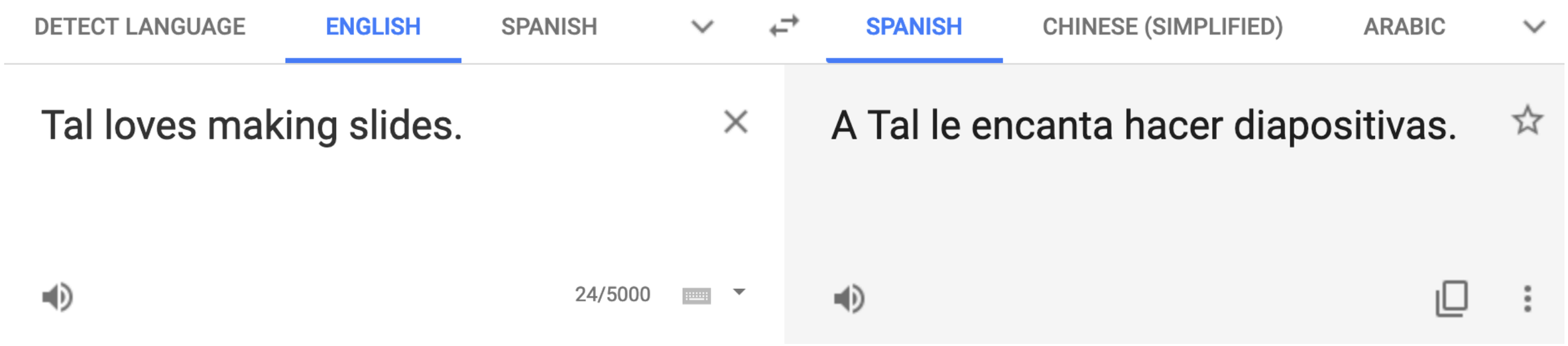
Conditioned language modeling for machine translation



- Spanish language model: $P(w_n \mid w_1 \ldots, w_{n-1})$
- Conditioned language model (Spanish given English): $P(w_n \mid x, w_1 \ldots, w_{n-1})$
- Goal in machine translation: find $\arg\max_{y \in Y} P(y \mid x)$

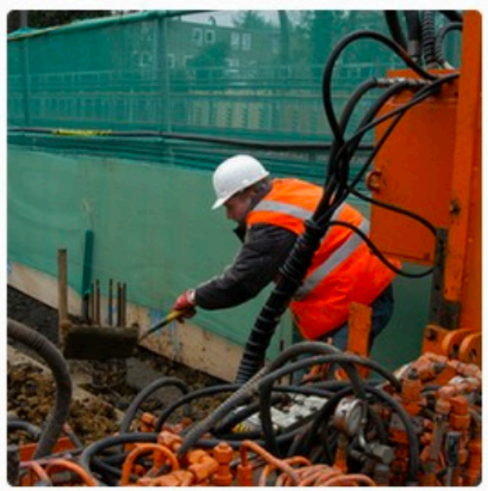
Training data for probabilistic machine translation: a parallel corpus

Source sentence	Reference translation
My dear Athos, we are enveloped in a network of spies.	Mon cher Athos, nous sommes enveloppés dans un réseau d’espions!
These two great thoroughfares, intersected by the two first, formed the canvas upon which reposed, knotted and crowded together on every hand, the labyrinthine network of the streets of Paris.	Ces deux grandes voies, croisées avec les deux premières, formaient le canevas sur lequel reposait, noué and serré en tous sense, le réseau dédaléen des rues de Paris.
She flies, she is joyous, she is just born; she seeks the spring, the open air, liberty: oh, yes! But let her come in contact with the fatal network, and the spider issues from it, the hideous spider!	Elle vole, elle est joyeuse, elle vient de naître; elle cherche le printemps, le grand air, la liberté; oh! Oui, mais qu’elle se heurte à la rosace fatale, l’araignée en sort, l’araignée hideuse!

Conditioned language modeling: image captioning

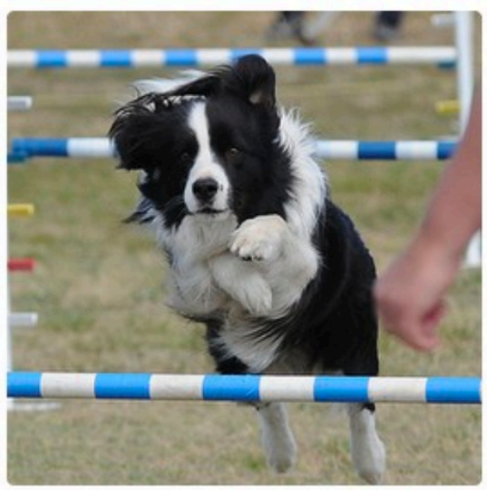
- Generate a caption given a image:

"man in black shirt is playing guitar."



"construction worker in orange safety vest is working on road."

"girl in pink dress is jumping in air."



"black and white dog jumps over bar."

