

Natural Language Processing with Deep Learning

CS224N/Ling284



Christopher Manning

Lecture 9: More on Gated Recurrent Units and NMT

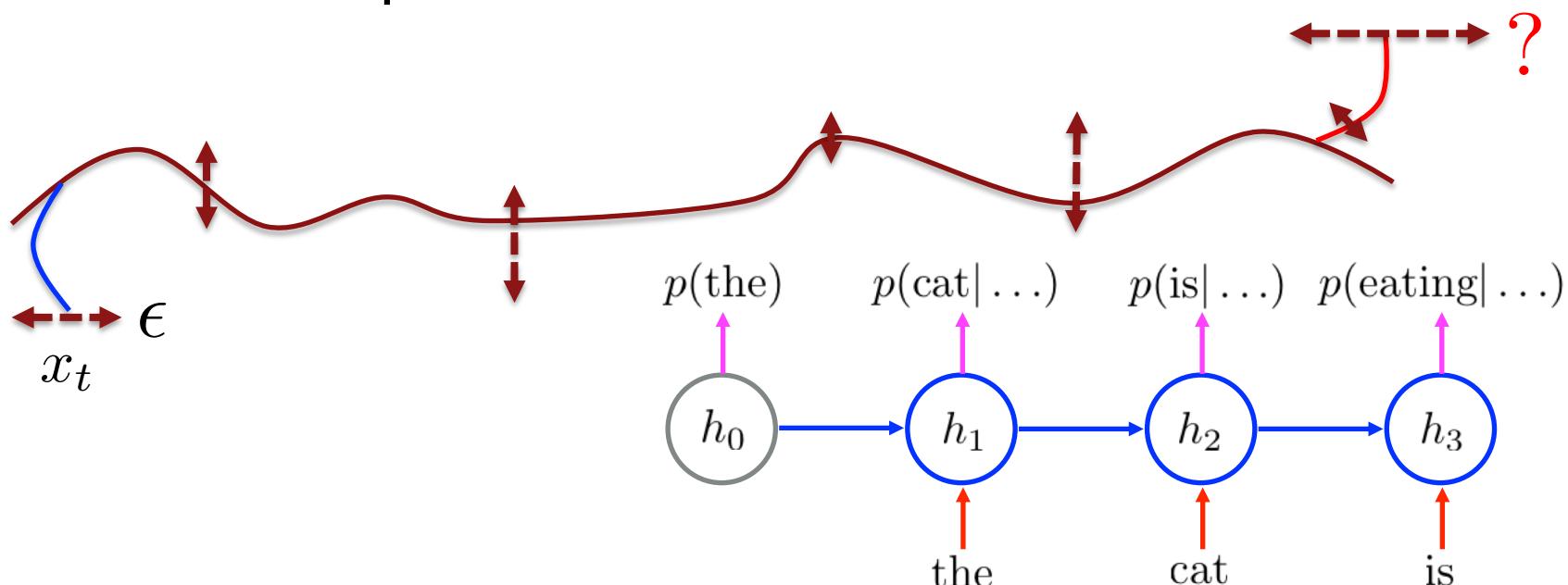
4. One more look at gated recurrent units and MT

Intuitively, what happens with RNNs?

1. Measure the influence of the past on the future

$$\frac{\partial \log p(x_{t+n} | x_{$$

2. How does the perturbation at t affect $p(x_{t+n} | x_{?$



Backpropagation through Time

Problem: Vanishing gradient is super-problematic

- When gradient goes to zero, we cannot tell whether
 1. No dependency between t and $t+n$ in data, or
 2. Wrong configuration of parameters (the vanishing gradient condition)
- Is the problem with the naïve transition function?

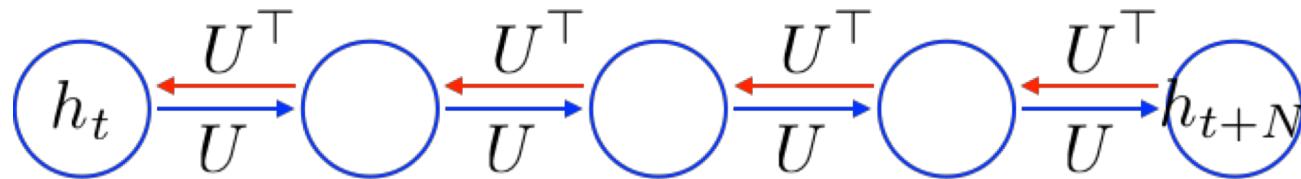
$$f(h_{t-1}, x_t) = \tanh(W [x_t] + Uh_{t-1} + b)$$

- With it, the temporal derivative leads to vanishing

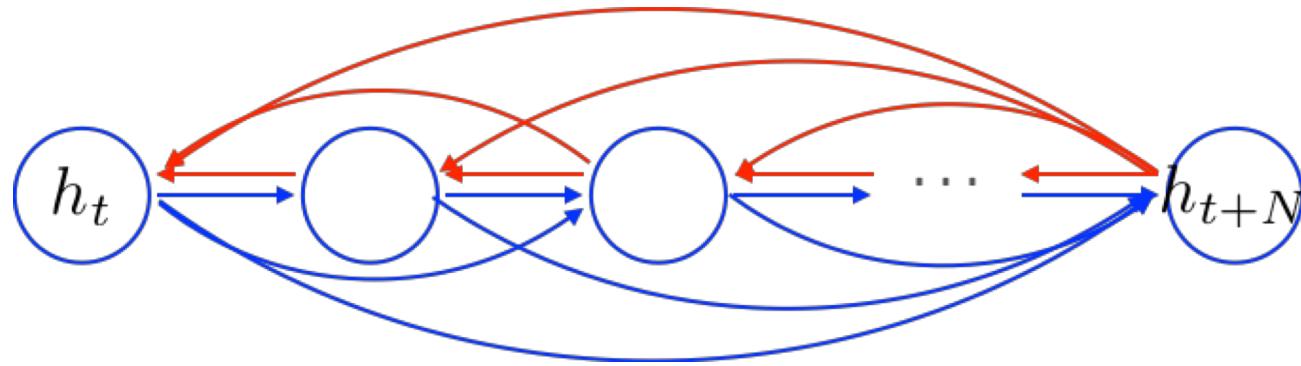
$$\frac{\partial h_{t+1}}{\partial h_t} = U^\top \frac{\partial \tanh(a)}{\partial a}$$

