Natural Language Processing

in the Era of Deep Learning

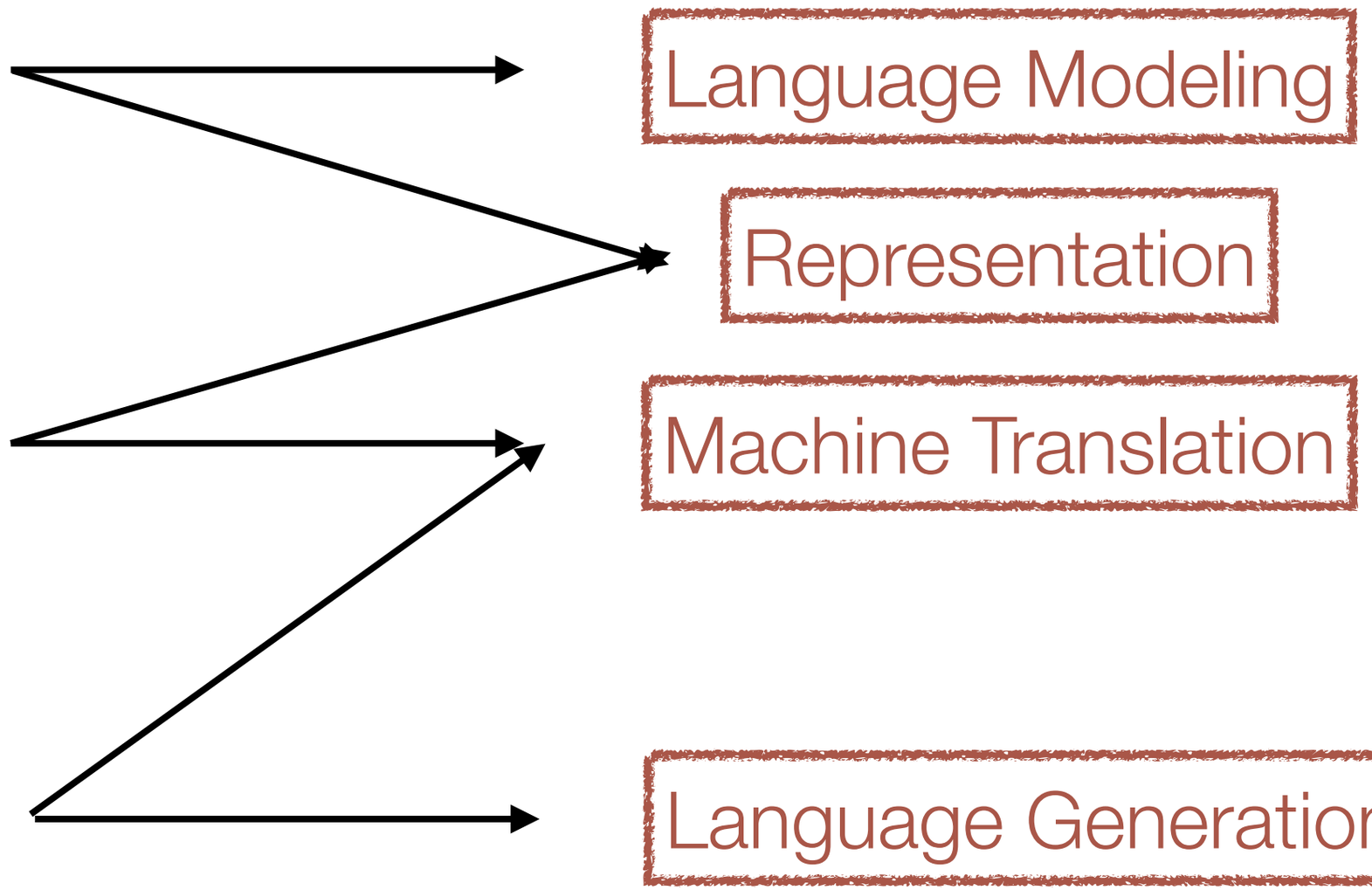
Yanran Li
The Hong Kong Polytechnic University

Language is Important

Why? How?

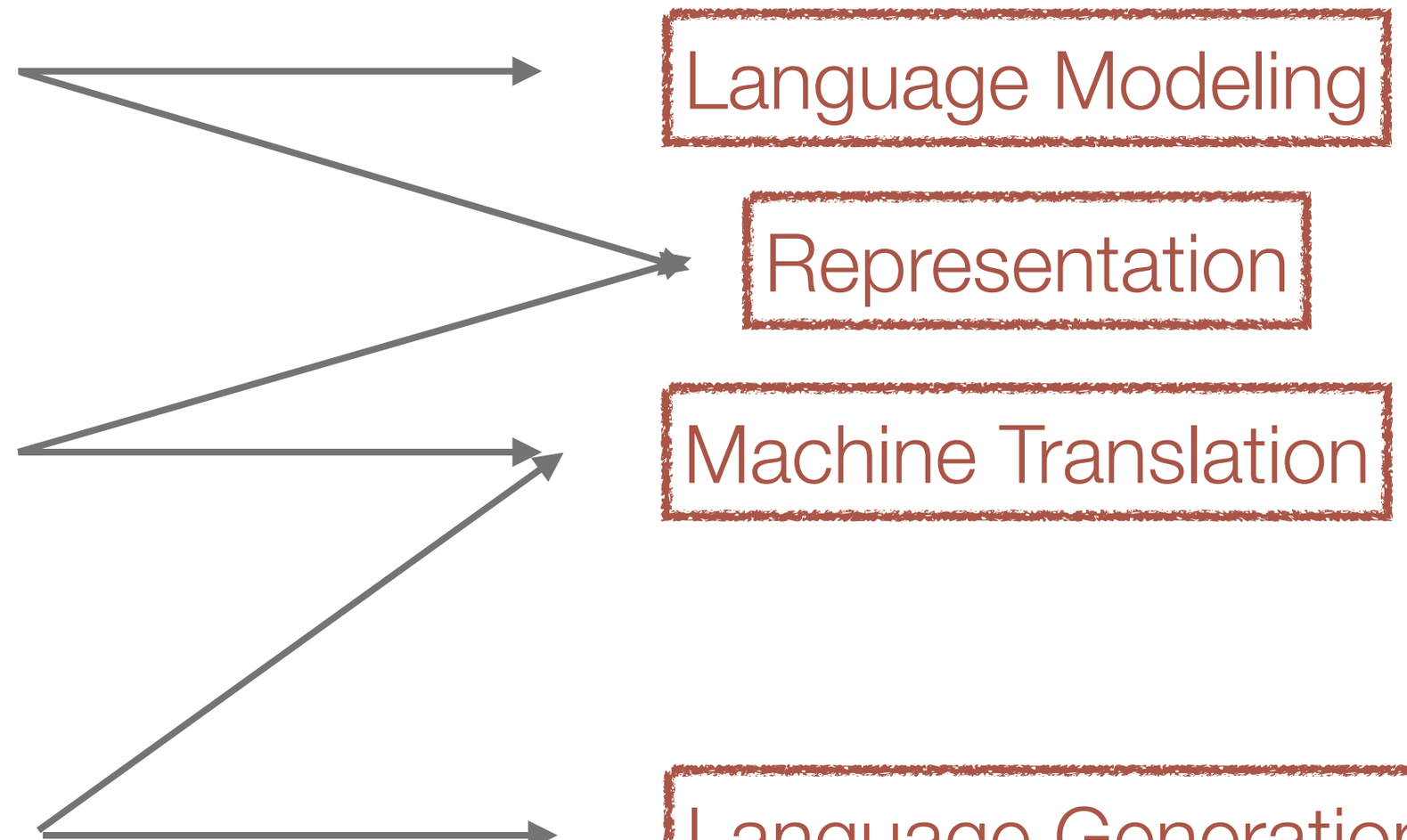
Language is important as we use it to

- ✦ Perceive
- ✦ Understand
- ✦ Communicate



A Roadmap of today's sharing…

- Perceive
- Understand
- Communicate



Language Modeling

What is it?

Language Modeling is all about

- measuring how likely a sentence is…
  - A sentence: “I would like to commend the rapporteur on his work”
  - can be modelled as: $(x_1, x_2, \ldots, x_L)$
  - How likely is this sentence?
  - It is then cast as: what is the probability of $(x_1, x_2, \ldots, x_L)$

N-gram Language Model

- is a past popular way to calculate $p(x_1, x_2, \ldots, x_L)$
  - it depends on n-th order Markov assumption
  - therefore, we only need to collect n-gram statistics from corpus.

N-gram Language Model

- e.g. *p(*i, would, like, to, …, ., </s>*)*
- Unigram:
  - *p(*i*)p(*would*)p(*like*)…p(*</s>*)*
- Bigram:
  - *p(*i*)p(*would|i*)p(*like|would*)…p(*</s>*)*
- Trigram:
  - *p(*i*)p(*would|i*)p(*like|i,would*)…*

word	unigram	bigram	trigram
i	6.684	3.197	3.197
would	8.342	2.884	2.791
like	9.129	2.026	1.031
to	5.081	0.402	0.144
commend	15.487	12.335	8.794
the	3.885	1.402	1.084
rapporteur	10.840	7.319	2.763
on	6.765	4.140	4.150
his	10.678	7.316	2.367
work	9.993	4.816	3.498
.	4.896	3.020	1.785
</s>	4.828	0.005	0.000
average	8.051	4.072	2.634
perplexity	265.136	16.817	6.206

…but it is limited as two issues

- Data Sparsity
  - # of n-gram statistics, sometimes = 0
- Lack of Generalization
  - (shining sun), (shining star)
  - (shining moon)?

before the era of Deep Learning

- Data Sparsity
  - # of n-gram statistics, sometimes = 0 + δ
  - we solve this problem by smoothing techniques
- Lack of Generalization
  - (shining sun), (shining star)
  - (shining moon)?
  - but we’re bothered with the second problem

Welcome to the Era of

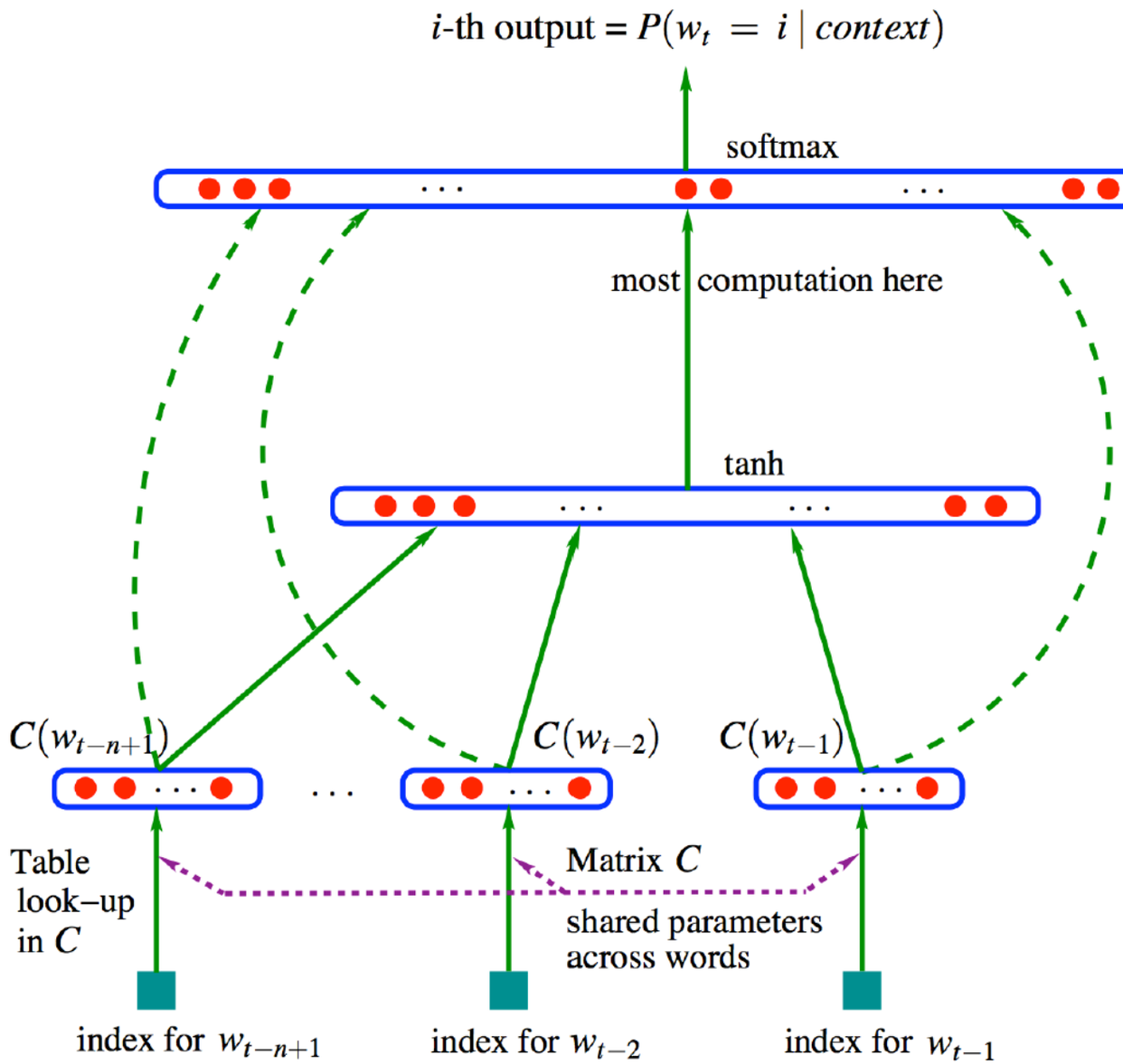
Deep Learning

Lack of Generalization is due to

- Discrete representation (1-hot representation)
  - star: [0 0 0 0 1 0 0 0 0 0]
  - sun: [0 0 0 1 0 0 0 0 0 0]
  - moon: [0 0 0 0 0 0 1 0 0 0]
  - # of n-gram statistics

It is *solved by Neural Language Modeling (2003)

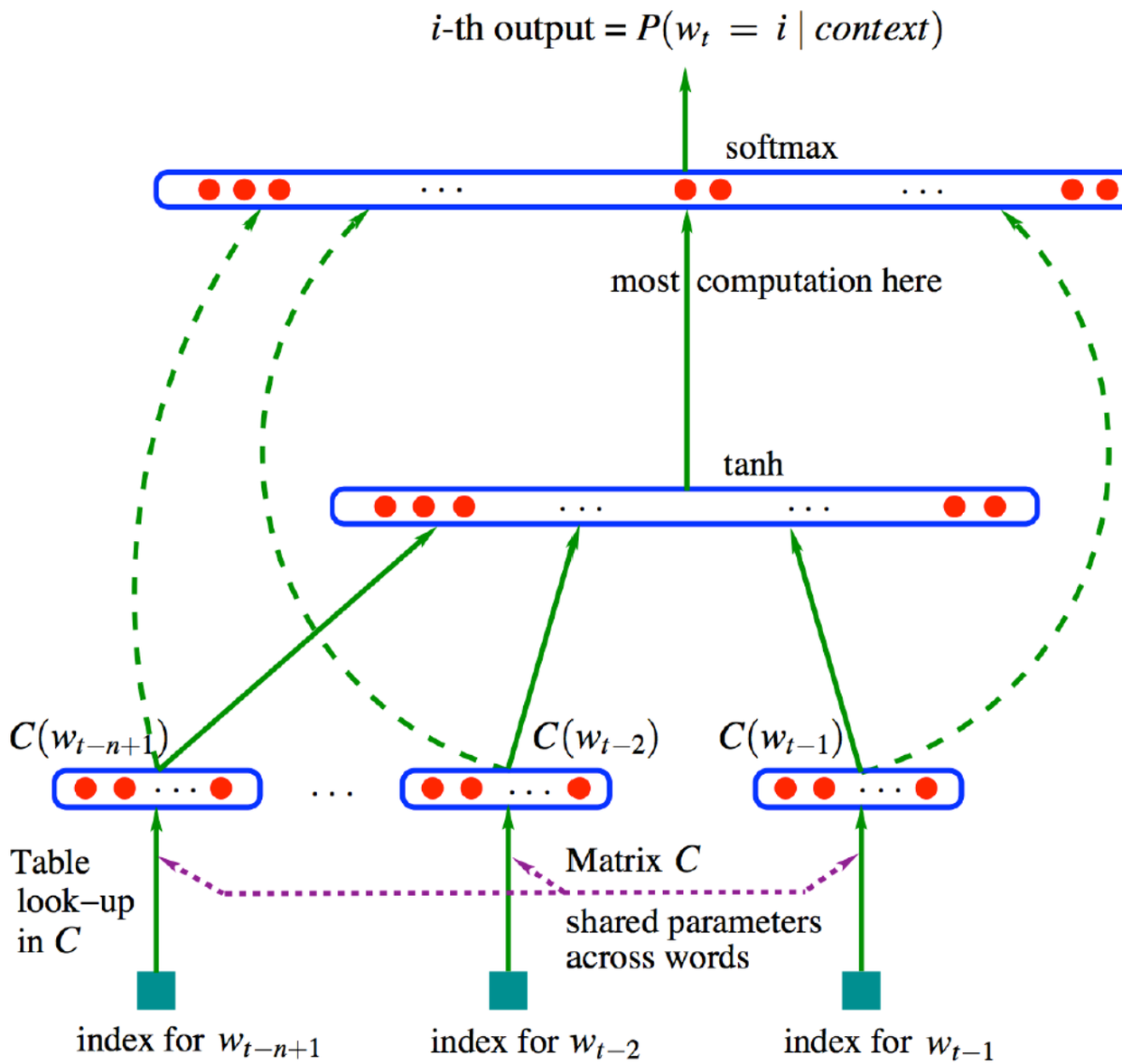
- ✦ Yoshua Bengio. A neural probabilistic language model.



Neural Language Model

It brings the powerful

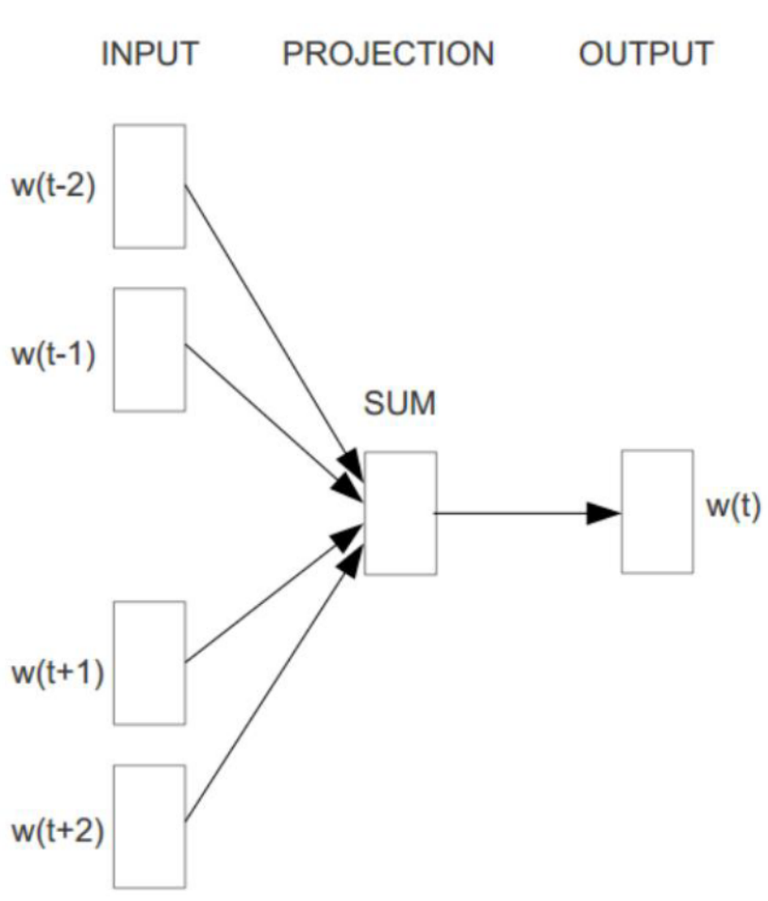
- ✦ Continuous-space word representation: embedding

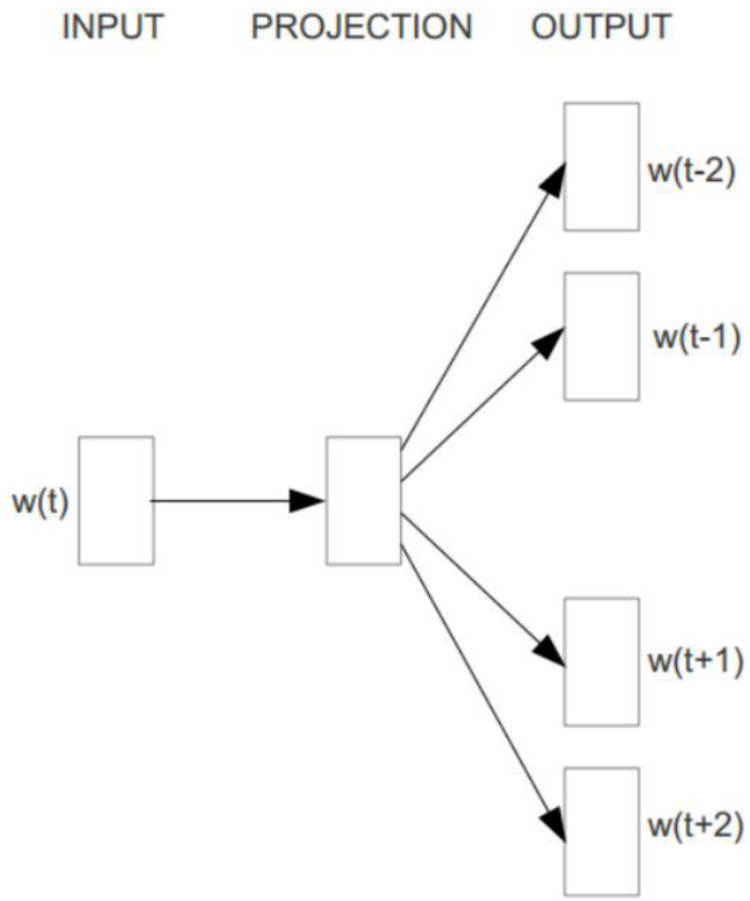


Neural Language Model

…and can be calculated by Word2Vec

- Tomas Mikolov et al. Distributed representations of words and phrases and their compositionality. NIPS 2013