Conditioned language modeling: summarization

- Generate a summary (or headline) given a document:

I(7): the white house on thursday warned iran of possible new sanctions after the un nuclear watchdog reported that tehran had begun sensitive nuclear work at a key site in defiance of un resolutions .
G: us warns iran of step backward on nuclear issue

I(11): russia 's gas and oil giant gazprom and us oil major chevron have set up a joint venture based in resource-rich northwestern siberia , the interfax news agency reported thursday quoting gazprom officials .
G: gazprom chevron set up joint venture

(Rush et al. 2015)

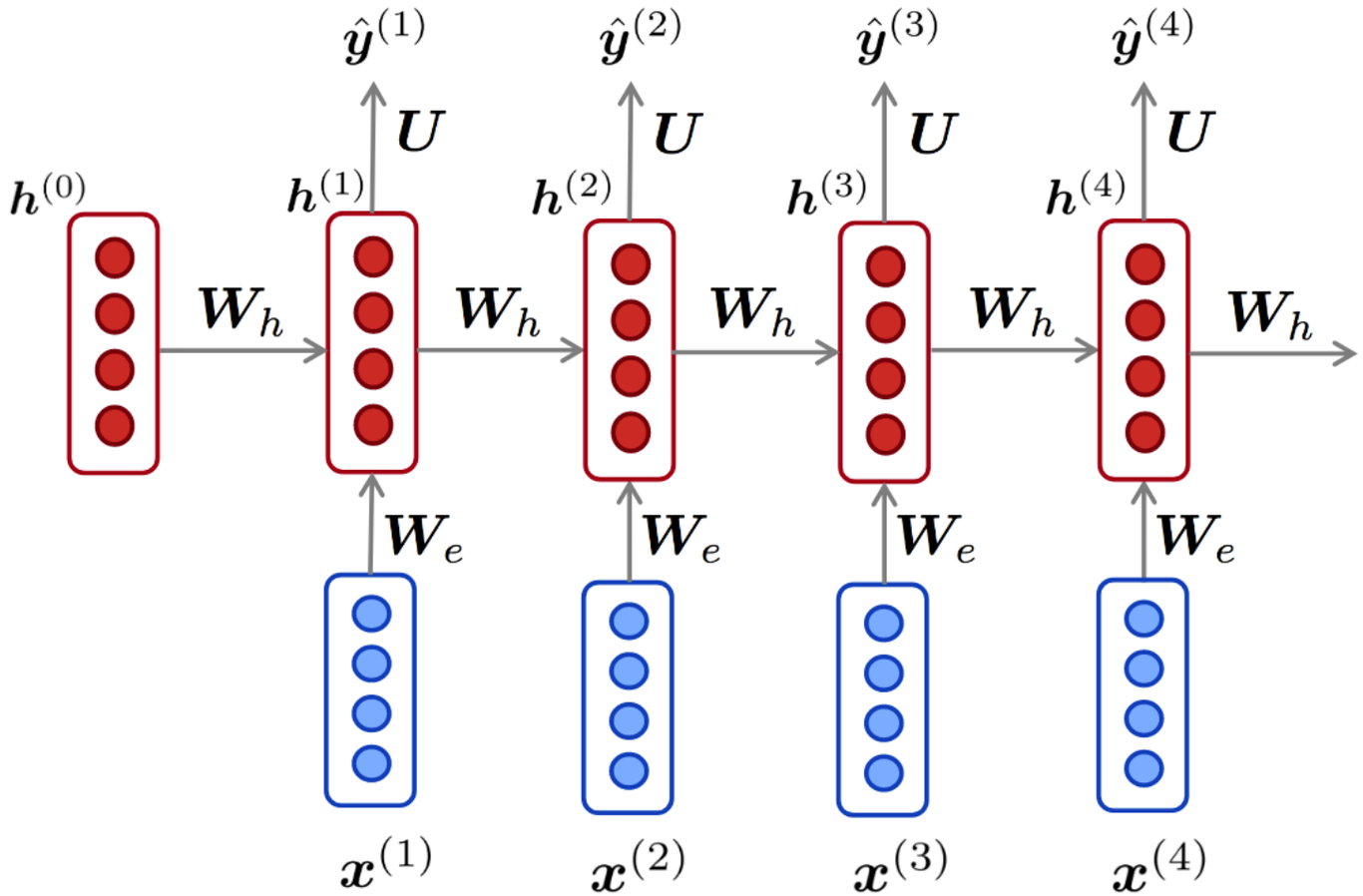
Conditioned language modeling: dialog response generation

- Generate the next turn in a dialog given the previous turns:

Role	response
User	Does money buy happiness?
Bot	Depends how much money you spend on it .
User	What is the best way to buy happiness ?
Bot	You just have to be a millionaire by your early 20s, then you can be happy .
User	This is so difficult !
Bot	You have no idea how hard it is to be a millionaire and happy . There is a reason the rich have a lot of money

