Neural NLP Tutorial– Part II

1 page recap

LM: Input X → Probability

Classification: Input X → $C=\{c_1, \ldots, c_n\}$

Generation:

Conditioned Language Models

- Not just generate text, generate text according to some specification

Input *X*	Output *Y* (**Text**)	Task
Structured Data	NL Description	NL Generation
English	Japanese	Translation
Document	Short Description	Summarization
Utterance	Response	Response Generation
Image	Text	Image Captioning
Speech	Transcript	Speech Recognition

Calculating the Probability of a Sentence

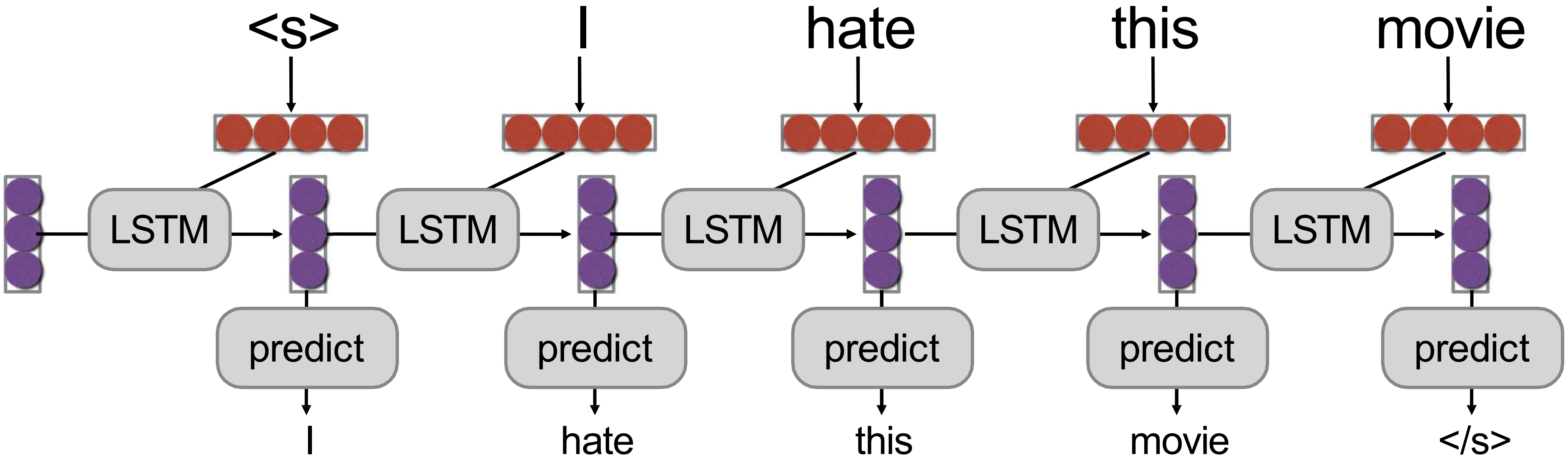
$$P(X) = \prod_{i=1}^{I} P(\underbrace{x_i}_{\text{Next Word}} \mid \underbrace{x_1, \ldots, x_{i-1}}_{\text{Context}})$$

Conditional Language Models

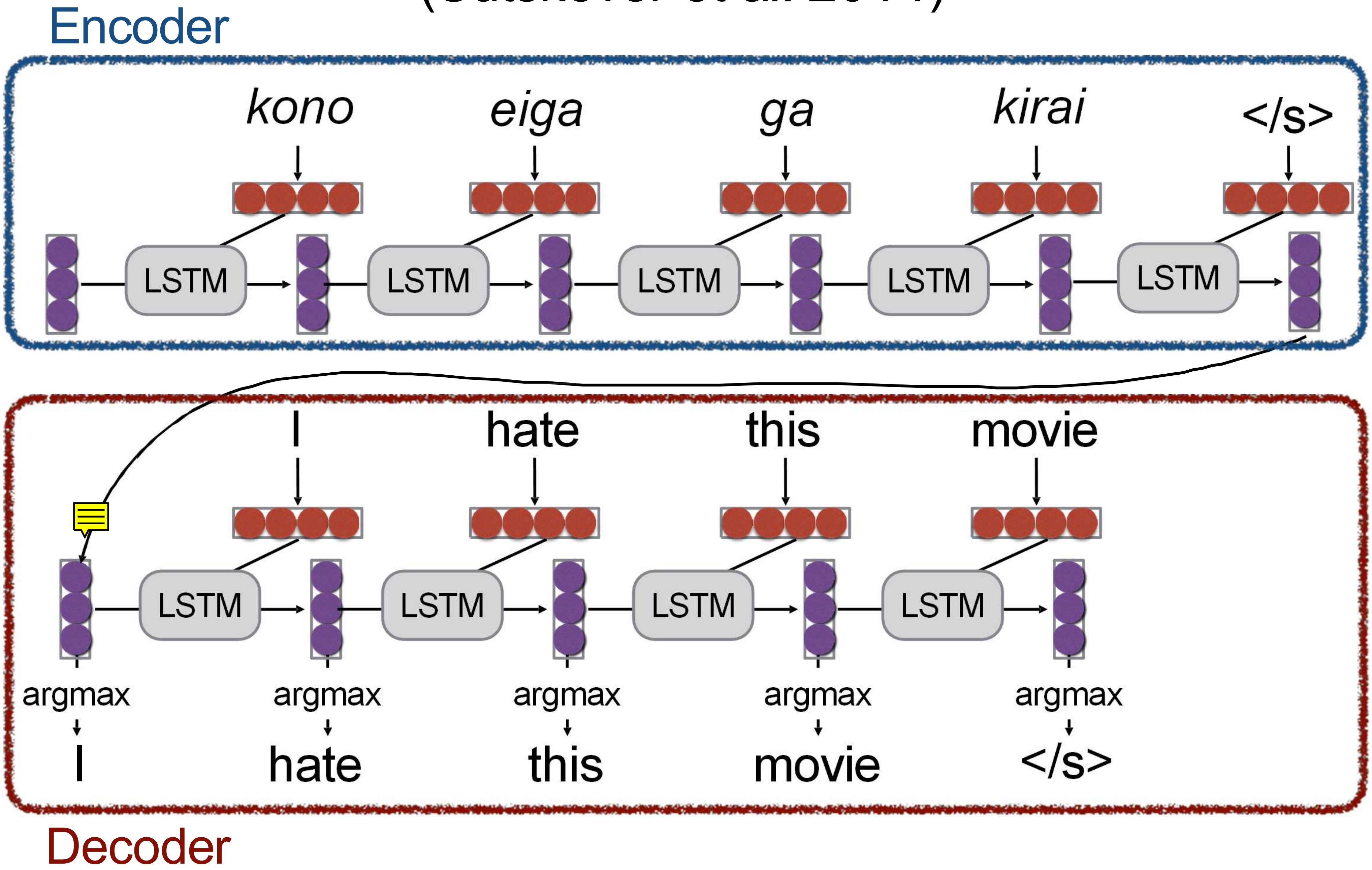
$$P(Y|X) = \prod_{j=1}^{J} P(y_j \mid \underline{X}, y_1, \ldots, y_{j-1})$$

Added Context!

