Natural Language Understanding

Kyunghyun Cho, NYU & U. Montreal

Fun Trivia

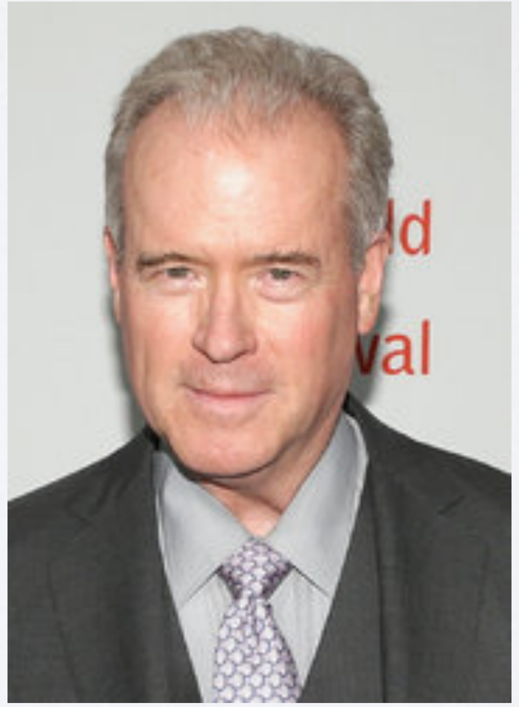
HISTORY OF MT RESEARCH

Topics: Two Most Important Moments in MT Research

- In 1949: Warren Weaver's Memorandum <Translation>

- In 1991-1993: Statistical MT from IBM

Vincent (left) and Stephen Della Pietra



Courant Institute of Mathematical Sciences
New York University

".. it is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the "Chinese code." If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation?"

- Weaver (1949)

Warren Weaver, 1894-1978

Warren Weaver Hall

The Mathematics of Statistical Machine Translation: Parameter Estimation

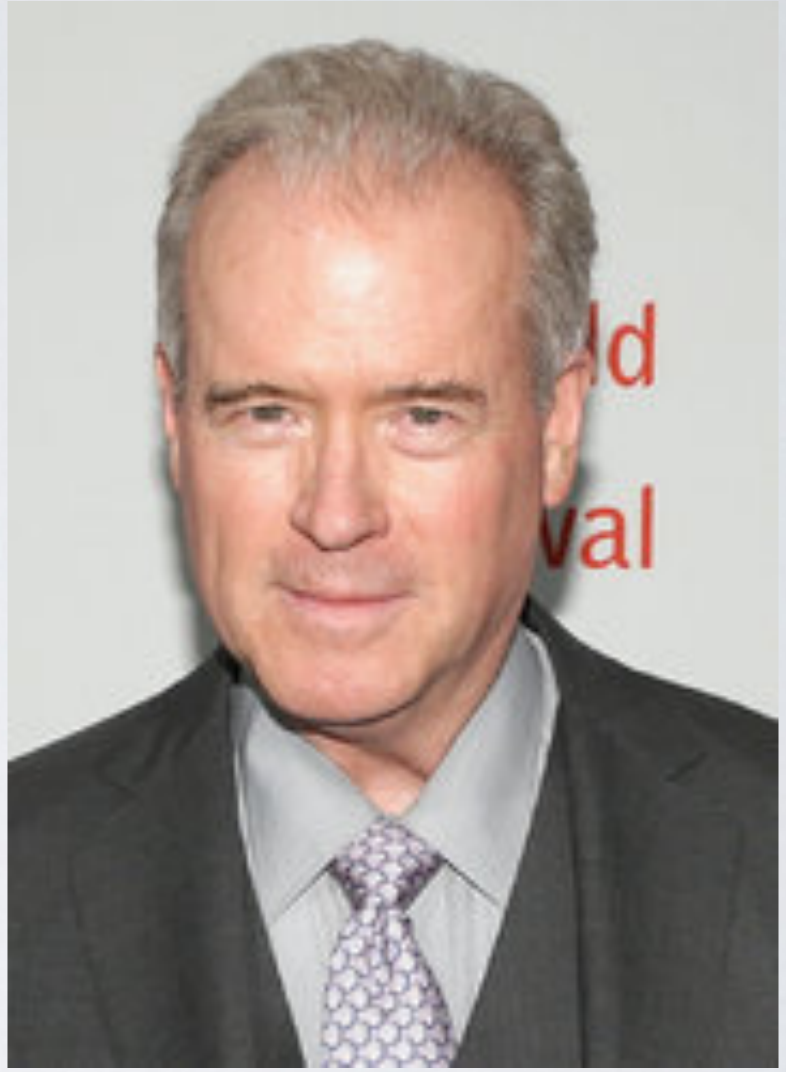
Peter F. Brown*
IBM T.J. Watson Research Center