Gated Recurrent Unit

- It implies that the error must backpropagate through all the intermediate nodes:

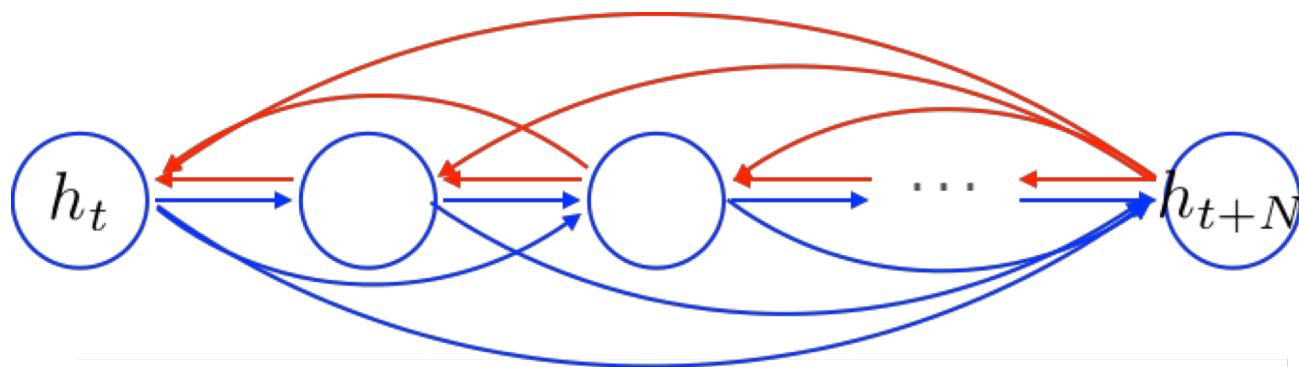


- Perhaps we can create shortcut connections.



Gated Recurrent Unit

- Perhaps we can create *adaptive* shortcut connections.

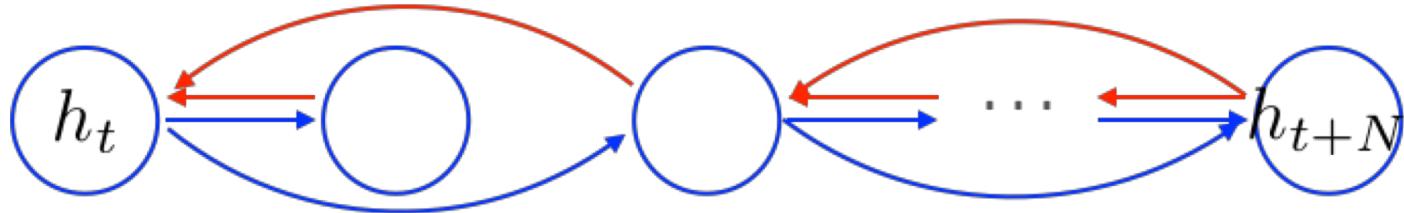


$$f(h_{t-1}, x_t) = u_t \odot \tilde{h}_t + (1 - u_t) \odot h_{t-1}$$

- Candidate Update $\tilde{h}_t = \tanh(W [x_t] + U h_{t-1} + b)$
- Update gate $u_t = \sigma(W_u [x_t] + U_u h_{t-1} + b_u)$

Gated Recurrent Unit

- Let the net prune unnecessary connections *adaptively*.