(One Type of) Language Model (Mikolov et al. 2011)



(One Type of) Conditional Language Model
(Sutskever et al. 2014)



The Generation Problem

- We have a model of $P(Y|X)$, how do we use it to generate a sentence?
- Two methods:
  - **Sampling:** Try to generate a *random* sentence according to the probability distribution.
  - **Argmax:** Try to generate the sentence with the *highest* probability.

Ancestral Sampling

- **Randomly generate** words one-by-one.

while y_{j-1} != “</s>”:
  $y_j \sim P(y_j \mid X, y_1, \ldots, y_{j-1})$

- An **exact method** for sampling from P(X), no further work needed.

Argmax Search

- **Greedy search:** One by one, pick the single highest-probability word

```
while y_{j-1} != "</s>":
  y_j = argmax P(y_j | X, y_1, …, y_{j-1})
```

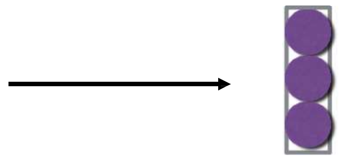
- **Beam search:** keep multiple hypotheses

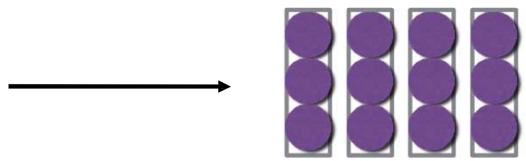
Representing Sentences as Vectors

Problem!

“You can’t cram the meaning of a whole %&!$ing sentence into a single $&!*ing vector!”
— Ray Mooney

- But what if we could use multiple vectors, based on the length of the sentence.

this is an example 

this is an example 

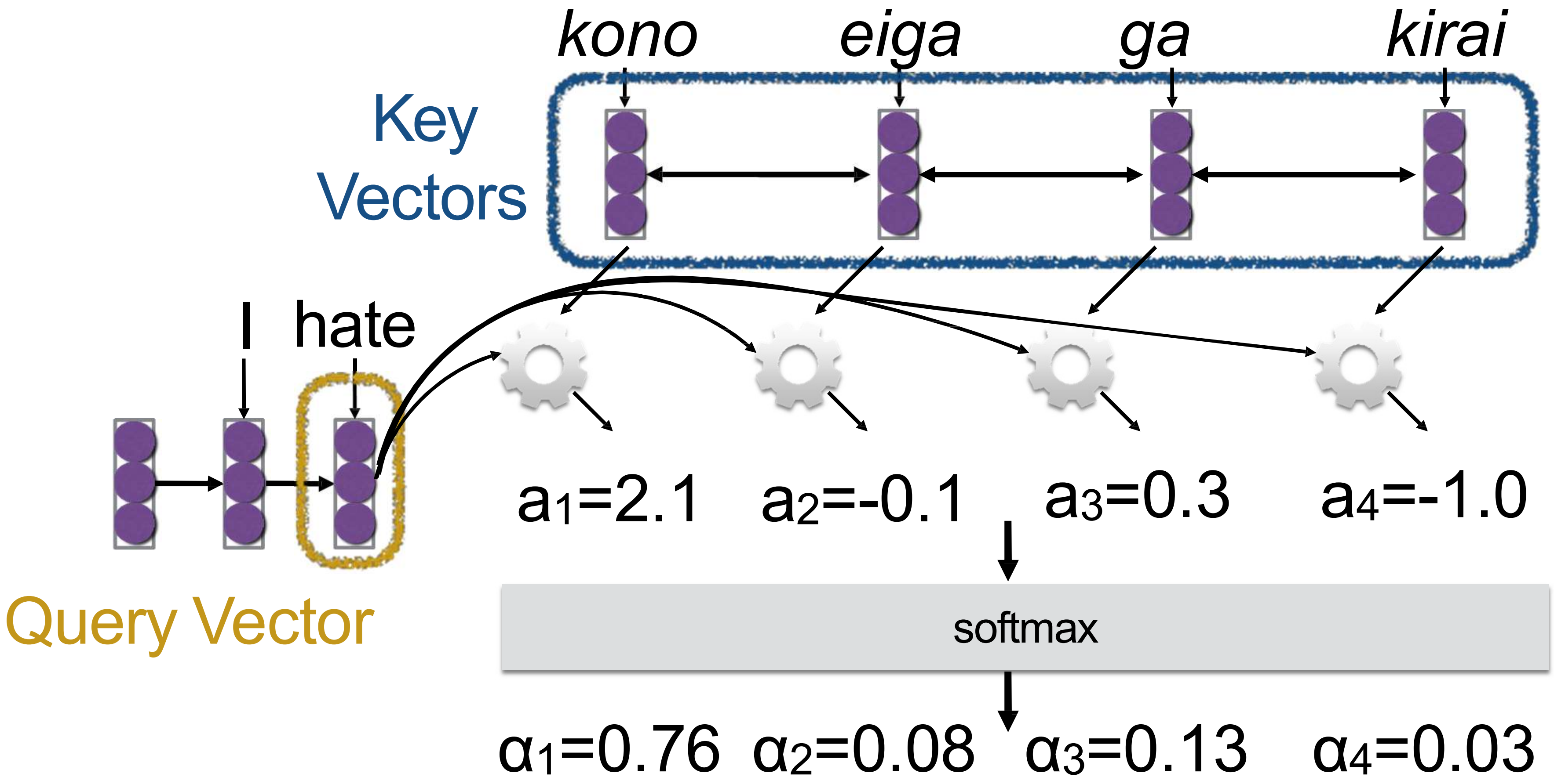
"Attention"!

