Artificial Intelligence, Machine Translation and Future Linguists Like You

Graham Neubig

@New York Circle of Translators Meeting (9/21/2020)

Carnegie Mellon University
Language Technologies Institute

翻訳とは、言語において、Aという形で表現されているものを、その意味に対応するBという形に置き換える行為をさす。具体的には、自然言語において、起点言語 による文章を、別の目標言語 による文章に変換する行為をさす。

↓ Google Translate

Translation refers to the act of replacing ~~what is expressed in~~ the ~~form of~~ B that corresponds to ~~that~~ meaning. Specifically, in natural language, it refers to the act of converting a sentence in ~~the~~ source language into a sentence in another target language.

a different form

an expression that takes

the same

a

田中花子は友人が少ないが、数少ない友人をとても大事にする。

↓

Hanako Tanaka has few friends, but she cherishes the few friends very much.

佐藤太郎は友人が少ないが、数少ない友人をとても大事にする。

↓

Taro Sato has few friends, but he cherishes a few friends very much.

田中花子は友人が少なく佐藤太郎は友人が多い。だが、田中は数少ない友人をとても大事にする。

↓

Hanako Tanaka has few friends and Taro Sato has many friends. However, Tanaka takes great care of his few friends.

コード
ko-do
(code, cord, chord)

電気のコードが切れた。
The electric cord has broken.

プログラムのコードを書いた。
I wrote the code for the program.

楽譜にコードを書いた。
I wrote the chord on the score.

プログラマーなのでコードを読むのが得意。
As a programmer, I am good at reading code.

ミュージシャンなのでコードを読むのが得意。
As a musician, I am good at reading ~~codes~~.

Javaで、ギターのコードを表示するコードを書いた。
In Java, I wrote a ~~chord~~ that displays the chord of the guitar.

Why Can Machines Translate?

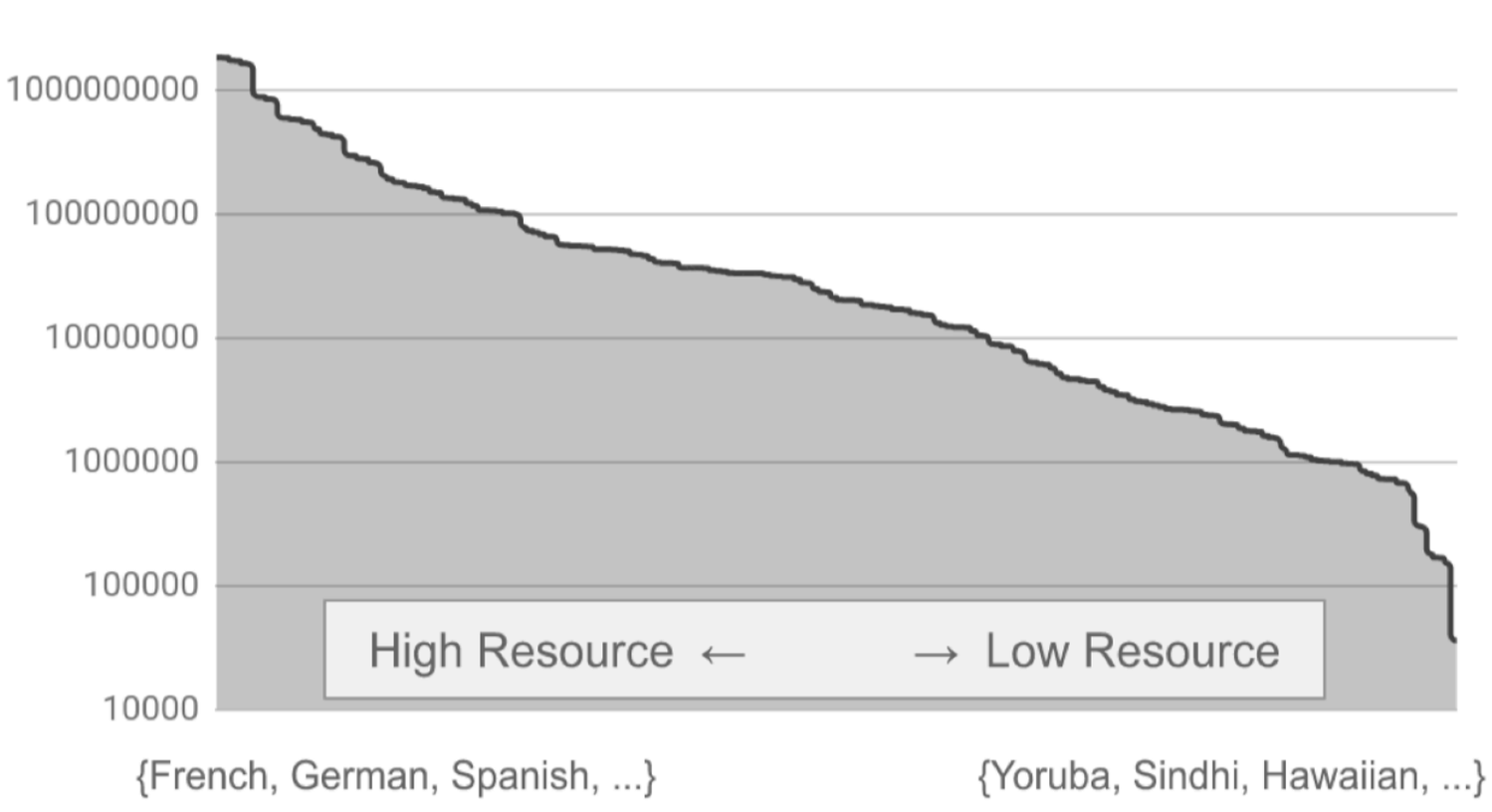
Translation Methods

- **Rule-based:** Do linguistic analysis, syntactic transformation, and word/phrase replacement
  - + Can be complicated/sophisticated in design
  - - Hard to handle every contingency, not enough linguists to build for every language
- **Translation memory:** Look up most similar sentence in database
  - + Easy to understand - look up most recent sentence
  - - Cannot generalize to new sentences
- **Phrase-based translation:** Look up translations for *small chunks* then combine the small chunks together
  - + Can generalize better as long as small chunks are covered
  - - Hard to generate fluent sentences, make mistakes on syntax
- **Neural machine translation:** Feed input into probabilistic model that predicts next word
  - + Better at syntactic correctness, fluency, and leveraging context
  - - Can make big mistakes (dropping/hallucinating content)

data-driven methods

Available Data

- **Parallel corpora:** Translated sentences in source/target language
- Built by translators or hobbyists, for government offices, newspapers, companies, end-users



Google's MT Data [Arivazhagan+ 19]

OPUS Open Source Corpus
http://opus.nlpl.eu/

Parallel Training Data

Source Language
(X)

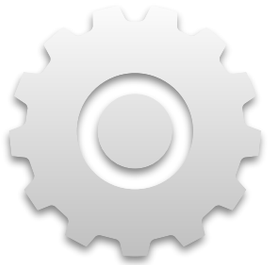
Target Language
(Y)

Training Process

Probabilistic
Translation Model
P(Y|X)

Source Language Input
(X)

Translation Process



Target Language Output
(Y)

Image Credit: pngtree

NMT Basic Idea: Predict Next Word's Probability and Choose Max

X = watashi wa kouen wo shiteimasu

$P(y_1=I|X) = 0.96$ $P(y_1=it|X) = 0.01$ → $y_1=I$

$P(y_1=talk|X) = 0.03$ ...

$P(y_2=am|X,y_1) = 0.90$ $P(y_2=will|X,y_1) = 0.02$ → $y_2=am$

$P(y_2=was|X,y_1) = 0.06$ ...

$P(y_3=giving|X,y_{1,2}) = 0.41$ $P(y_3=presenting|X,y_{1,2}) = 0.15$ → $y_3=giving$

$P(y_3=talking|X,y_{1,2}) = 0.20$ ...

$P(y_4=a|X,y_{1\text{-}3}) = 0.85$ $P(y_4=my|X,y_{1\text{-}3}) = 0.04$ → $y_4=am$

$P(y_4=the|X,y_{1\text{-}3}) = 0.10$ ...

$P(y_5=talk|X,y_{1\text{-}4}) = 0.51$ $P(y_5=presentation|X,y_{1\text{-}4}) = 0.12$ → $y_5=talk$

$P(y_5=lecture|X,y_{1\text{-}4}) = 0.30$ ...

$P(y_6=\text{<done>}|X,y_{1\text{-}5}) = 0.51$ $P(y_6=with|X,y_{1\text{-}5}) = 0.12$ → $y_6=\text{<done>}$

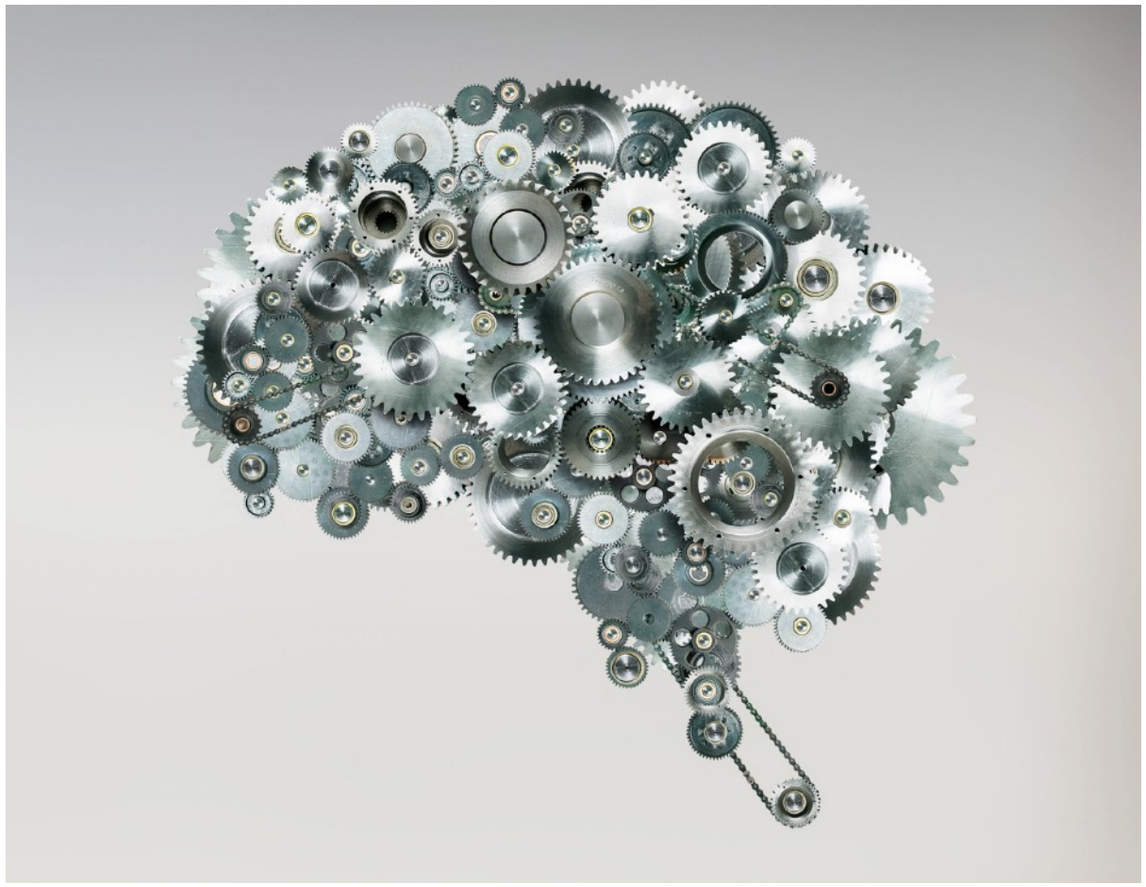
$P(y_6=and|X,y_{1\text{-}5}) = 0.30$ ...

How to Make Difficult Predictions?

Neural Networks

An Aside: What is "Artificial Intelligence?"

From Willyam Bradberry



From "Wired"

An Aside: What is "Artificial Intelligence?"