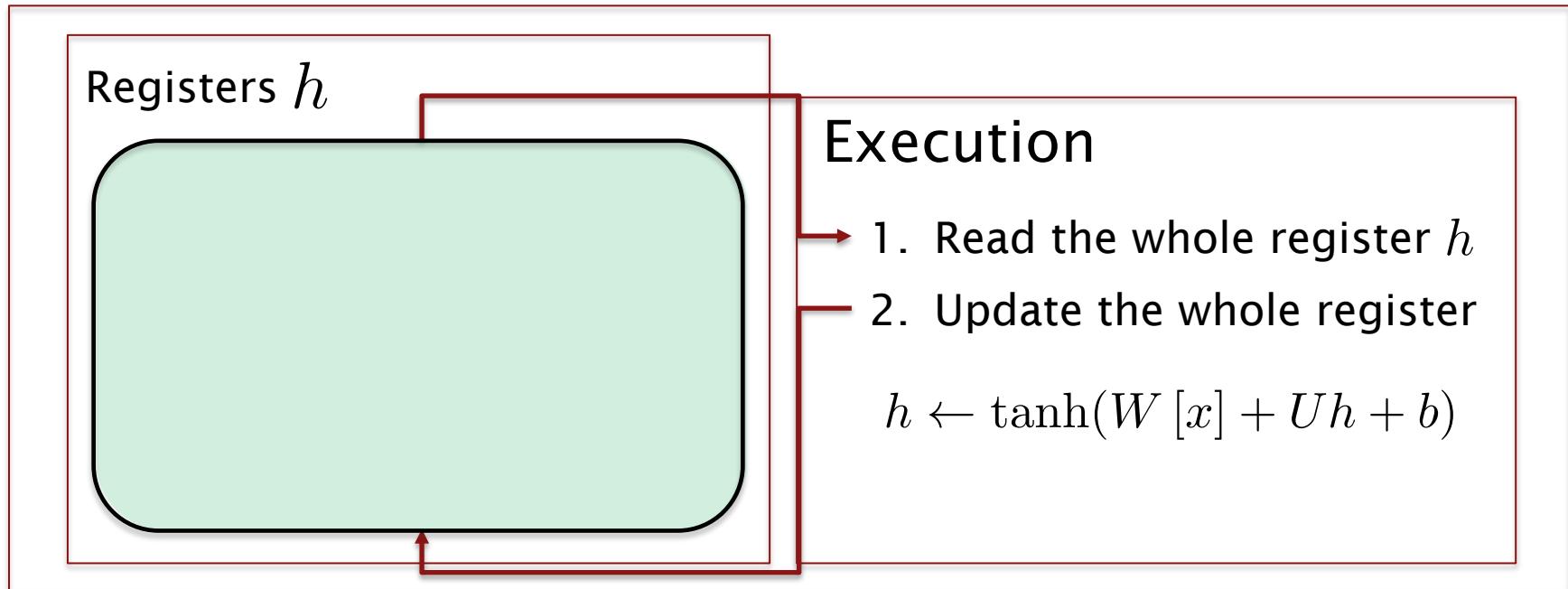


$$f(h_{t-1}, x_t) = u_t \odot \tilde{h}_t + (1 - u_t) \odot h_{t-1}$$

- Candidate Update $\tilde{h}_t = \tanh(W [x_t] + U(r_t \odot h_{t-1}) + b)$
- Reset gate $r_t = \sigma(W_r [x_t] + U_r h_{t-1} + b_r)$
- Update gate $u_t = \sigma(W_u [x_t] + U_u h_{t-1} + b_u)$

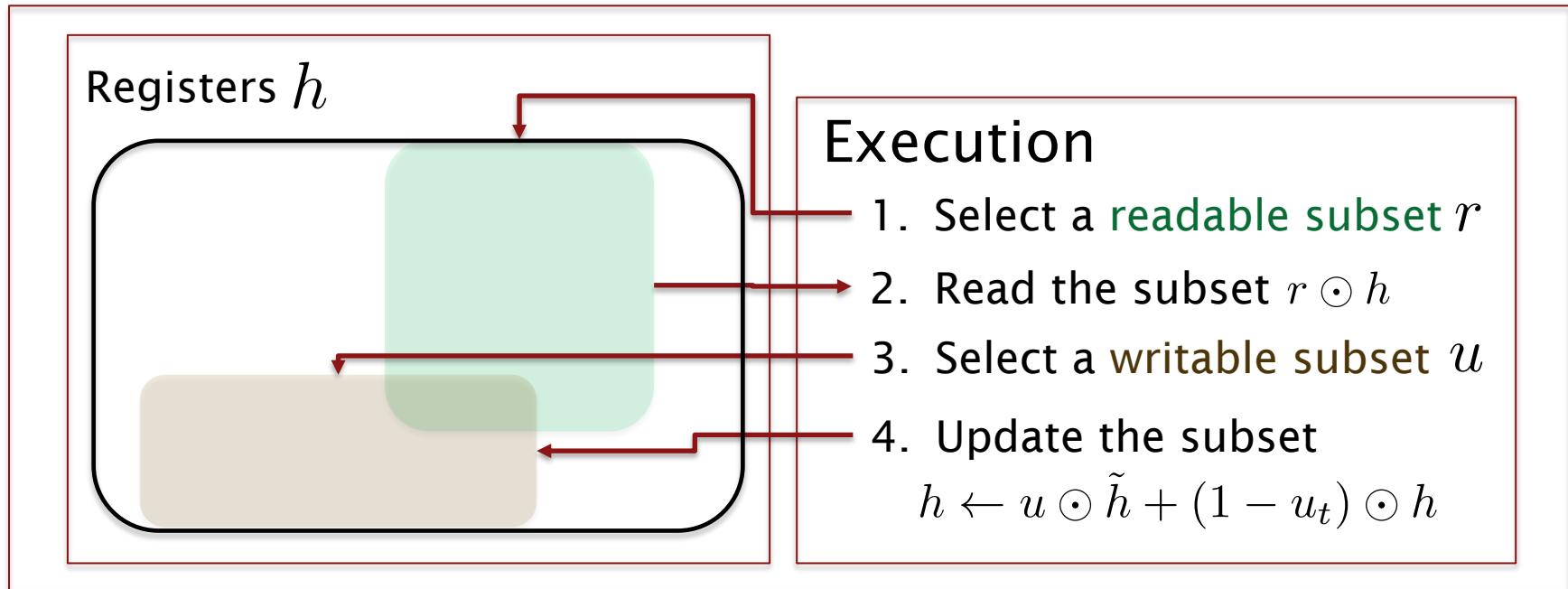
Gated Recurrent Unit

tanh-RNN



Gated Recurrent Unit

GRU ...