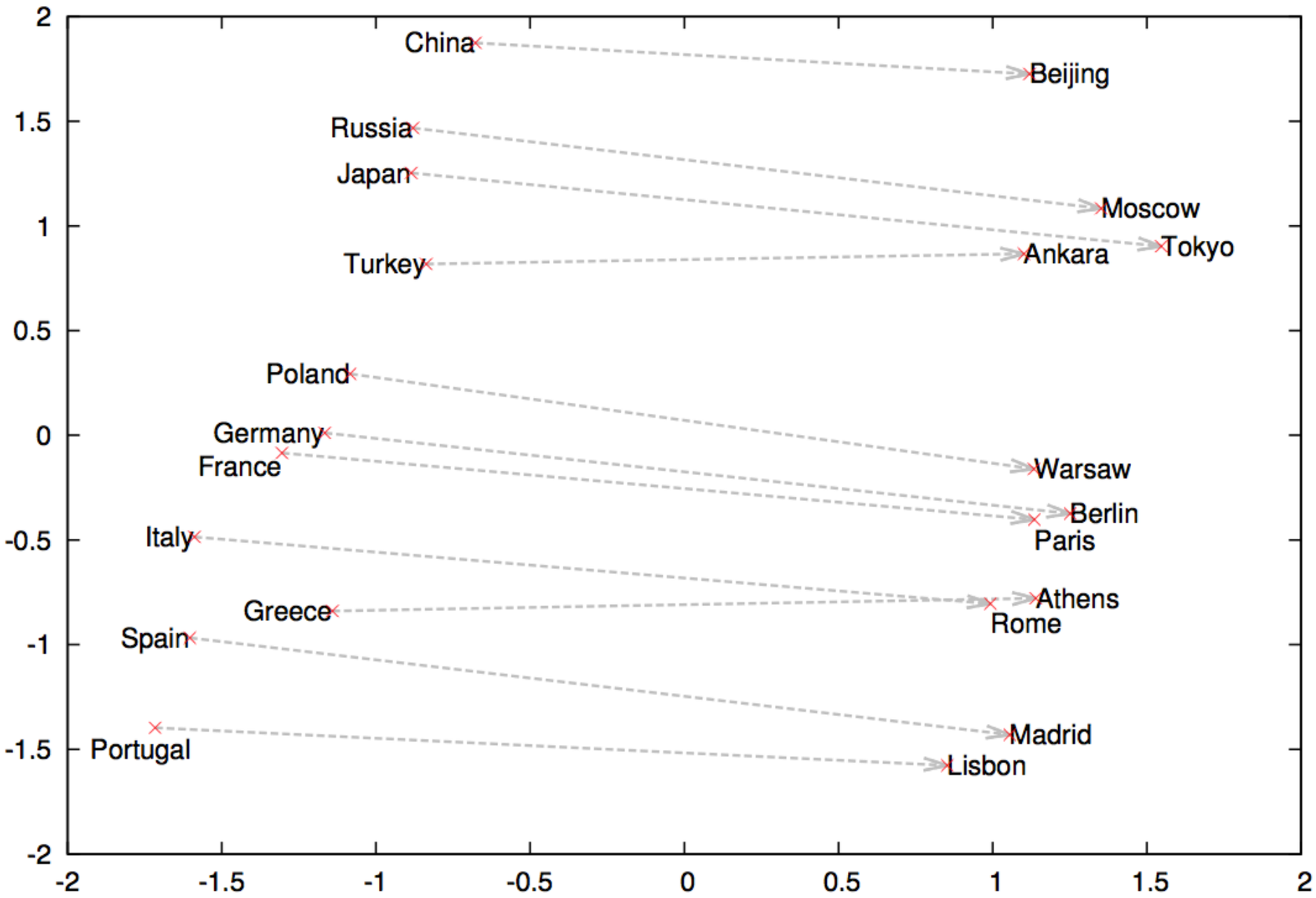
So, what do embeddings help?

- ✦ Word Similarity Calculation[11]

ANCE	JESUS	XBOX	REDDISH	SCRATCHED	MEGABITS
STRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
LGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
LMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
TALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
EECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
EDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
RWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
ROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
NGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
LERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES

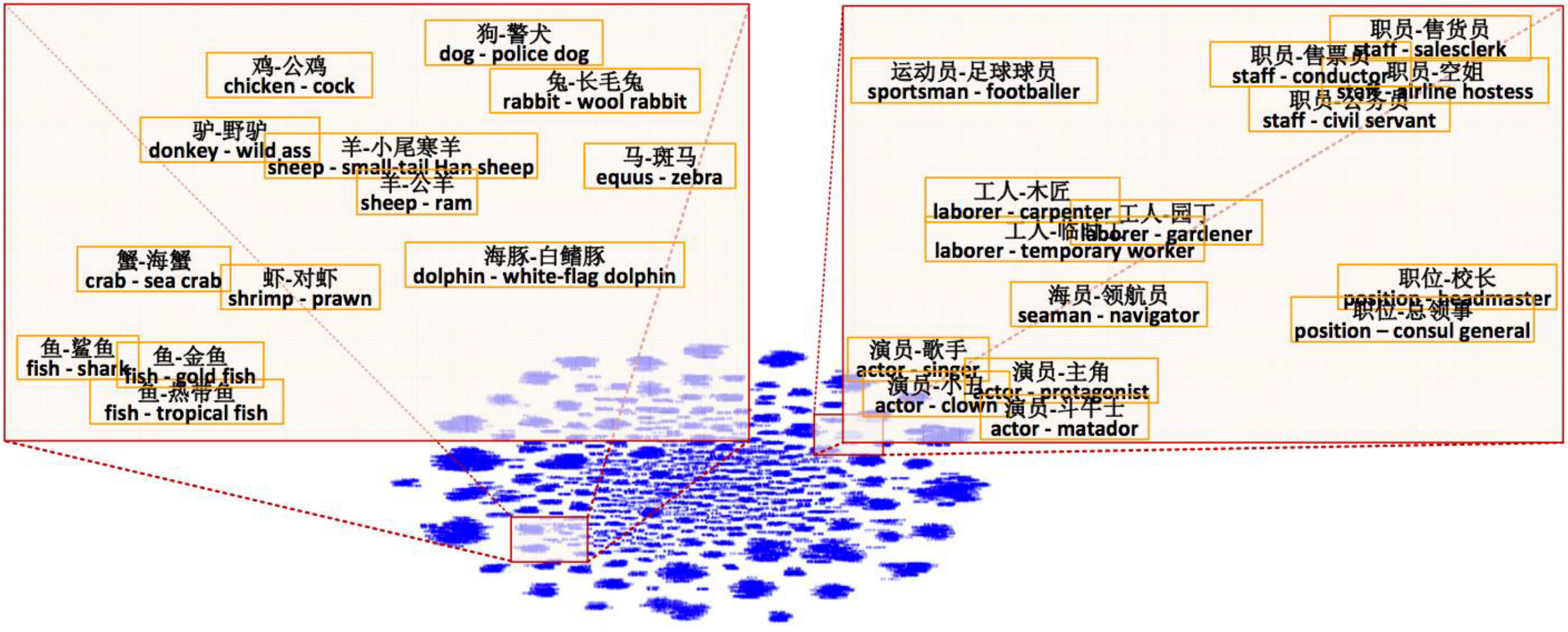
So, what do embeddings help?

- Linguistic Regularity: Properties



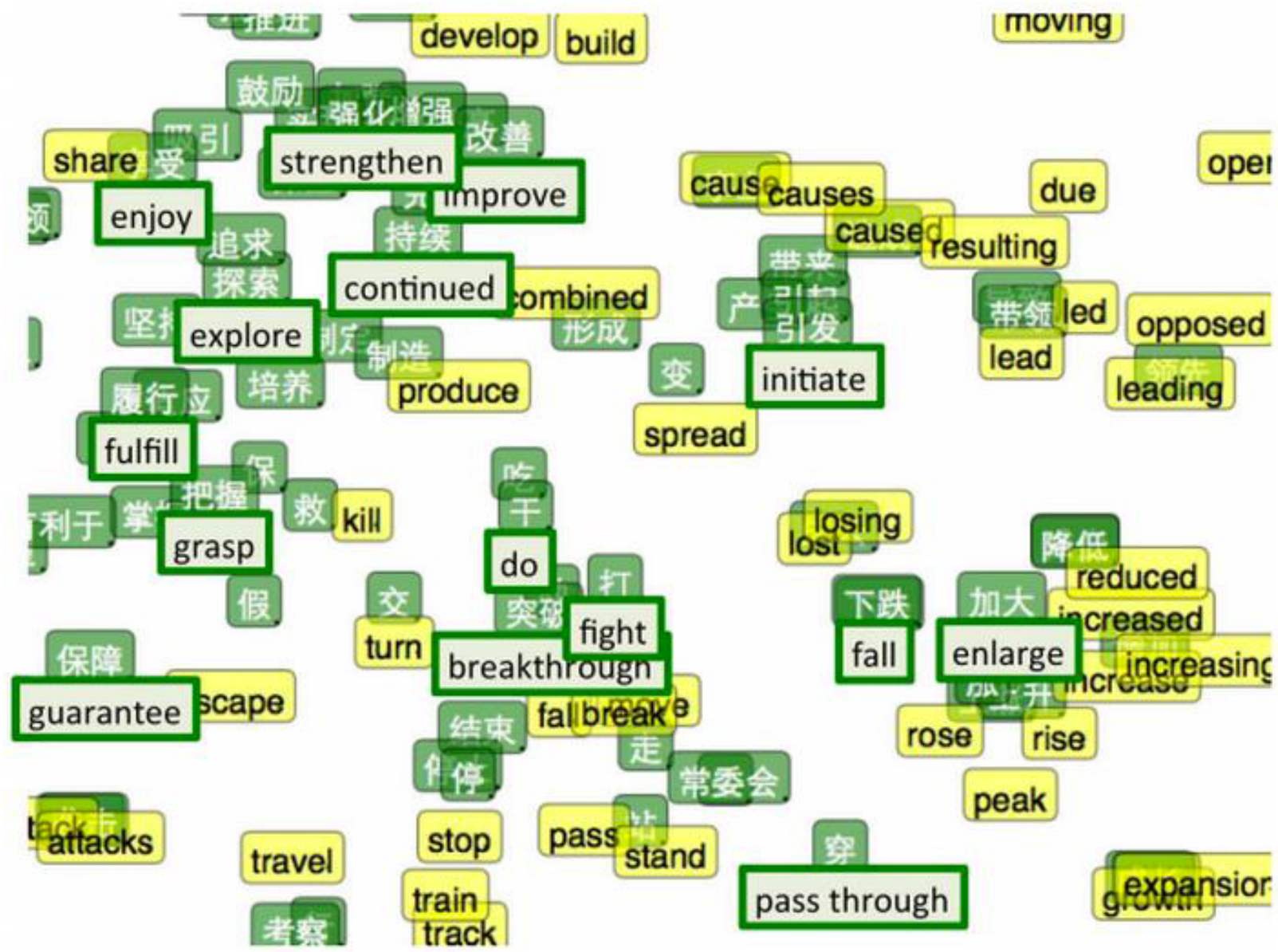
So, what do embeddings help?

- ✦ Word Relations: A specific kind of linguistic regularity[10]



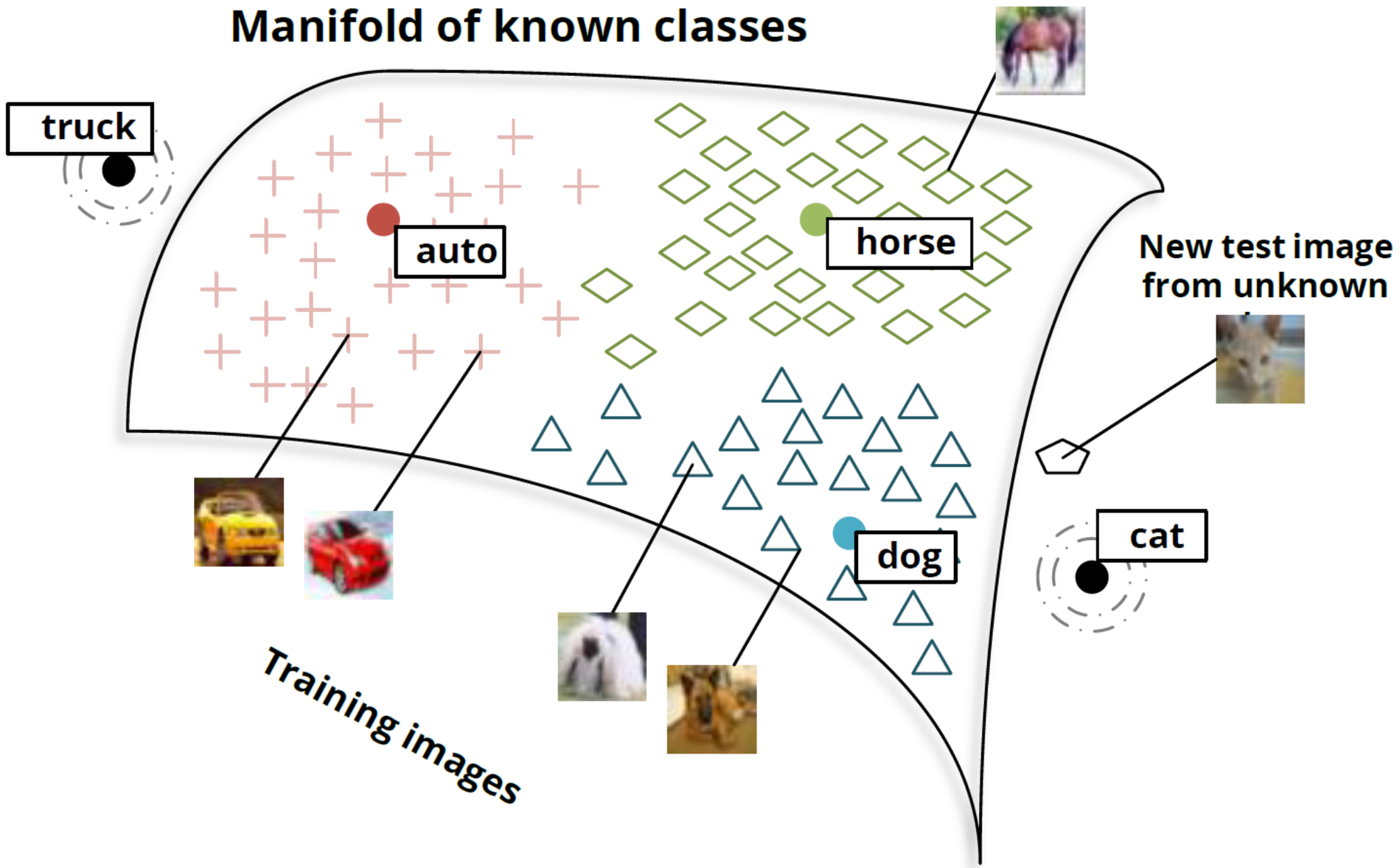
So, what do embeddings help?

- ✦ Word Relations: in multi-lingual space[12]



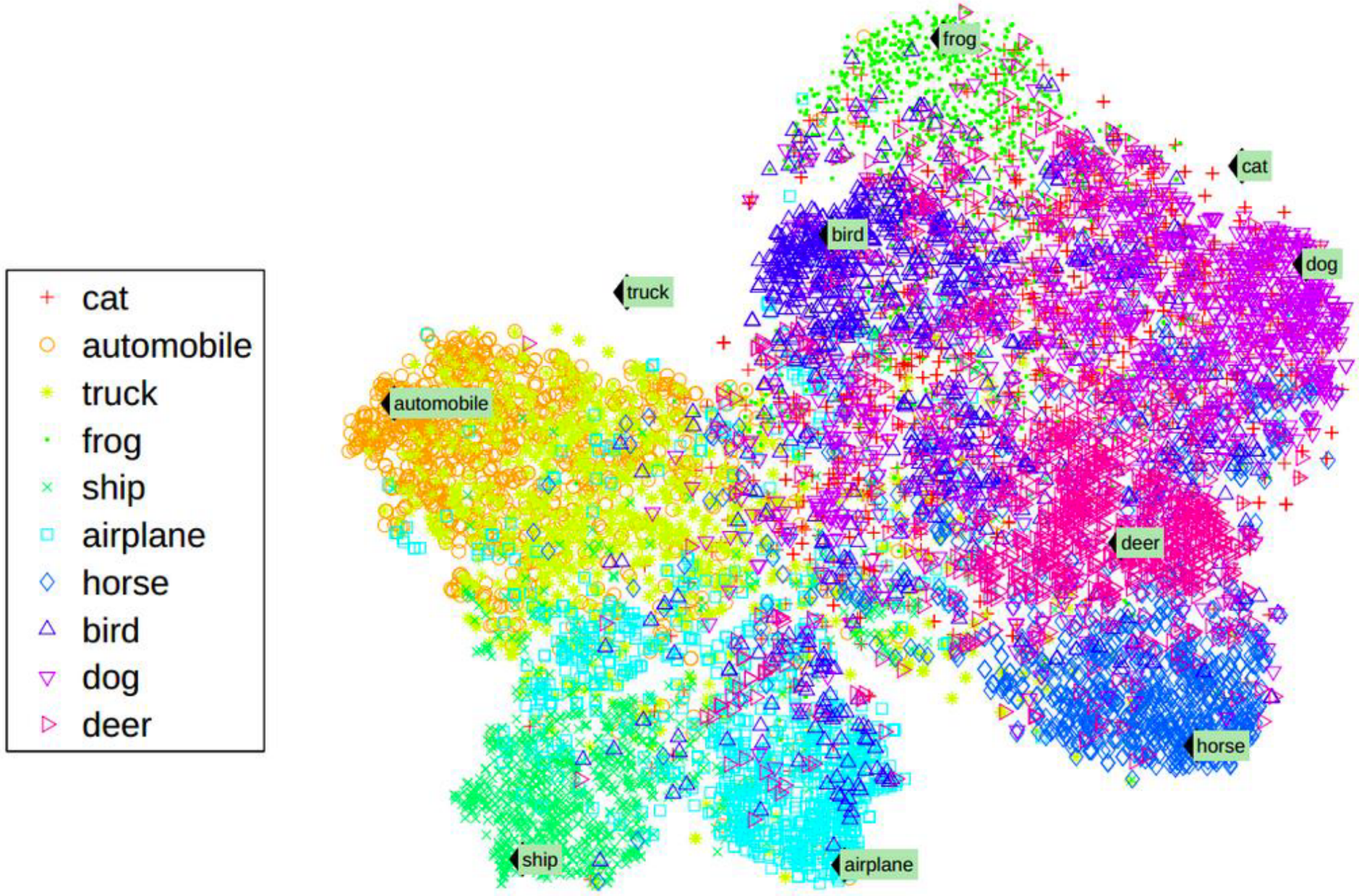
So, what do embeddings help?

- ✦ Word Relations: in multi-modal space



So, what do embeddings help?

- ✦ Word Relations: in multi-modal space[13]



Language is Important

Why? How?