(Bahdanau et al. 2015)

- Encode each word in the sentence into a vector
- When decoding, perform a linear combination of these vectors, weighted by “attention weights”
- Use this combination in picking the next word

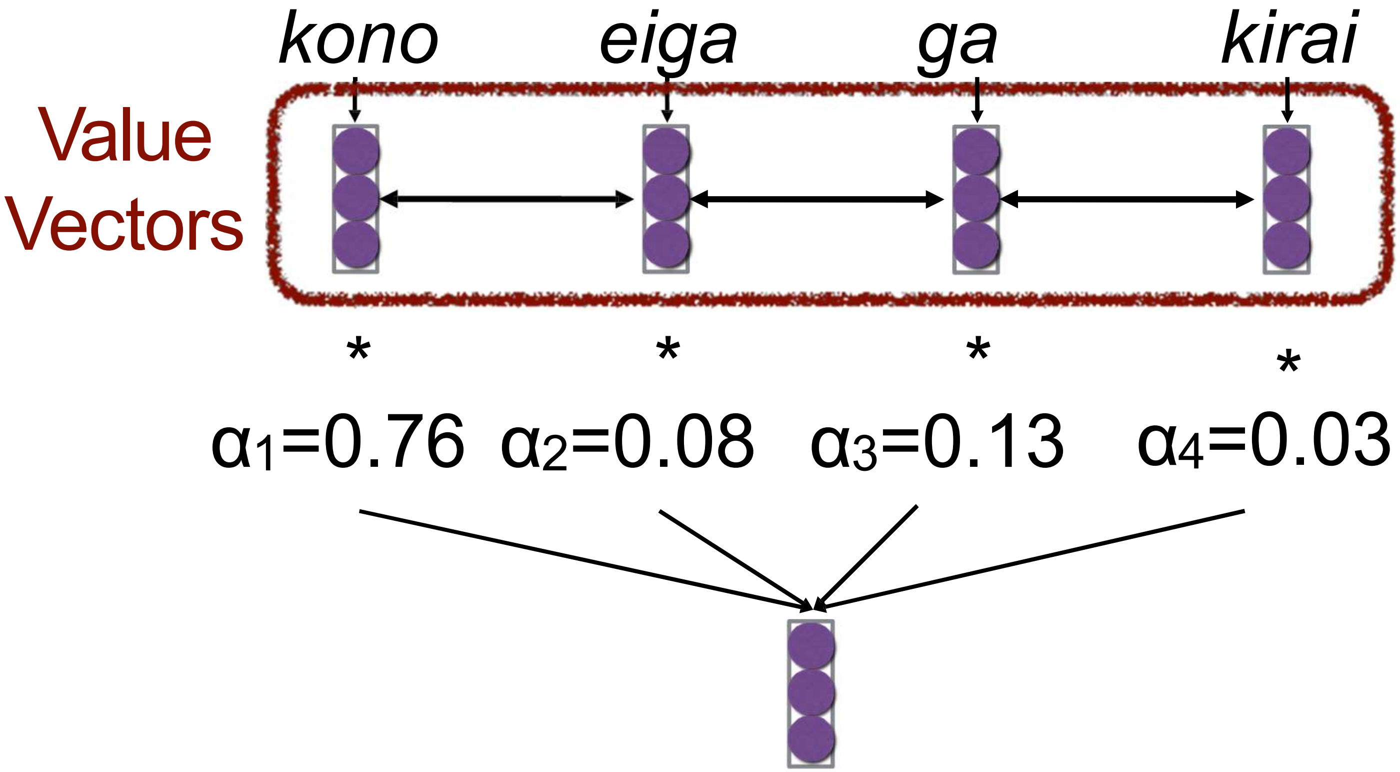
Calculating Attention (1)

- Use "query" vector (decoder state) and "key" vectors (all encoder states)
- For each query-key pair, calculate weight
- Normalize to add to one using softmax



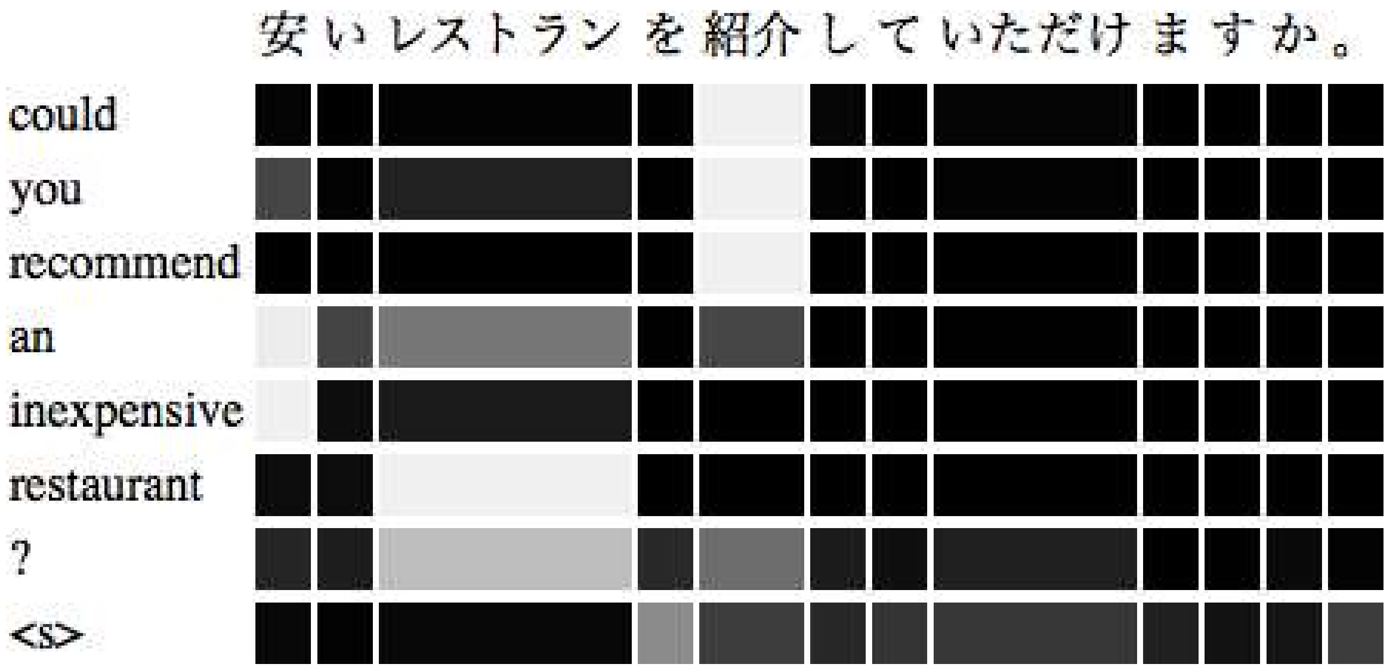
Calculating Attention (2)

- Combine together value vectors (usually encoder s tates, like key vectors) by taking the weighted sum



- Use this in any part of the model you like

Work also as Explanation



Evaluating Generation

How good is a translation?
Problem: no single right answer

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.