Gated recurrent units are much more realistic!
Note that there is some overlap in ideas with attention

Gated Recurrent Units

Two most widely used gated recurrent units: GRU and LSTM

Gated Recurrent Unit

[Cho et al., EMNLP2014;
Chung, Gulcehre, Cho, Bengio, DLUFL2014]

$$h_t = u_t \odot \tilde{h}_t + (1 - u_t) \odot h_{t-1}$$

$$\tilde{h}_t = \tanh(W [x_t] + U(r_t \odot h_{t-1}) + b)$$

$$u_t = \sigma(W_u [x_t] + U_u h_{t-1} + b_u)$$

$$r_t = \sigma(W_r [x_t] + U_r h_{t-1} + b_r)$$

Long Short-Term Memory

[Hochreiter & Schmidhuber, NC1999;
Gers, Thesis2001]

$$h_t = o_t \odot \tanh(c_t)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

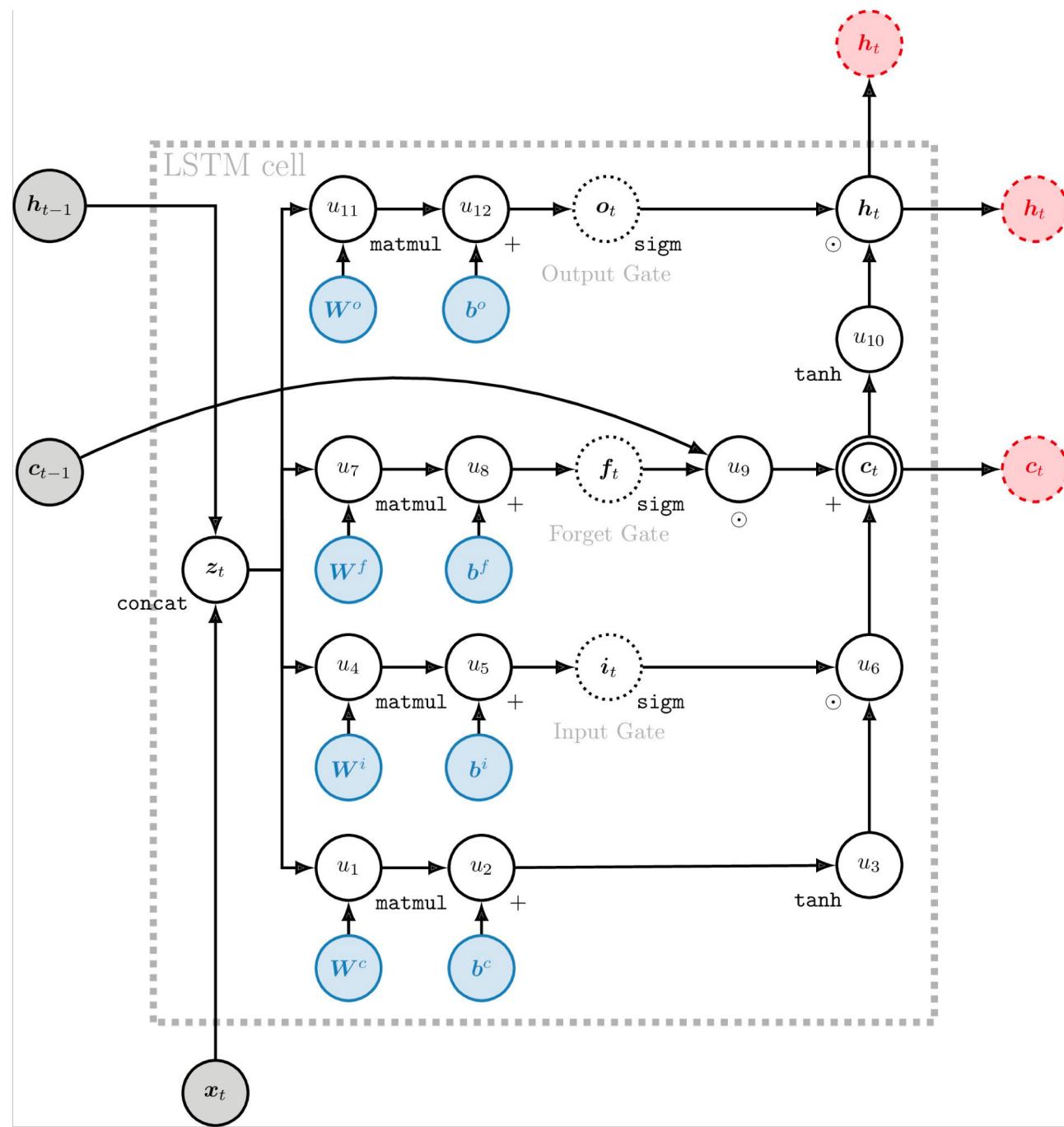
$$\tilde{c}_t = \tanh(W_c [x_t] + U_c h_{t-1} + b_c)$$