Language is important and we have tools

- *“The use of word representations… has become a key “secret sauce” for the success of many NLP systems in recent years, across tasks including named entity recognition, part-of-speech tagging, parsing, and semantic role labeling.”*

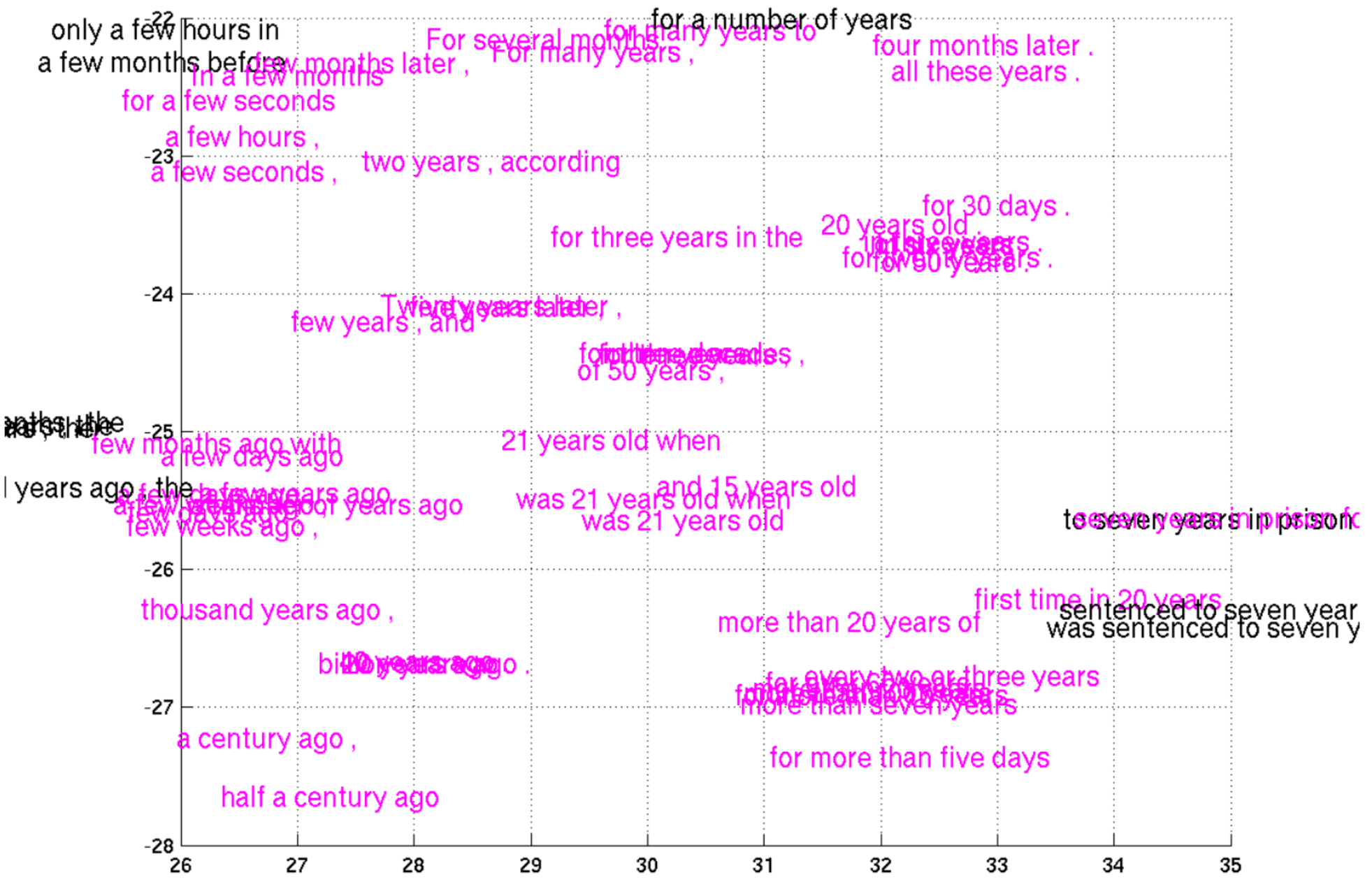
- Language takes as a key and complementary role for us to understand the world.

Word representations

What's next?

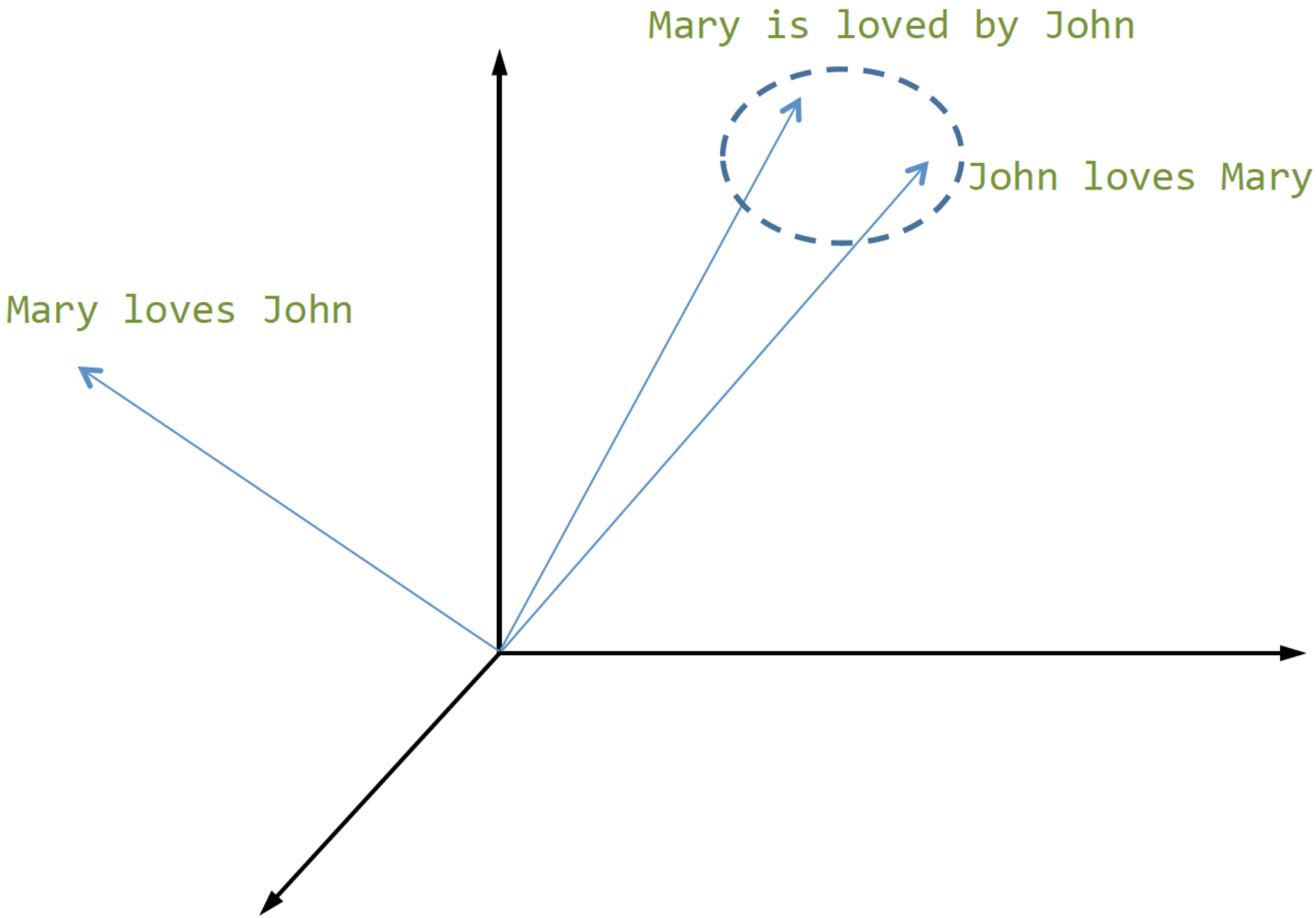
We still have make progress on

- Phrase Representations[14]



We still have make progress on

- Sentence Representations

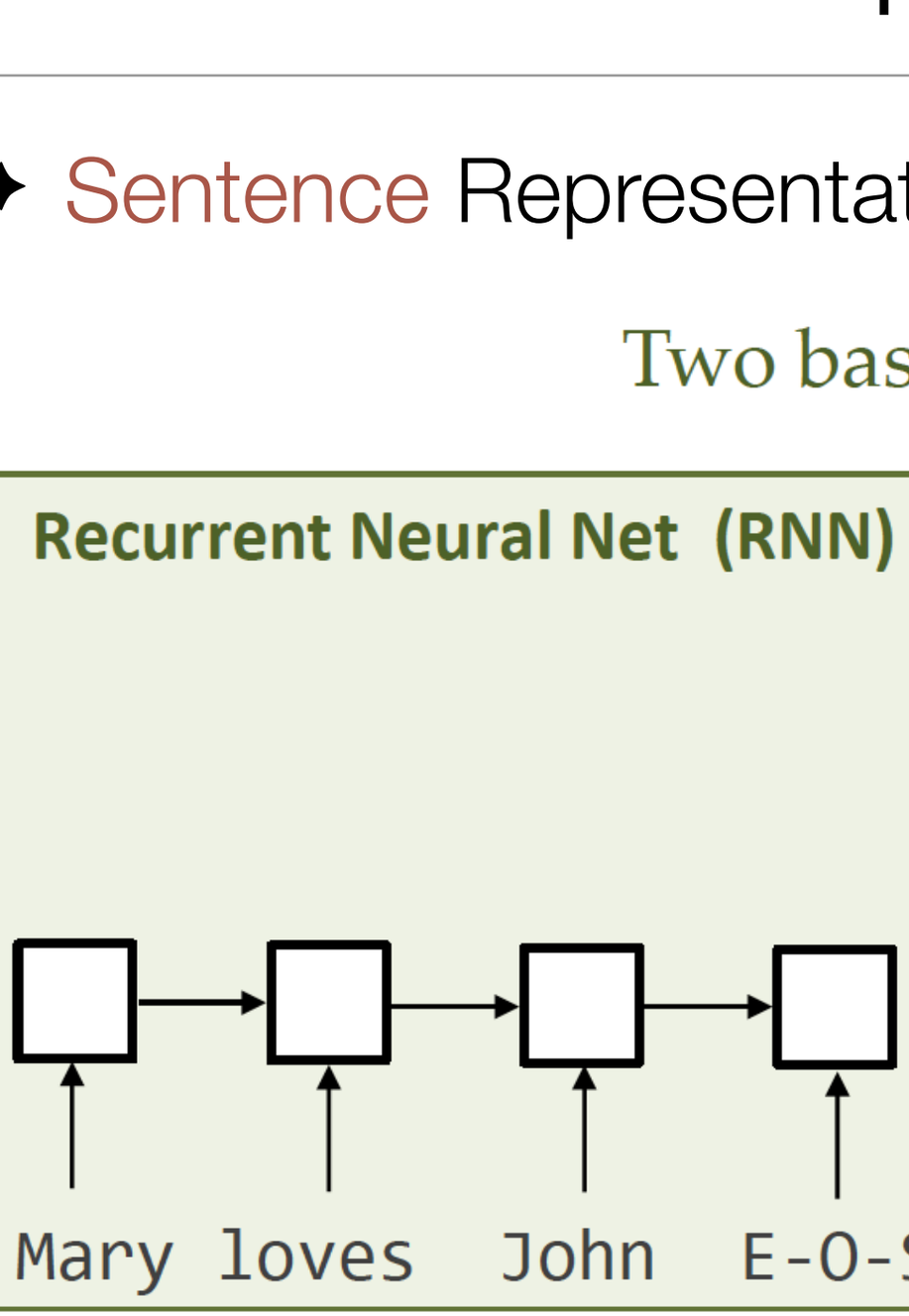


We still have make progress on

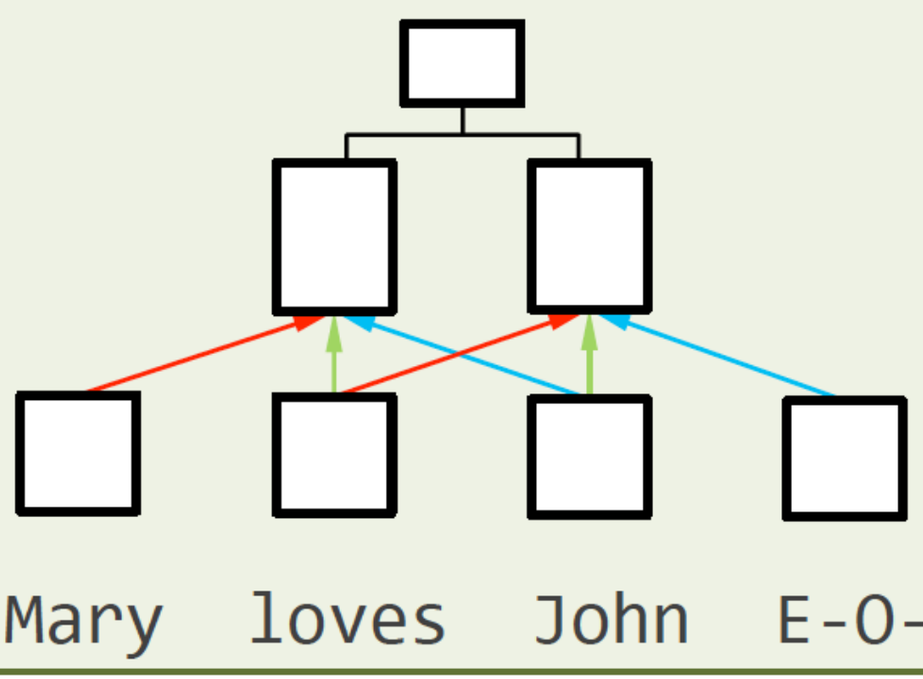
- ✦ Sentence Representations by

Two basic architectures

Recurrent Neural Net (RNN)

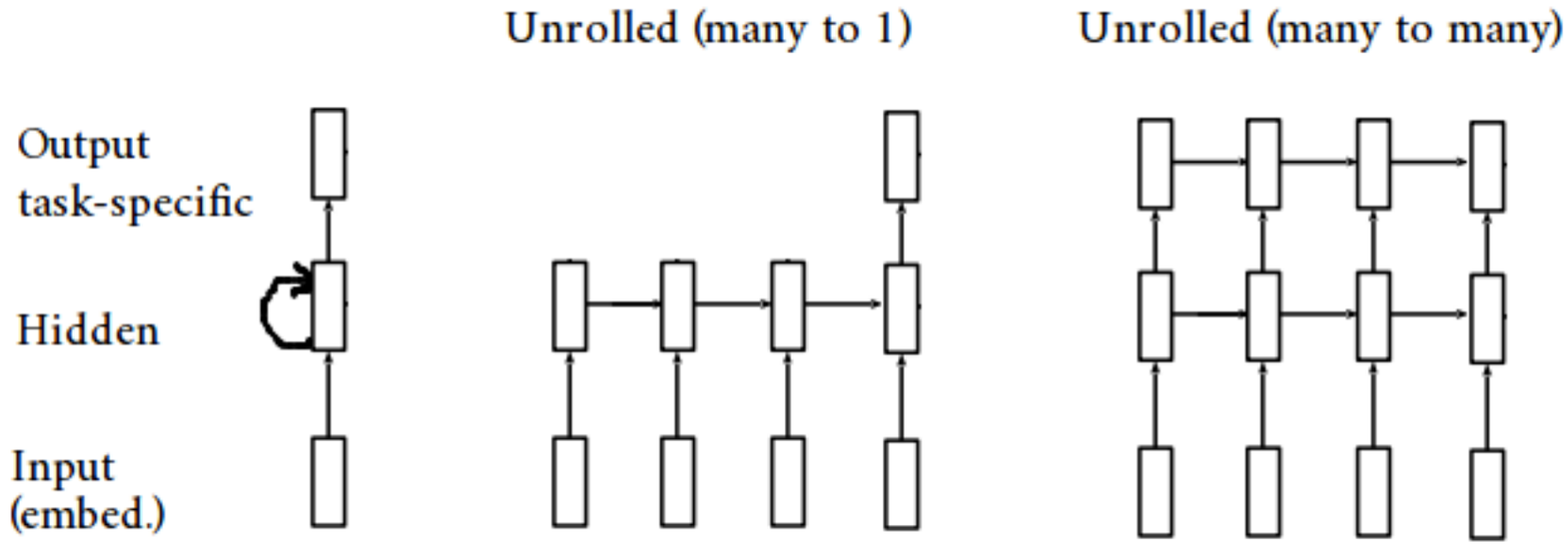


Convolutional Neural Net (CNN)

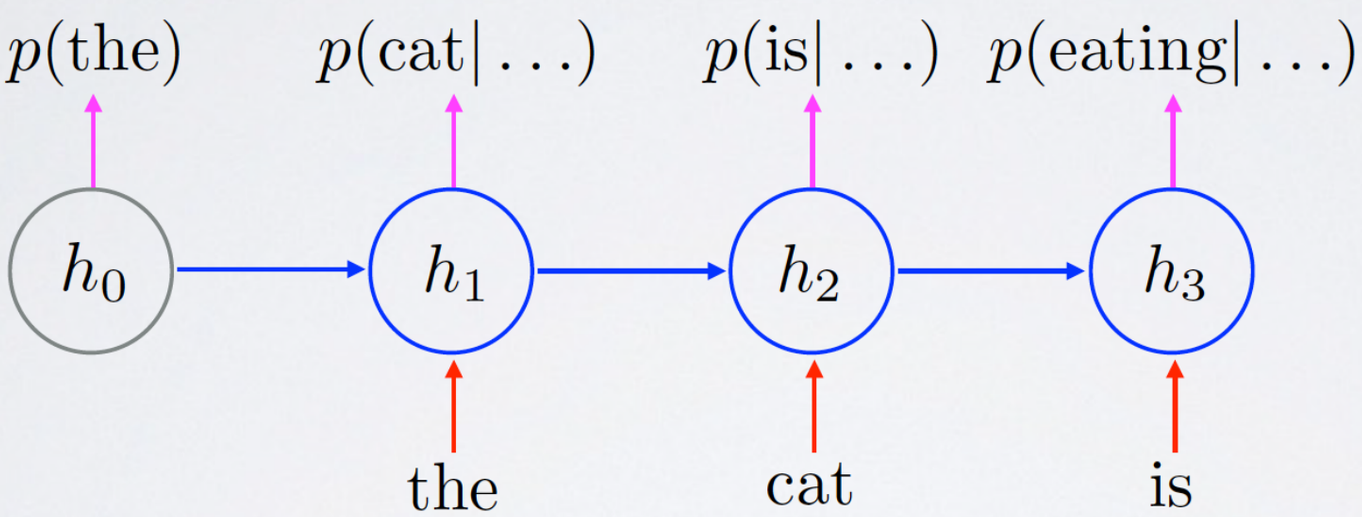


Recurrent Neural Networks (RNN)

- Sentence as sequence of words
- Recursively

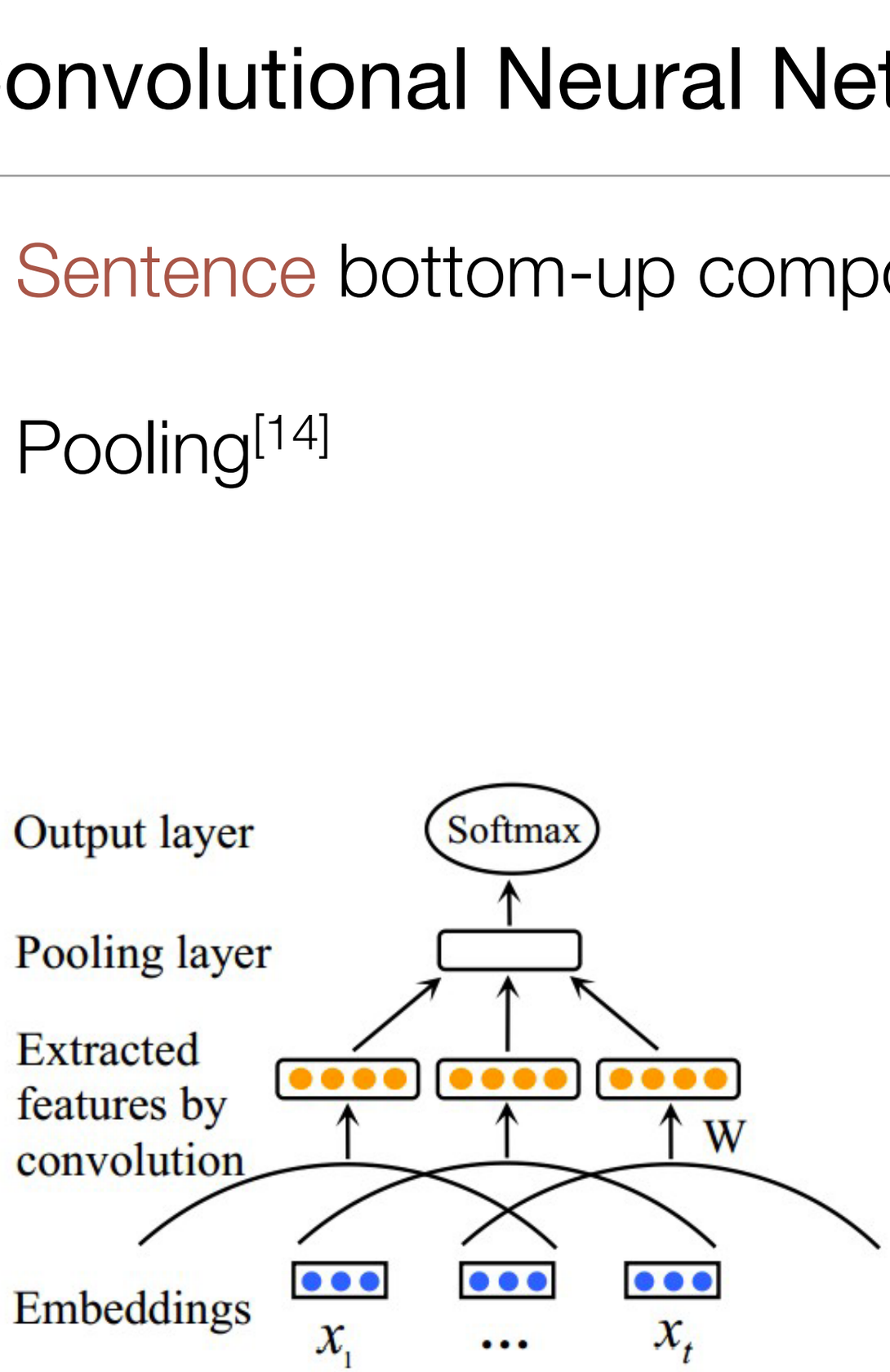


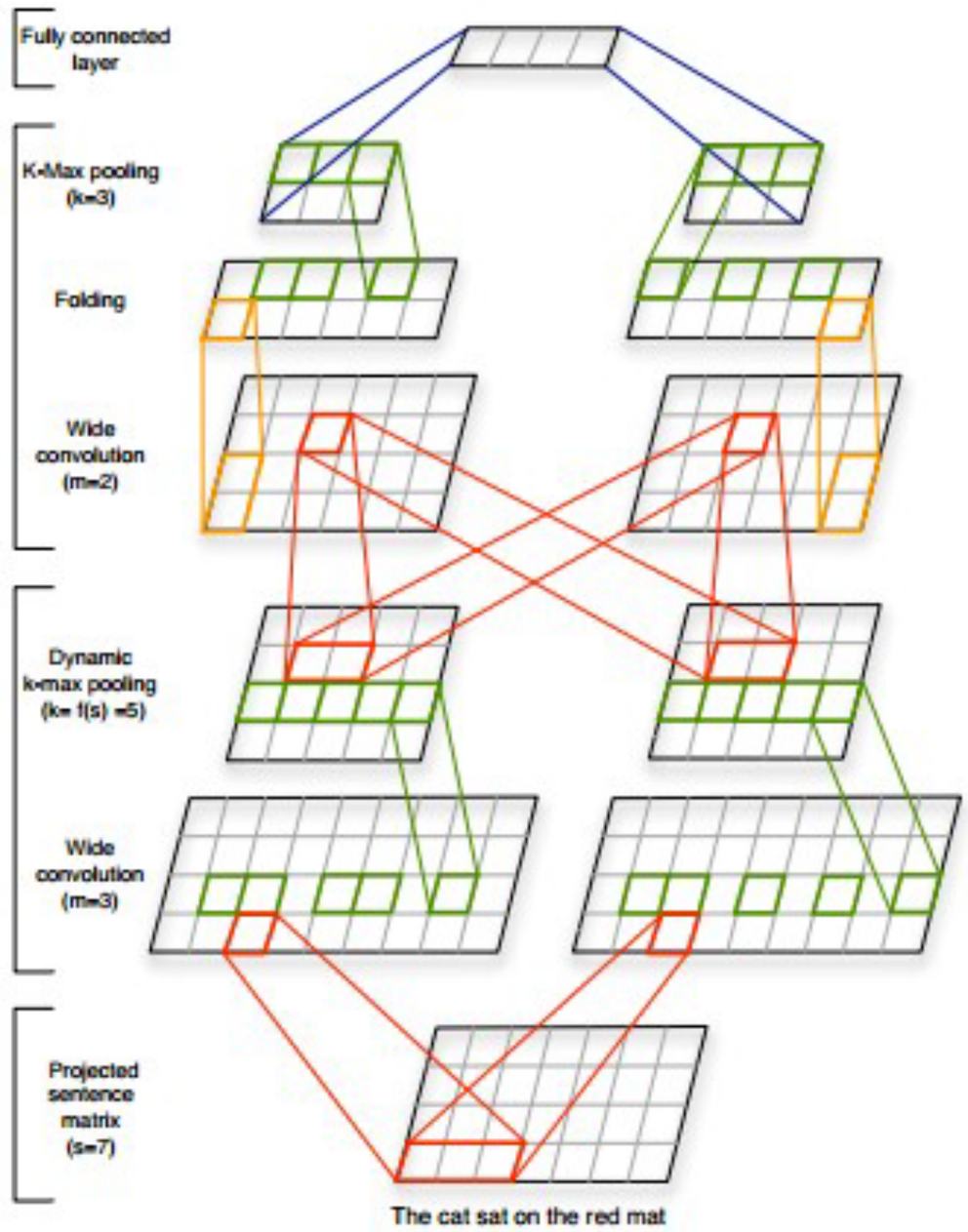
Example: $p(\text{the}, \text{cat}, \text{is}, \text{eating})$