Evaluation

- How good is a given machine translation system?
- Many different translations acceptable
- Evaluation metrics
  - Subjective judgments by human evaluators
  - Automatic evaluation metrics
  - Task-based evaluation

Adequacy and Fluency

- Human judgment
  - Given: machine translation output
  - Given: input and/or reference translation
  - Task: assess quality of MT output

- Metrics
  - **Adequacy:** does the output convey the meaning of the input sentence? Is part of the message lost, added, or distorted?
  - **Fluency:** is the output fluent? Involves both grammatical correctness and idiomatic word choices.

Fluency and Adequacy: Scales

Adequacy	
5	all meaning
4	most meaning
3	much meaning
2	little meaning
1	none

Fluency	
5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible

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

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

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

Recall

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

F-measure

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

Precision and Recall of Words

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

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible

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

flaw: no penalty for reordering

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)

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

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

Synonym, order, roughly reflected

BLEU examples

SYSTEM A: Israeli officials responsibility of airport safety

2-GRAM MATCH (Israeli officials) — 1-GRAM MATCH (airport)

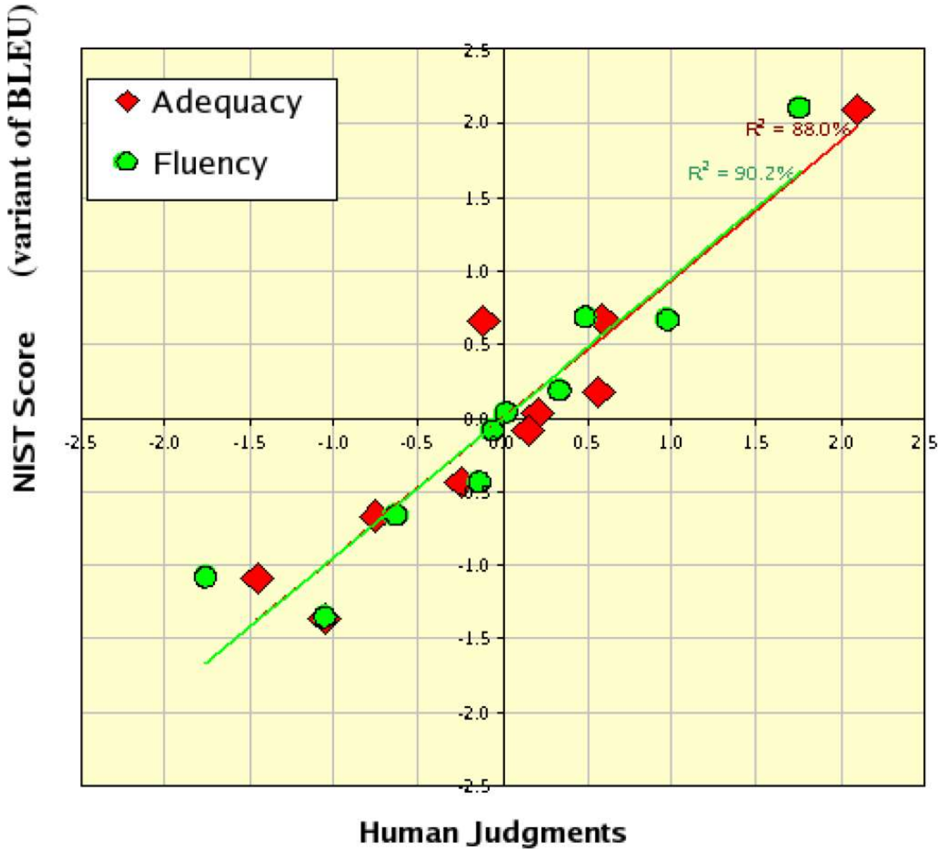
REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible

2-GRAM MATCH (airport security) — 4-GRAM MATCH (Israeli officials are responsible)

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%

automatic metrics such as BLEU correlate with human judgement



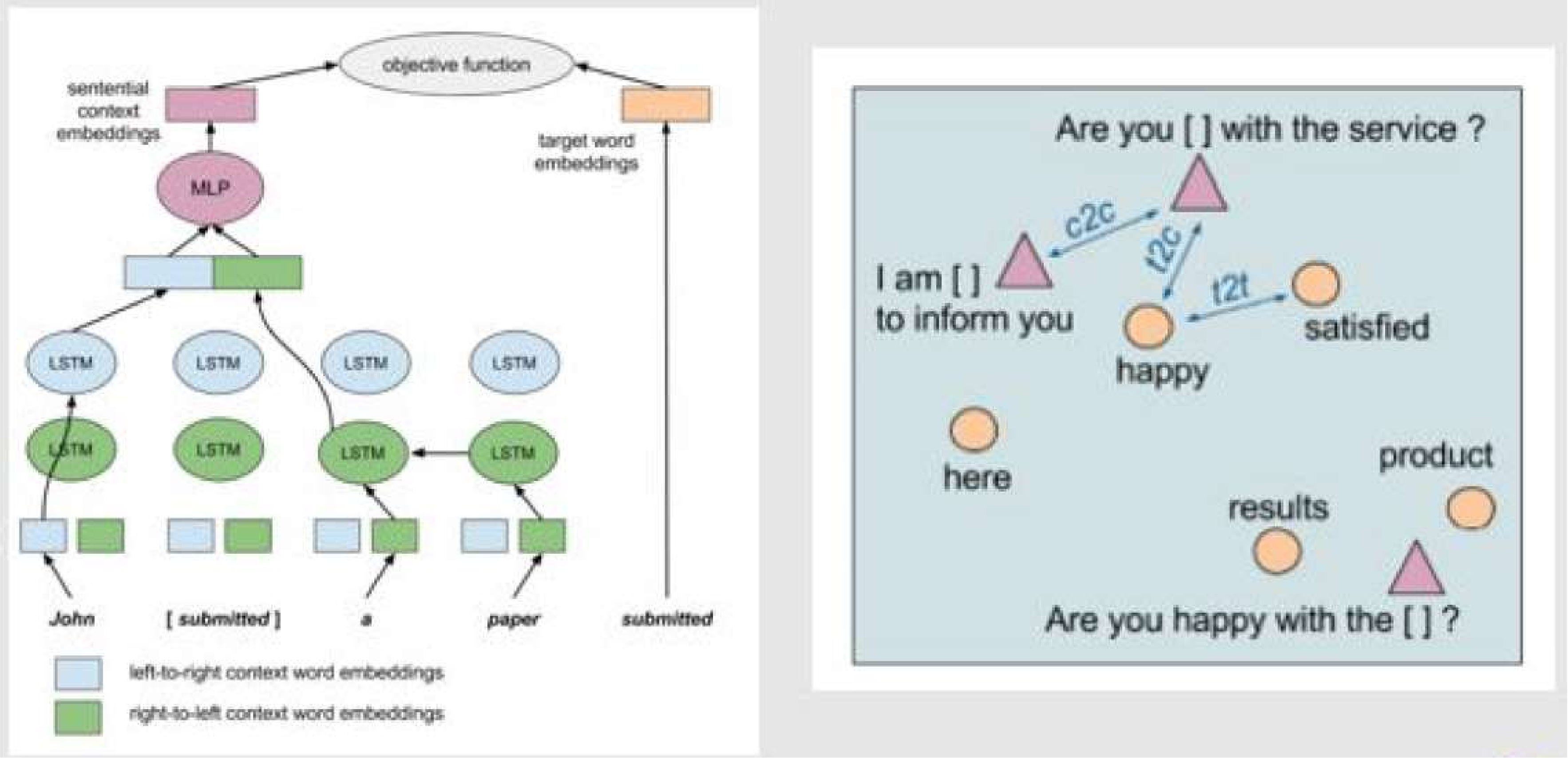
ROUGE - a recall-based counterpart to BLEU