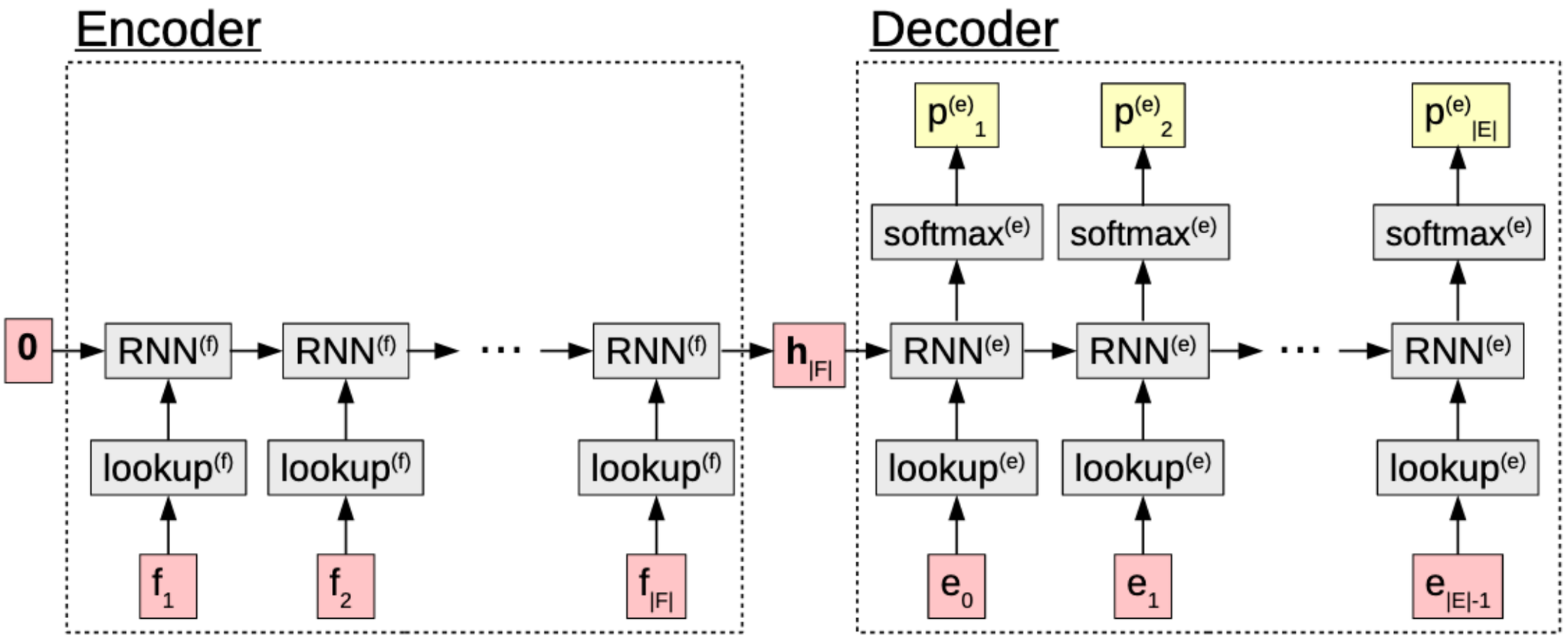
(Zhang et al. 2019)

Unconditional RNN LM



(Neubig 2017)

Encoder/decoder RNN



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

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

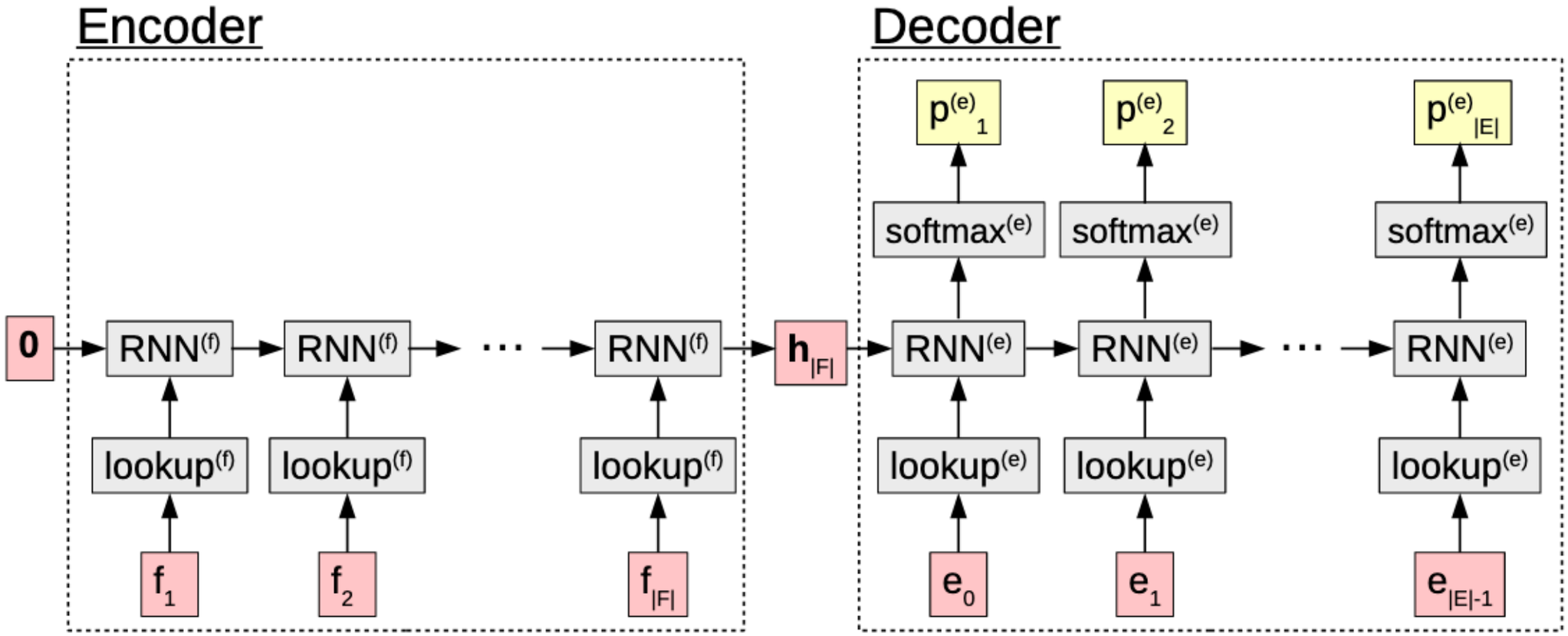
$$\boldsymbol{m}_t^{(e)} = M_{\cdot,e_{t-1}}^{(e)}$$