- **Artificial Intelligence:** Technology that does something that seems intelligent

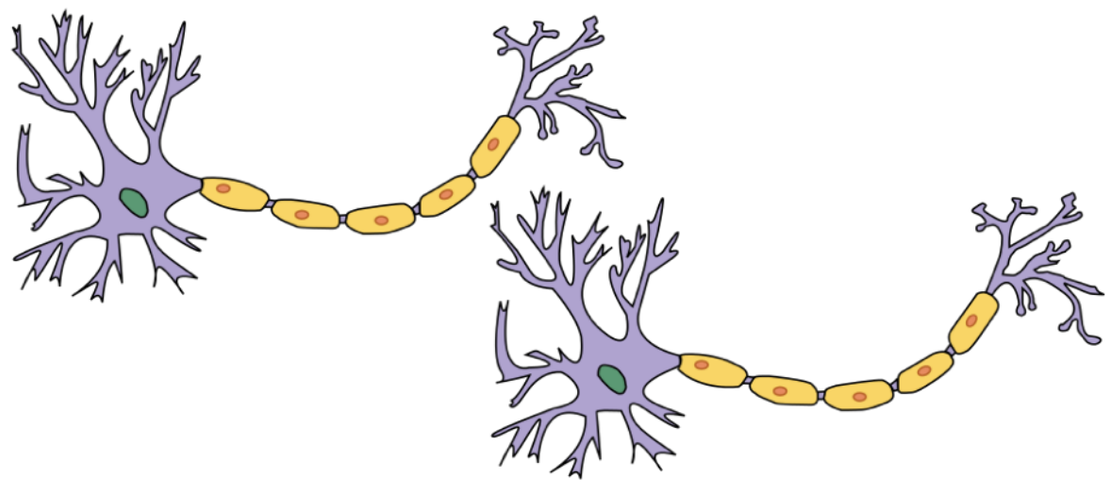
- **Machine Learning:** Technology that learns from data to do something that cannot be done easily otherwise

  - (translation is a good example)

- In it's current actualization: mostly just *computation graphs* with parameters trained on lots of data

“Neural” Networks

Original Motivation: Neurons in the Brain



Current Conception: Computation Graphs

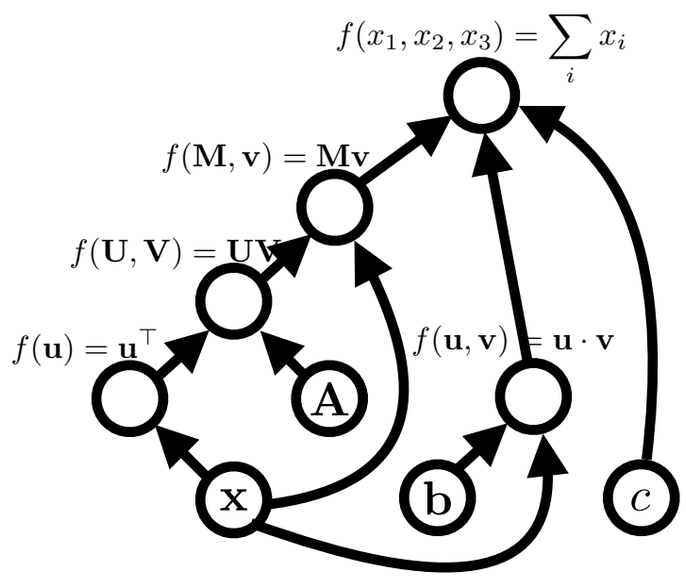


Image credit: Wikipedia

expression:

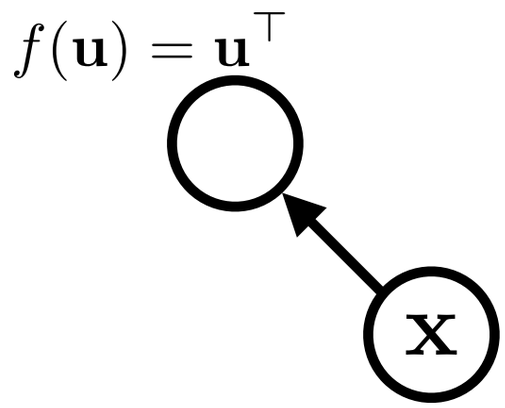
$$\mathbf{x}$$

graph:

expression:

$$\mathbf{x}^\top$$

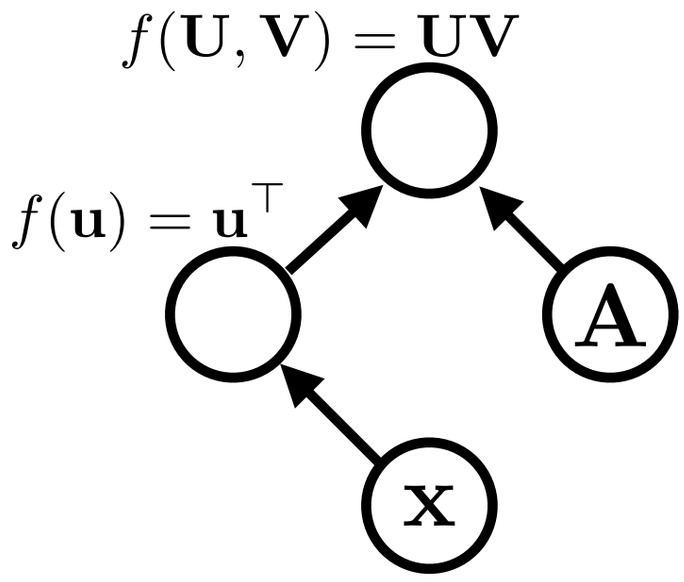
graph:



expression:

$$\mathbf{x}^\top \mathbf{A}$$

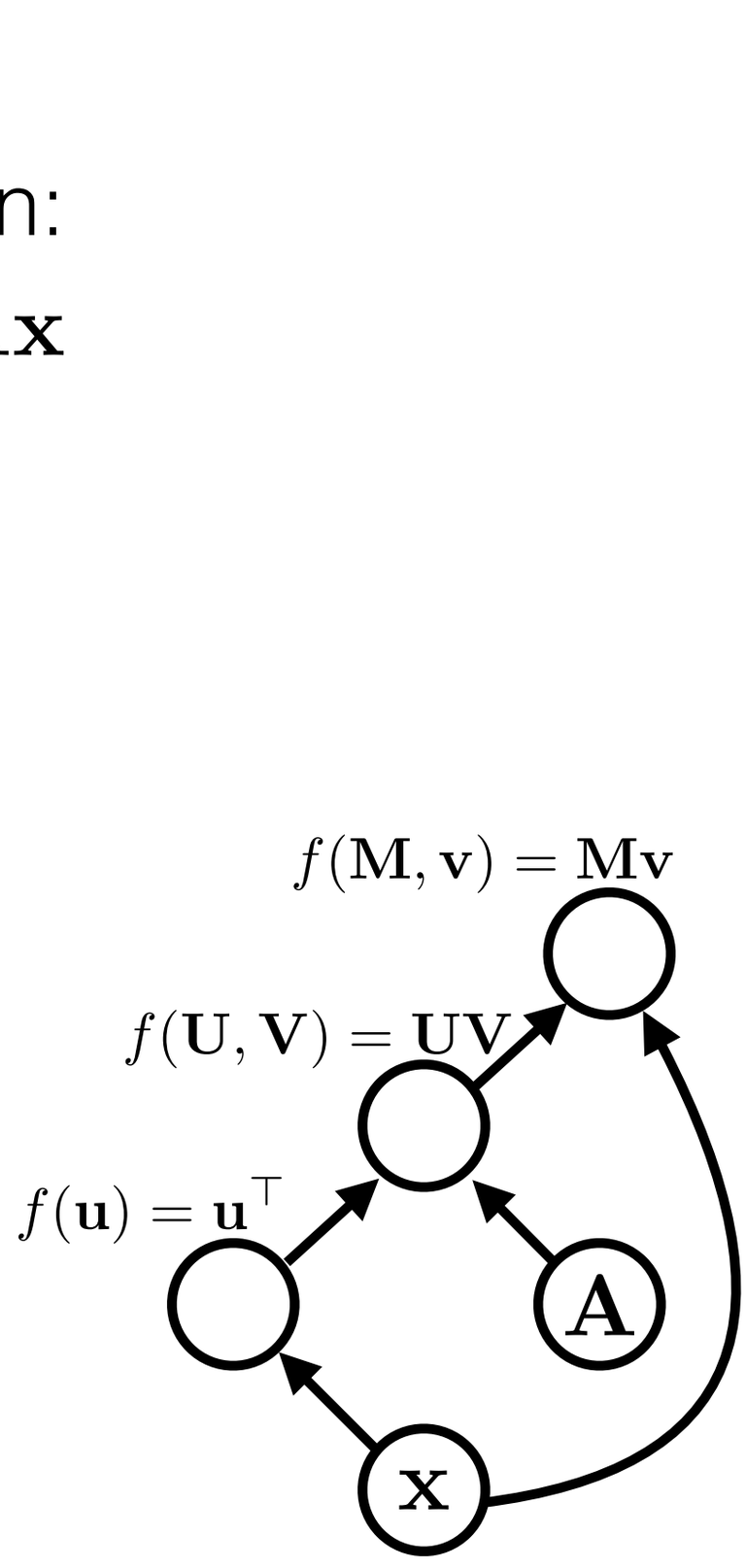
graph:



expression:

$$\mathbf{x}^\top \mathbf{A} \mathbf{x}$$

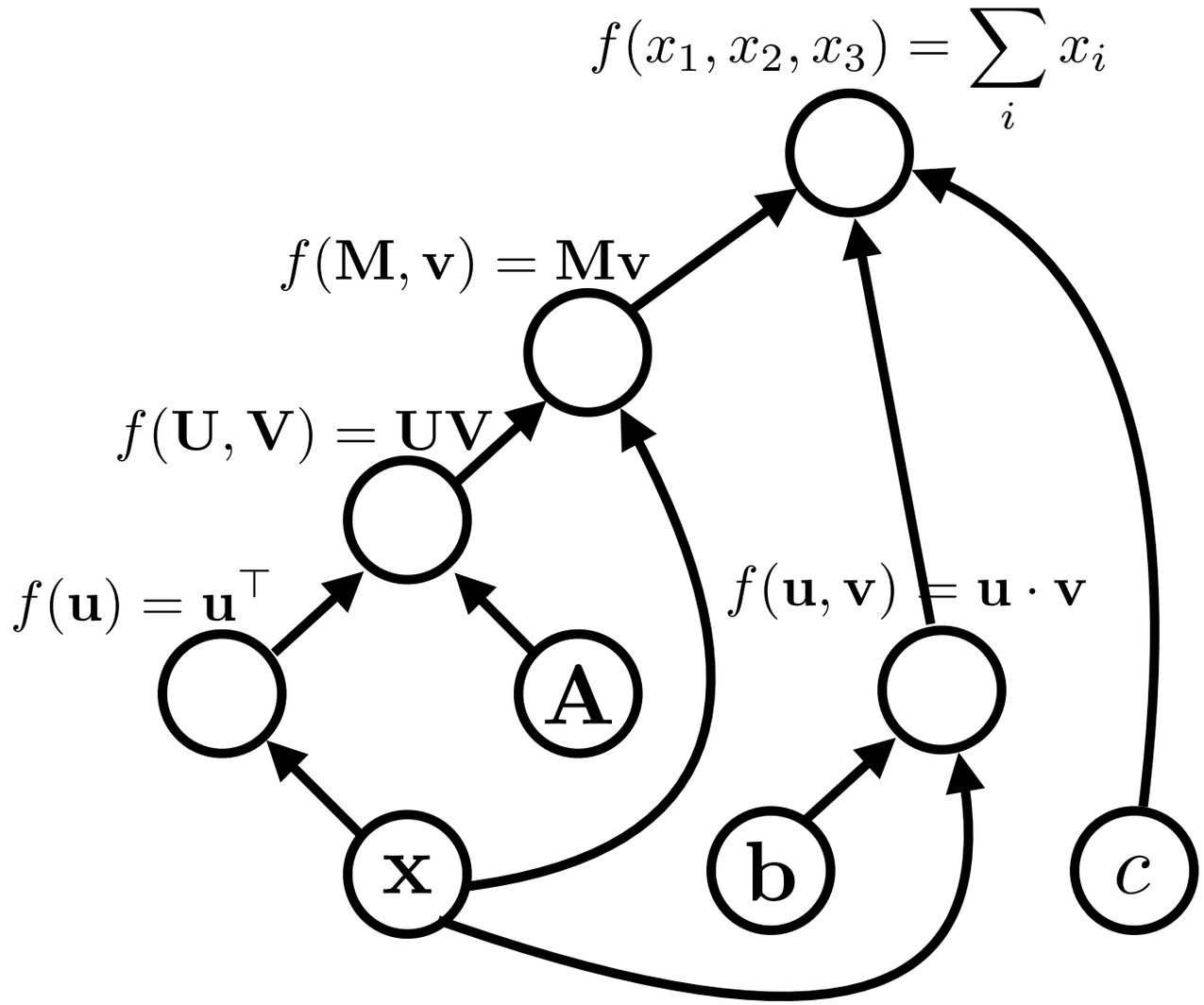
graph:



expression:

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b} \cdot \mathbf{x} + c$$

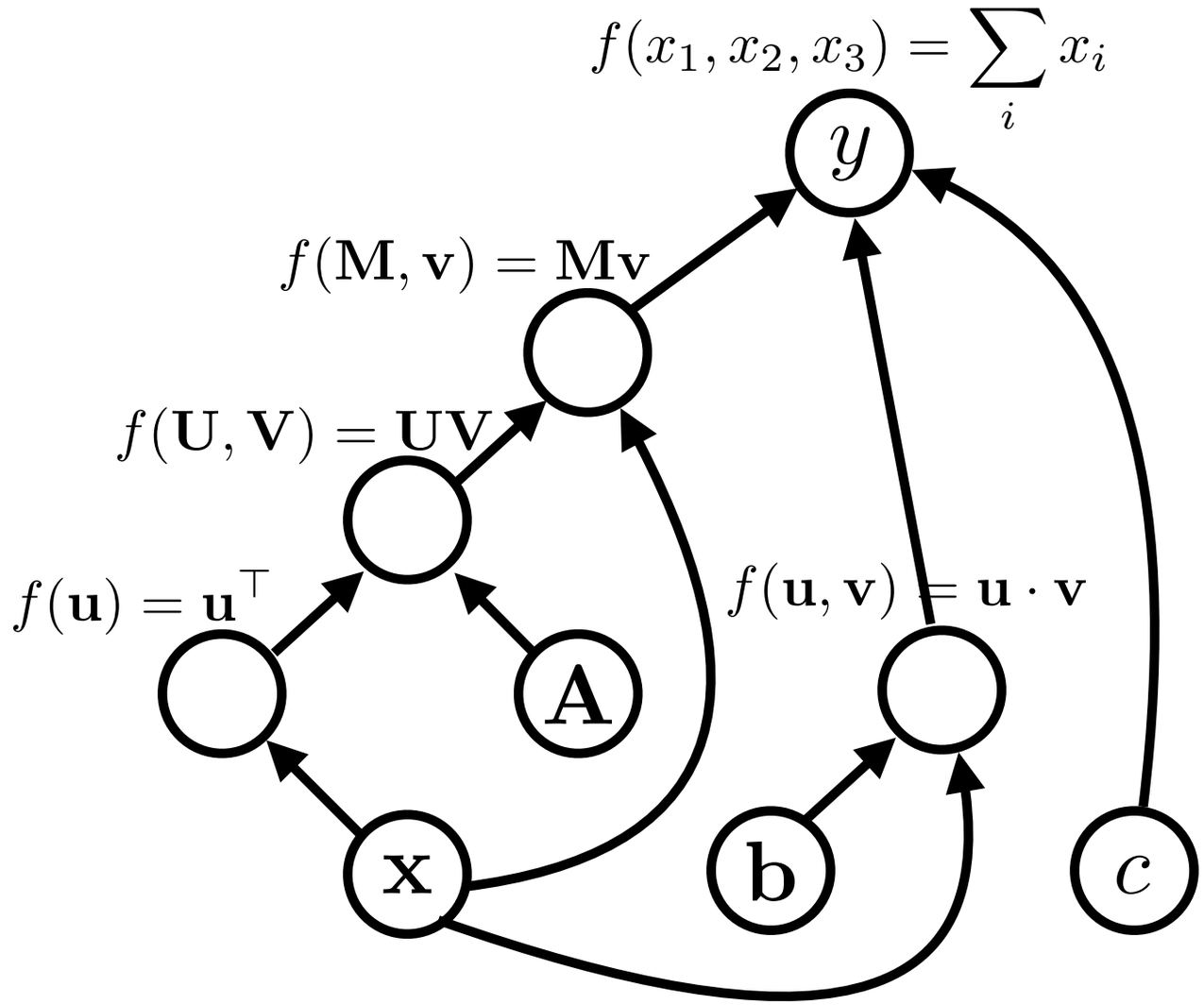
graph:



expression:

$$y = \mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b} \cdot \mathbf{x} + c$$

graph:



"Learning" Neural Networks

- **Build a graph** to calculate something you want to maximize/minimize (e.g. probability of translation given an input)

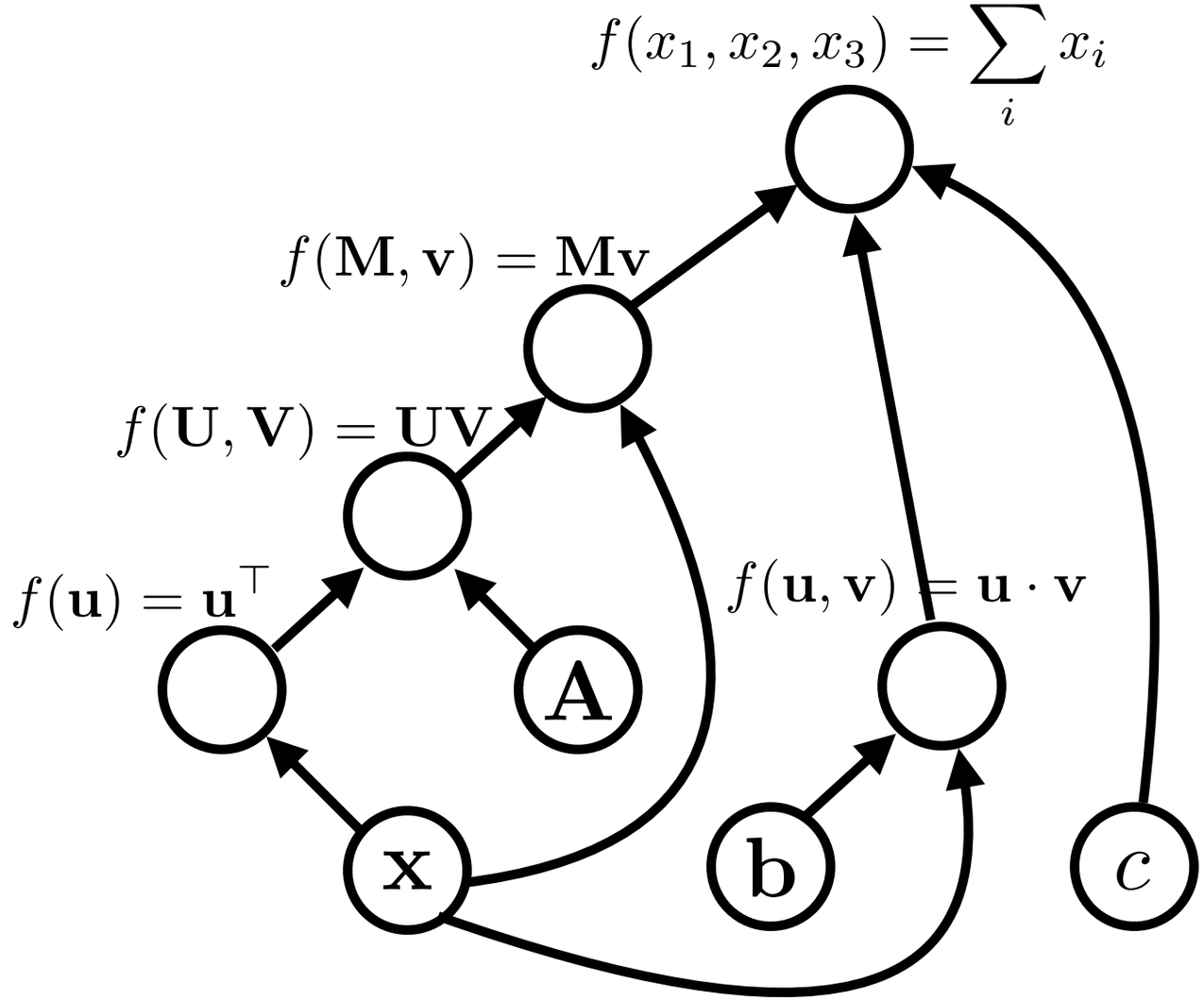
- **Calculate the value itself**

- **Calculate the derivatives** of model parameters given the value (back propagation)

- **Update the model parameters** to increase/decrease the value

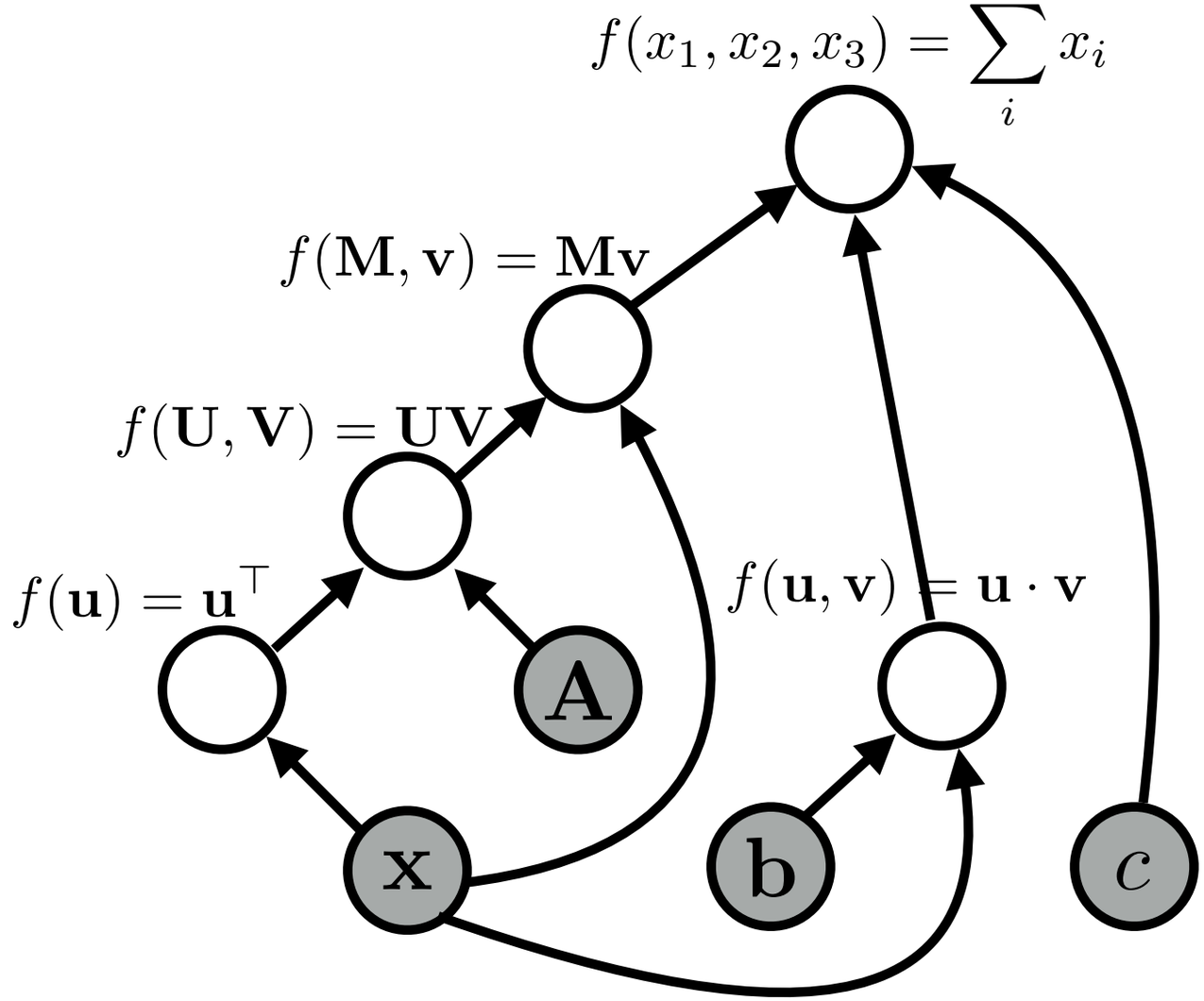
Forward Propagation

graph:



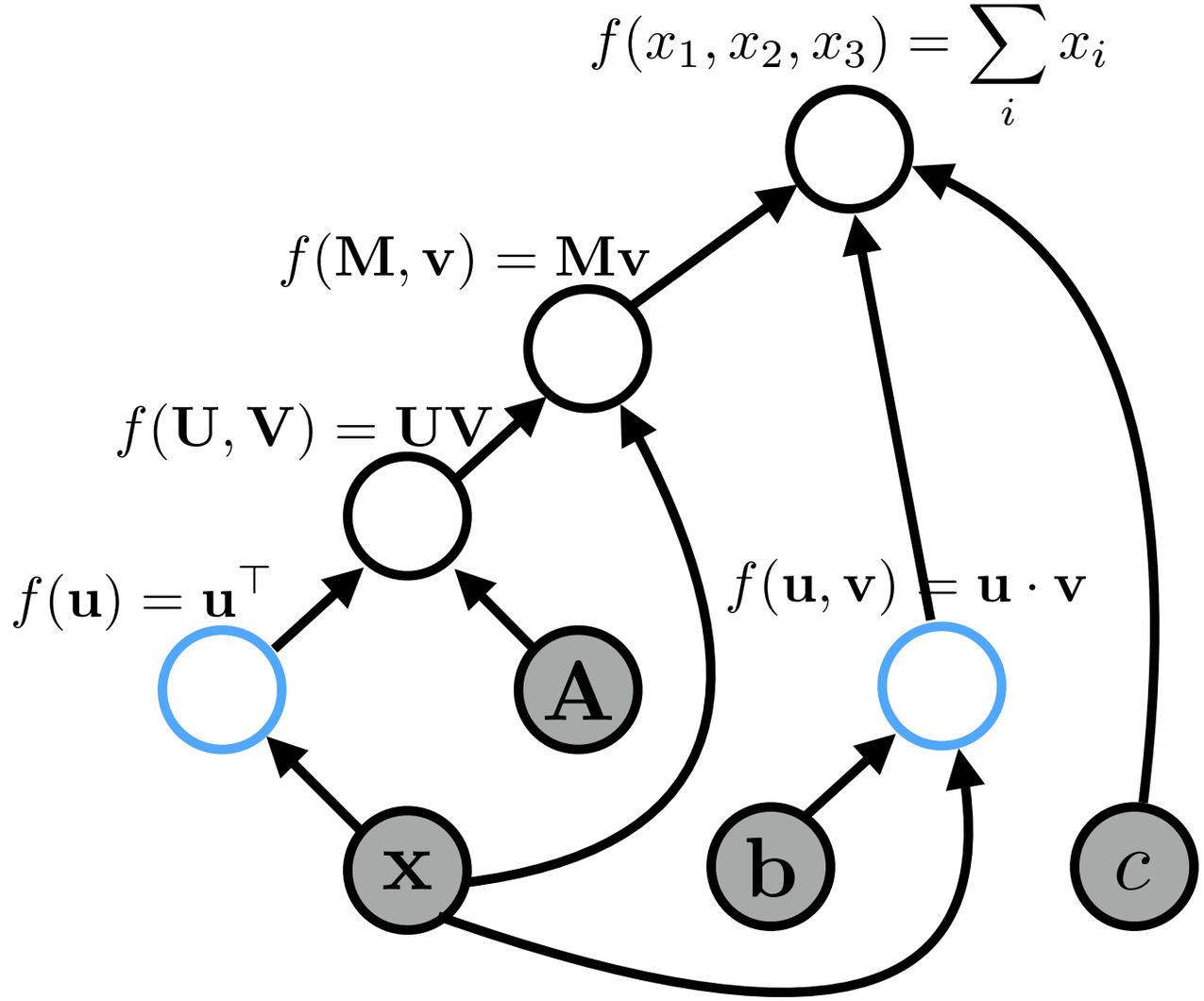
Forward Propagation

graph:



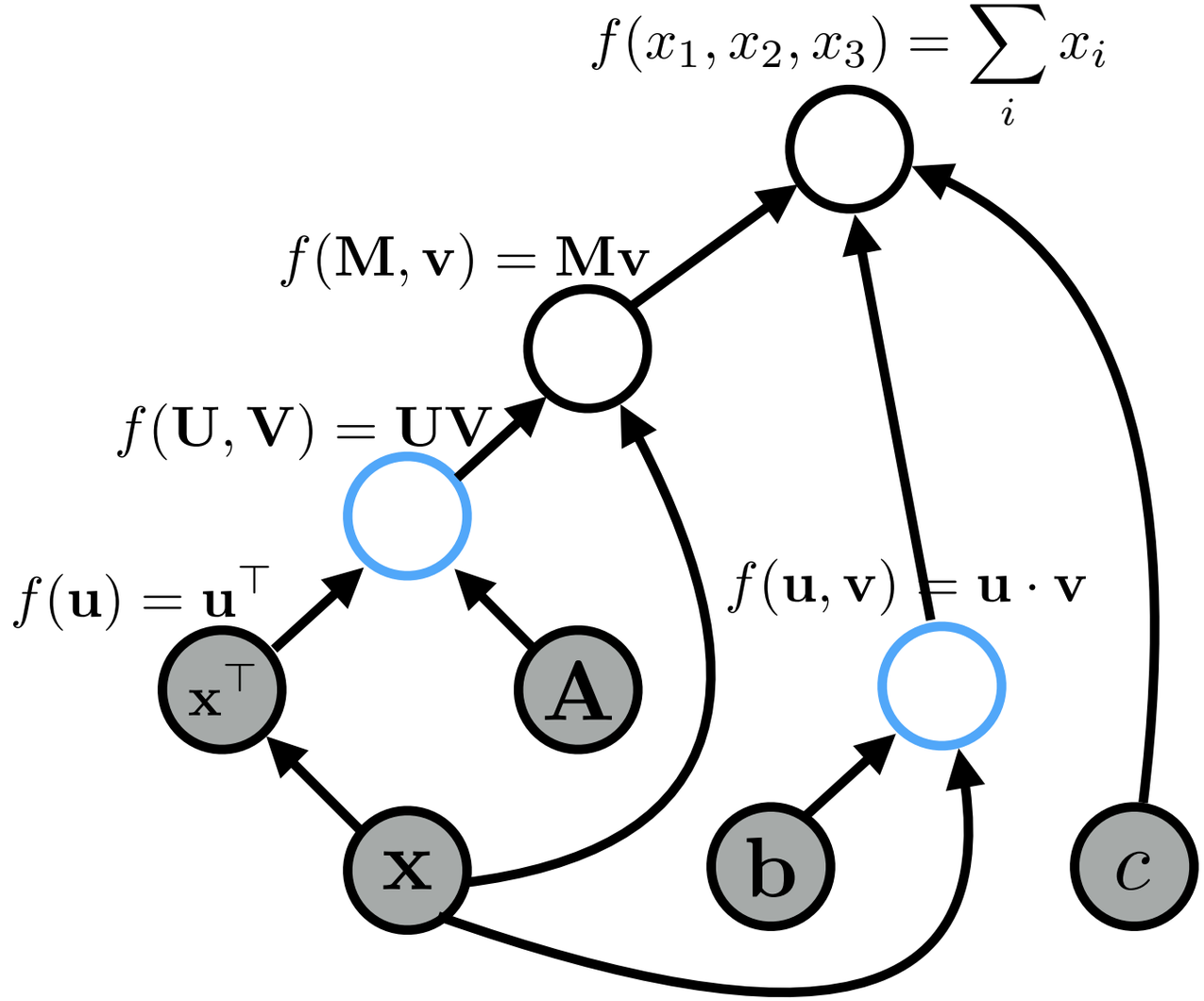
Forward Propagation

graph:



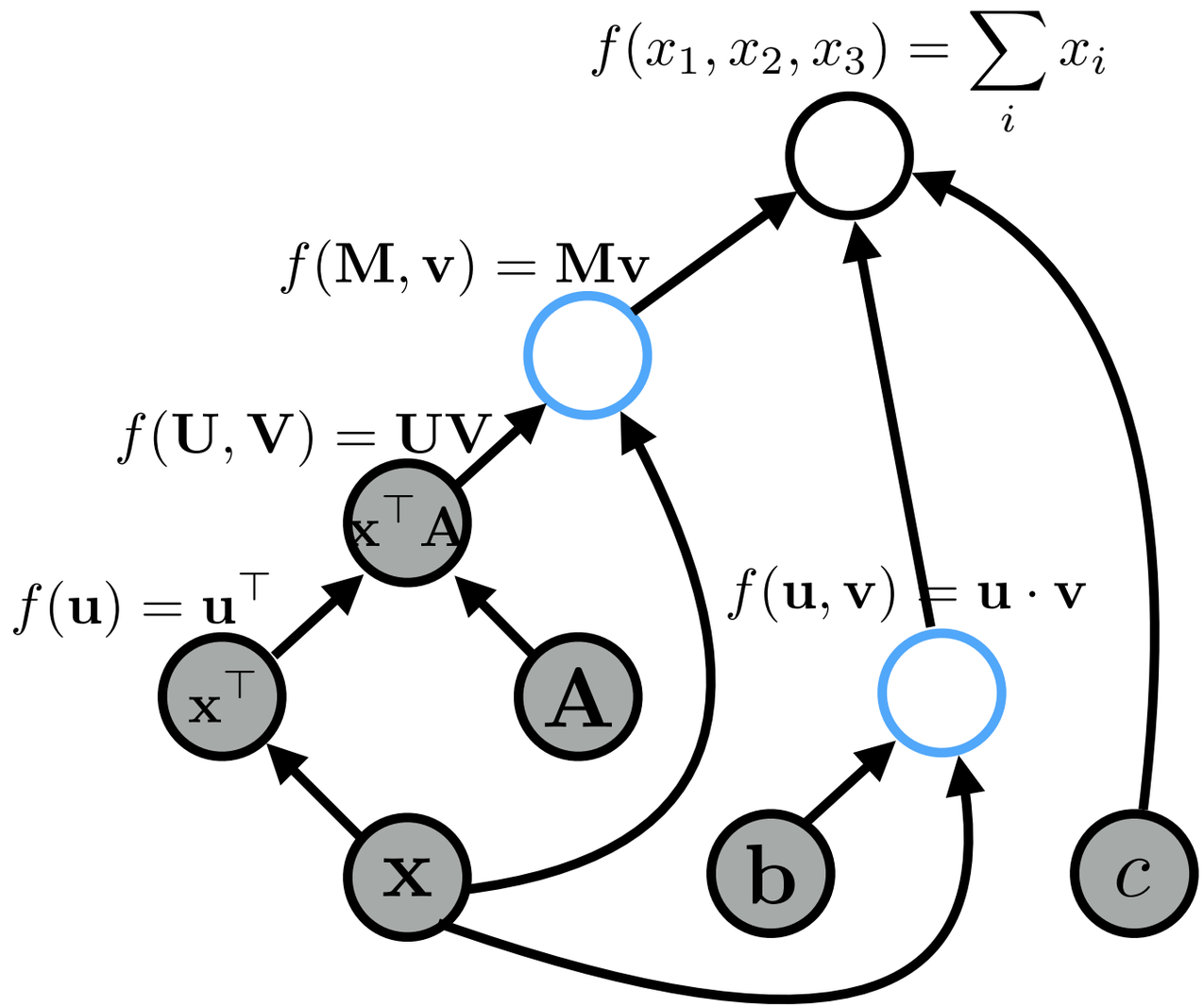
Forward Propagation

graph:



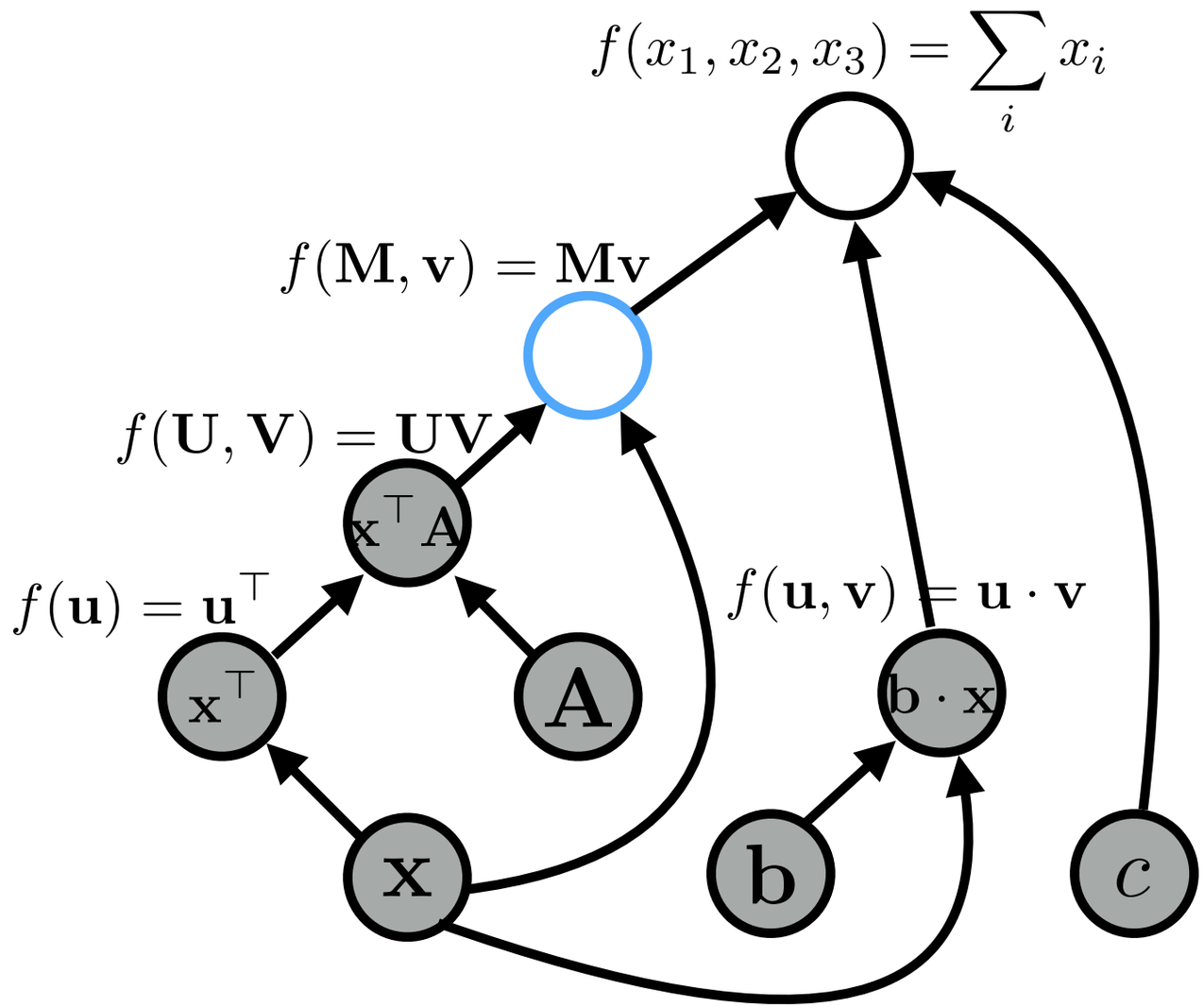
Forward Propagation

graph:



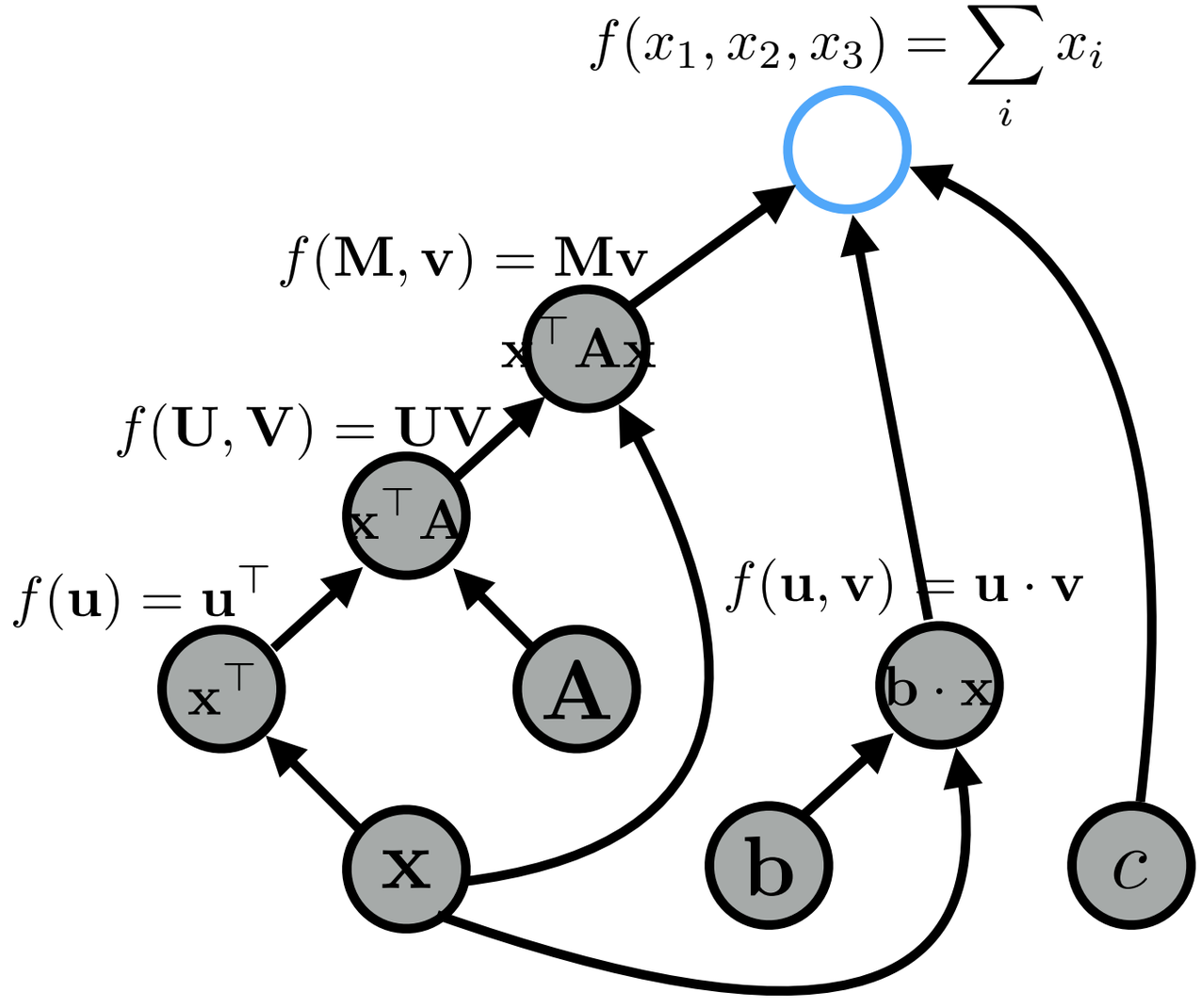
Forward Propagation

graph:



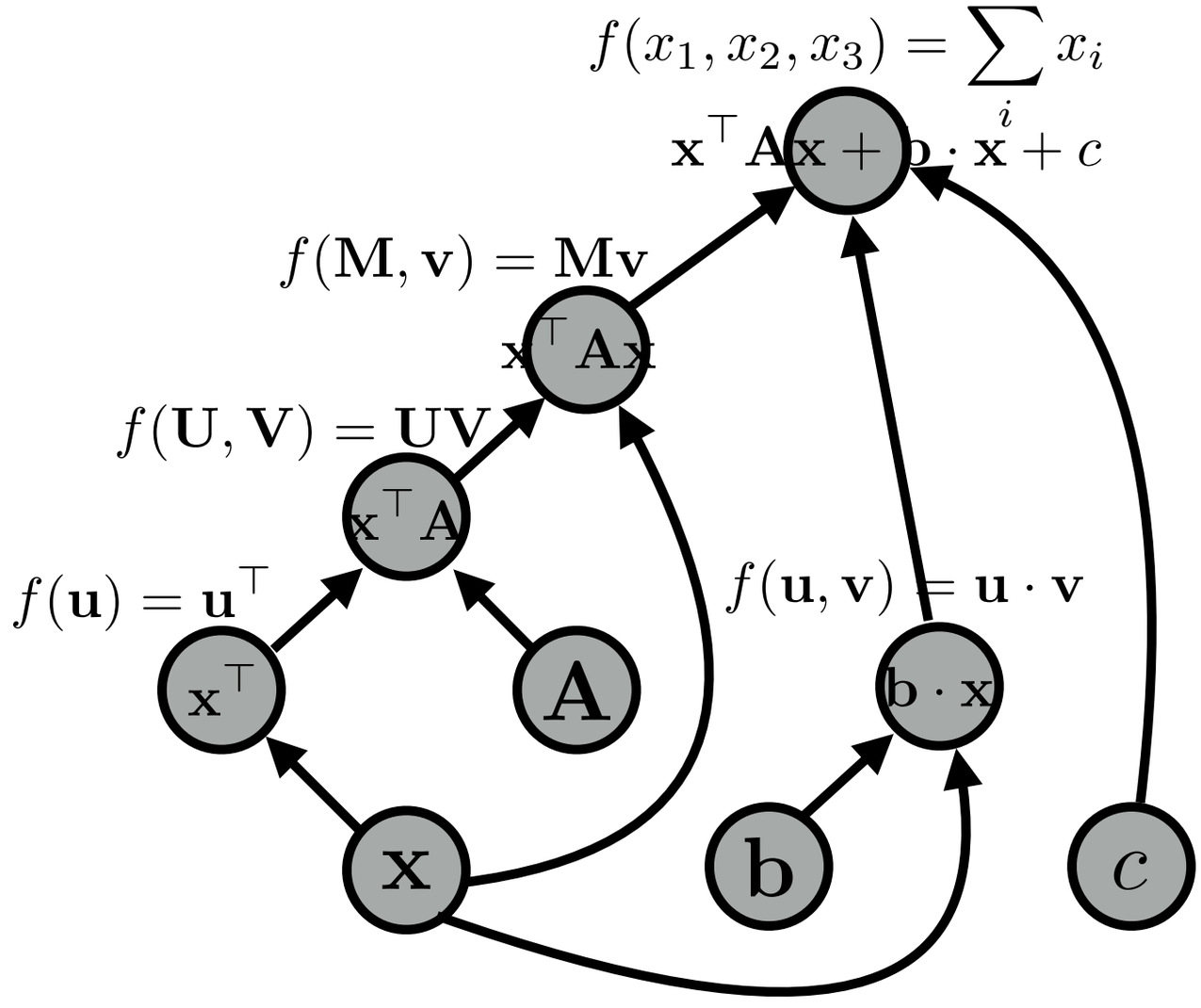
Forward Propagation

graph:



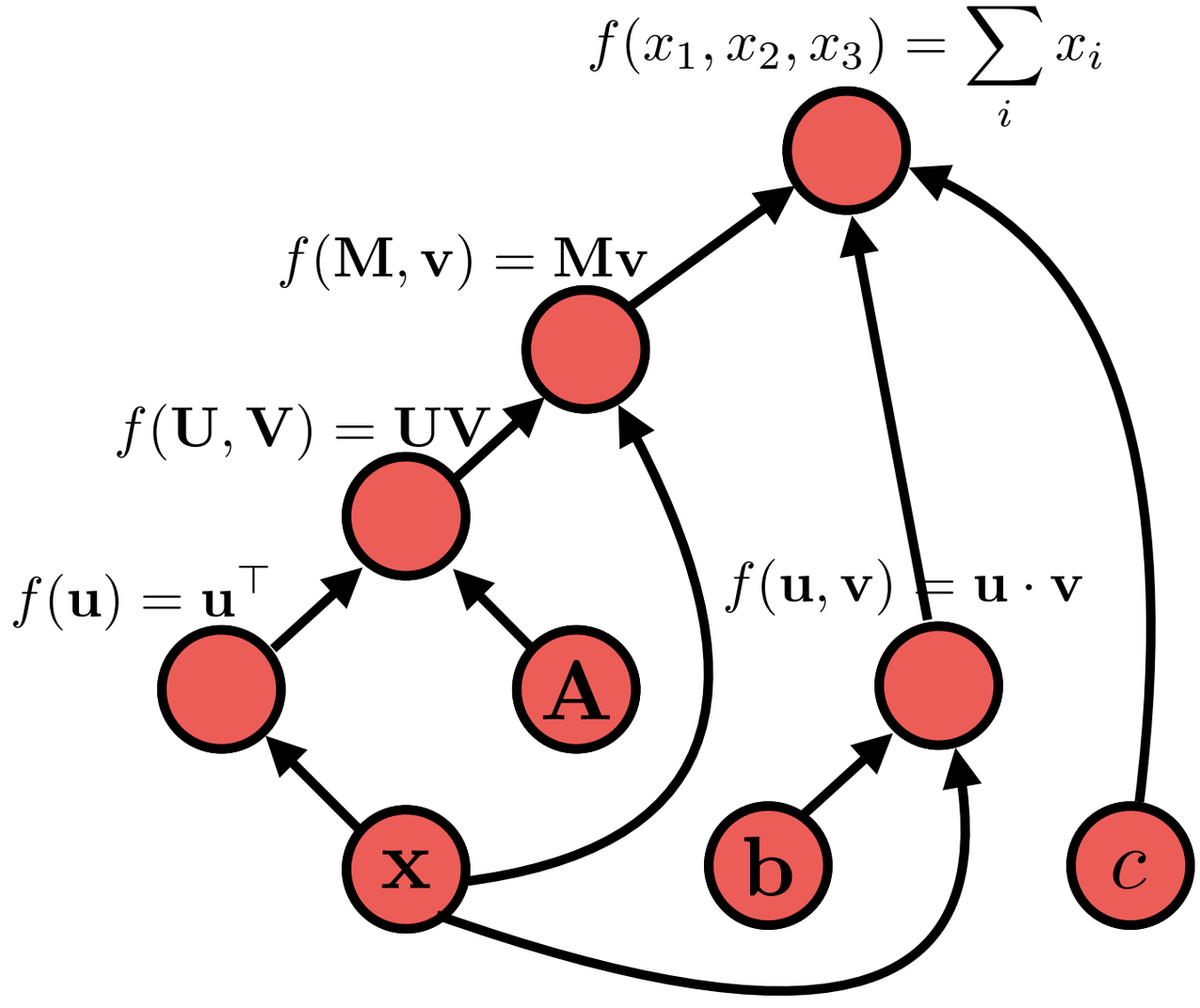
Forward Propagation

graph:



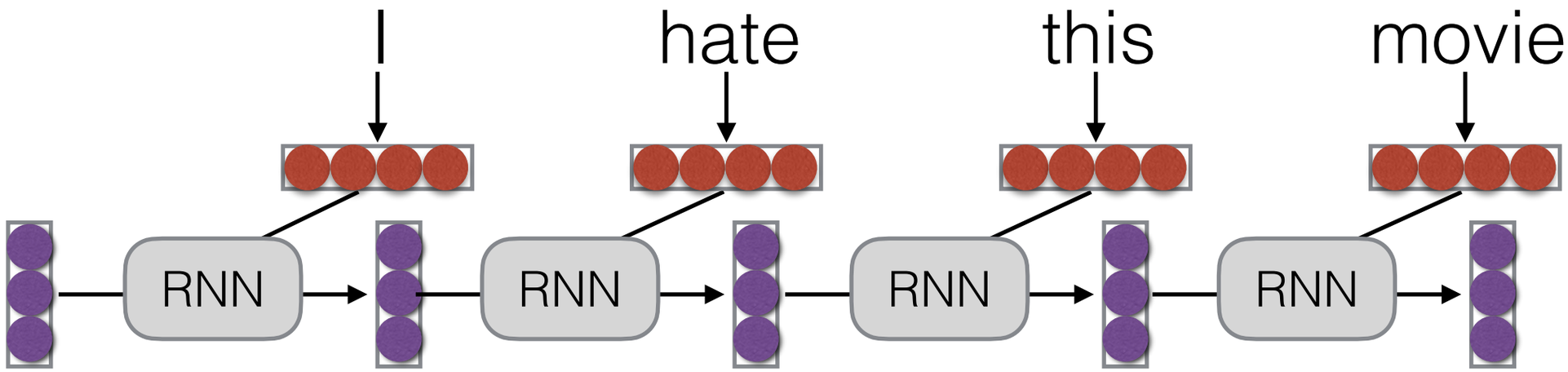
Back Propagation

graph:



(A Gross Oversimplification of) Neural Machine Translation

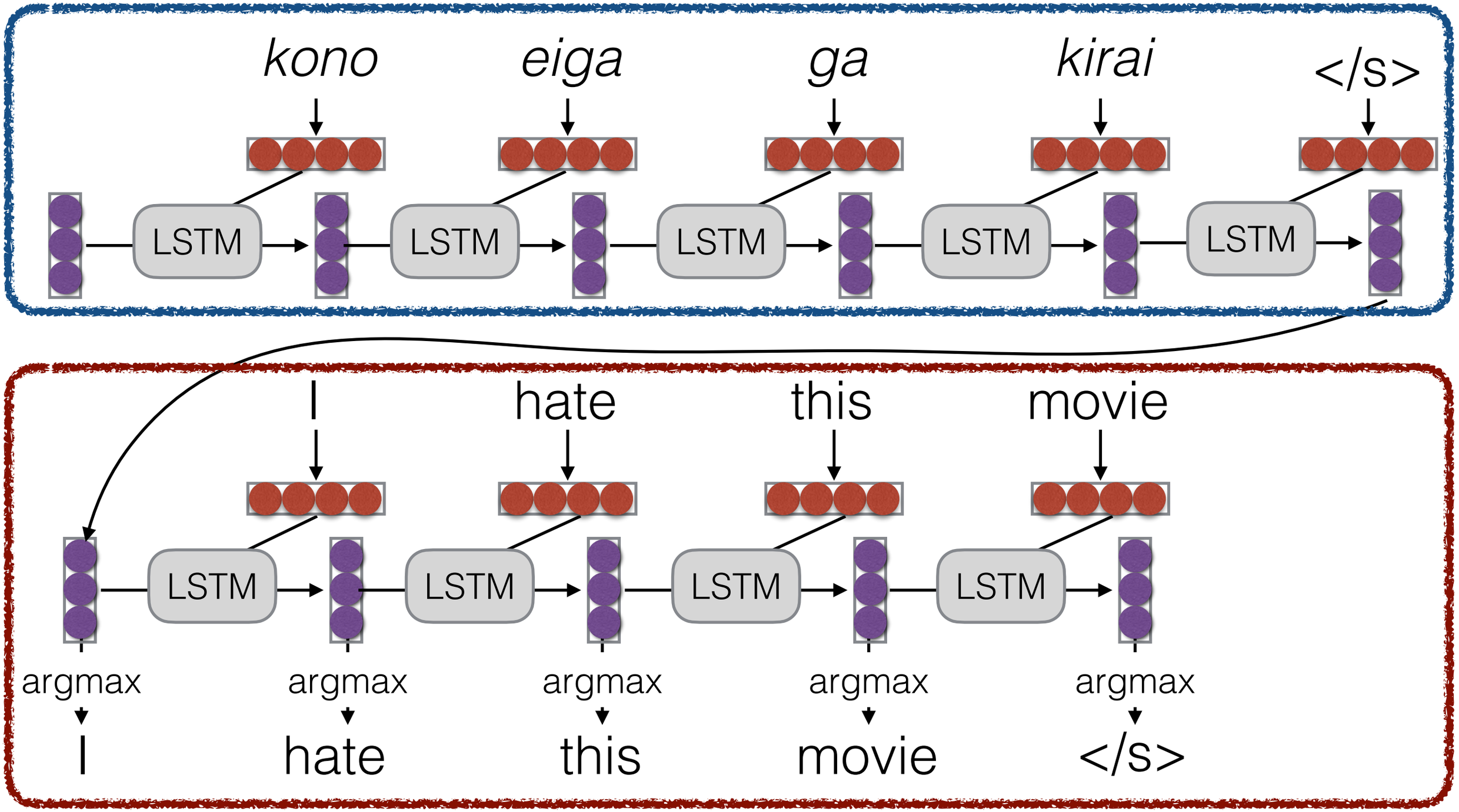
Recurrent Neural Networks



- Read in one word at a time and extract "features**"**
- **Word embeddings (red):** features of each individual word
- **Hidden words (purple):** features of *all* words seen so far

Encoder- Decoders

Encoder

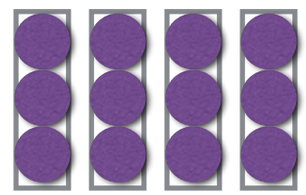


Decoder

Sentence Representations

this is an example 

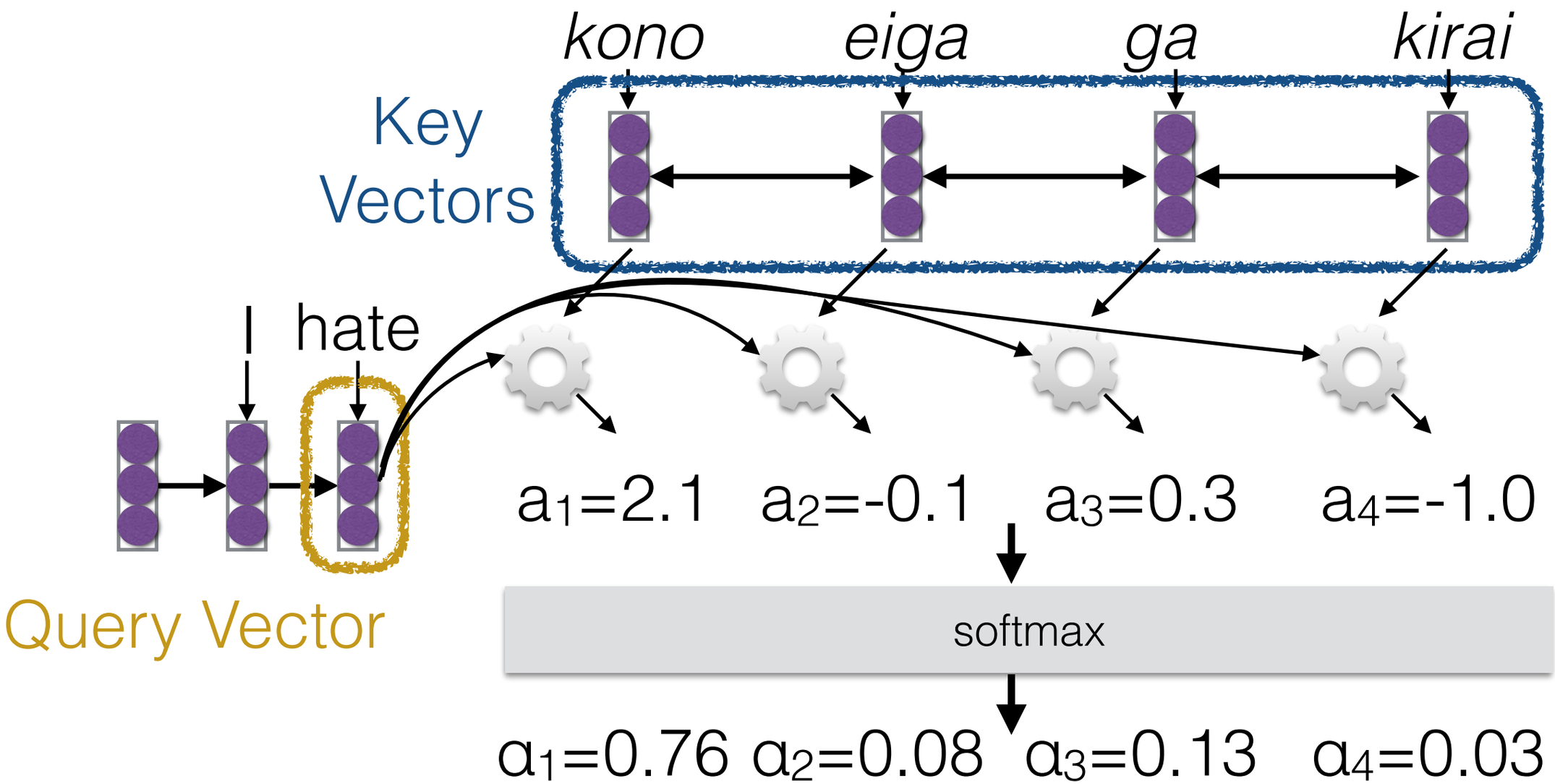
- "read this document, remember it, and translate it"

this is an example 

- "read this document, reference it while translating"

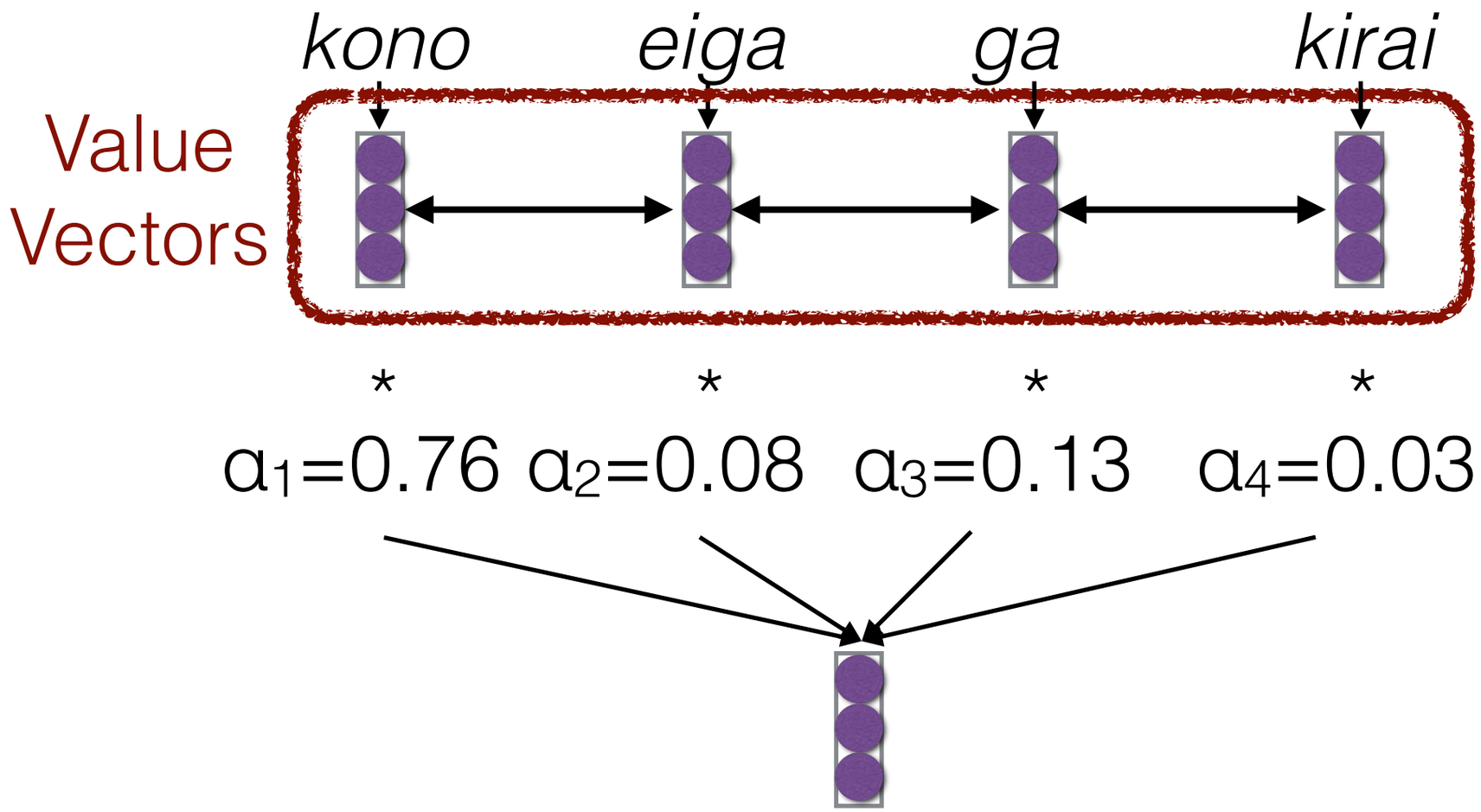
Calculating Attention (1)

- 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 states, like key vectors) by taking the weighted sum