- Idea: what % of the words or n-grams in the **reference** occur in the **generated output**?

- ROUGE and its variants are often used to evaluate *text summarization* systems

Contextual Embedding

Context2Vec

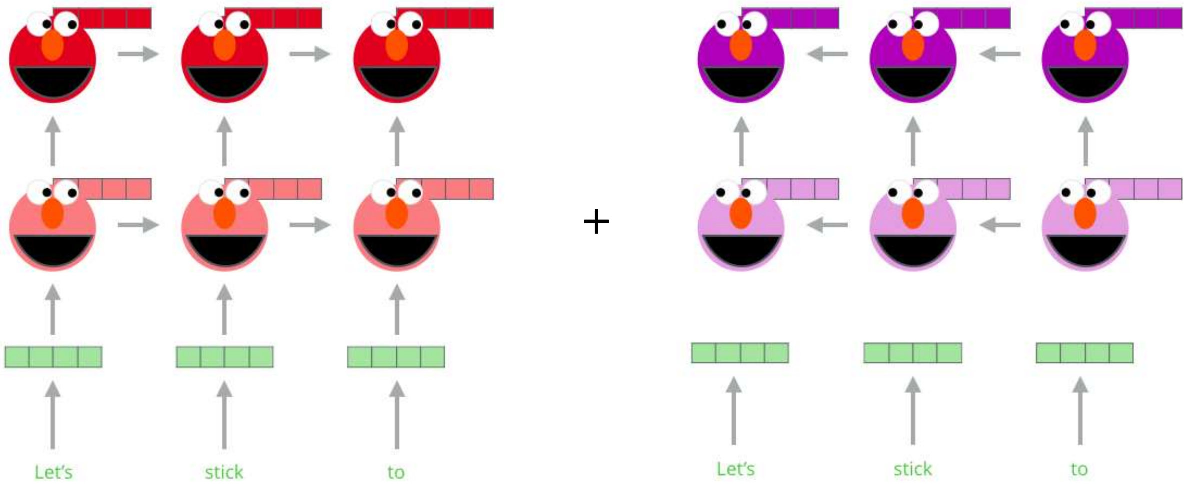


ELMo (Embeddings from Langauge Model): Task-specific Contextual Embedding

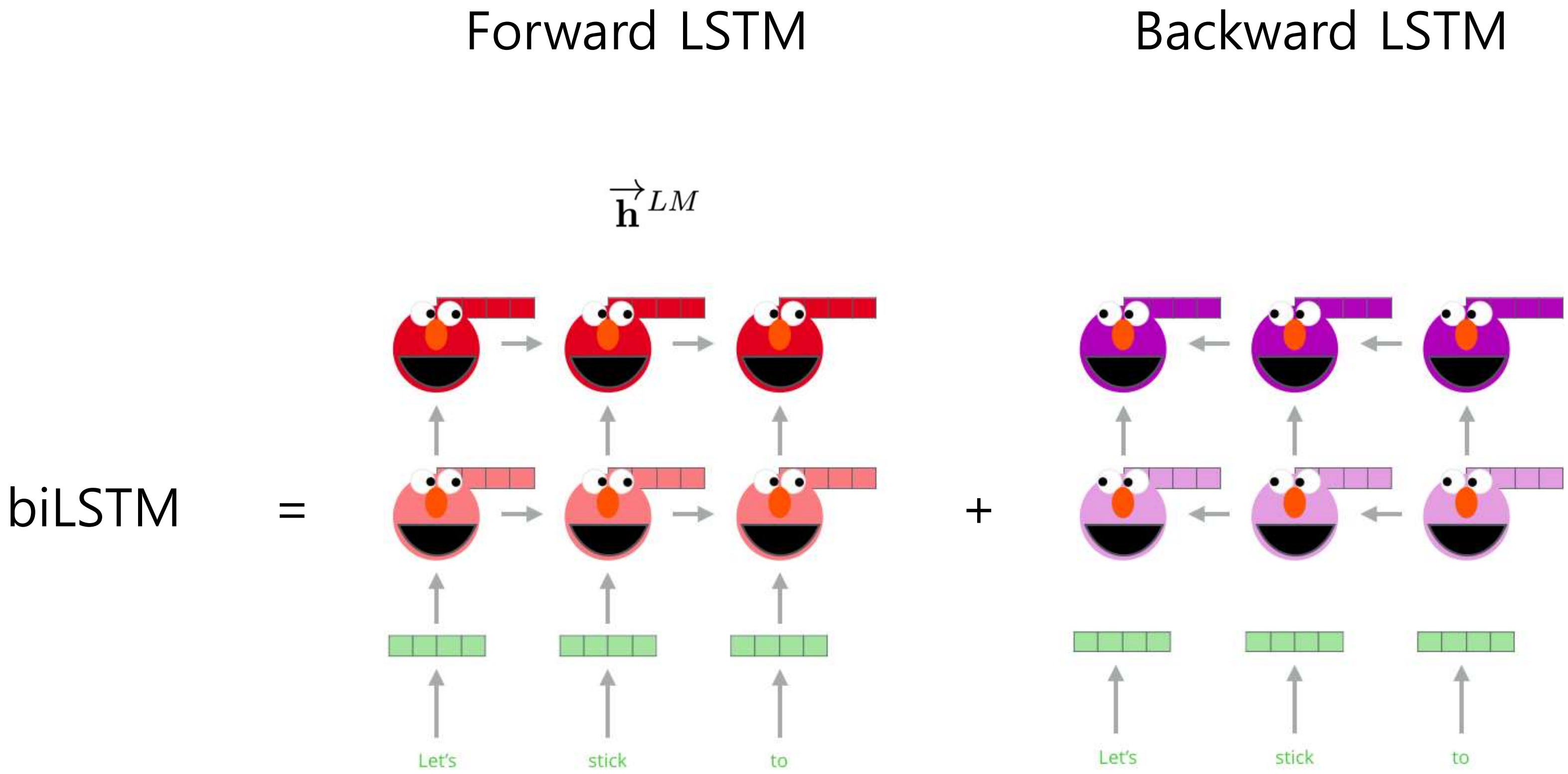
Forward LSTM

Backward LSTM

biLSTM =



+


Forward LSTM
Backward LSTM
$\overrightarrow{\mathbf{h}}^{LM}$
biLSTM
=
+
Let's
stick
to
Let's
stick
to