Convolutional Neural Networks (CNN)

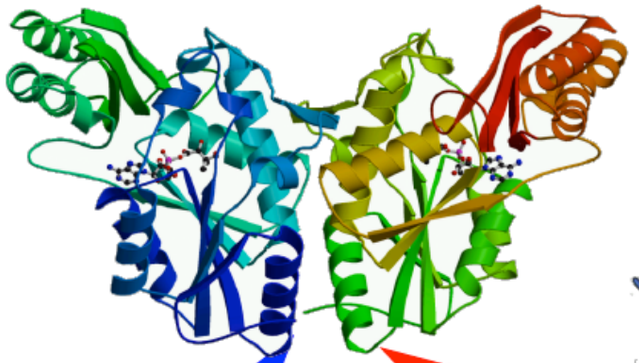
- Sentence bottom-up compositional structures
- Pooling[14]





Despite these advantages and promising results, there is a major limitation in NMT compared to the existing phrase-based approach

Despite these advantages and promising results, there is a major limitation in NMT compared to the existing phrase-based approach. That is, the number of target words must be limited. This is mainly because the complexity of training and using an NMT model increases as the number of target words increases. Also, the parametric nature of the neural machine translation makes it difficult for the model to estimate the conditional probabilities of rare words well.



On Using Very Large Target Vocabulary for Neural Machine Translation

Sébastien Jean Université de Montréal
Kyunghyun Cho Université de Montréal
Roland Memisevic Université de Montréal
Yoshua Bengio Université de Montréal CIFAR Senior Fellow

Abstract

Neural machine translation, a recently proposed approach to machine translation based purely on neural networks, has shown promising results compared to the existing approaches such as phrase-based statistical machine translation. Despite its recent success, neural machine translation has its limitation in handling a larger vocabulary, as training complexity as well as decoding complexity increase proportionally to the number of target words. In this paper, we propose a method based on importance sampling that allows us to use a very large target vocabulary without increasing training complexity. We show that decoding can be efficiently done even with the model having a very large target vocabulary by selecting only a small subset of the whole target vocabulary. The models trained by the proposed approach are empirically found to outperform the baseline models with a small vocabulary as well as the LSTM-based neural machine translation models. Furthermore, when we use the ensemble of a few models with very large target vocabularies, we achieve the state-of-the-art translation performance (measured by BLEU) on the English→German translation and almost as high performance as state-of-the-art English→French translation system.

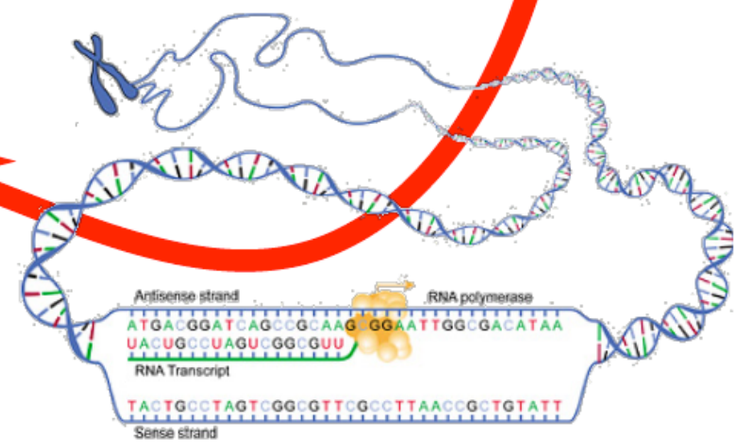
1 Introduction

Neural machine translation (NMT) is a recently introduced approach to solving machine translation (Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2014; Sutskever et al., 2014). In neural machine translation, one builds a single neural network that reads a source sentence and generates its translation. The whole neural network is jointly trained to maximize the conditional probability of a correct translation given a source sentence, using the bilingual corpus. The NMT models have shown to perform as well as the most widely used conventional translation systems (Sutskever et al., 2014; Bahdanau et al., 2014).

Neural machine translation has a number of advantages over the existing statistical machine translation system, specifically, the phrase-based system (Koehn et al., 2003). First, NMT requires a minimal set of domain knowledge. For instance, all of the models proposed in (Sutskever et al., 2014), (Bahdanau et al., 2014) or (Kalchbrenner and Blunsom, 2013) do not assume any linguistic property in both source and target sentences except that they are sequences of words. Second, the whole system is jointly tuned to maximize the translation performance, unlike the existing phrase-based system which consists of many feature functions that are tuned separately. Lastly, the memory footprint of the NMT model is often much smaller than the existing system which relies on maintaining large tables of phrase pairs.

Despite these advantages and promising results, there is a major limitation in NMT compared to the existing phrase-based approach. That is, the number of target words must be limited. This is mainly because the complexity of training and using an NMT model increases as the number of target words increases. Also, the parametric nature of the neural machine translation makes it difficult for the model to estimate the conditional probabilities of *rare* words well.

A usual practice is to construct a target vocabulary of the k most frequent words (a so-called shortlist), where k is often in the range of 30, 000 (Bahdanau et al., 2014) to 80, 000 (Sutskever et al., 2014). Any word not included in this vocabulary is mapped to a special token representing an *unknown* word [UNK]. This approach works well when there are only a few unknown words in the target sentence, but it has been observed that the translation performance degrades rapidly as the number of unknown words increases (Cho et al., 2014a; Bahdanau et al., 2014).



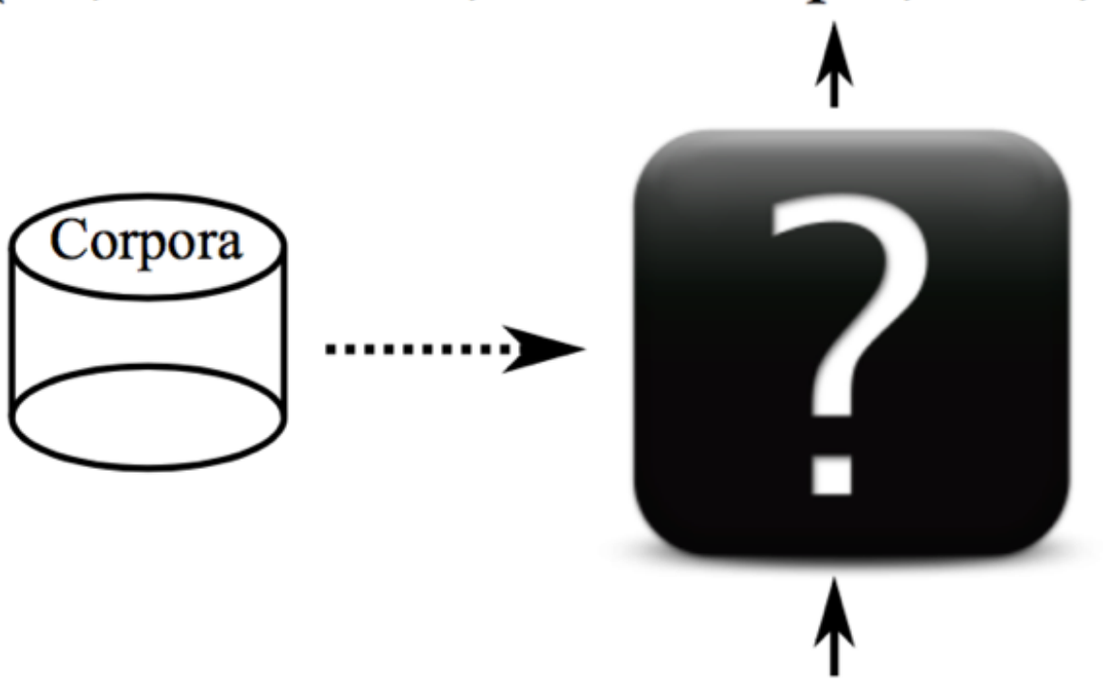
Sentence representation

What can we do?

Machine Translation

- It is a kind of communication on sentence level.

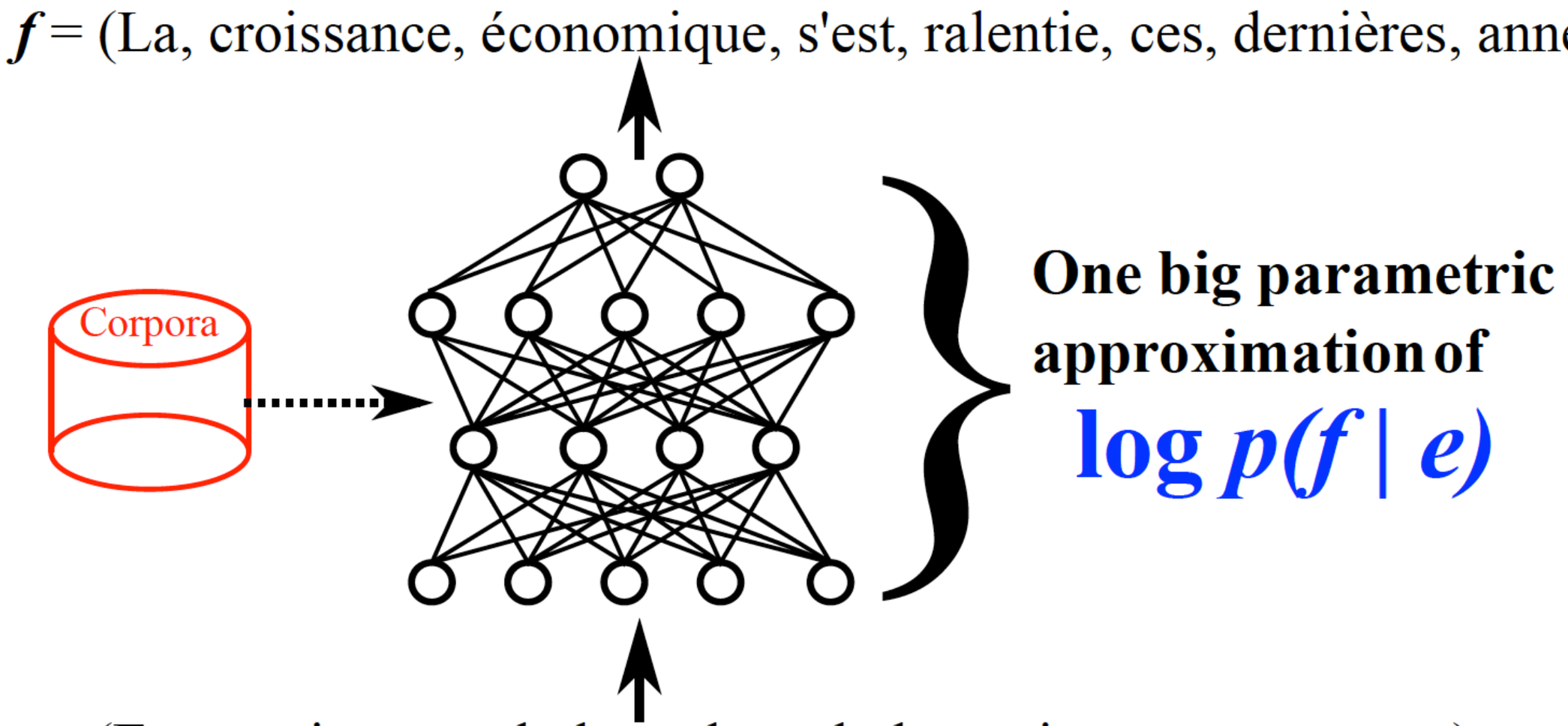
$\boldsymbol{f}$ = (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)



$\boldsymbol{e}$ = (Economic, growth, has, slowed, down, in, recent, years, .)

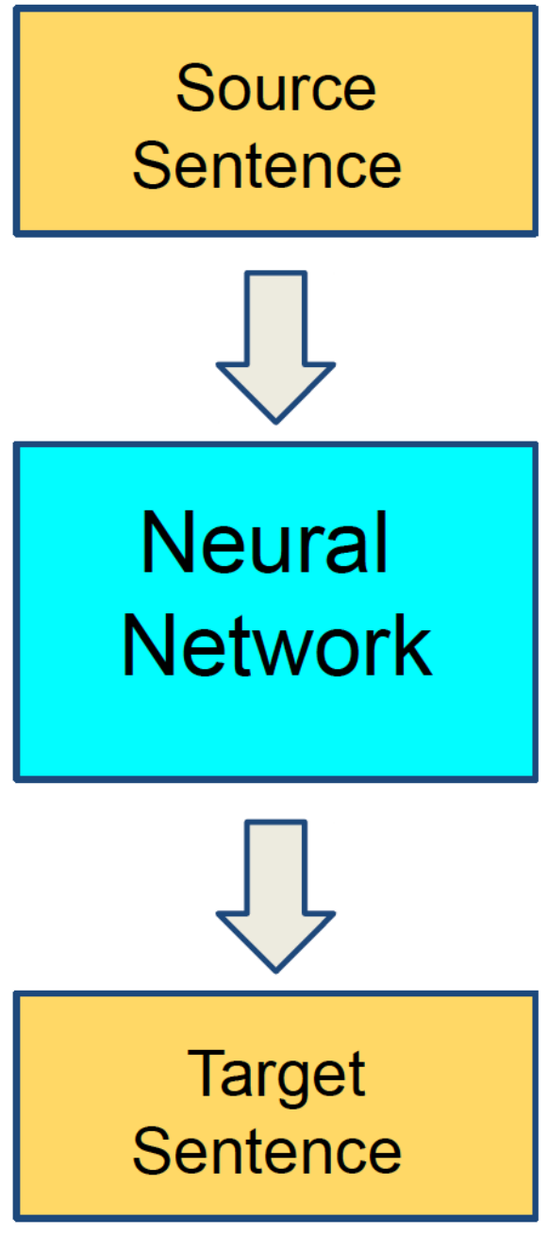
Neural Machine Translation

- Replace the whole MT system with a single neural network.



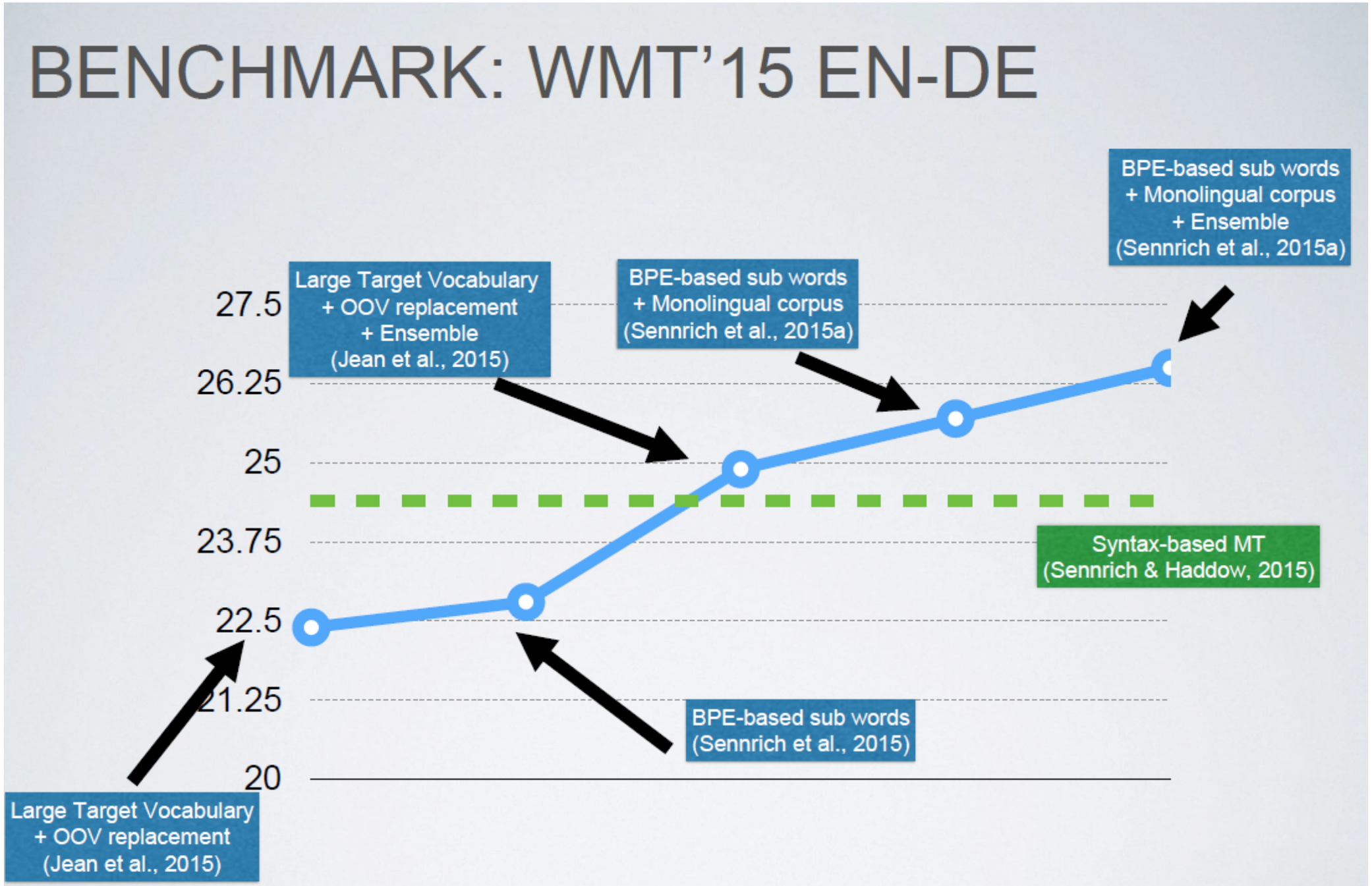
Neural Machine Translation

- Pros
  - No more manual feature engineering
  - End-to-End: Every subcomponent of a network is tuned explicitly to maximize translation quality