$$o_t = \sigma(W_o [x_t] + U_o h_{t-1} + b_o)$$

$$i_t = \sigma(W_i [x_t] + U_i h_{t-1} + b_i)$$

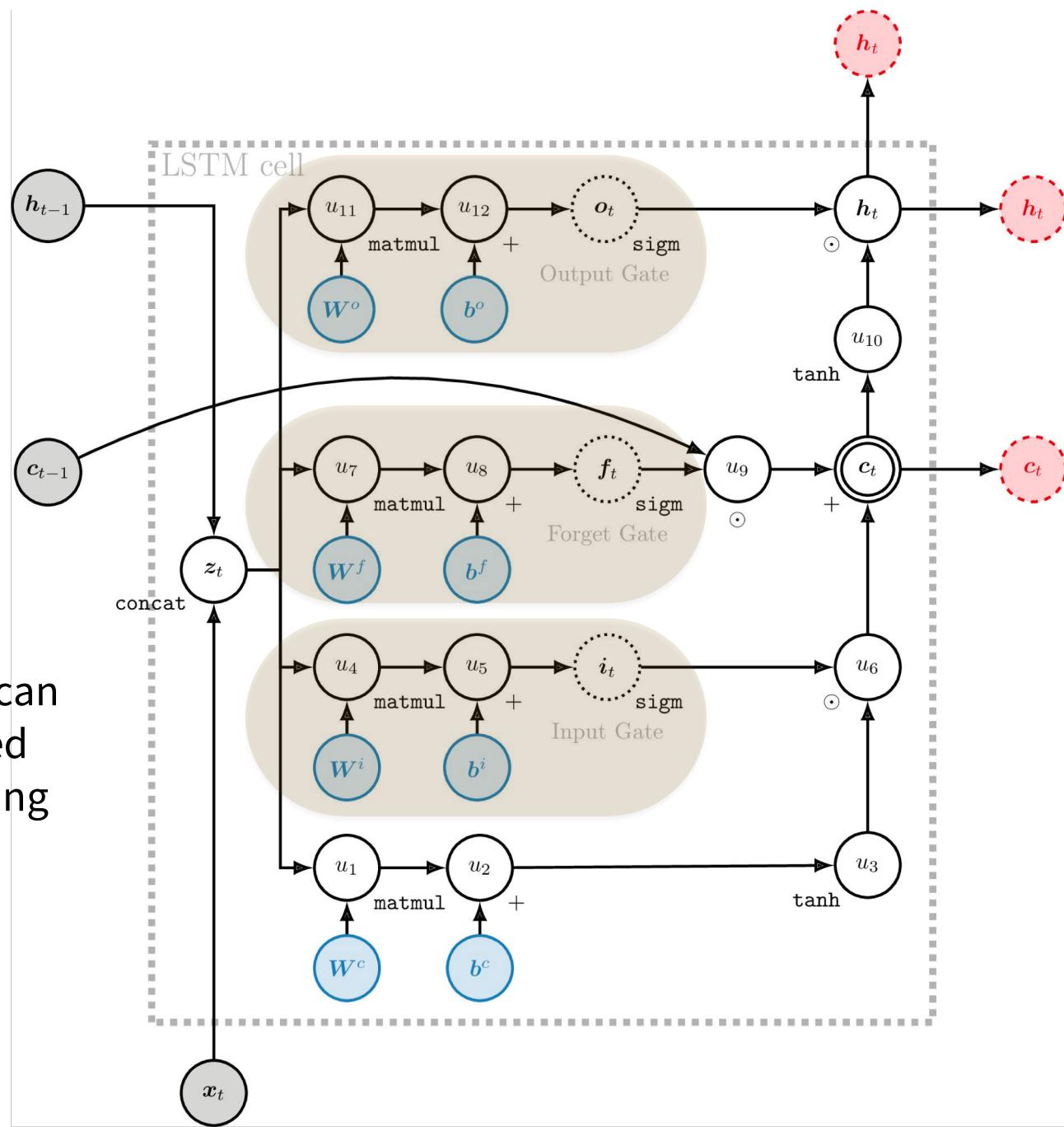
$$f_t = \sigma(W_f [x_t] + U_f h_{t-1} + b_f)$$