- Use this in any part of the model you like

A Graphical Example

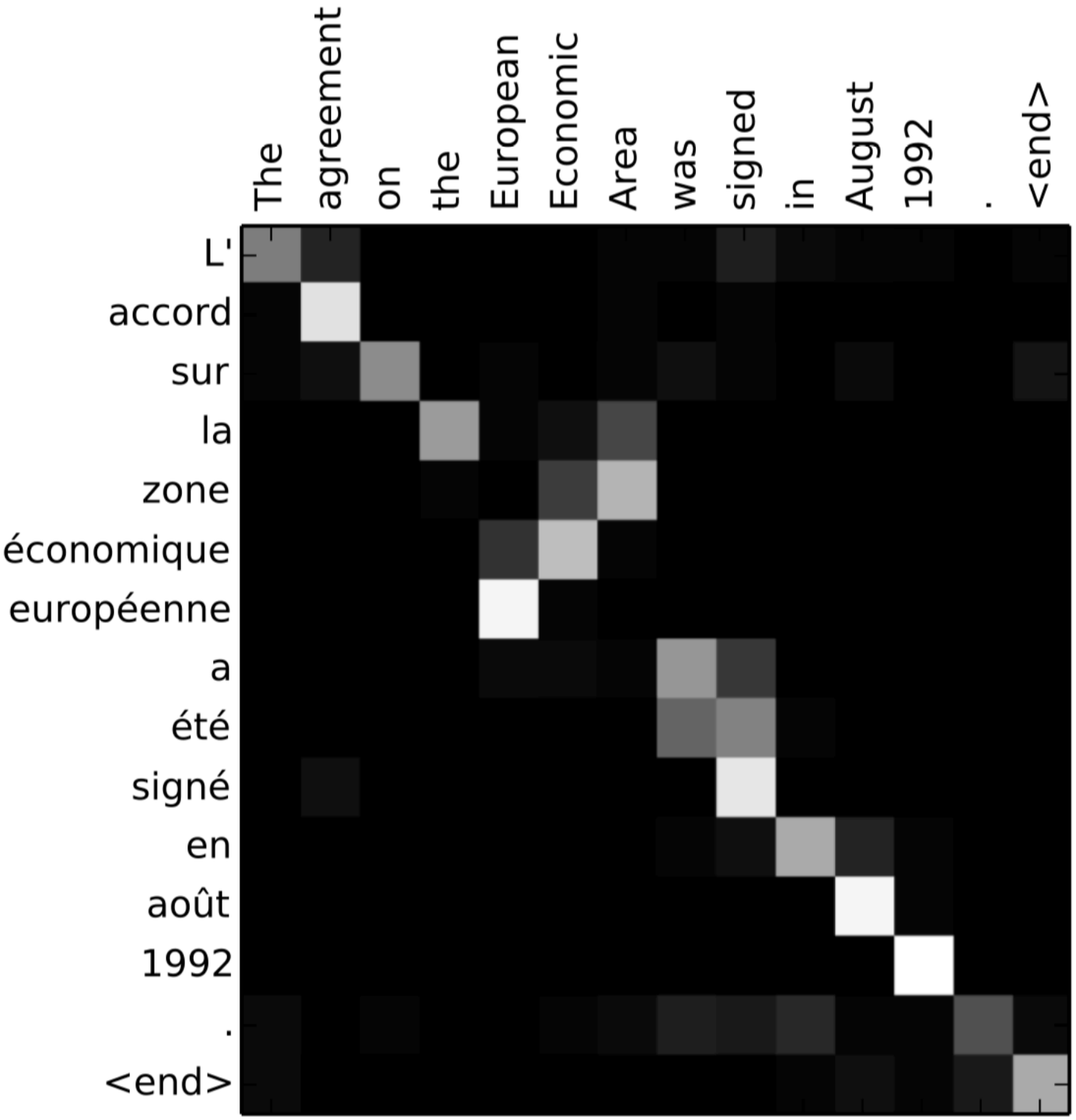
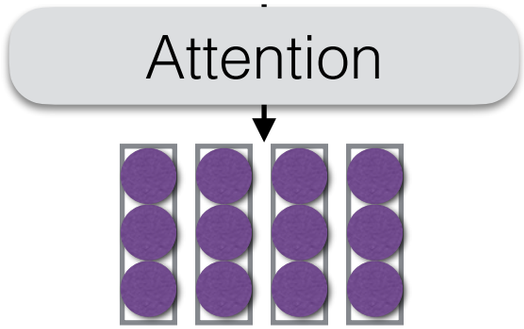


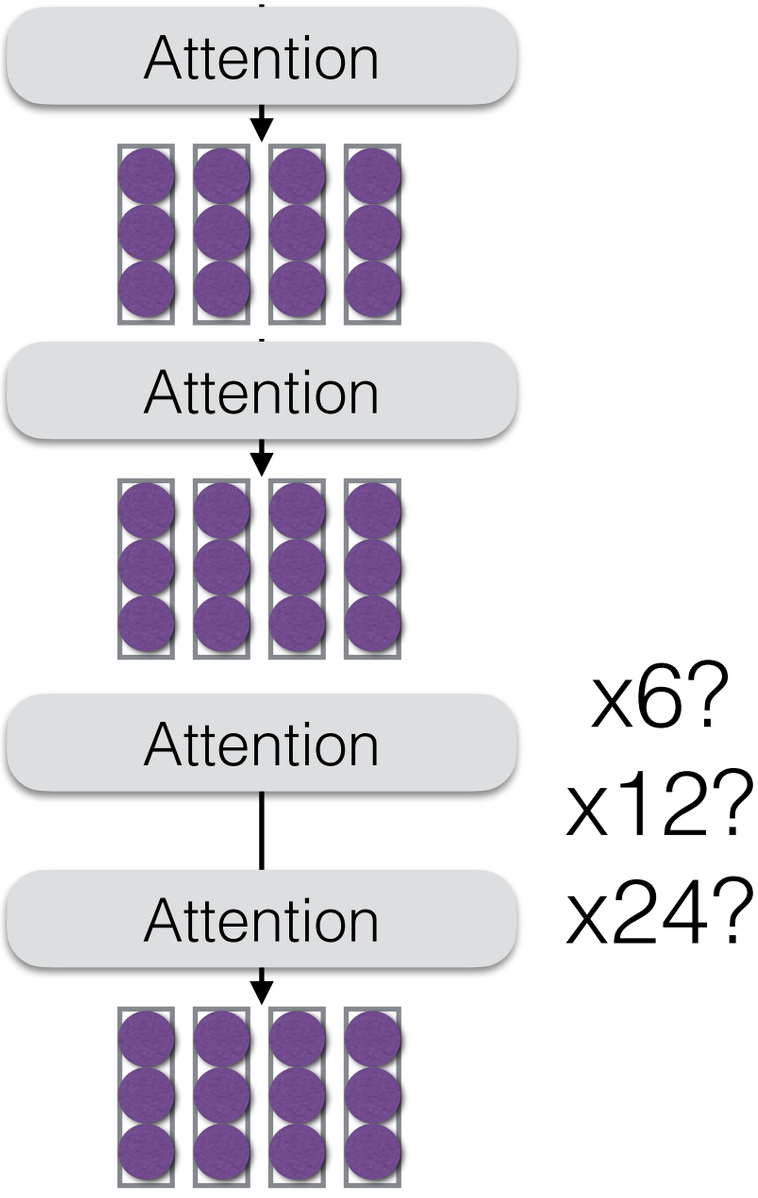
Image from Bahdanau et al. (2015)

"Deep" Learning

this is an example



this is an example



x6?
x12?
x24?

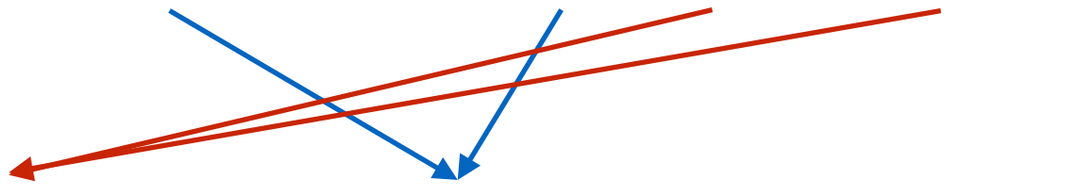
Again, Why Can/Can't Machines Translate?

What Can/Can't Machine Translation Currently Do?

- **Theoretically:** Neural networks can model arbitrarily complicated functions -- i.e. they can do anything
- **In Practice:**
  - MT is very good at modeling direct associations
  - MT is pretty good at modeling associations that require capturing word similarity
  - MT is not great at capturing phenomena that require more than one step of "reasoning"
  - MT cannot learn things that are not included in its data

プログラム の コード を 書いた 。

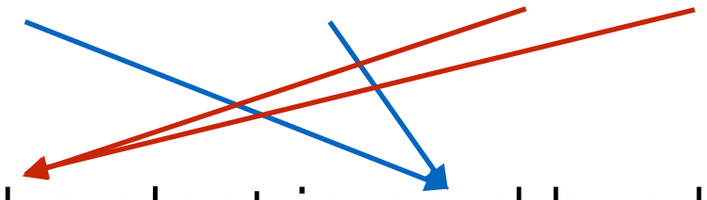
program of ko-do OBJ wrote .



I wrote the code for the program.

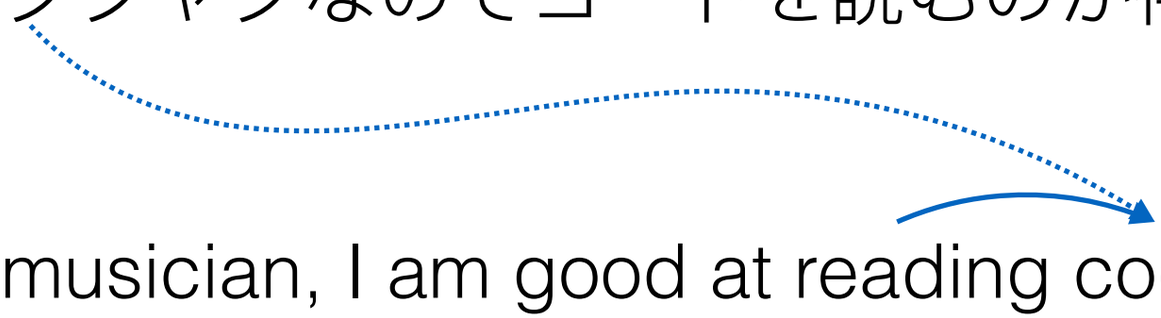
電気 の コード が 切れた 。

electric of ko-do SUBJ broke .



The electric cord has broken.

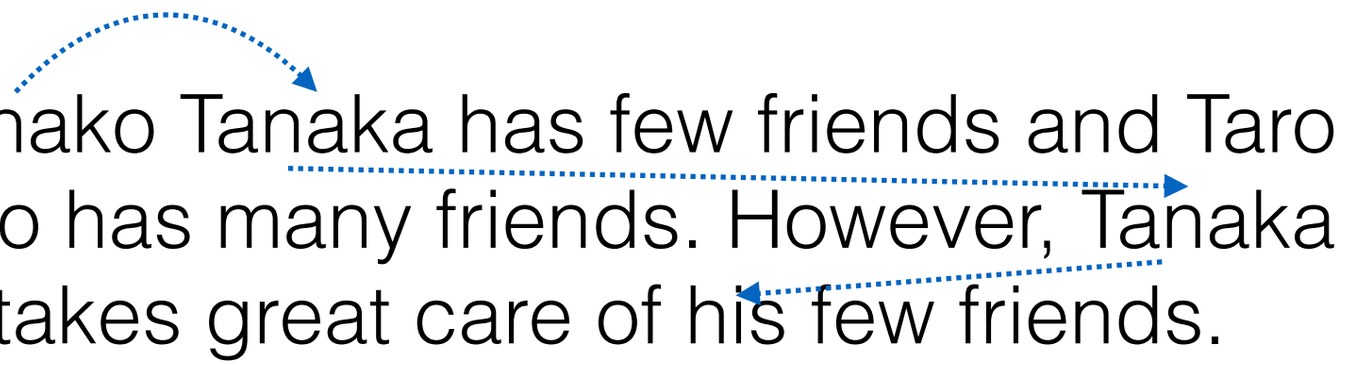
ミュージシャンなのでコードを読むのが得意。



As a musician, I am good at reading codes.

"reading" more strongly associated with "codes" than "ミュージシャン" with "chords"?

