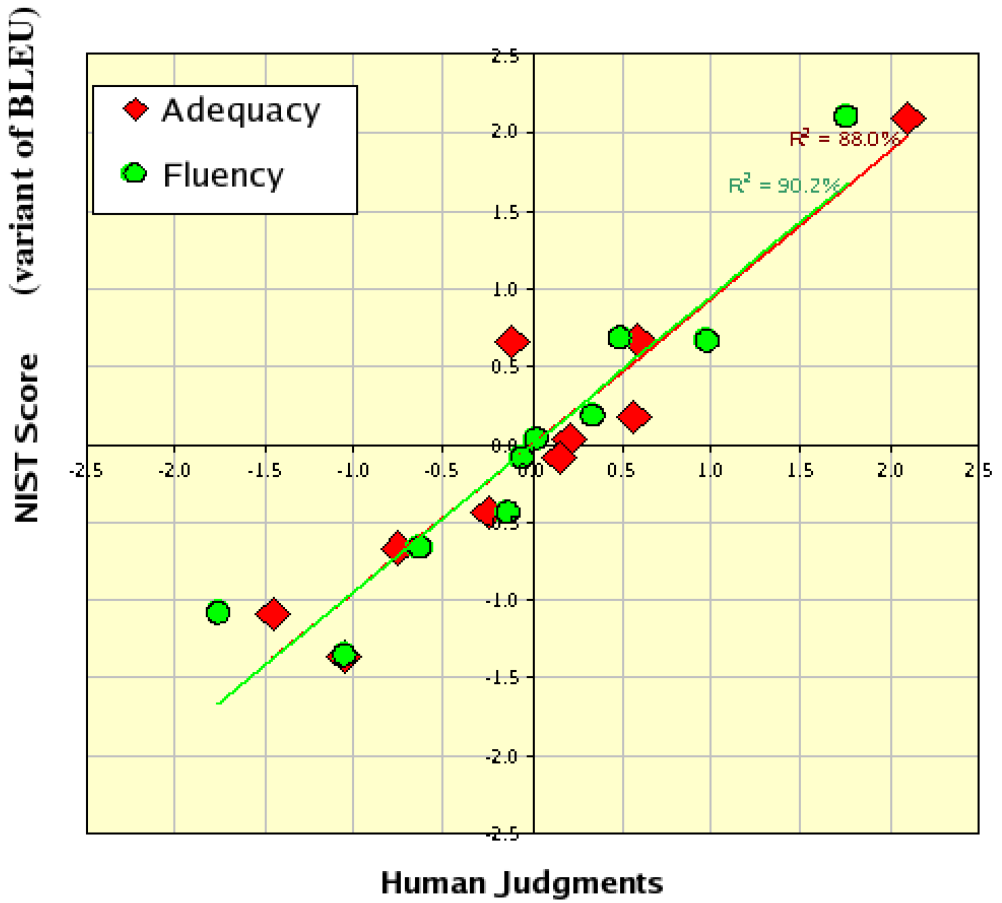
SYSTEM	Jim walks home
REFERENCE	Joe goes home

Use of paraphrases

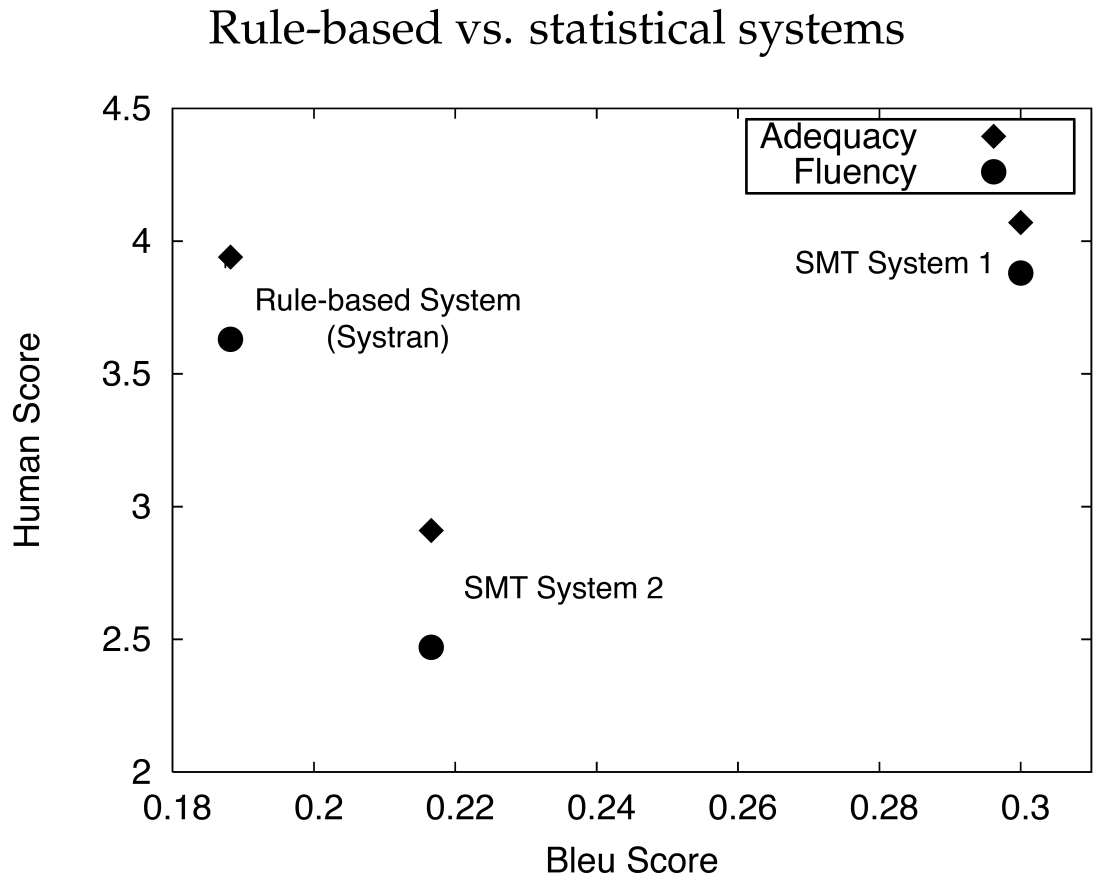
Drawbacks of Automatic Metrics

- All words are treated as equally relevant
- Operate on local level
- Scores are meaningless (absolute value not informative)
- Human translators score low on BLEU

Yet automatic metrics such as BLEU correlate with human judgement



Caveats: bias toward statistical systems



Automatic metrics

- Essential tool for system development
- Use with caution: not suited to rank systems of different types
- Still an open area of research
  - Connects with semantic analysis

Extensions of encoder-decoder

- Tree-structured outputs
- Additional mechanisms:
  - Copy
  - Coverage
- Alternative attentions:
  - Attend to multiple sentences
  - Attend to a sentence and image

Today

- Sequence generation when no 1-1 correspondence between input and output
- Encoder-decoder
- Attention
- Evaluation
- Later directions