The LSTM

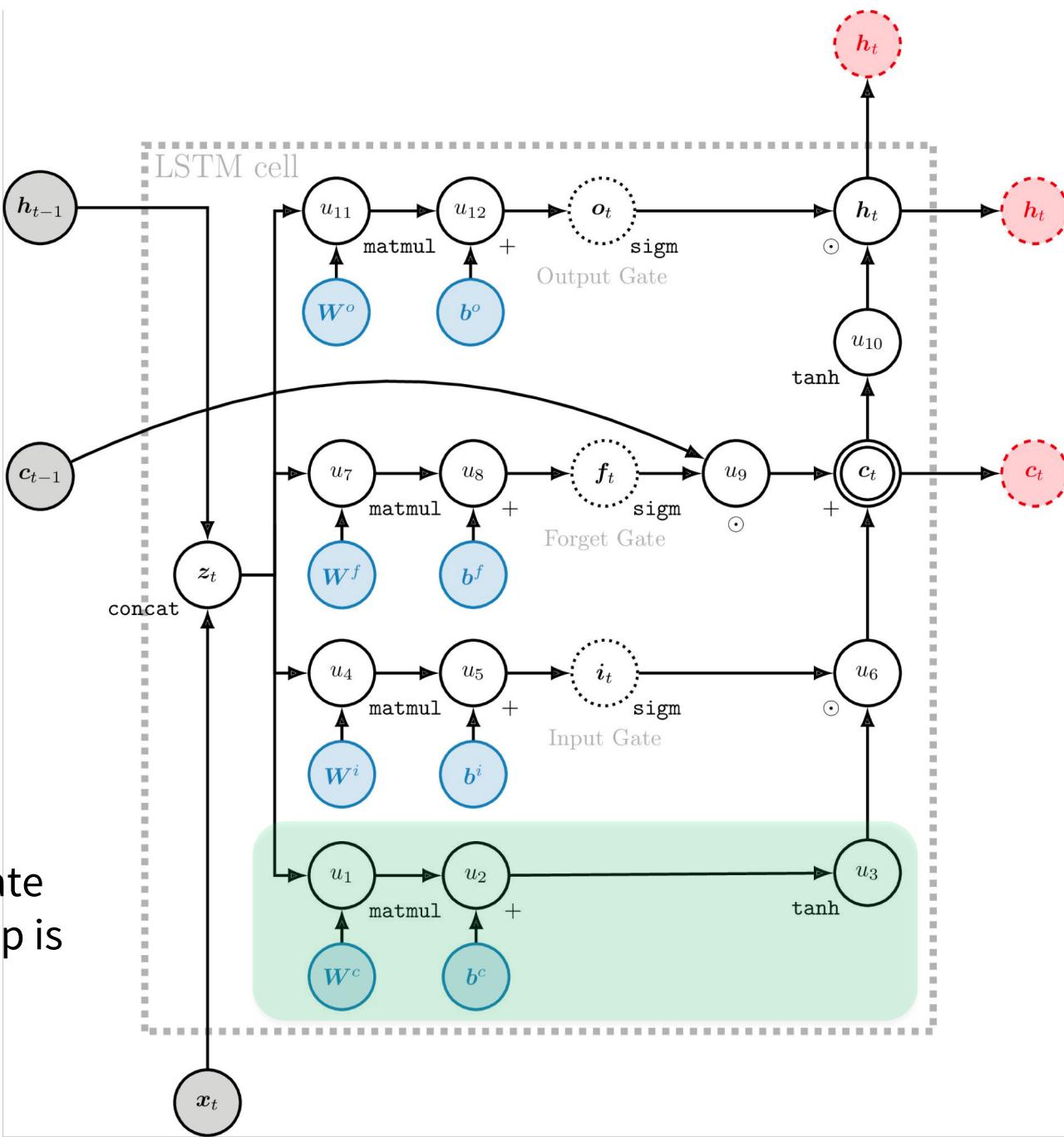


The LSTM

The LSTM gates all operations so stuff can be forgotten/ignored rather than it all being crammed on top of everything else