Stephen A. Della Pietra*
IBM T.J. Watson Research Center

Vincent J. Della Pietra*
IBM T.J. Watson Research Center

Robert L. Mercer*
IBM T.J. Watson Research Center

Mercer St.



Robert L. Mercer
(Hedge Fund Magnate)*

251 **Mercer** Street
New York, N.Y. 10012-1185

* NY Times

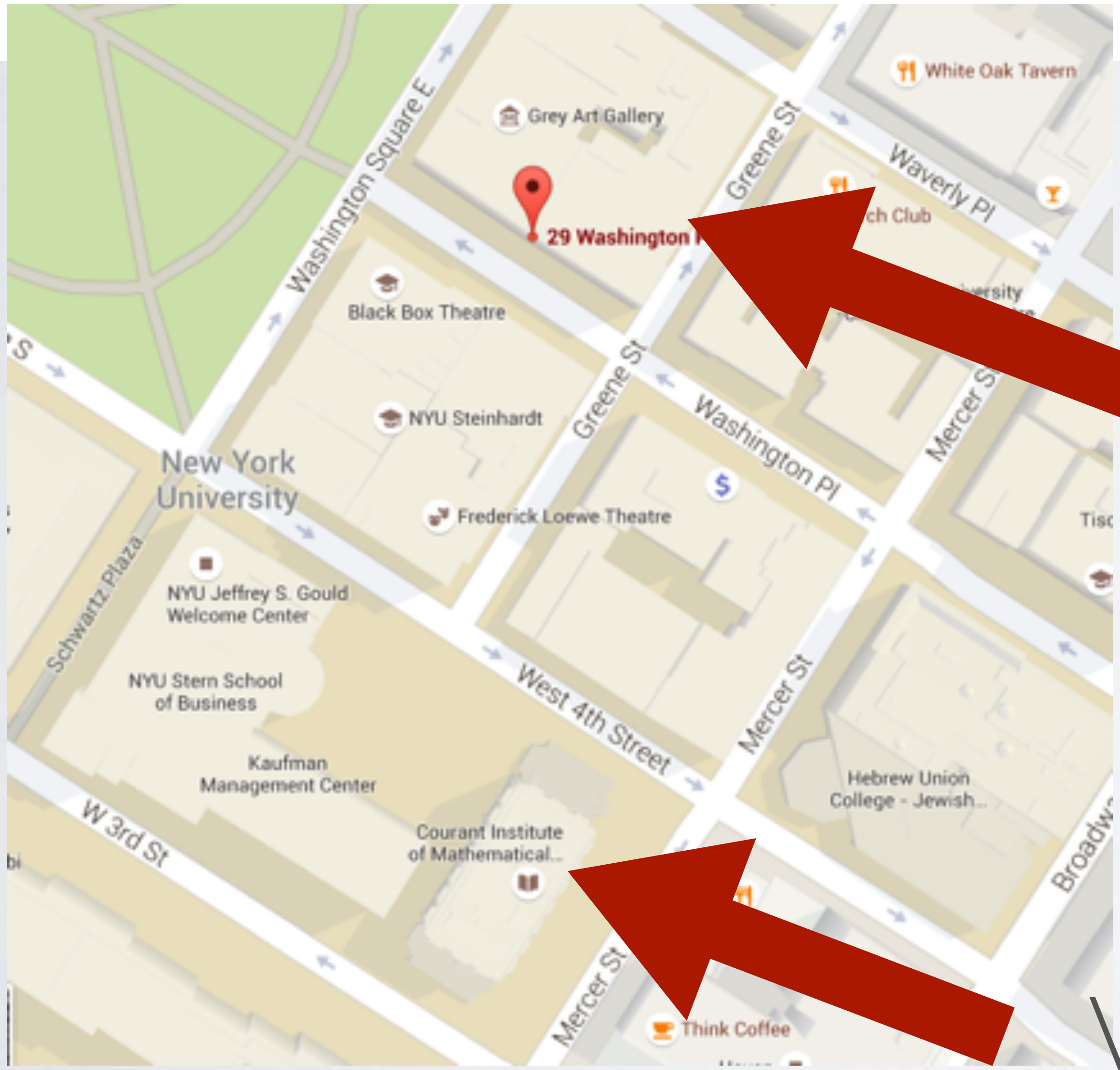
The Mathematics of Statistical Machine Translation: Parameter Estimation