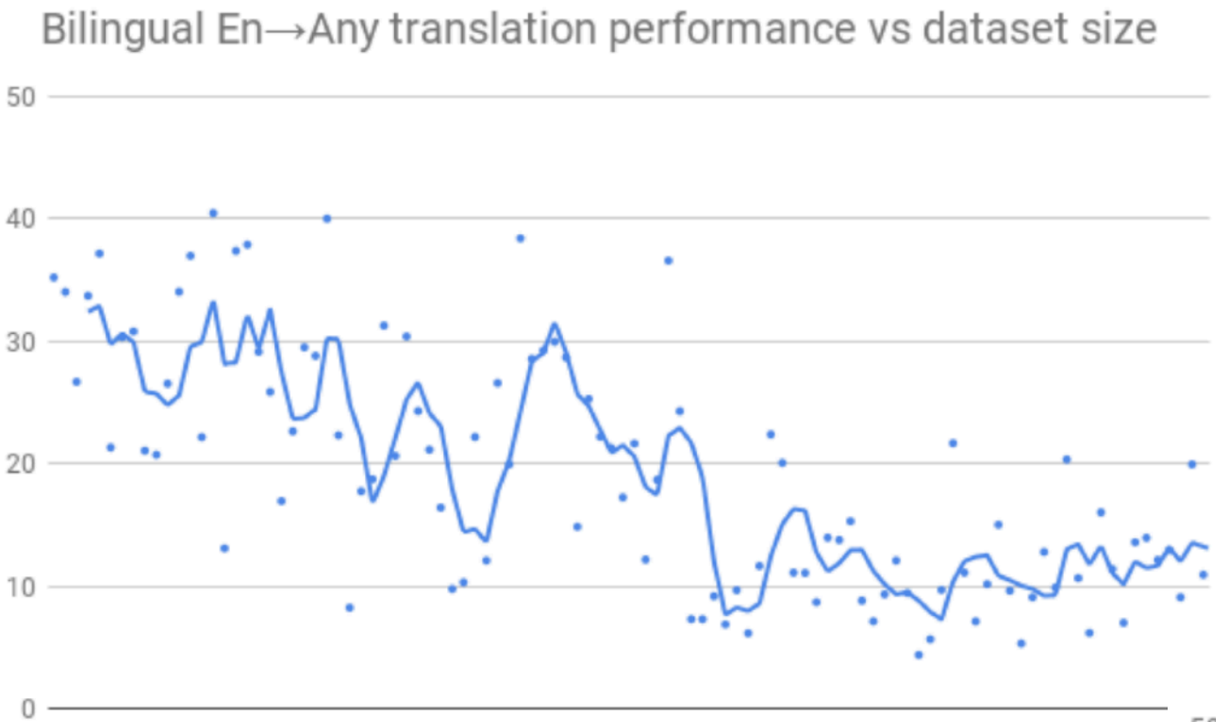
田中花子は友人が少なく佐藤太郎は友人が多い。
だが、田中は数少ない友人をとても大事にする。

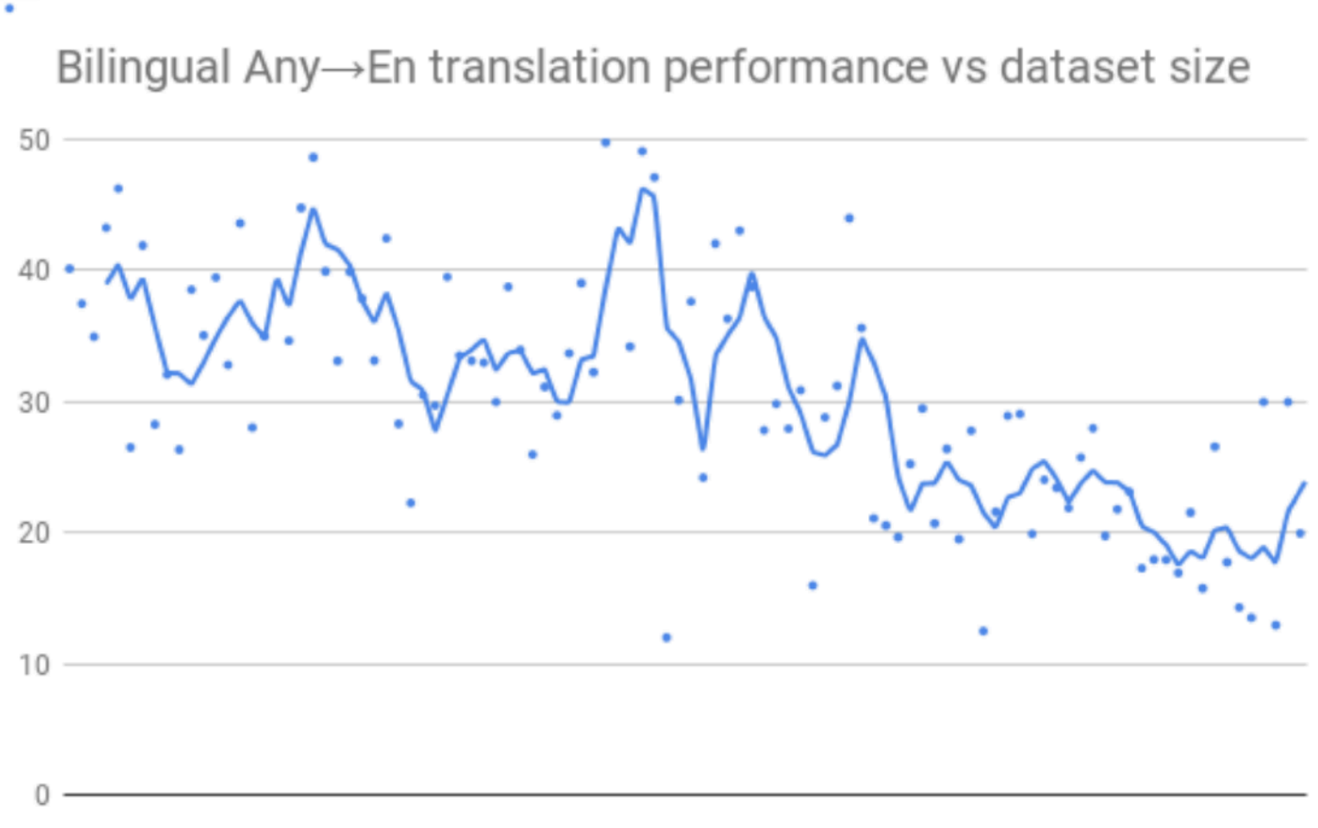


Hanako Tanaka has few friends and Taro Sato has many friends. However, Tanaka takes great care of his few friends.

Too many steps of between stereotypically female name "Hanako" and the final prediction of a gendered pronoun.

Data Efficiency





MT Performance by Data [Arivazhagan+ 19]

What Does this Mean for Future Linguists like You?

Overall Implications

- Machine translation is pretty good, and it will continue to get better
- However, it's not perfect
  - Especially in *domains or languages where little data exists*
  - Especially in *non-literal translations*
  - Especially in any situation where you really have to *think*
- Barring a major technological breakthrough, the above weak points will improve gradually, not be solved any time soon

Shared Strengths

- Humans and machines are better at different things
  - *Machines:* memory, speed
  - *Humans:* reasoning, estimating uncertainty, communication with clients
- For high-quality translation, the answer will likely not be human translation *or* machine translation, but how we can best combine both

Case Study: Chinese-English News Translation

Error Category	Errors			Significance		
	H_A	H_B	MT_1	$H_A - H_B$	$H_A - MT_1$	$H_B - MT_1$
Incorrect Word	51	52	85		***	***
Semantics	33	36	48			
Grammaticality	18	16	37		**	**
Missing Word	37	69	56	***	*	
Semantics	22	62	34	***		***
Grammaticality	15	7	22			**
Named Entity	16	19	30		*	
Person	1	10	10	*	*	
Location	5	4	6			
Organization	4	4	8			
Event	1	1	3			
Other	5	1	7			
Word Order	1	4	17		***	**
Factoid	1	1	6			
Word Repetition	2	4	4			
Collocation	15	18	27			
Unknown Words/Misspellings	0	1	0			
Context (Register, Coreference, etc.)	6	9	12			
Any	81	103	118	*	***	
Total	129	177	237	**	***	**

More mistaken words by MT

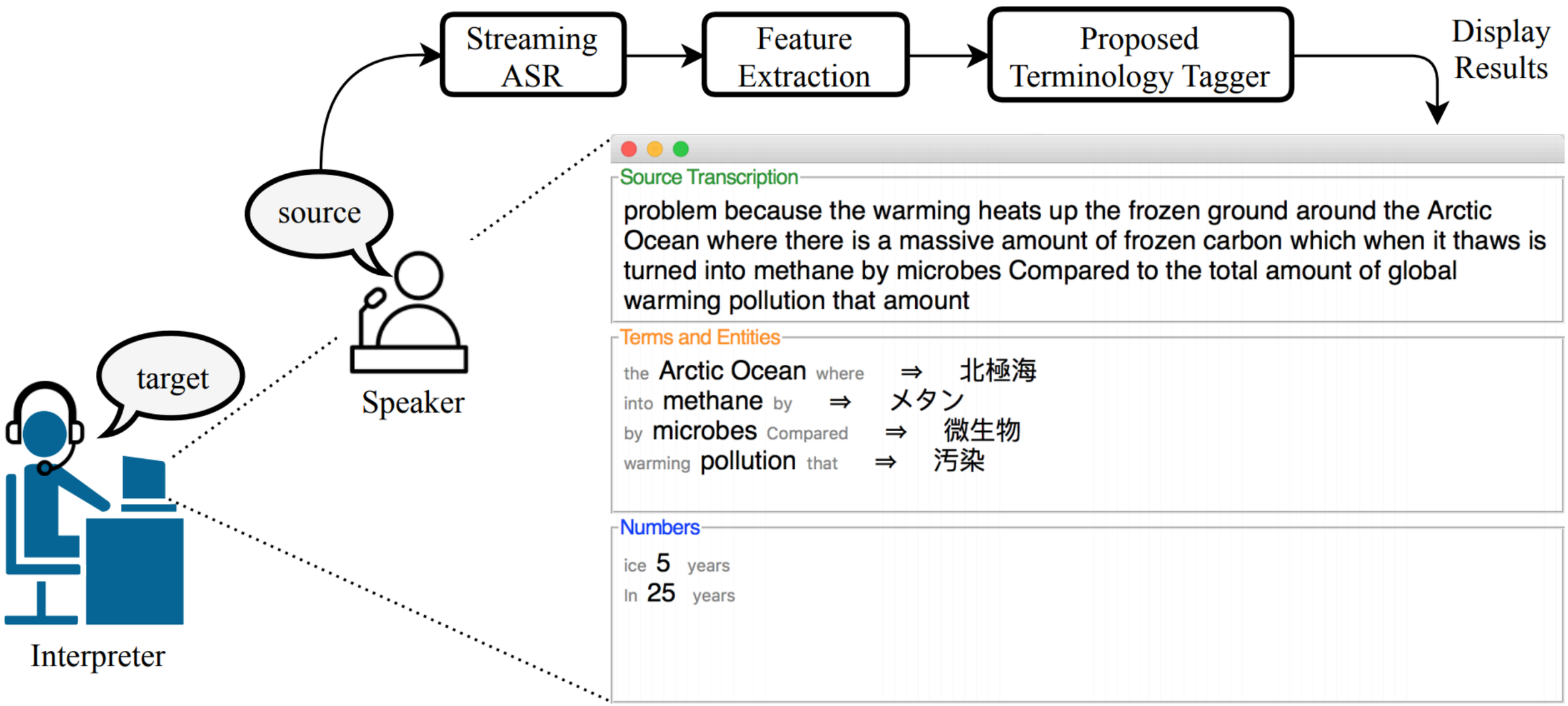
More missing words by human translator B

Many more grammatical errors by MT

Samuel Laubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, Antonio Toral. dA Set of Recommendations for Assessing Human–Machine Parity in Language Translation. Journal of Artificial Intelligence Research (JAIR). February 2020.

An Example of Symbiosis: Terminology Assistance for Interpreters

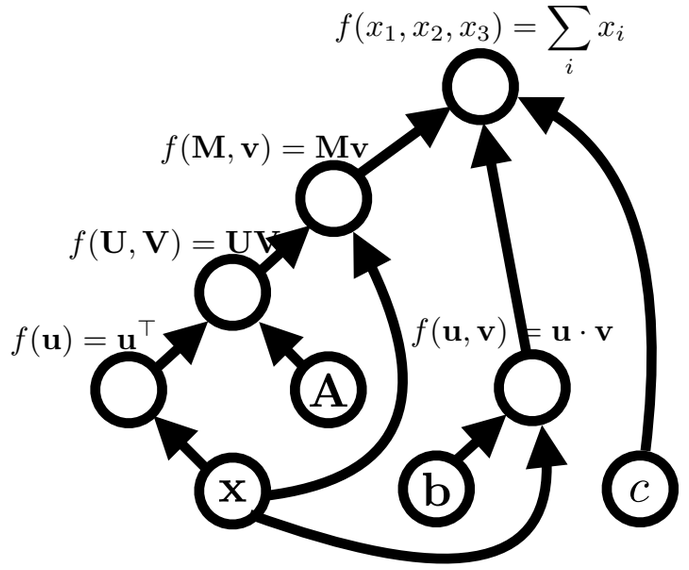
- Simultaneous interpretation is an strenuous task, and term recall is difficult
- One solution: leave interpreting to humans, but have machines assist in term recall



- Nikolai Vogler, Craig Stewart, Graham Neubig.
  Lost in Interpretation: Predicting Untranslated Terminology in Simultaneous Interpretation.
  Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL). June 2019.

Conclusion

- Modern MT can seem like magic at times, but it's not



- It does lots of things well, and lots of things poorly
- Understanding it and working with it is likely the path forward

Thank You! Questions?