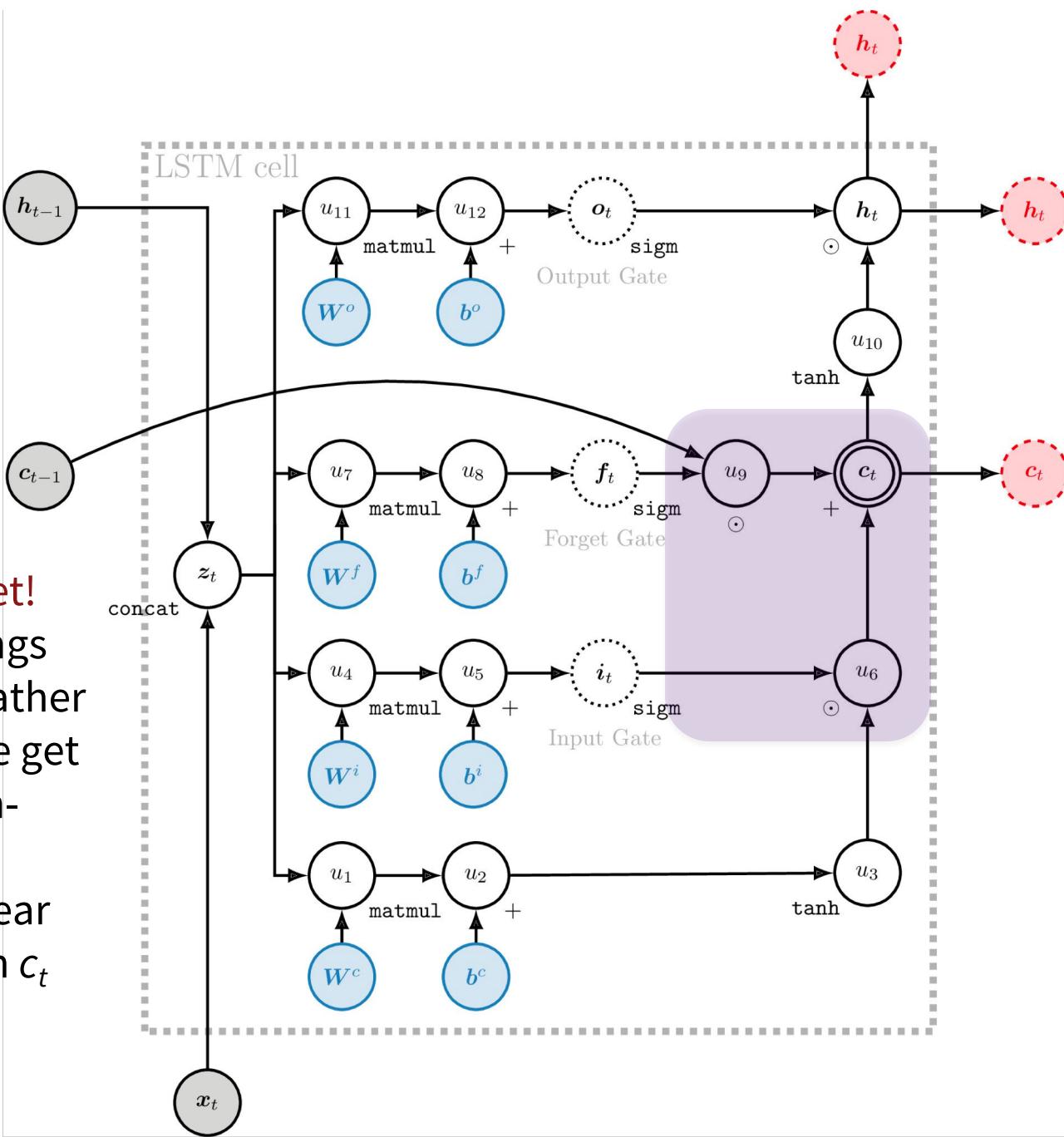
The LSTM



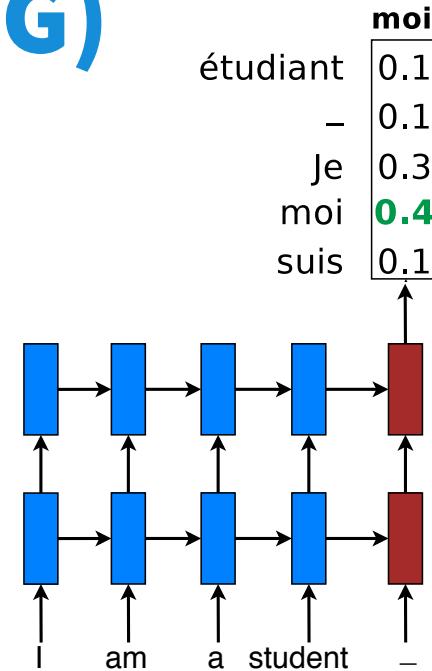
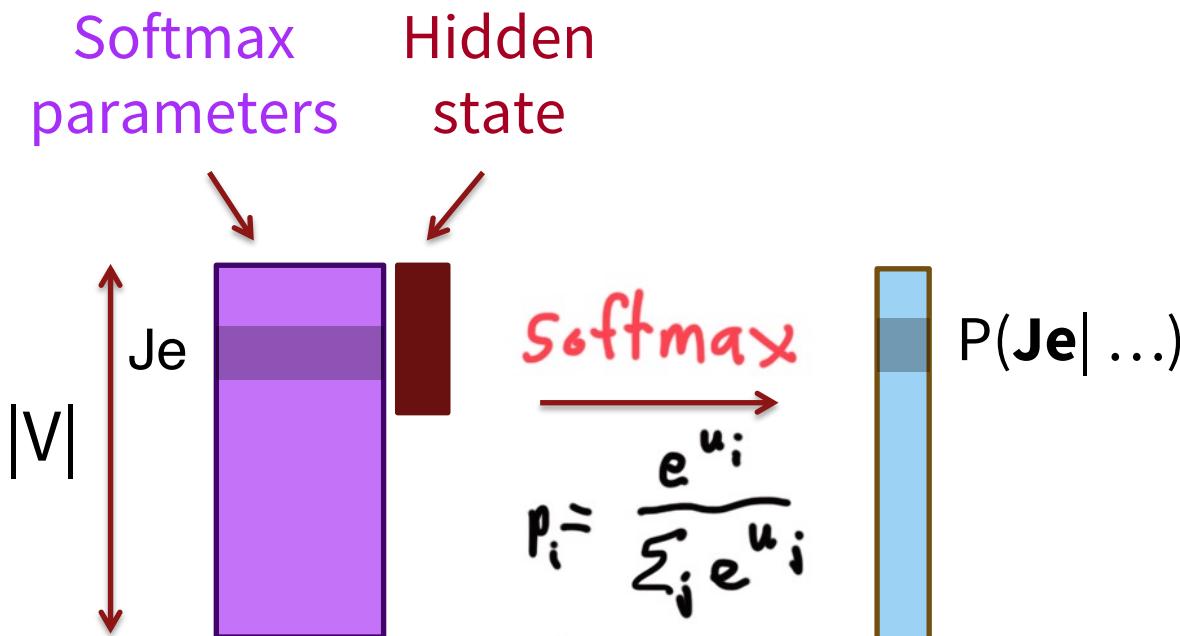
The non-linear update
for the next time step is
just like an RNN

The LSTM

This part is the secret!
(Of other recent things like ResNets too!) Rather than multiplying, we get c_t by adding the non-linear stuff and c_{t-1} !
There is a direct, linear connection between c_t and c_{t-1} .



5. The large output vocabulary problem in NMT (or all NLG)



Softmax computation is expensive.

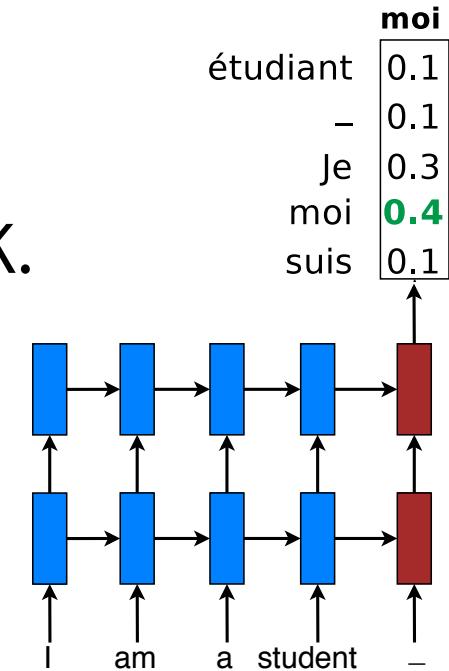
The word generation problem

- Word generation problem
 - Vocabs used are usually modest: 50K.

The ecotax portico in Pont-de-Buis
Le portique écotaxe de Pont-de-Buis



The <unk> portico in <unk>
Le <unk> <unk> de <unk>



Possible approaches for output

- *Hierarchical softmax*: tree-structured vocabulary
- *Noise-contrastive estimation*: binary classification
- *Train* on a subset of the vocabulary at a time;
test on a smart on the set of possible translations
 - Jean, Cho, Memisevic, Bengio. ACL2015
- *Use attention to work out what you are translating*:
You can do something simple like dictionary lookup
- *More ideas we will get to*: Word pieces; char. models

MT Evaluation – an example of eval

- Manual (the best!?):
 - **Adequacy and Fluency** (5 or 7 point scales)
 - Error categorization
 - **Comparative ranking of translations**
- Testing in an application that uses MT as one sub-component
 - E.g., question answering from foreign language documents
 - May not test many aspects of the translation (e.g., cross-lingual IR)
- Automatic metric:
 - **BLEU (Bilingual Evaluation Understudy)**
 - Others like TER, METEOR, ...