$$\boxed{\overrightarrow{\mathbf{h}}_{k,j}^{LM}}$$

k : k-th token

j : j-th forward LSTM layer

Let's ------------ 1st token

stick ------------ 2nd token

to ------------ 3rd token

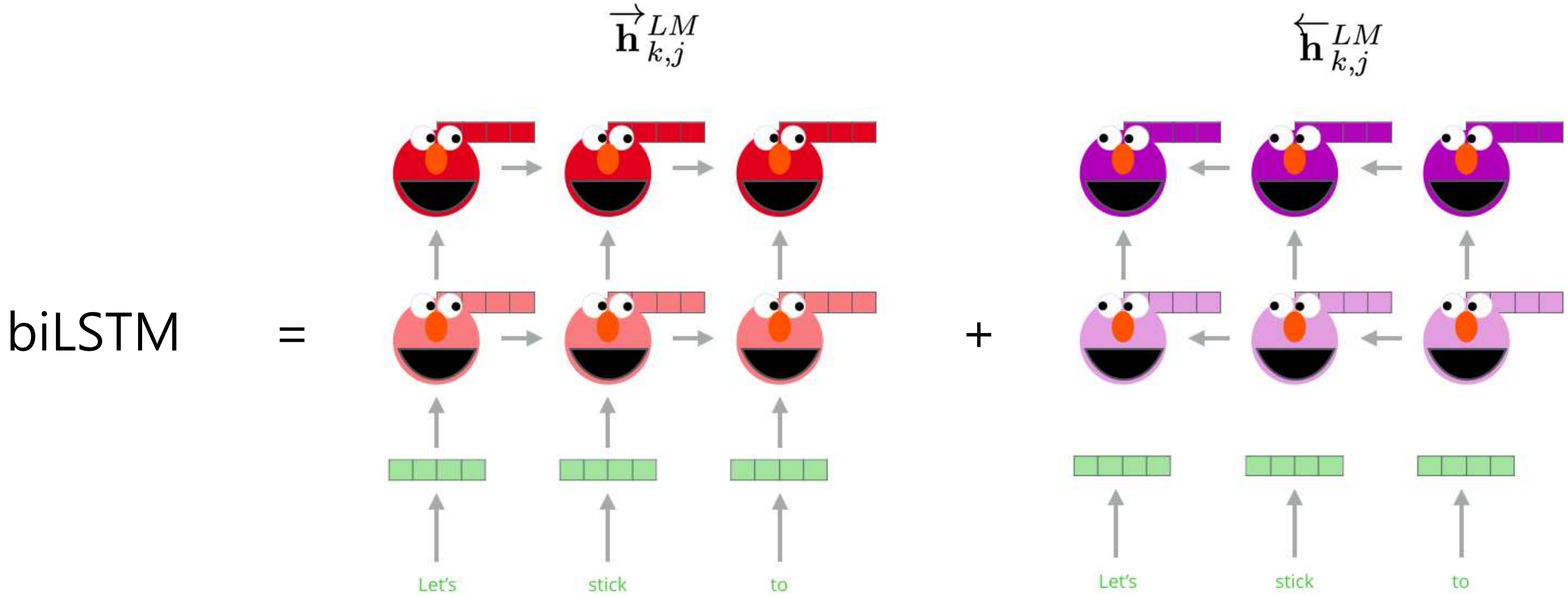
 ------------ 1st LSTM layer

 ------------ 2nd LSTM layer

 ------------ 3rd LSTM layer

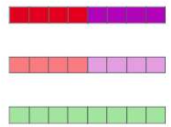
Forward LSTM

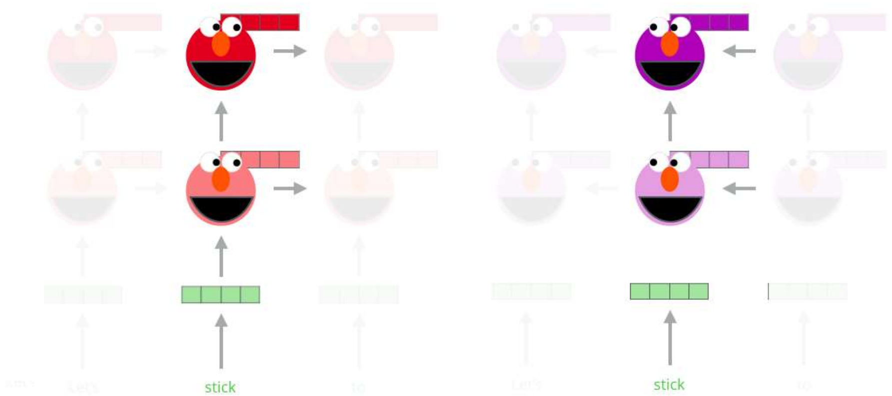
Backward LSTM



$$\mathbf{h}_{k,j}^{LM} = [\overrightarrow{\mathbf{h}}_{k,j}^{LM}; \overleftarrow{\mathbf{h}}_{k,j}^{LM}]$$