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

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

(Neubig 2017)

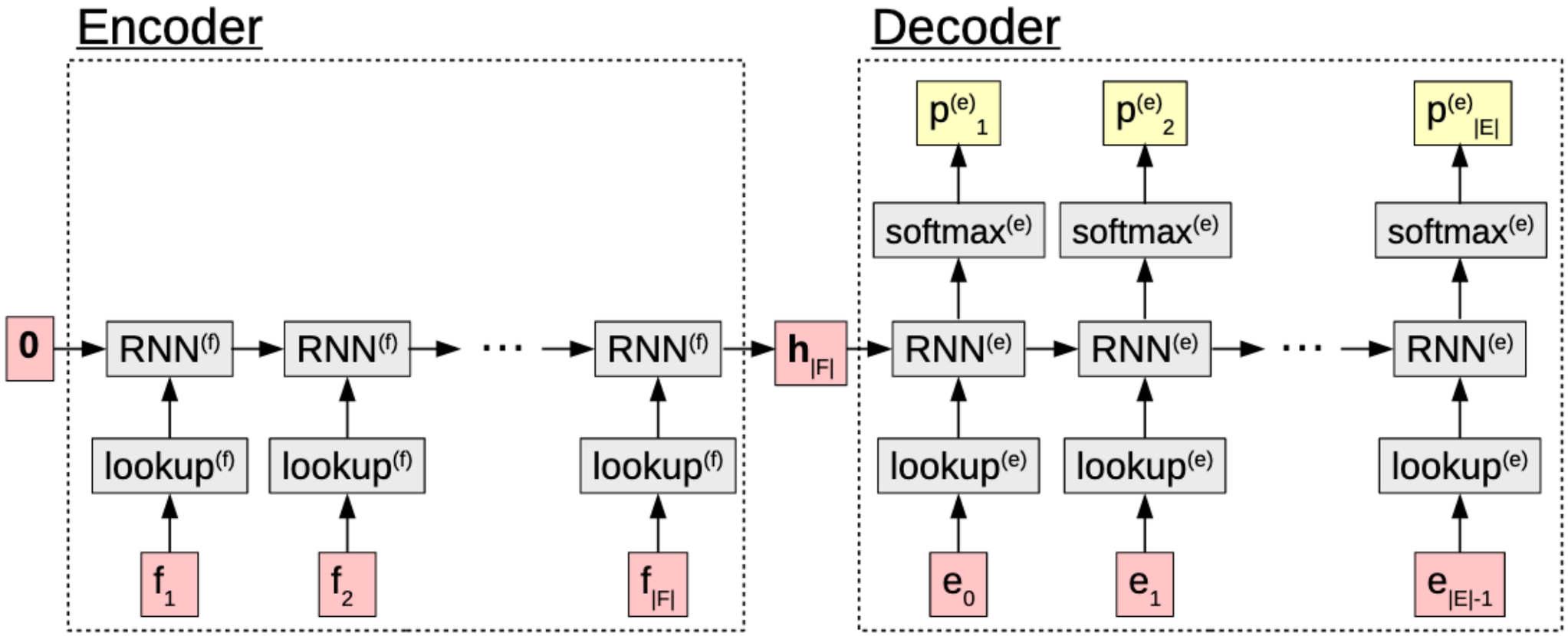
Decoding strategies: random sampling



Sample from $P(y \mid x)$:

Sample a word from $p_t^{(e)}$ until we sample </s>

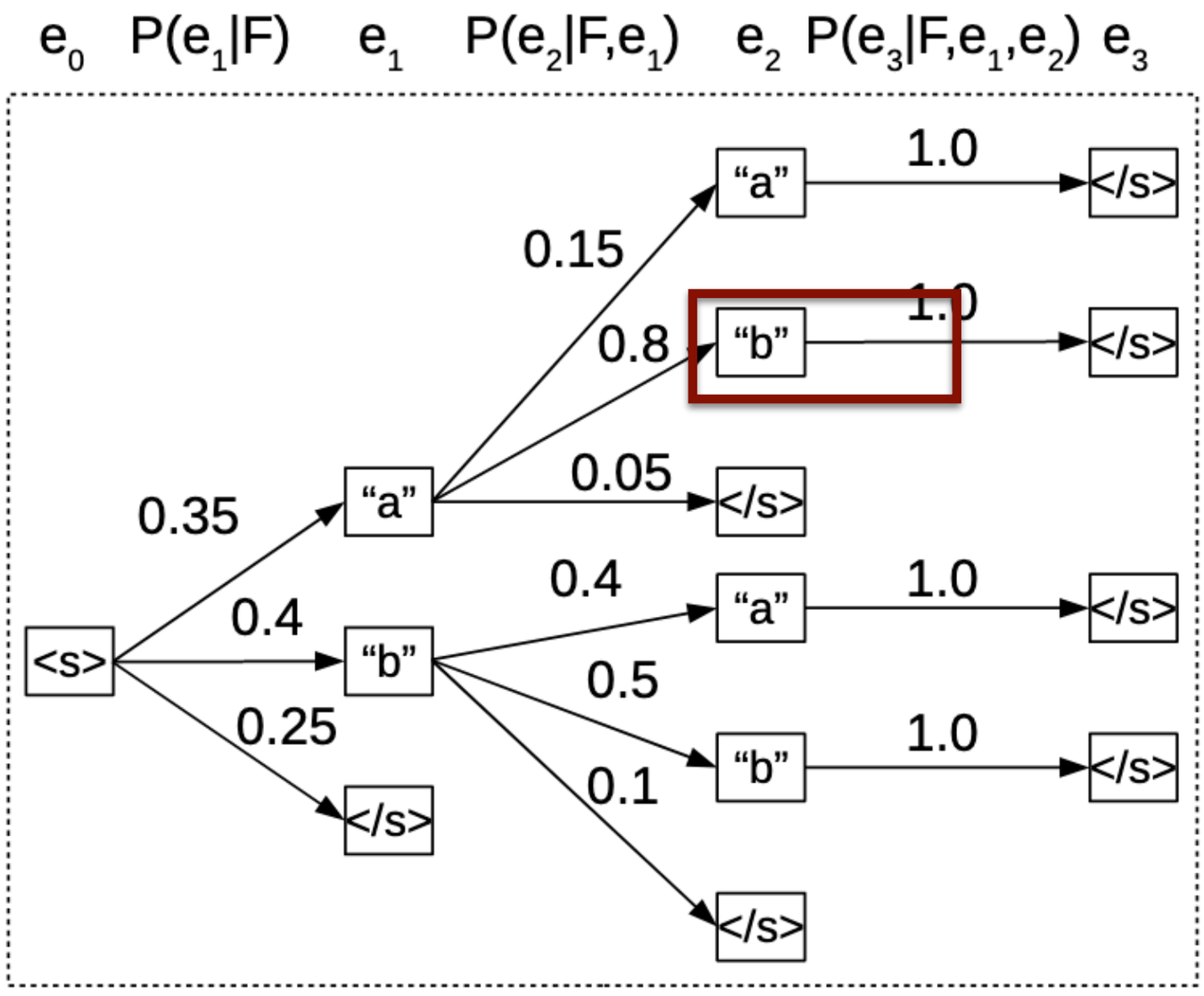
Decoding strategies: greedy decoding



The output at time step t is $\arg\max_{1 \le i \le |V|} p_{t,i}^{(e)}$

(still until we sample </s>)