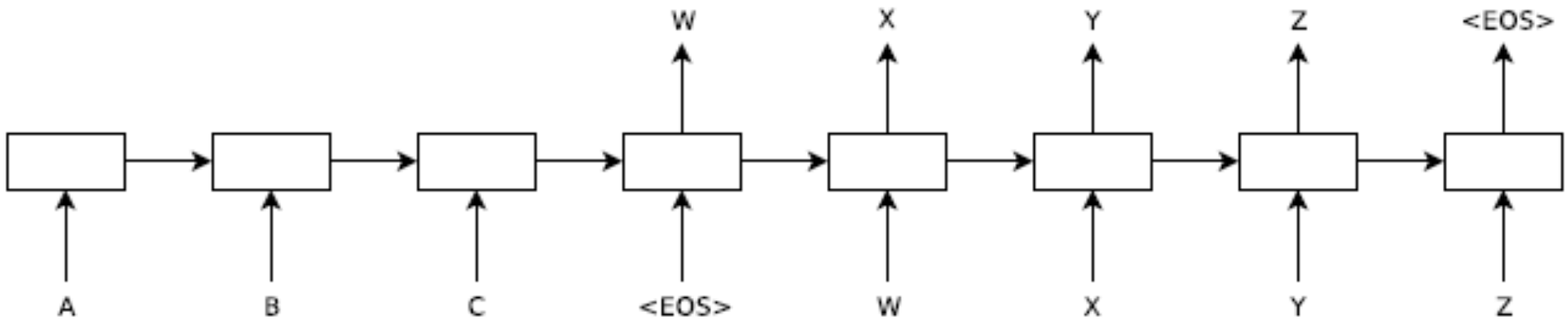
Neural MT

Neural Machine Translation



Sequence to sequence learning with

- ✦ The encoder-decoder framework

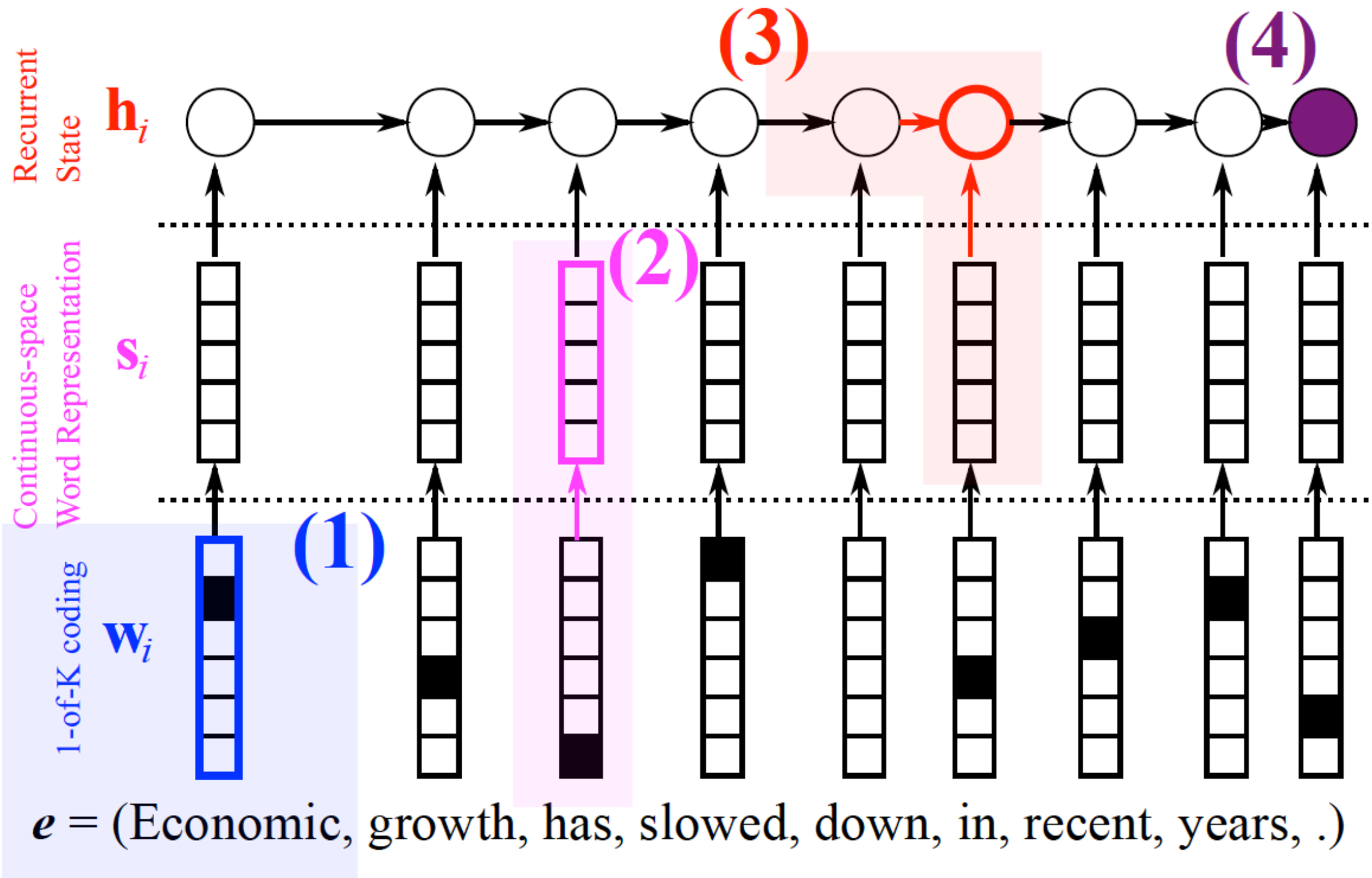


- ✦ Encoder reads sentence “A B C” in source language
- ✦ Decoder generates sentence “W X Y Z” in target language

The encoder-decoder framework

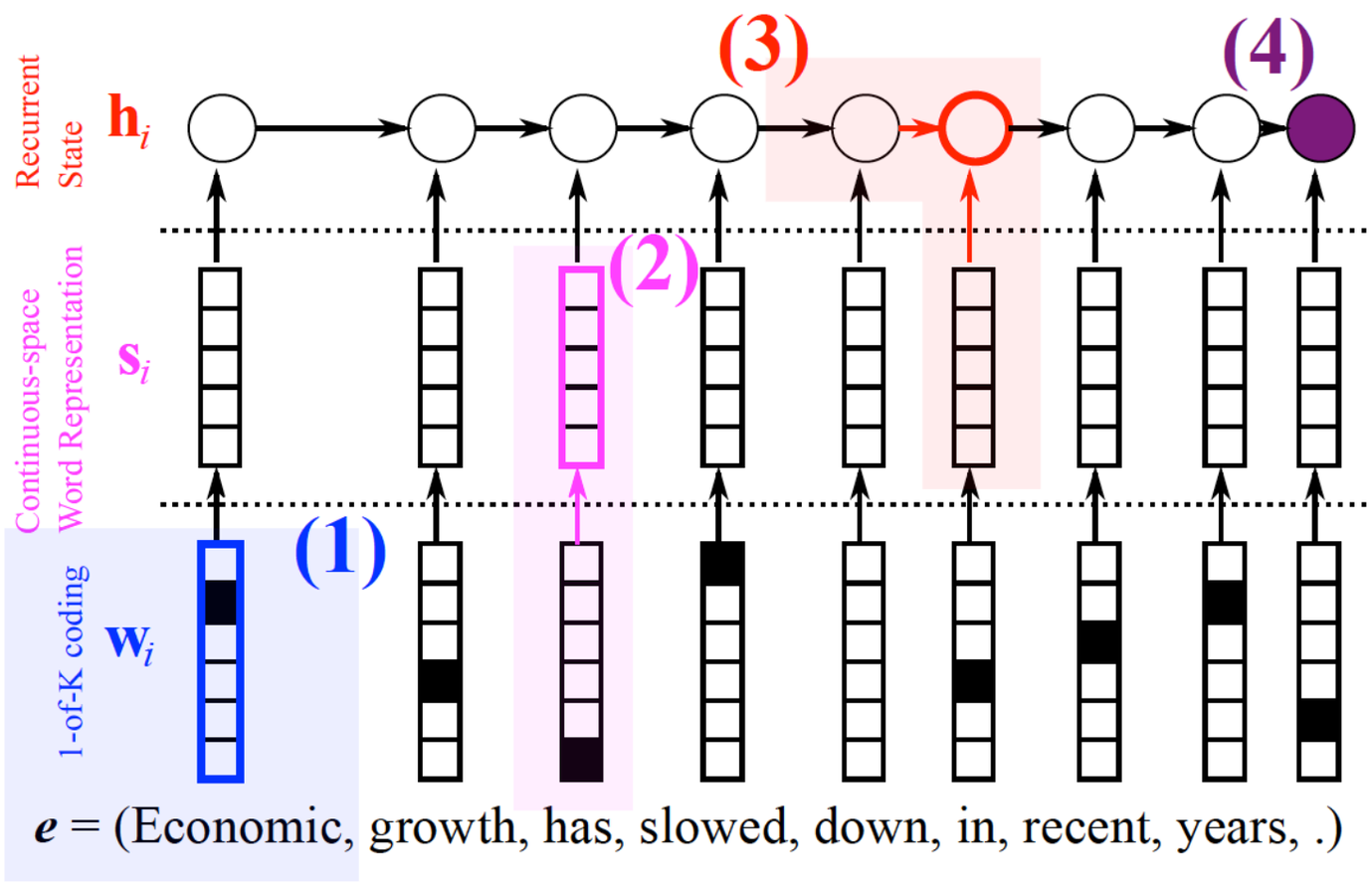
- Comprehensive
  - It unites content understanding and language modelling in a single architecture.
  - Thus, it allows the end-to-end automatic generation.
- Flexible
  - It allows to condition on anything, i.e. images, videos, words, etc.
- Scalable
  - Able to supplement the language model with external corpora.

The encoder-decoder framework — Encoder

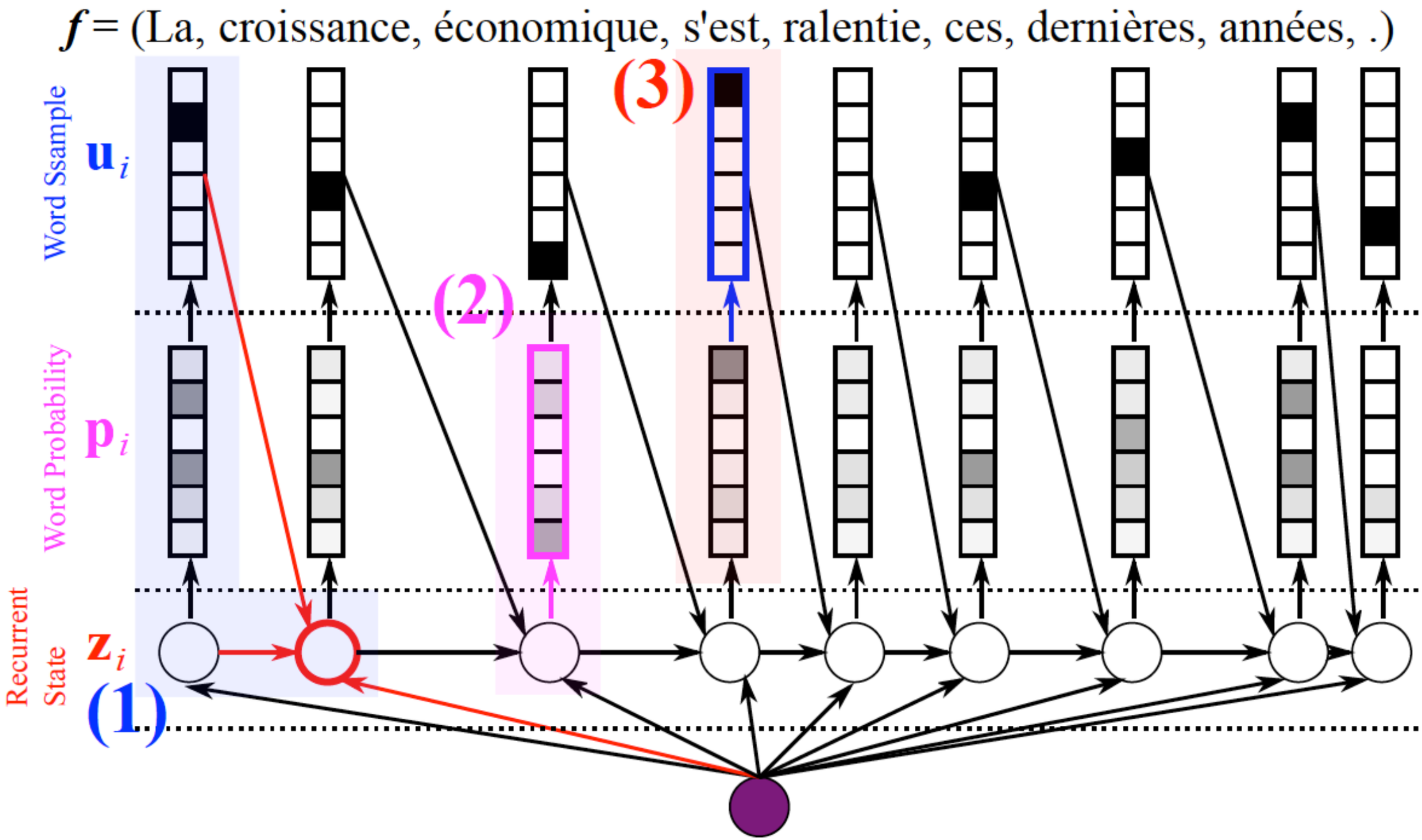


The encoder-decoder framework — Encoder

- ✦ 1-of-K coding of source words
- ✦ Continuous-space representation
- ✦ Recursively read words

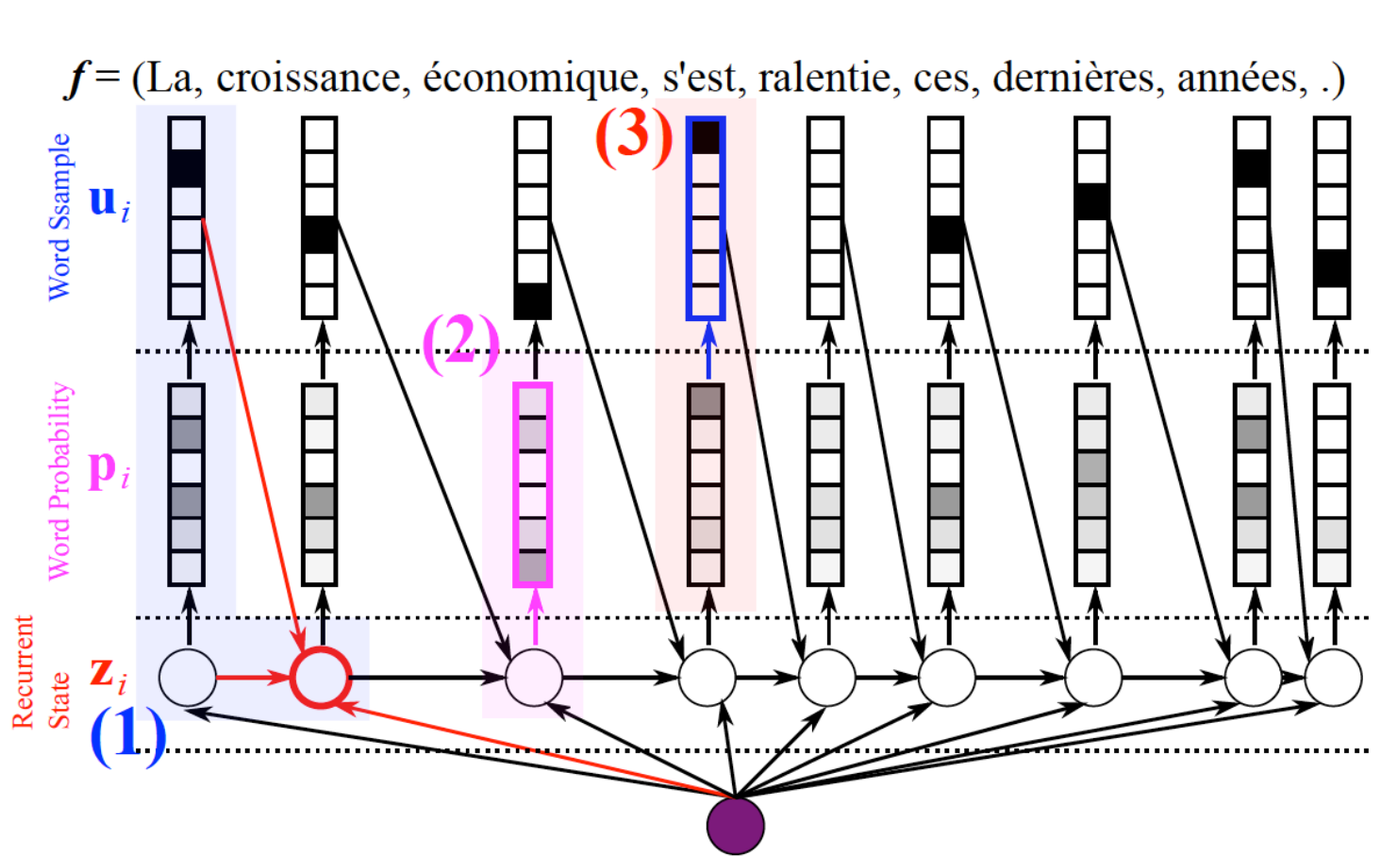


The encoder-decoder framework — Decoder



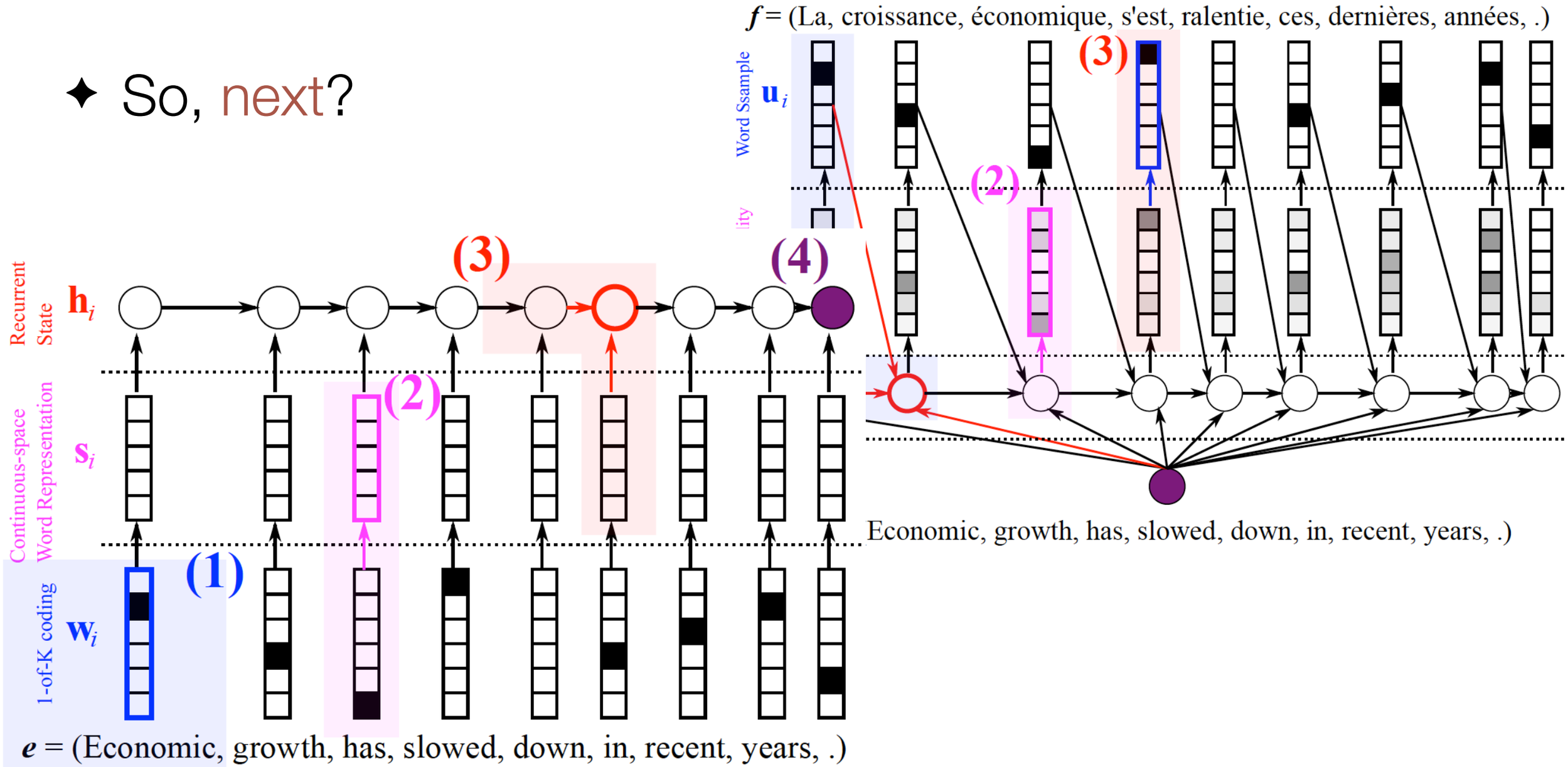
The encoder-decoder framework — Decoder

- ✦ Recursively update the memory
- ✦ Compute the next word probability
- ✦ Sample a next word



It looks fantastic, but…

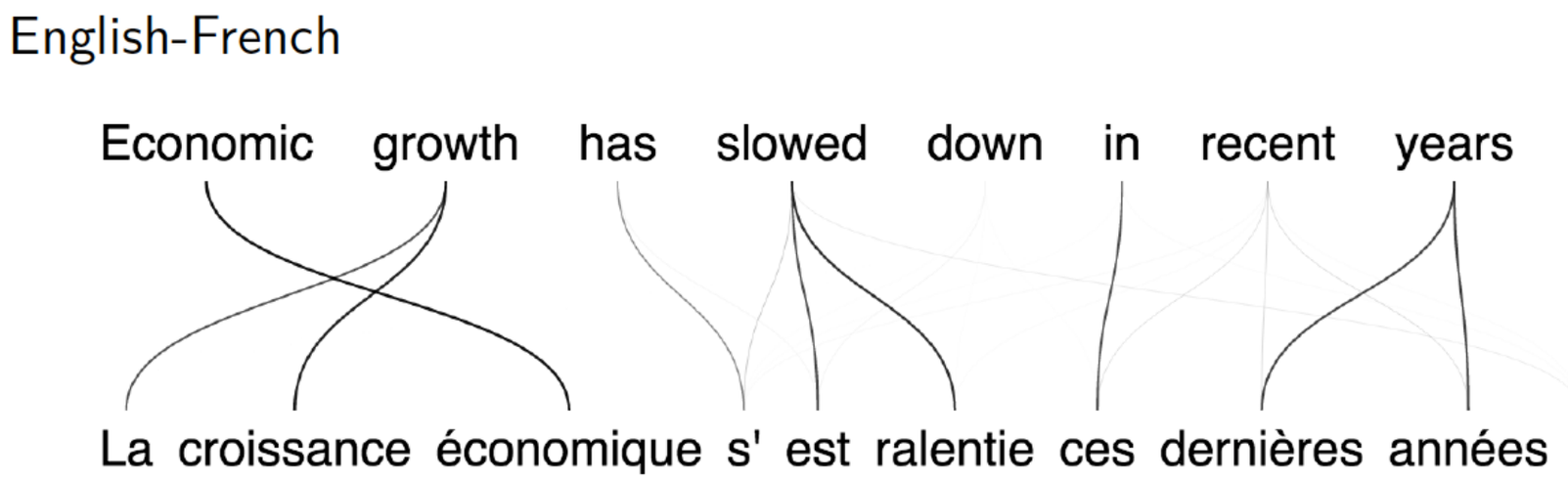
- ✦ “You can’t cram the meaning of a whole %&!$# sentence into a single $&!#* vector!”
- ✦ So, next?