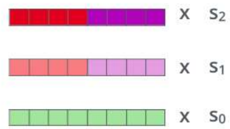
1- Concatenate hidden layers





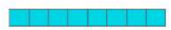
$$s_j^{task}\mathbf{h}_{k,j}^{LM}$$

2- Multiply each vector by a weight based on the task



$$\mathbf{ELMo}_k^{task} = \gamma^{task}\sum_{j=0}^{L} s_j^{task}\mathbf{h}_{k,j}^{LM}$$

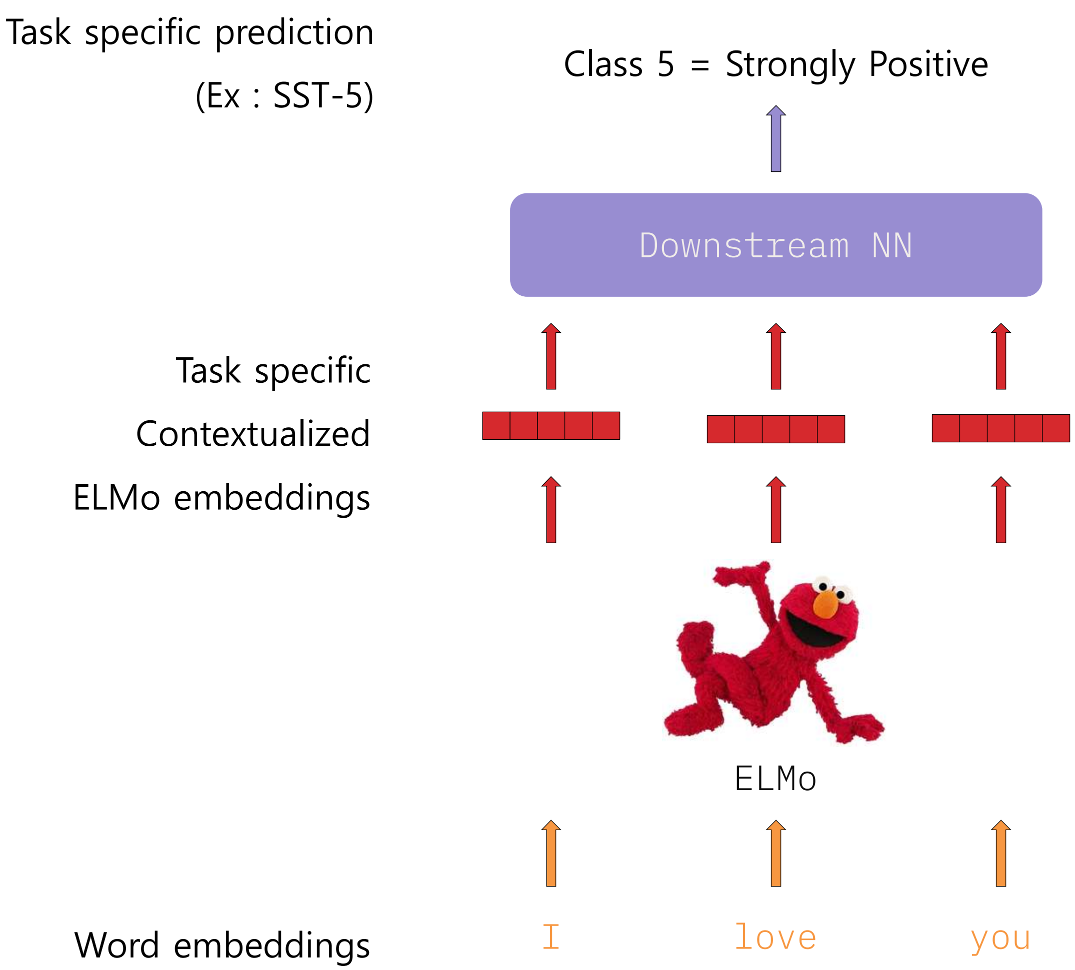
3- Sum the (now weighted) vectors

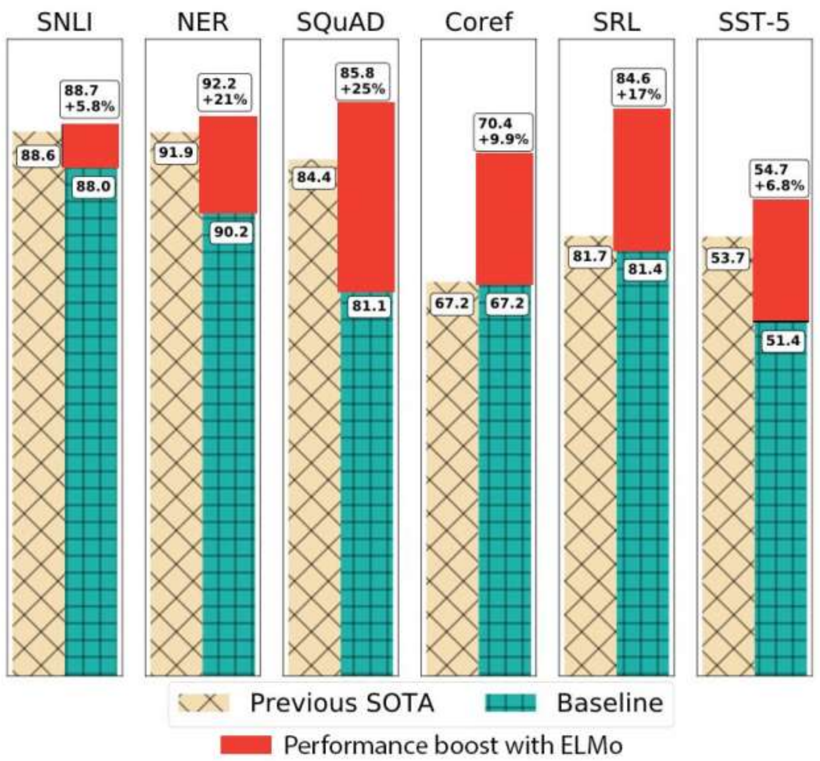


ELMo embedding of "stick"

for this **task** in this **context**

Downstream tasks





TASK	PREVIOUS SOTA		ELMO + BASELINE
SQuAD	Liu et al. (2017)	84.4	85.8
SNLI	Chen et al. (2017)	88.6	88.7 ± 0.17
SRL	He et al. (2017)	81.7	84.6
Coref	Lee et al. (2017)	67.2	70.4
NER	Peters et al. (2017)	91.93 ± 0.19	92.22 ± 0.10
SST-5	McCann et al. (2017)	53.7	54.7 ± 0.5