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?
 - An n-gram is a sequence of n words
 - Not allowed to match same portion of reference translation twice at a certain n-gram level (two MT words *airport* are only correct if two reference words *airport*; can't cheat by typing out “the the the the”)
 - Do count unigrams also in a bigram for unigram precision, etc.

Brevity Penalty

- Brevity Penalty
 - Can't just type out single word “the” (precision 1.0!)
- It was thought quite hard to “game” the system (i.e., to find a way to change machine output so that BLEU goes up, but quality doesn't)

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.

- BLEU is a weighted geometric mean, with a brevity penalty factor added.
 - Note that it's precision-oriented
- BLEU4 formula
(counts n-grams up to length 4)

$$\exp \left(0.5 * \log p_1 + 0.25 * \log p_2 + 0.125 * \log p_3 + 0.125 * \log p_4 - \max(\text{words-in-reference} / \text{words-in-machine} - 1, 0) \right)$$

p_1 = 1-gram precision

p_2 = 2-gram precision

p_3 = 3-gram precision

p_4 = 4-gram precision

Note: only works at corpus level (zeroes kill it); there's a smoothed variant for sentence-level

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 al by the police .	#7
the ringer is killed by the police .	#8
police killed the gunman .	#9
	#10

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

Multiple Reference Translations

Reference translation 1:

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.

Reference translation 2:

Guam International Airport and its offices are maintaining a high state of alert after receiving an e-mail that was from a person claiming to be the wealthy Saudi Arabian businessman Bin Laden and that threatened to launch a biological and chemical attack on the airport and other public places.

Machine translation:

The American [?] international airport and its 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.

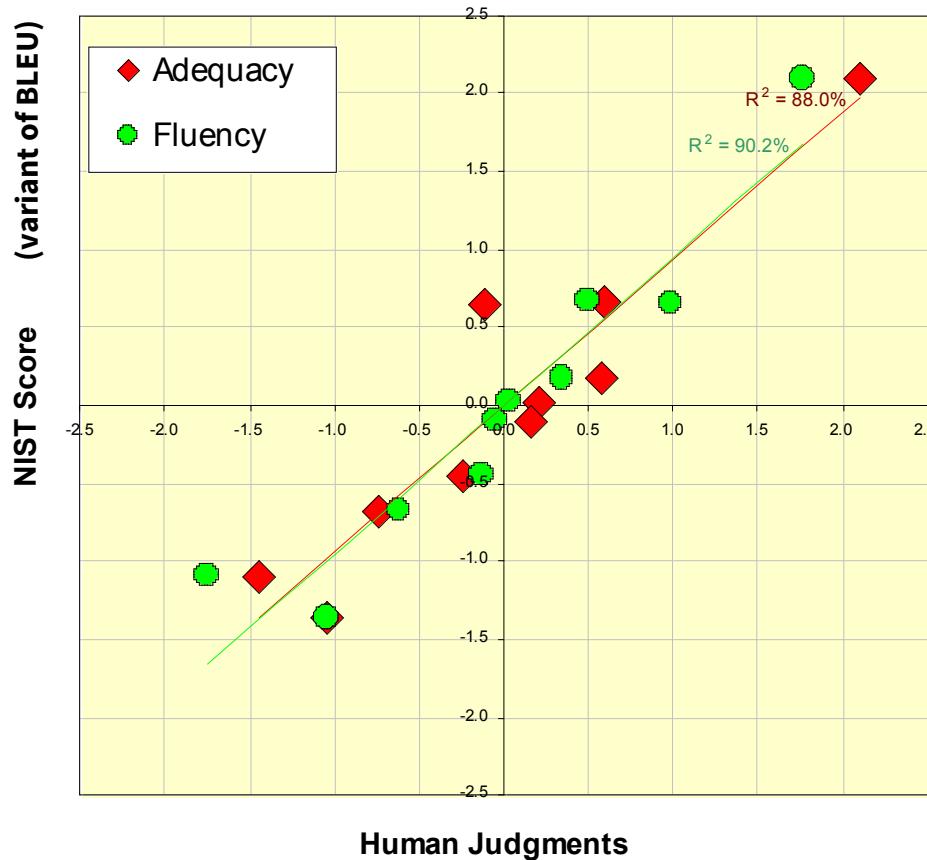
Reference translation 3:

The US International Airport of Guam and its office has received an email from a self-claimed Arabian millionaire named Laden , which threatens to launch a biochemical attack on such public places as airport . Guam authority has been on alert .

Reference translation 4:

US Guam International Airport and its office received an email from Mr. Bin Laden and other rich businessman from Saudi Arabia . They said there would be biochemistry air raid to Guam Airport and other public places . Guam needs to be in high precaution about this matter .

Initial results showed that BLEU predicts human judgments well



slide from G. Doddington (NIST)

Automatic evaluation of MT

- People started optimizing their systems to maximize BLEU score
 - BLEU scores improved rapidly
 - The correlation between BLEU and human judgments of quality went way, way down
 - MT BLEU scores now approach those of human translations but their true quality remains far below human translations
- Coming up with automatic MT evaluations has become its own research field
 - There are many proposals: TER, METEOR, MaxSim, SEPIA, our own RTE-MT
 - TERpA is a representative good one that handles some word choice variation.
- MT research **requires** some automatic metric to allow a rapid development and evaluation cycle.