Decoding strategies: greedy decoding

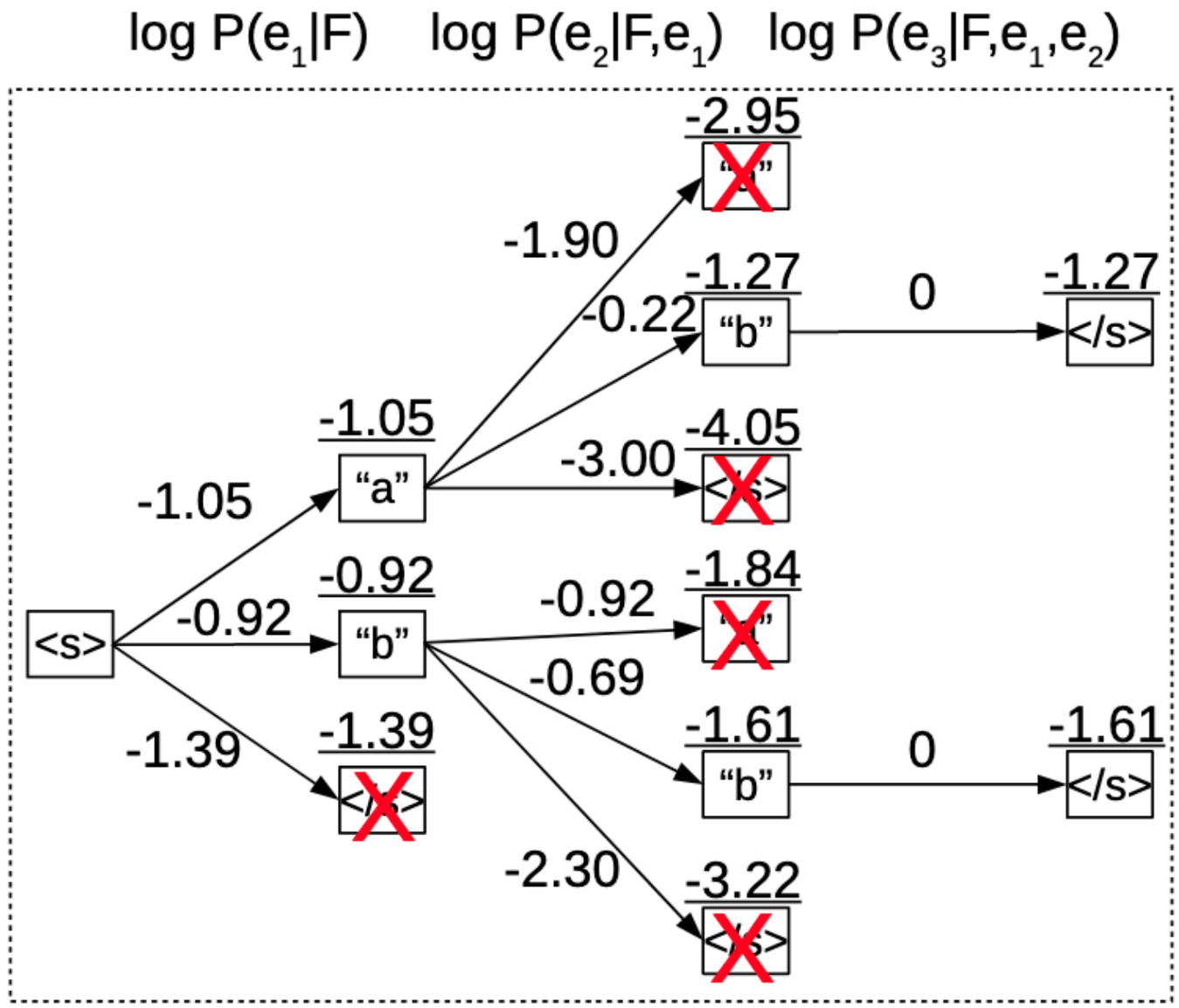


$0.35 \times 0.8 = 0.28$

$0.4 \times 0.5 = 0.2$

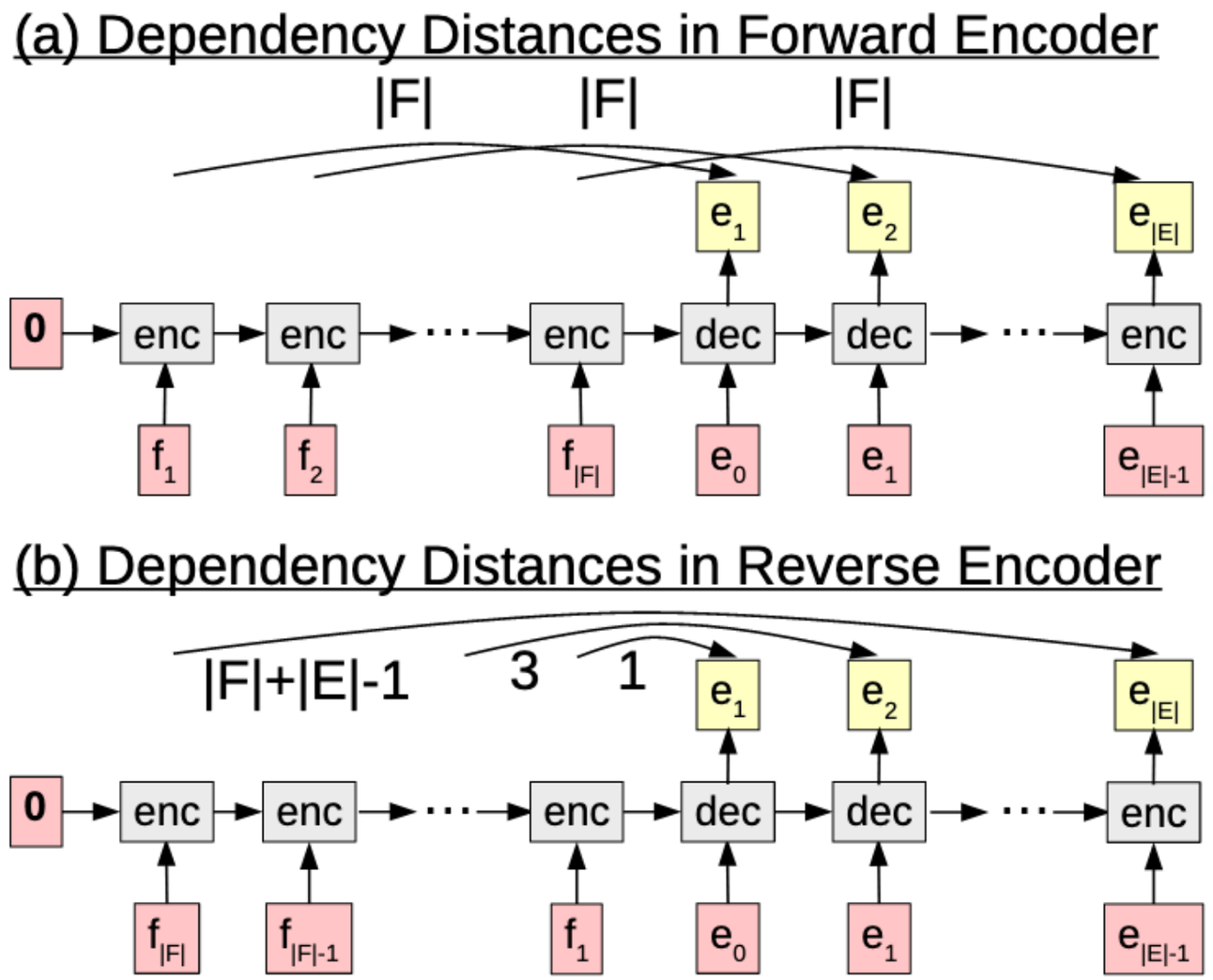
(Neubig, 2017)

Decoding strategies: beam search



(Neubig, 2017)

Reversing the encoder



(Sutskever et al. 2014)

Evaluating machine translation

这个 机场 的 安全 工作 由 以色列 方面 负责 .

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.
Israel is responsible for the airport's security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport's security is the responsibility of the Israeli security officials.

(Koehn, 2010)

Evaluating machine translation

- Exact match with reference is not realistic
- Human annotations: gold standard
- Automatic metrics are much cheaper, but may be distorted in various ways

(Koehn, 2010)