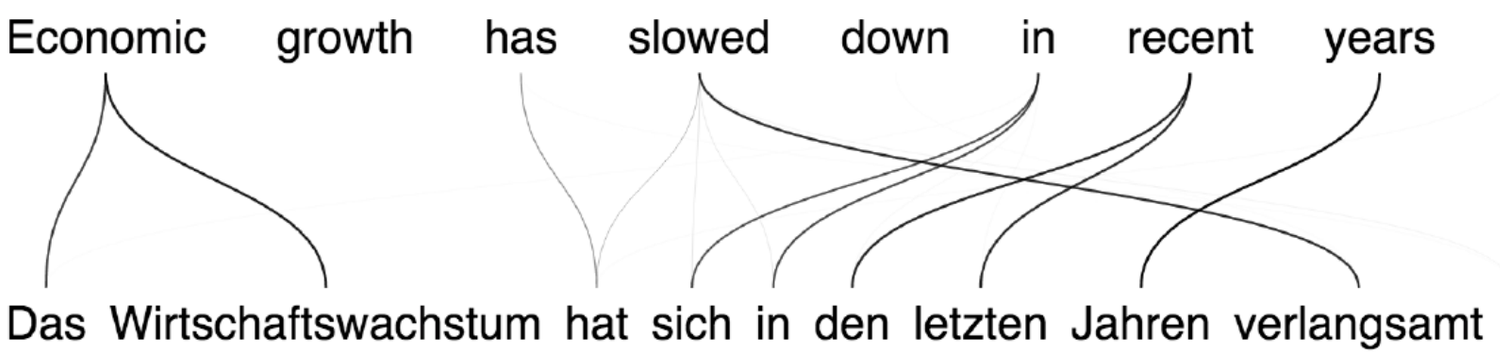
Attention-based Neural Machine Translation

- ✦ fixed representation of source sentence —> soft and dynamic[15]

English-French

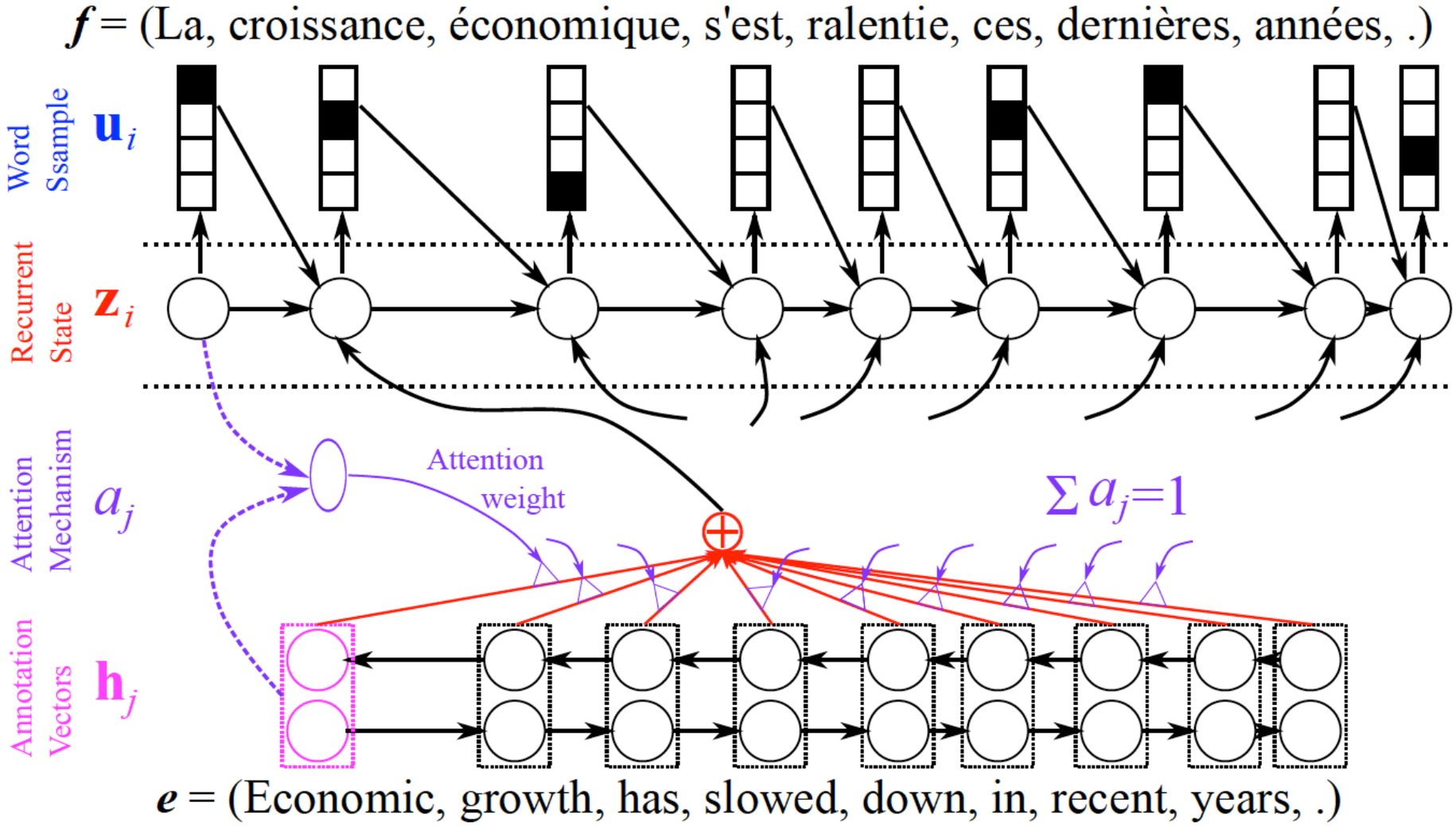


English-German



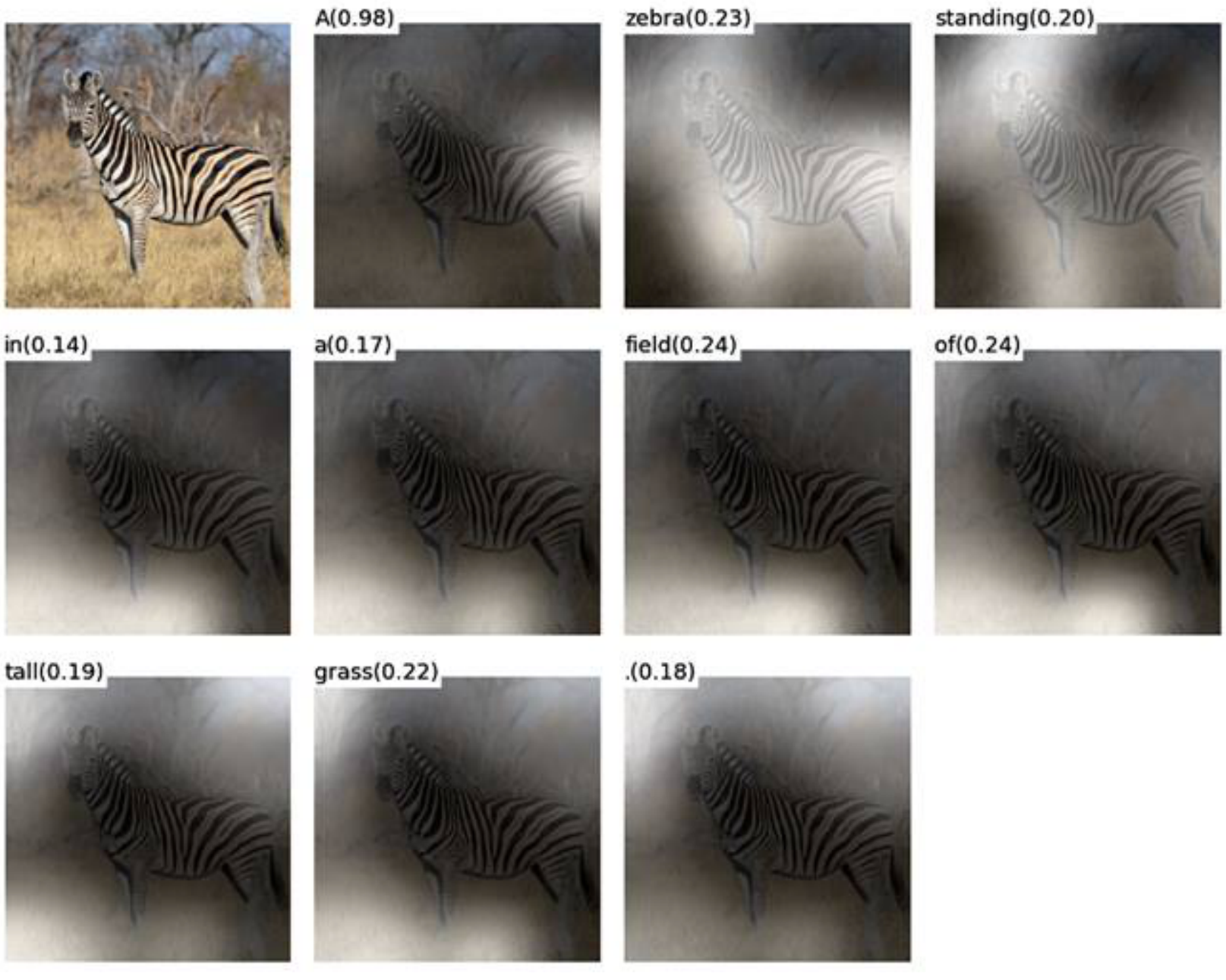
Attention-based Neural Machine Translation

- ✦ fixed representation of source sentence —> soft and dynamic[15]



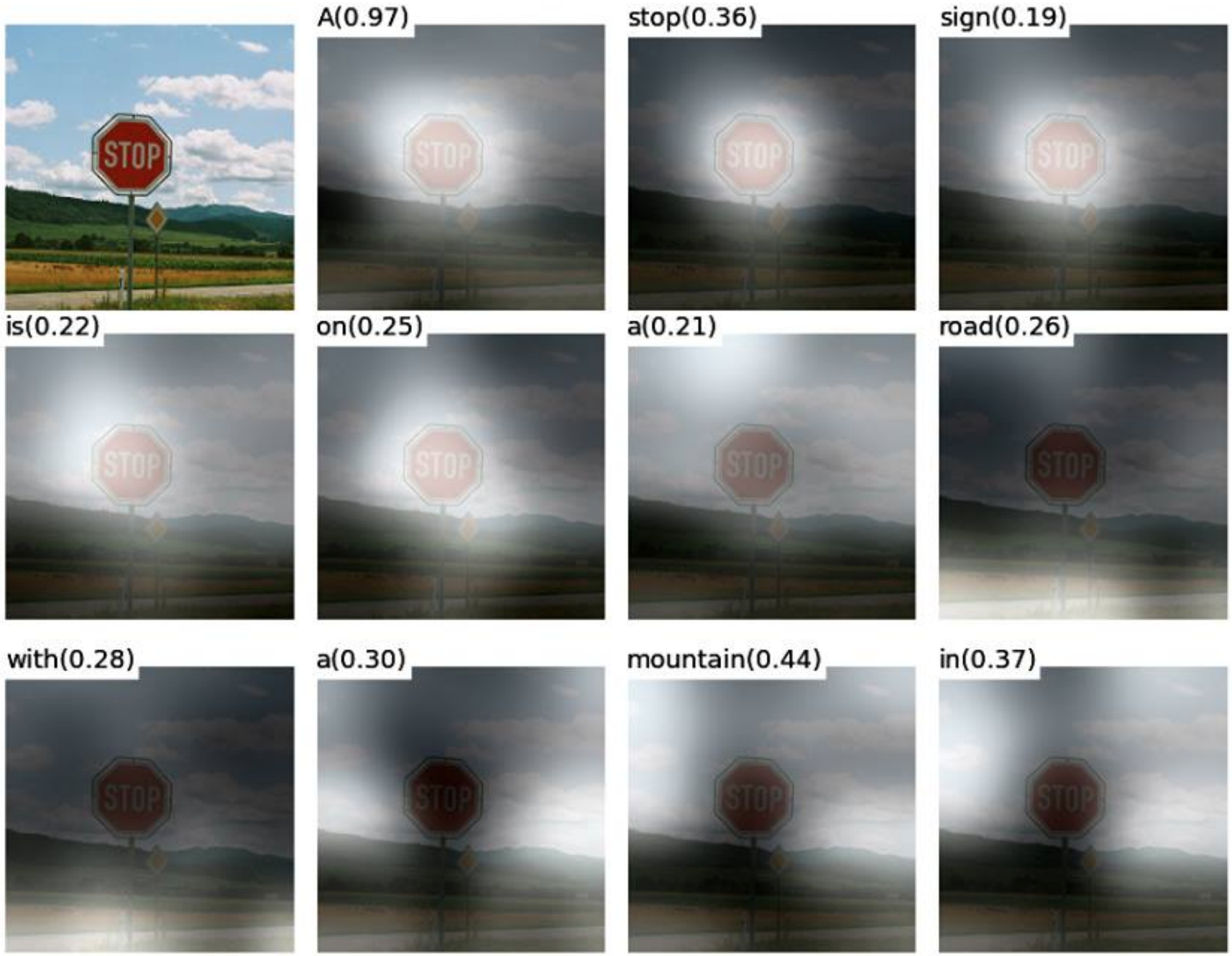
Attention is everywhere

- ✦ Going beyond language: image



Attention is everywhere

- ✦ Going beyond language: image

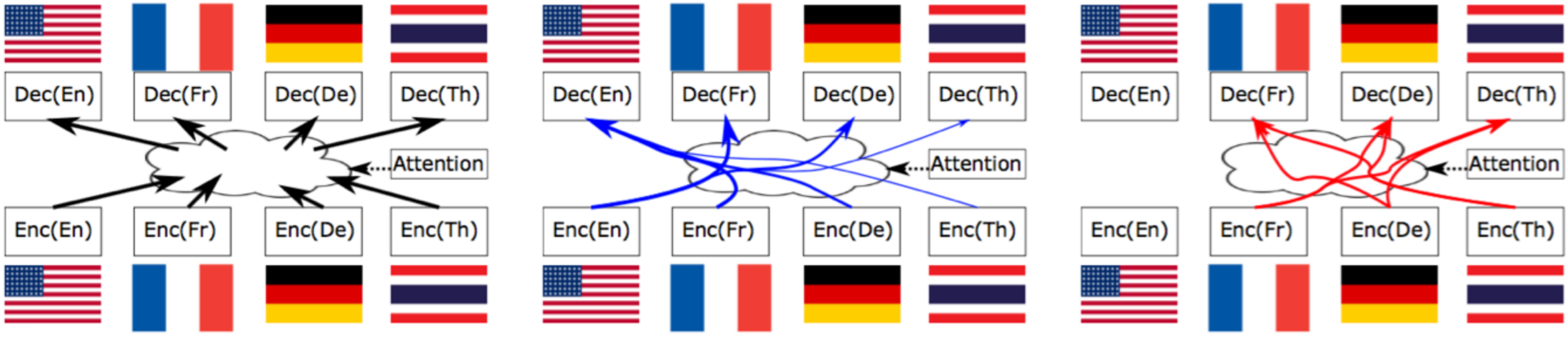


Attention is everywhere

- ✦ Going beyond language: video[17]

Multi-lingual Neural Machine Translation

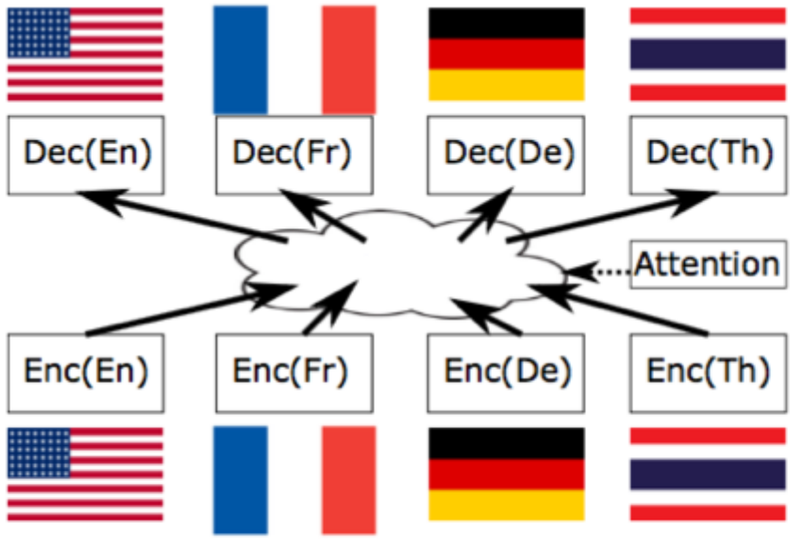
- ✦ Language Transfer



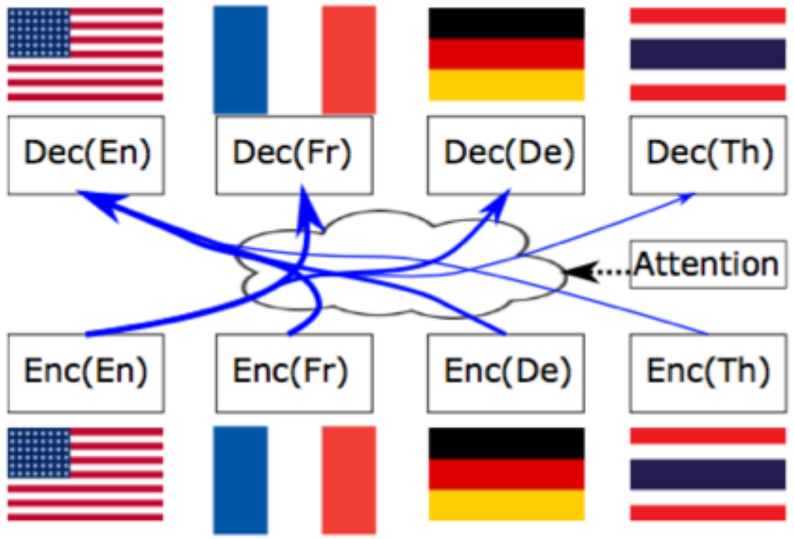
(a) Multilingual Neural MT (b) Available Parallel Corpora (c) Zero-Resource Language Pairs

Multi-lingual Neural Machine Translation

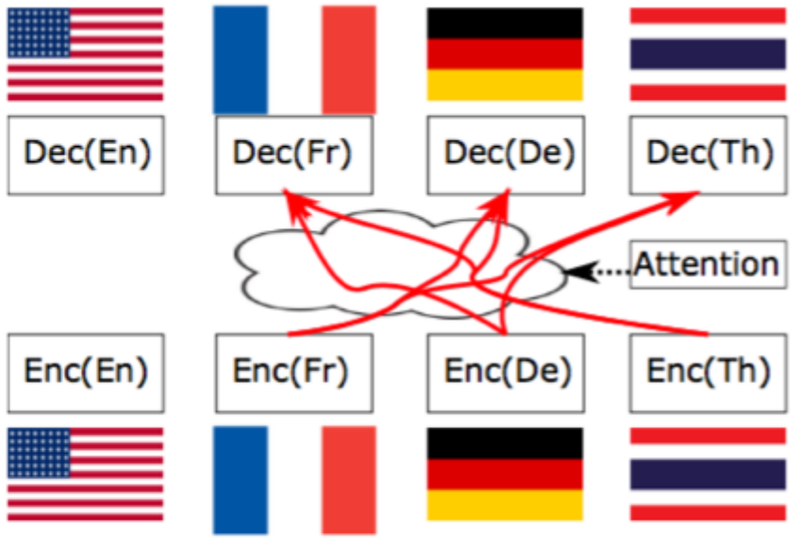
- Language Transfer
  - specific or universal?
  - attention?