Masked LM (MLM)

- Input: the man [MASK1] to [MASK2] store
- Label: [MASK1] = went; [MASK2] = store

Next Sentence Prediction (NSP)

- Input: the man went to the store [SEP] he bought a gallon of milk
- Label: IsNext

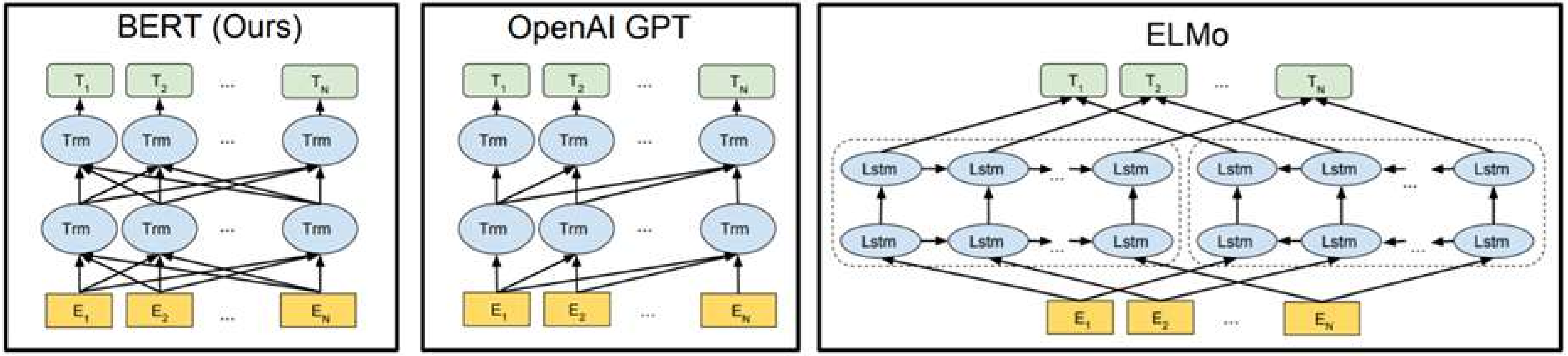


Figure 1: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks. Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

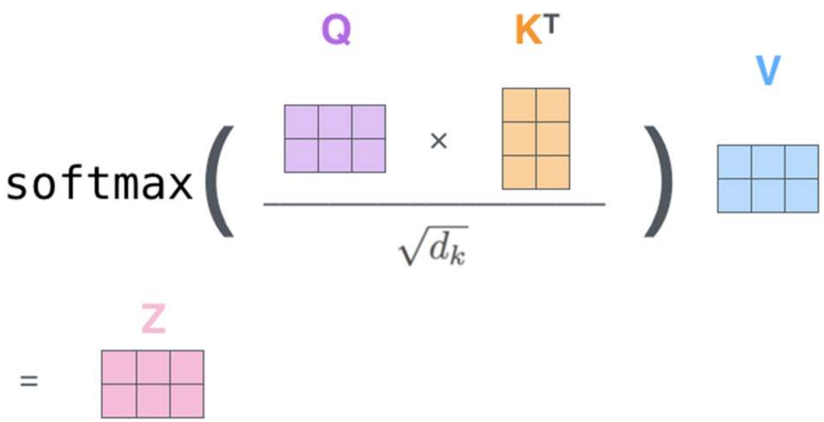
QKV

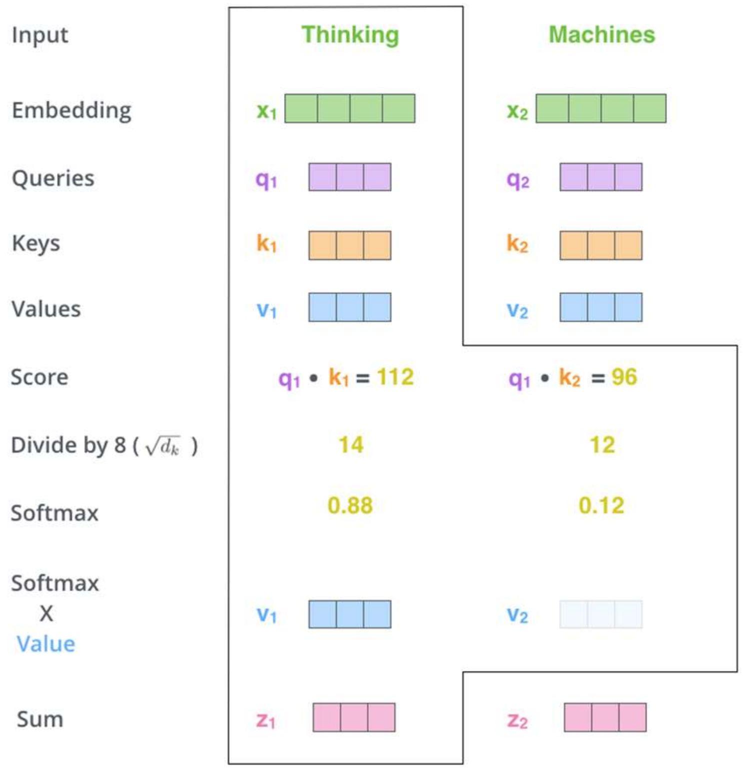
Q=SWH

K	V
SWH	100

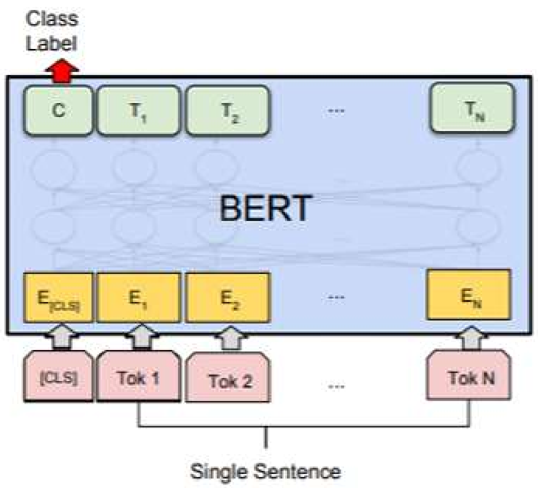
DB: If Q=K then output V

Transformer: output sim(Q,K)*V (0<=sim<=1)





Self-attention: Q=V



What about other tasks? SQuAD? RACE?
Are there better pretraining tasks than MLM/NSP? RoBERTa? Electra?
Can BERT contribute to replace BLEU/ROUGE? BertScore? BartScore?