Human evaluation

Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

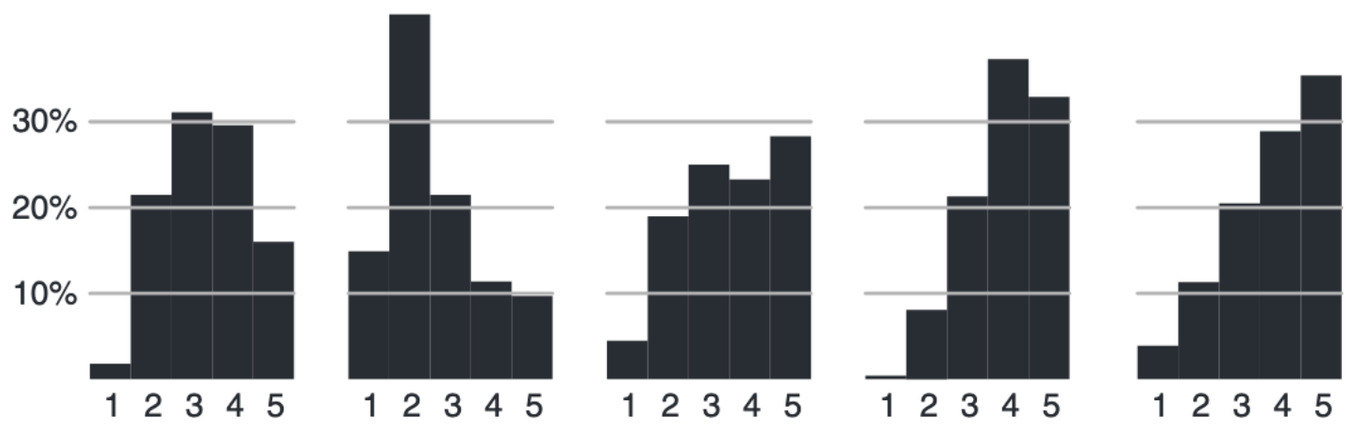
Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English		Annotate
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Challenges with human evaluation

- The meaning of the output of the system may only make sense if you first read the reference, but then you might not be rating adequacy correctly anymore

- Different annotators may use the scale differently

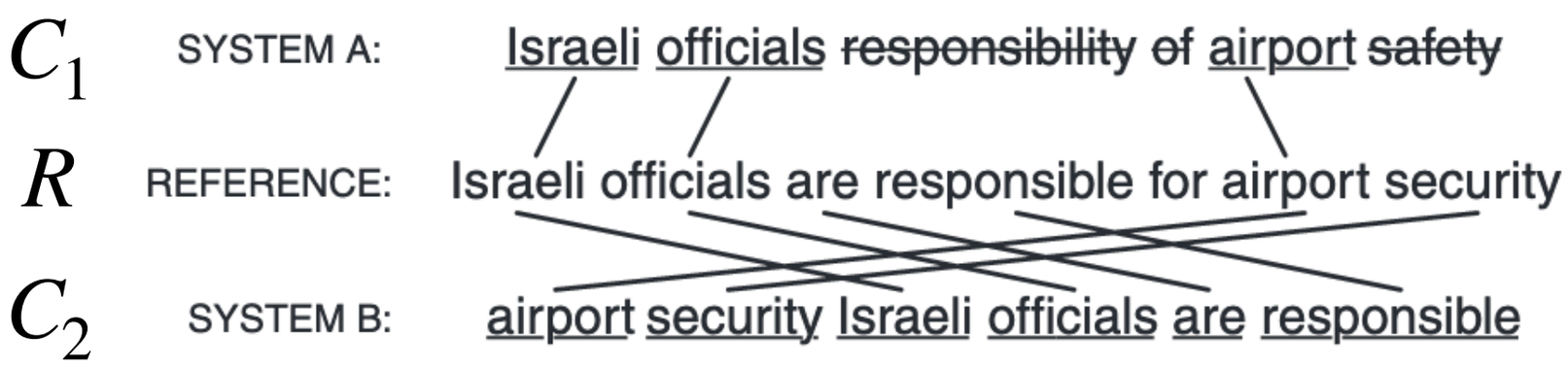
- Ranking two systems (which one is better?) may be easier



Some desiderata from an evaluation metric

- Consistency: inter-annotator agreement
- Cost: expert translators > bilingual > monolingual > fully automatic
- Correctness: correlation with best human judgments

Precision and recall



- Precision: $\frac{|C \cap R|}{|C|}$ Recall: $\frac{|C \cap R|}{|R|}$
- F-score: $\frac{|C \cap R|}{(|C| + |R|)/2}$
- Not sensitive to word order

BLEU

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

System A: 1-gram precision 3/6, 2-gram precision 1/5, 3-gram precision 0/4, 4-gram precision 0/3.
System B: 1-gram precision 6/6, 2-gram precision 4/5, 3-gram precision 2/4, 4-gram precision 1/3.

(Papineau et al. 2002)

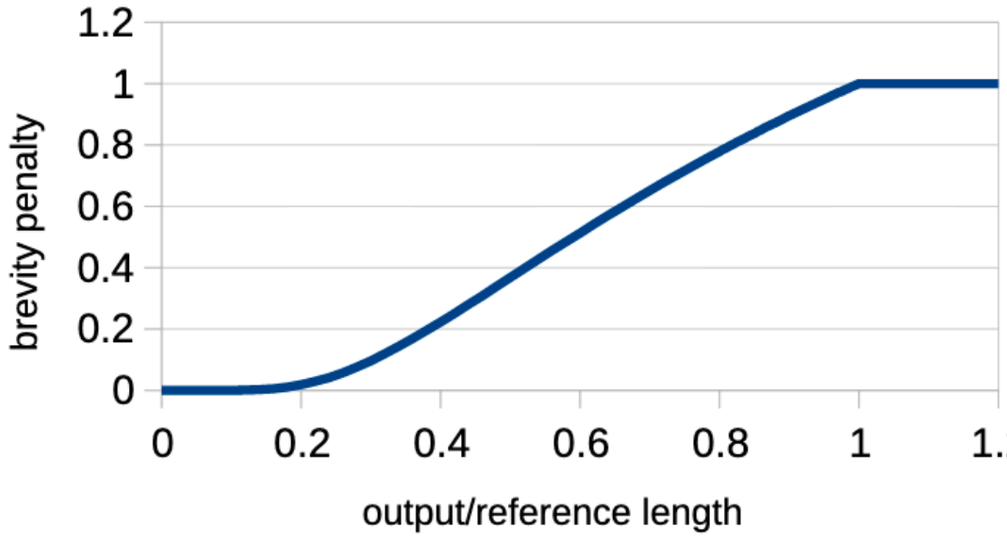
Using multiple references

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

- Having more references may increase the candidate translation's precision (and definitely won't decrease it)
- Recall cannot be used: we can't expect the candidate to contain n-grams from all references