(a) Multilingual Neural MT



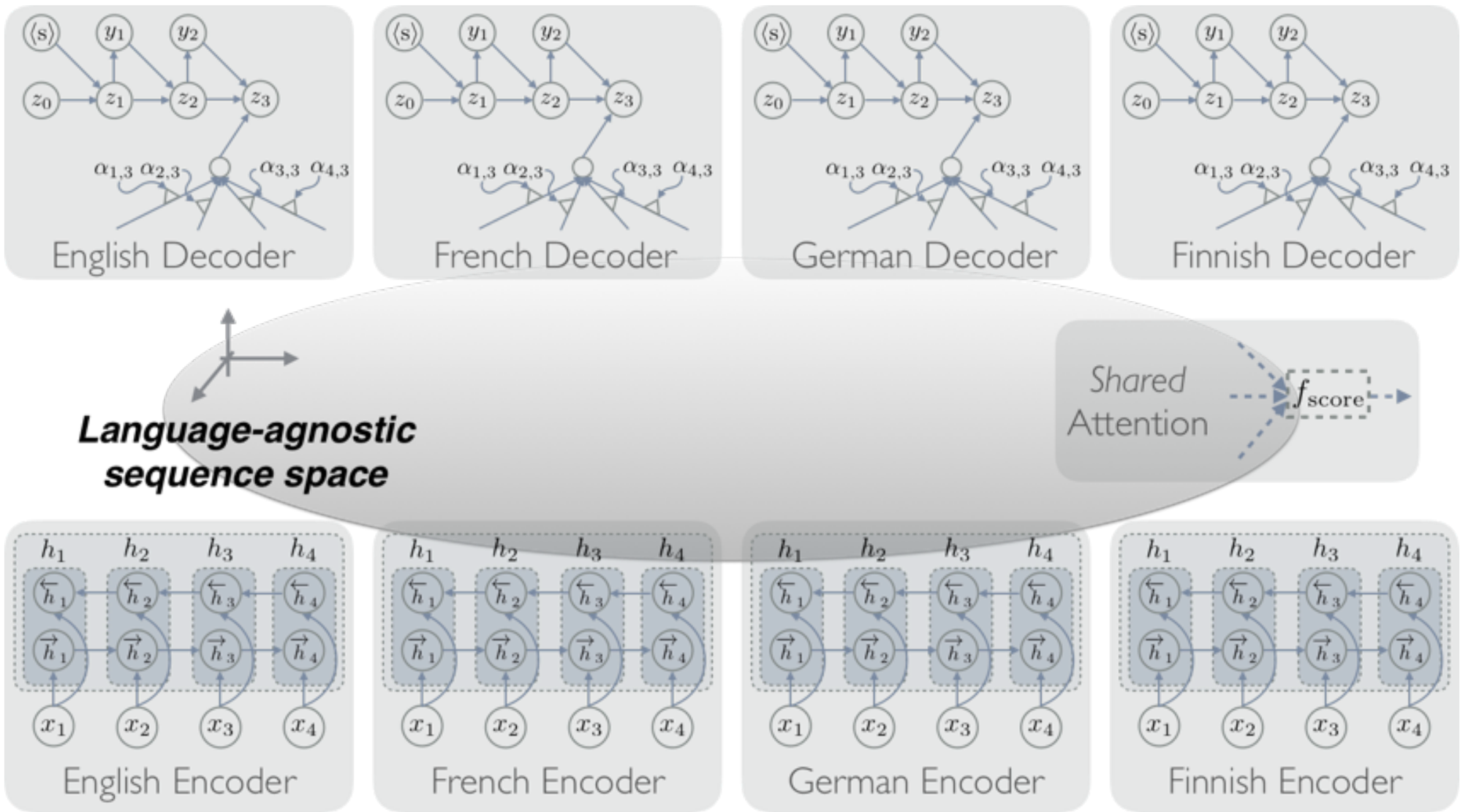
(b) Available Parallel Corpora



(c) Zero-Resource Language Pairs

Language Transfer

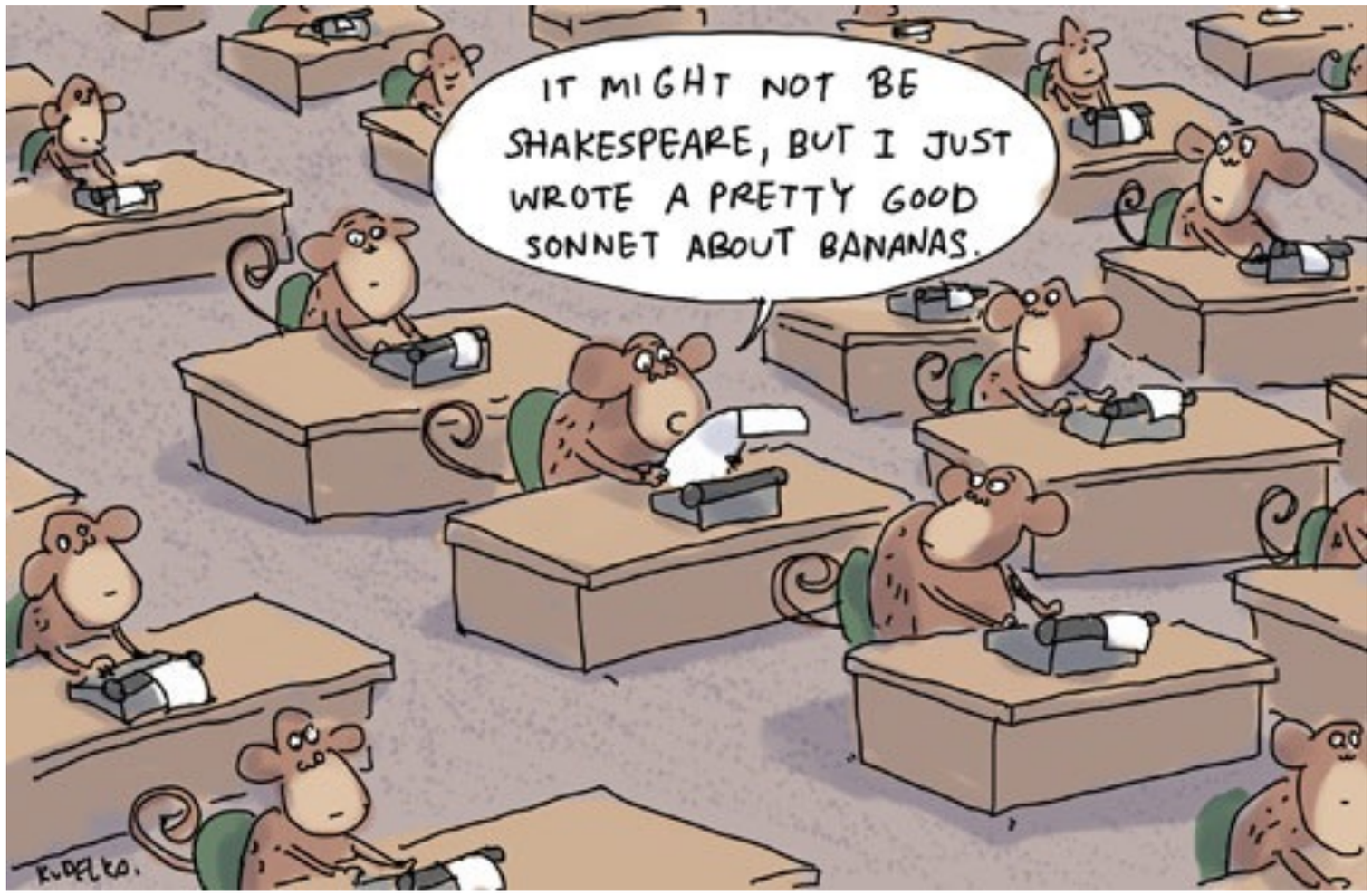
- Is the attention mechanism universal or specific to a language pair?[16]



Universal Attention Mechanism

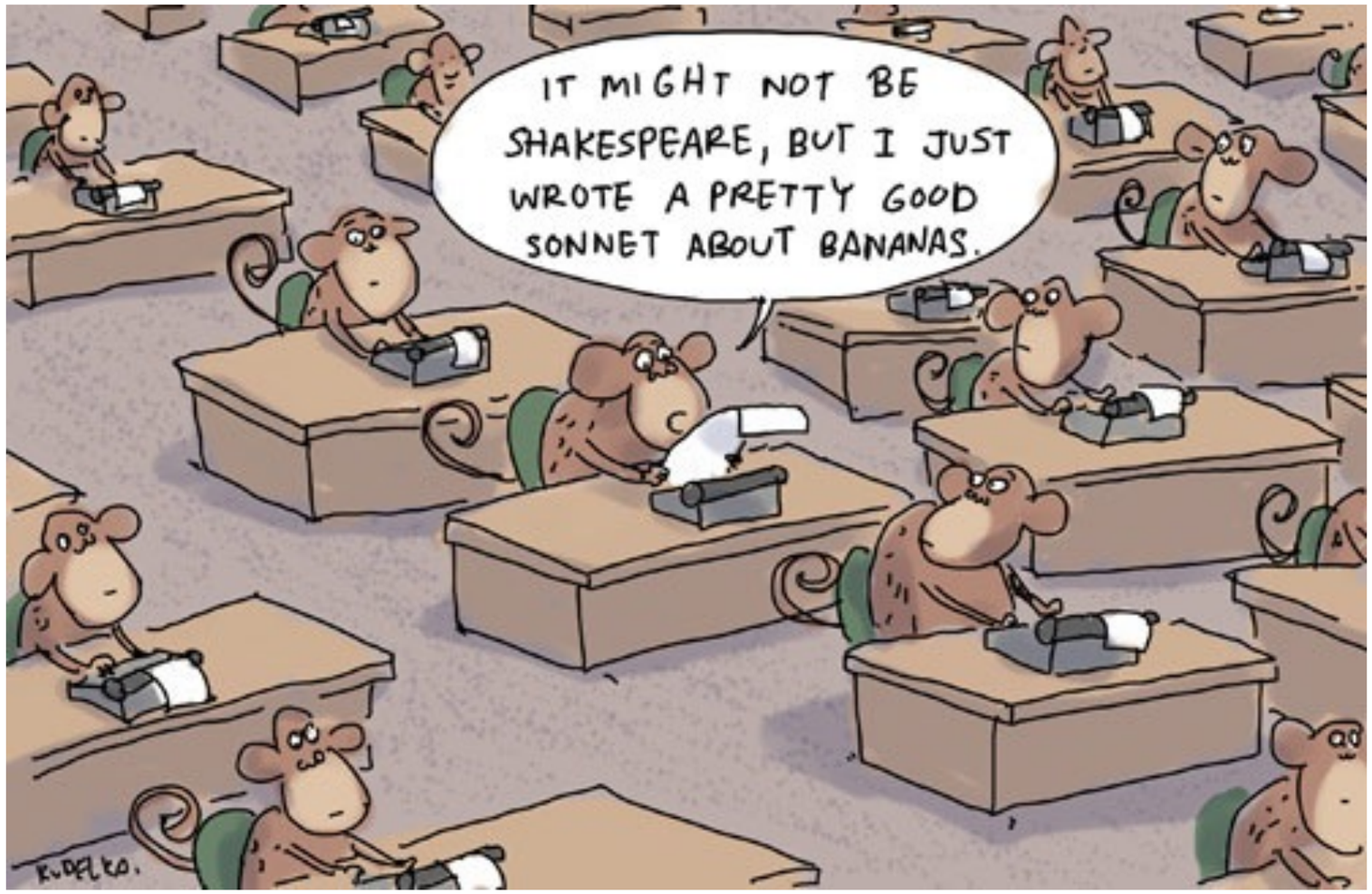
- ✦ Is the attention mechanism universal or specific to a language pair?[16] —> Yes!
  - ✦ Especially, effective for low-resource language pairs.

	Size	Single	Single+DF	Multi
En→Fi	100k	5.06/3.96	4.98/3.99	6.2/**5.17**
	200k	7.1/6.16	7.21/6.17	8.84/**7.53**
	400k	9.11/7.85	9.31/8.18	11.09/**9.98**
	800k	11.08/9.96	11.59/10.15	12.73/**11.28**
De→En	210k	14.27/13.2	14.65/13.88	16.96/**16.26**
	420k	18.32/17.32	18.51/17.62	19.81/**19.63**
	840k	21/19.93	21.69/20.75	22.17/**21.93**
	1.68m	23.38/23.01	23.33/22.86	23.86/**23.52**
En→De	210k	11.44/11.57	11.71/11.16	12.63/**12.68**
	420k	14.28/14.25	14.88/15.05	15.01/**15.67**
	840k	17.09/17.44	17.21/17.88	17.33/**18.14**
	1.68m	19.09/19.6	19.36/20.13	19.23/**20.59**



Do we understand?

Generation



Monkey v.s. Shakespeare

Infinite Monkey Theorem

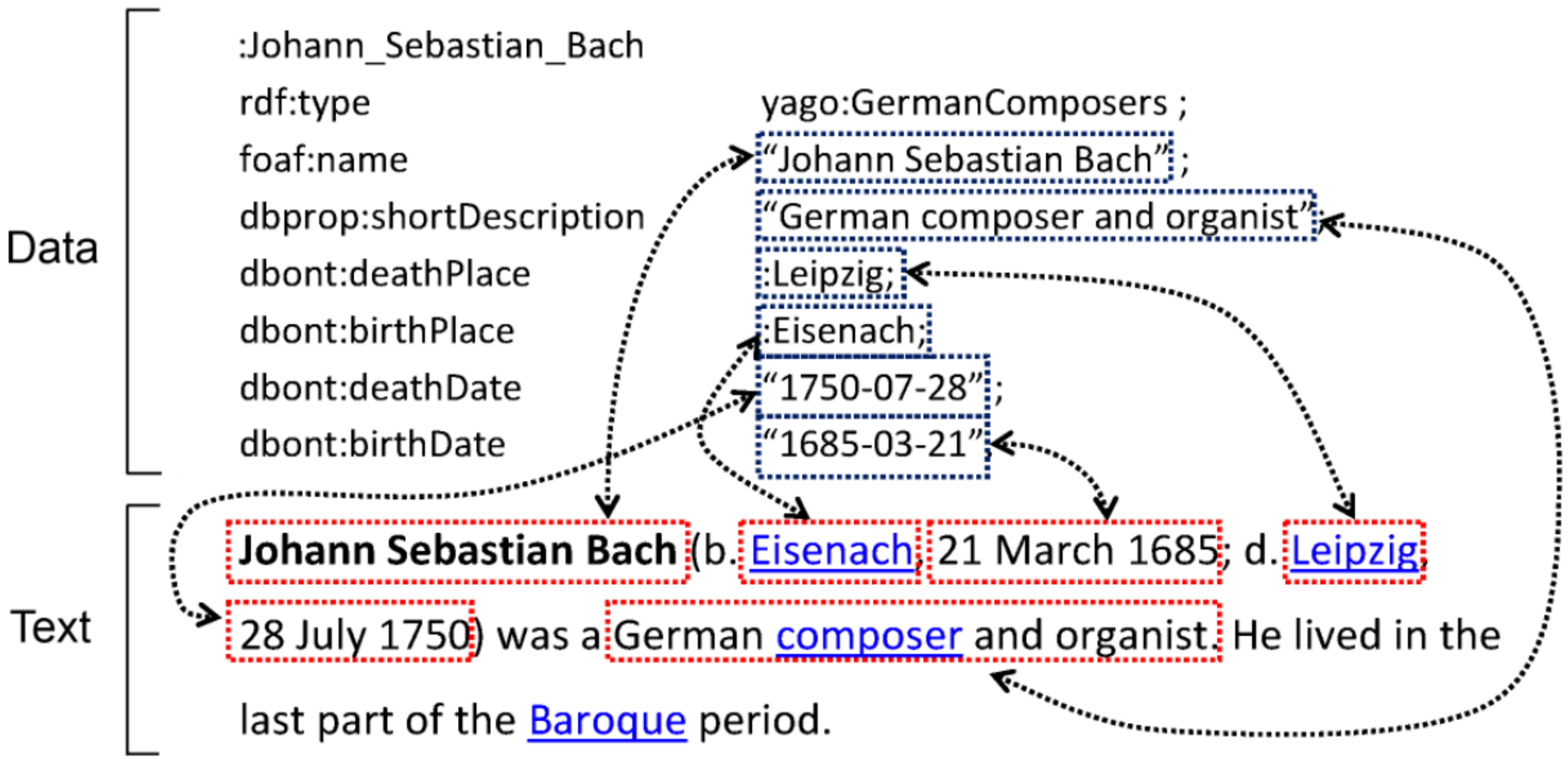
The Problem to Solve

- Can thousands of monkeys write out the Shakespeare's or the Dickens'?
- Can machines do?
- How can they do?
- What about other types/sorts/ genres of text?
- …
- How do machines generate natural language?

NLG used to be template-based (1993–)

- Filling up the templates by aligning data and text strings.

[1]



…but it is heavily limited because

- Number of templates grows quickly
  - No syntactic generalization. e.g. different templates for singular and plural versions
- Not much variation in output
  - No planning of text, just concatenation
- No understanding of text
  - Far from real scenario, far from “natural”

Automatic Generation

- Once aware of the communicative goals, machines can understand the meaning of the content (what is important, what is what), and automatically generate natural, meaningful and plausible language.

- The above definition is said myself.

The Automatic Age thanks to Deep Learning

- Machines are able to automatically generate
  - Weather Report
  - Chinese Couplet[2,3]
  - English Poetry, Chinese Poetry and Song lambrics[4]
  - Image Caption, Image and Video Description 🔥
  - Story (various genres) 🔥
  - Dialog 🔥

Image Caption, Image and Video Description

- Input:

- Output: a group of young girls standing next to each other on the beach.
- This requires[5]:
  - Identifying and detecting objects, scenes, people…
  - Reasoning about spatial relationships and properties of objects
  - Combining several sources of information into a coherent sentence

Image Caption, Image and Video Description

- Input:

- Output: a group of young girls standing next to each other on the beach.

- This requires[5]:
  - Identifying and detecting o… Representation …le…
  - Reasoning about spatial rela… Understand …operties of objects
  - Combining several source… Language Model …coherent sentence

It is somehow like machine translation

ConvNet
(encoder)

Neural LM
(decoder)

a group of young girls standing next to each other on the beach.

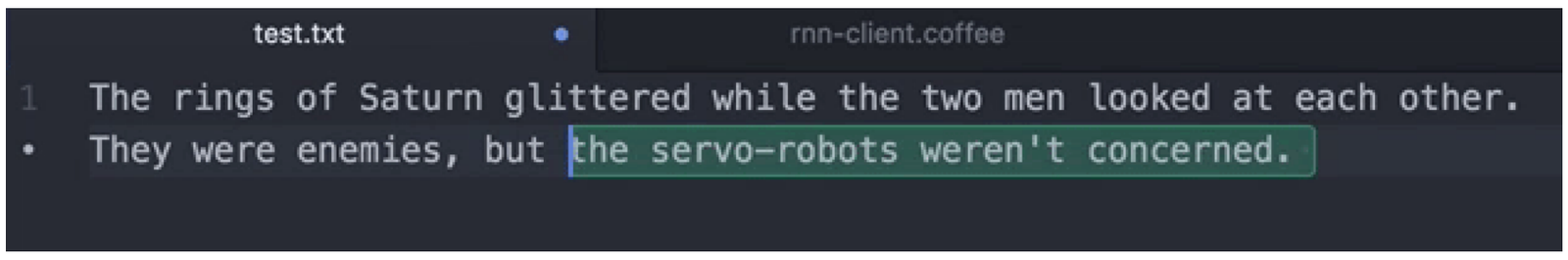
Joint Embedding Space
of Image and Text[6]

- The encoder-decoder framework is widely adopted

Story

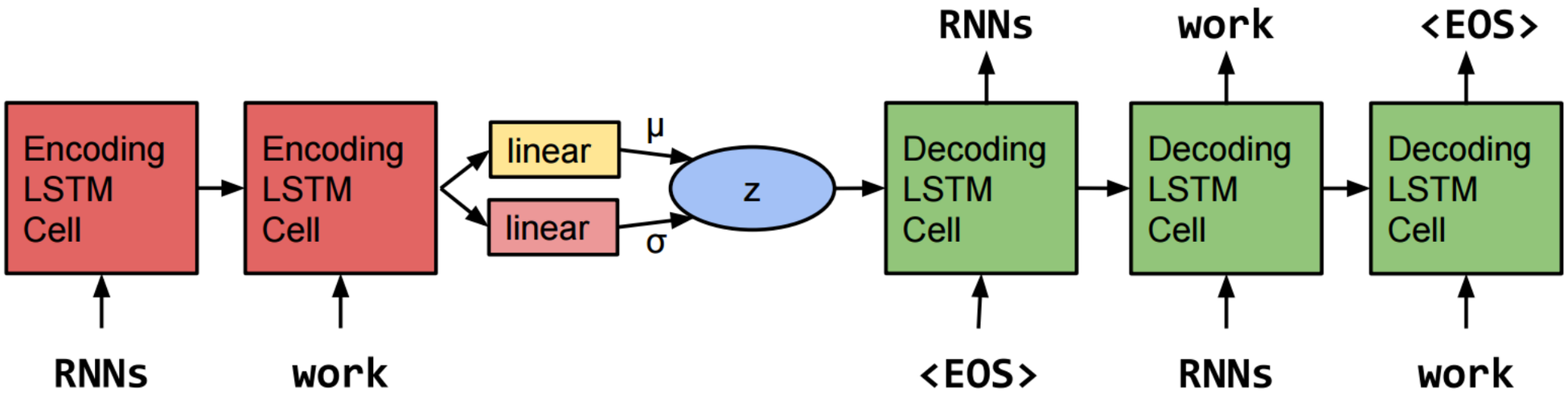
- Various genres
  - News, Romantics, Thrillers[7]…
- Storyteller, Story completion[8]

- Writing with machine[9]