Adding a brevity penalty



$$\text{BP} = \min(1, e^{1-r/c})$$

$$\text{BLUE} = \text{BP}\exp(\frac{1}{N}\sum_{n=1}^{N}\log p_n) \qquad \text{(e.g. } N = 4\text{)}$$

Alternatively, can weight different ns differently

BLEU details

- Scale is not interpretable: 32 on English-Chinese isn't comparable 25 on German-Spanish

- Sensitive to tokenization, number of references… Comparison are only meaningful if the evaluation setup is exactly the same

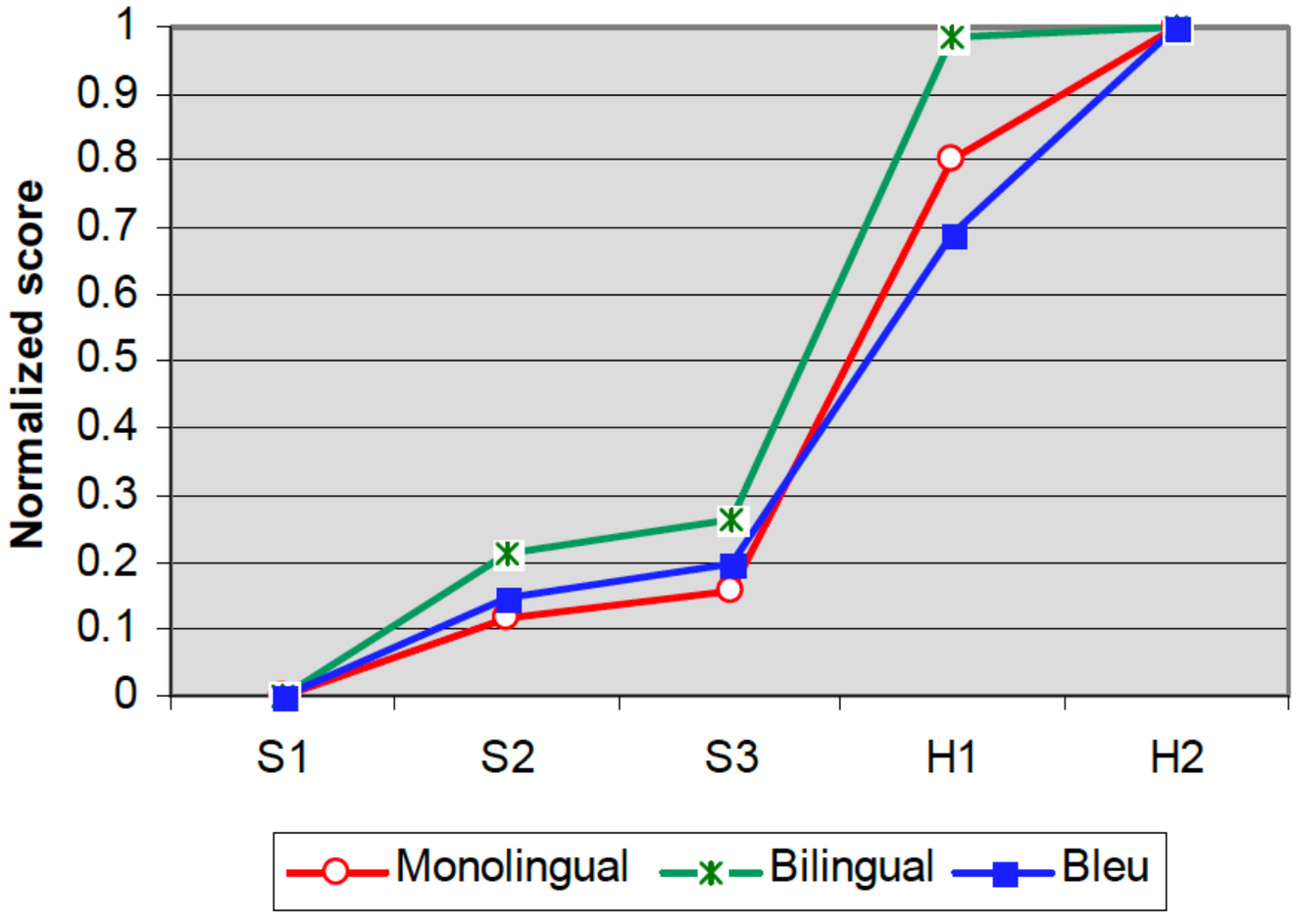
- "Modified" precision: repeated n-grams are counted multiple times, but only up to the number of times they occur in the reference, e.g. here the precision is 2/4:

  Candidate: Officials officials officials officials
  Reference: Israeli officials coordinate with Chinese officials

Pros and cons?

- Adding “not” to a translation radically changes its meaning, but not its BLEU…

- But, at least sometimes correlated with human judgments:



(Papineau et al. 2002)