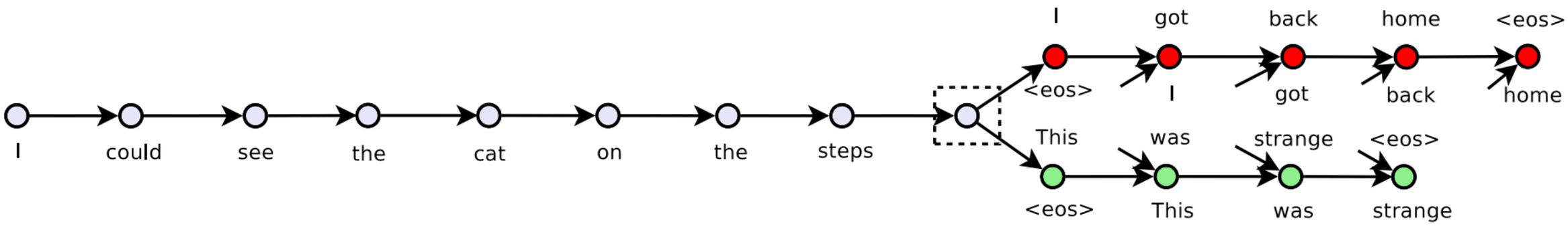


Story generation largely rely on RNN/LSTM

- Completion is still based on encoder-decoder framework where both are RNN/LSTM[8]

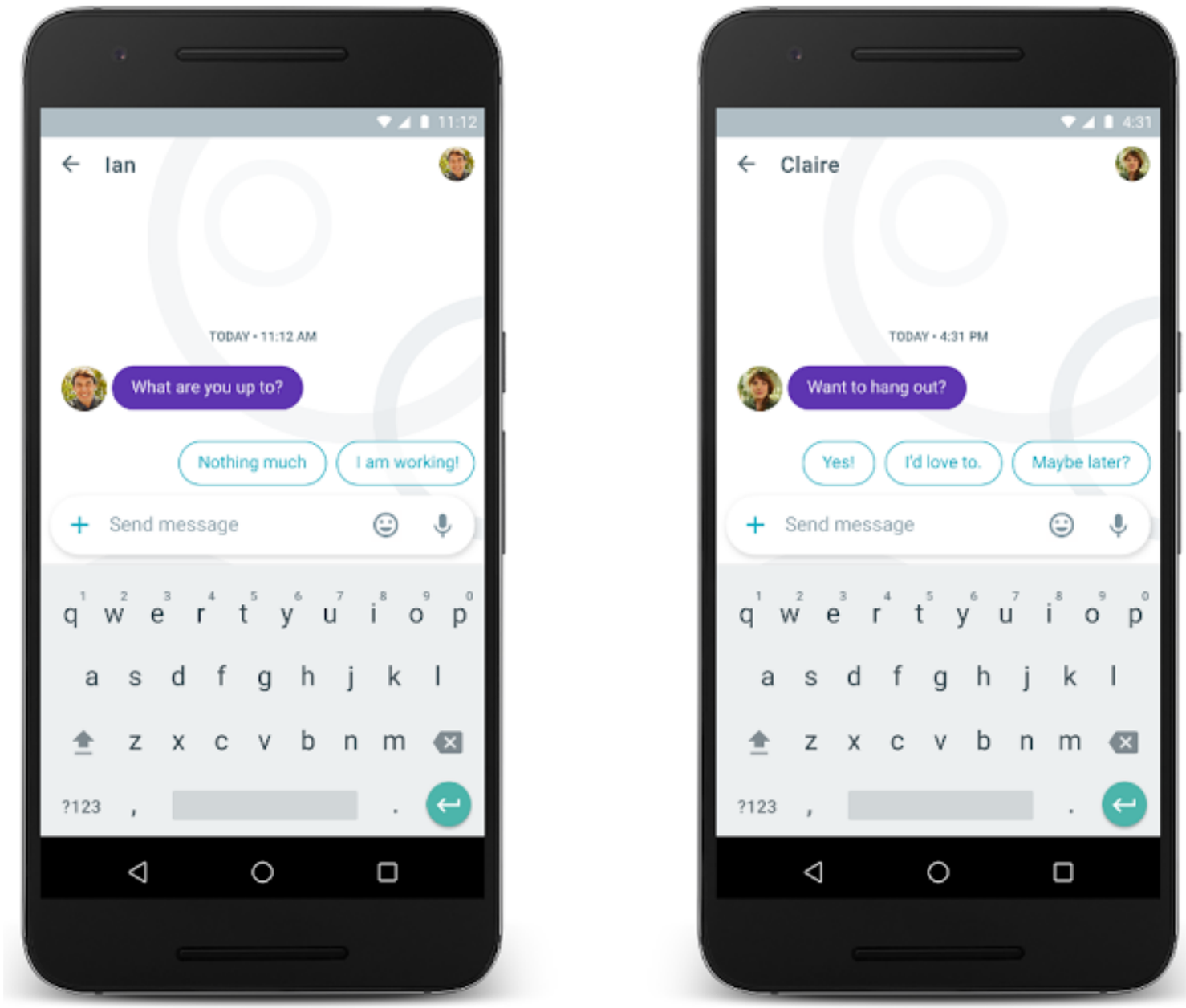


- Generation with no input is sampling, but we can make use of those embeddings[7]



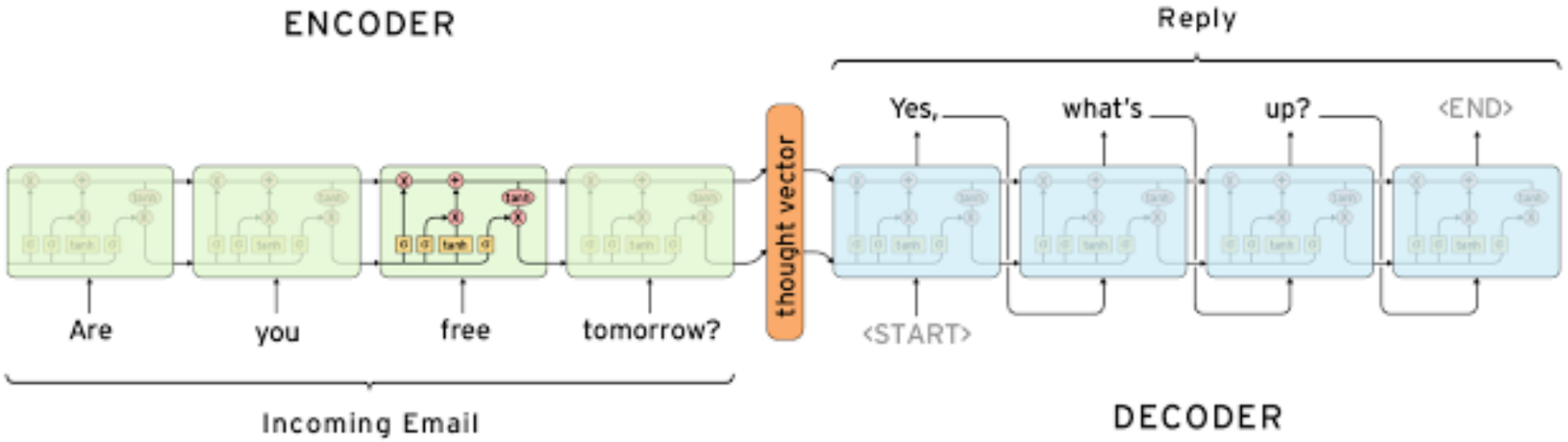
Dialog

- The hottest NLG application with vital commercial potentials
  - Siri (Apple)
  - Xiaobing/Tay (Microsoft)
  - Allo (Google)



Dialog generation involves a lot of difficulties

- ✦ Similar with other generation tasks



- ✦ Unique for itself

Difficulties for all generation tasks

- RNN/LSTM's problem
  - impotent for too long sentences (semantic drift)
  - grammatically problematic
  - highly frequent pattern domination (less variation)

Difficulties for all generation tasks

- RNN/LSTM's problem[8]

i went to the store to buy some groceries .
i store to buy some groceries .
i were to buy any groceries .
horses are to buy any groceries .
horses are to buy any animal .
horses the favorite any animal .
horses the favorite favorite animal .
horses are my favorite animal .

Table 1: Sentences produced by greedily decoding from points between two sentence encodings with a conventional autoencoder. The intermediate sentences are not plausible English.

" i want to talk to you . "
"i want to be with you . "
"i do n't want to be with you . "
i do n't want to be with you .
she did n't want to be with him .
he was silent for a long moment .
he was silent for a moment .
it was quiet for a moment .
it was dark and cold .
there was a pause .
it was my turn .

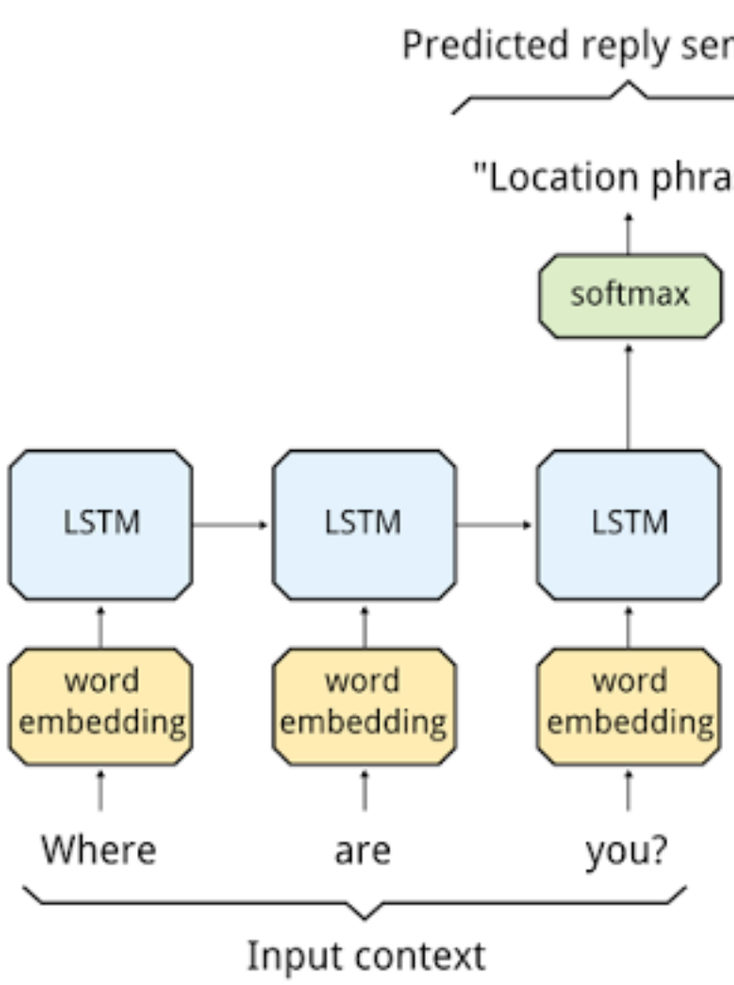
Table 8: Paths between pairs of random points in VAE space: Note that intermediate sentences are grammatical, and that topic and syntactic structure are usually locally consistent.

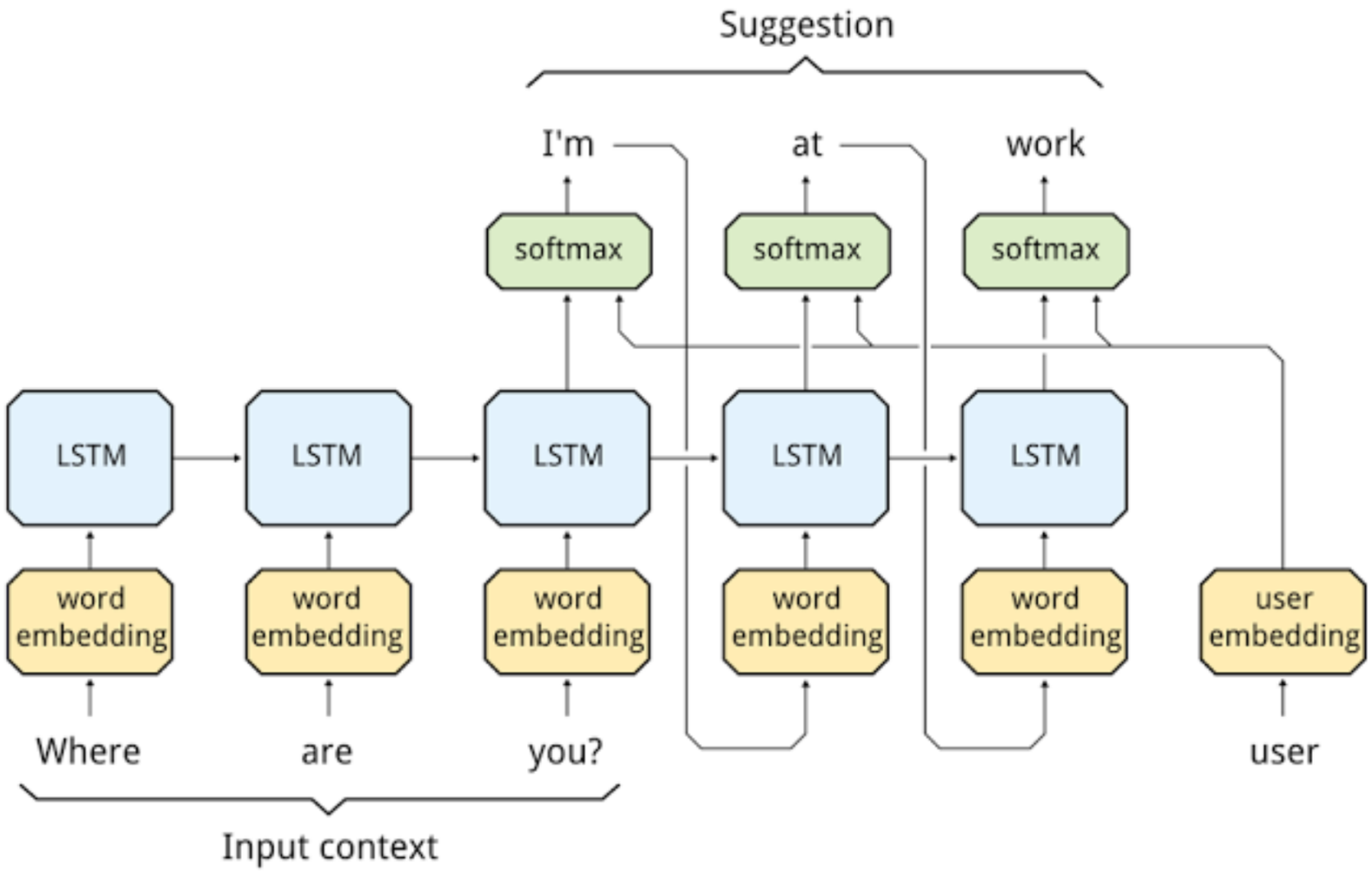
Unique difficulties for dialog generation

- Unique 1: “Reply Panacea”
  - e.g. “Thanks.” “Great!” “Awesome!”
  - debug: semantic drift + highly frequent pattern domination
  - fix: mixed response strategy

Architecture of Allo (Google)

- ✦ Mixed response strategy by semantic classification





Mixed Response Strategy

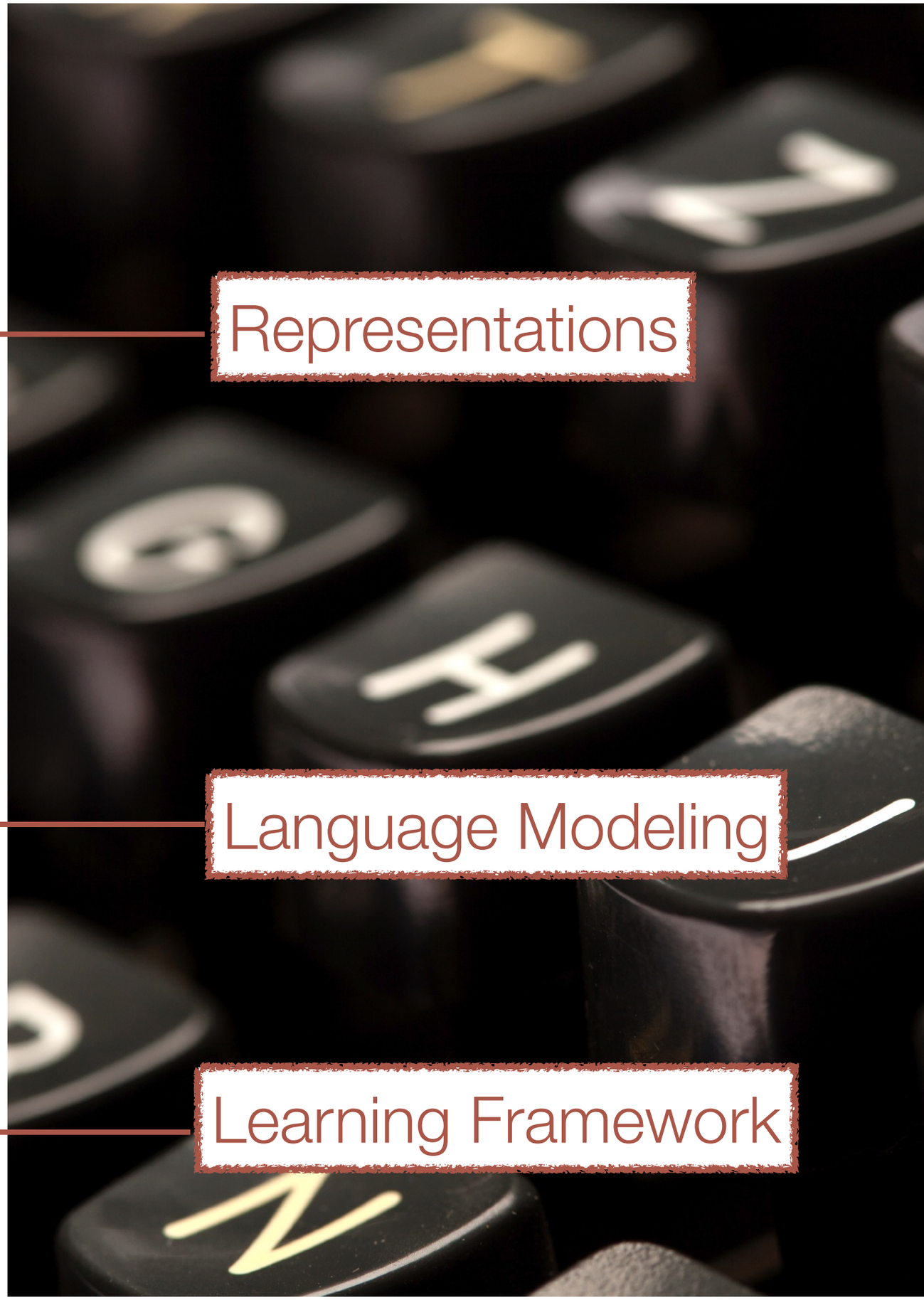
- Strategy Type 1: Knowledge Retrieval
  - Knowledge base, inference, relation, expert domain…
- Strategy Type 2: Usage Scenario
  - Location/Weather/…
  - Largely based on Type 1, “generate” based on A

Unique difficulties for dialog generation

- Unique 2: Language Style
  - debug: answers retrieved from knowledge base are often too long and stereotype…
  - fix: summarize, sentiment, natural (say words like a human)

Summary

- Continuous-space representations
  - word
  - sentence
  - even characters, bytes (crazy!)
- Language Modeling:
  - RNN
  - CNN
- End-to-End, Sequence-to-Sequence, Encoder-Decoder



Summary

- Continuous-space representations
  - word
  - sentence
  - even characters, bytes (crazy!)
- Language Modeling:
  - RNN
  - CNN
- End-to-End, Sequence-to-Sequence, Encoder-Decoder

We're all born in
the Era of Deep Learning

Thanks for your attention!

References

1. Daniel Duma and Ewan Klein. “Generating Natural Language from Linked Data: Unsupervised template extraction”. ACL 2013.

2. Rui Yan. “Chinese Couplet Generation with Neural Network Structures”. ACL 2016.

3. Long Jiang and Ming Zhou. “Generating Chinese Couplets using a Statistical MT Approach”. COLING 2008.

4. Qixin Wang, Tianyi Luo, Dong Wang, Chao Xing. “Chinese Song Iambics Generation with Neural Attention-based Model”. IJCAI 2016.

5. Ryan Kiros. “Generating image captions with neural networks”. CIFAR NCAP - Summer School 2014.

6. Ryan Kiros, Ruslan Salakhutdinov, Richard Zemel. “Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models”. TACL 2015.

7. Ryan Kiros et al. “Skip-Thought Vectors”. NIPS 2015.

8. Samuel R. Bowman et al. “Generating Sentences from a Continuous Space”. arXiv preprint 2015.

9. Writing with the machine. https://www.robinsloan.com/notes/writing-with-the-machine/

References

10. Fu, Ruiji, et al. “Learning semantic hierarchies via word embeddings”. ACL 2014.

11. Ronan Collobert et al. “Natural Language Processing (almost) from Scratch”. 2011.

12. Zou, Will Y., et al. “Bilingual word embeddings for phrase-based machine translation”. EMNLP 2013.

13. Richard Socher,et al. “Zero-Shot Learning Through Cross-Modal Transfer”. ICLR 2013.

14. Nal Kalchbrenner et al. “A Convolutional Neural Network for Modelling Sentences”. ACL 2014.

15. Dzmitry Bahdanau et al. “Neural machine translation by jointly learning to align and translate”. ICLR 2015.

16. Orhan Firat et al. “Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism”. NAACL 2016.

17. Li Yao et al. “Describing Videos by Exploiting Temporal Structure”. 2015.