Peter F. Brown*
IBM T.J. Watson Research Center

Stephen A. Della Pietra*
IBM T.J. Watson Research Center

Vincent J. Della Pietra*
IBM T.J. Watson Research Center

Robert L. Mercer*
IBM T.J. Watson Research Center



Warren Weaver Hall

Peter F. Brown

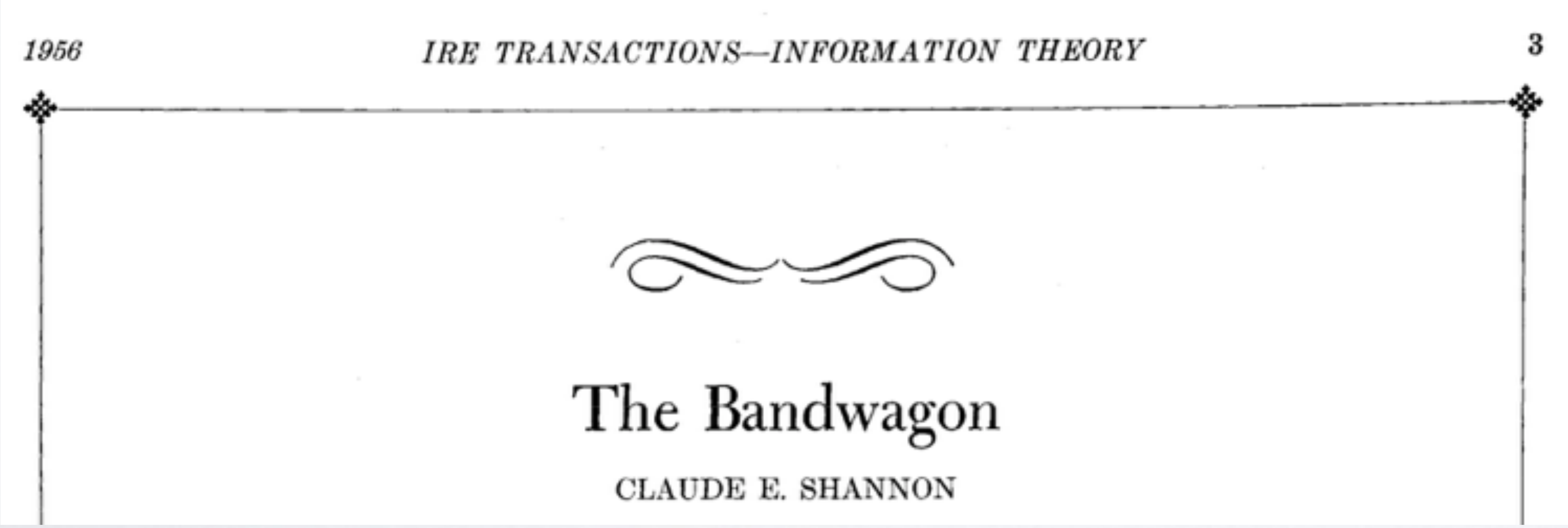
Maybe, there is something about CIMS, NYU with machine translation…

if you find a double della-pietra i'll be super impressed :)

Warning

"It will be all too easy for our somewhat artificial prosperity to collapse overnight when it is realized that the use of a few exciting words like information, entropy, redundancy, do not solve all our problems"

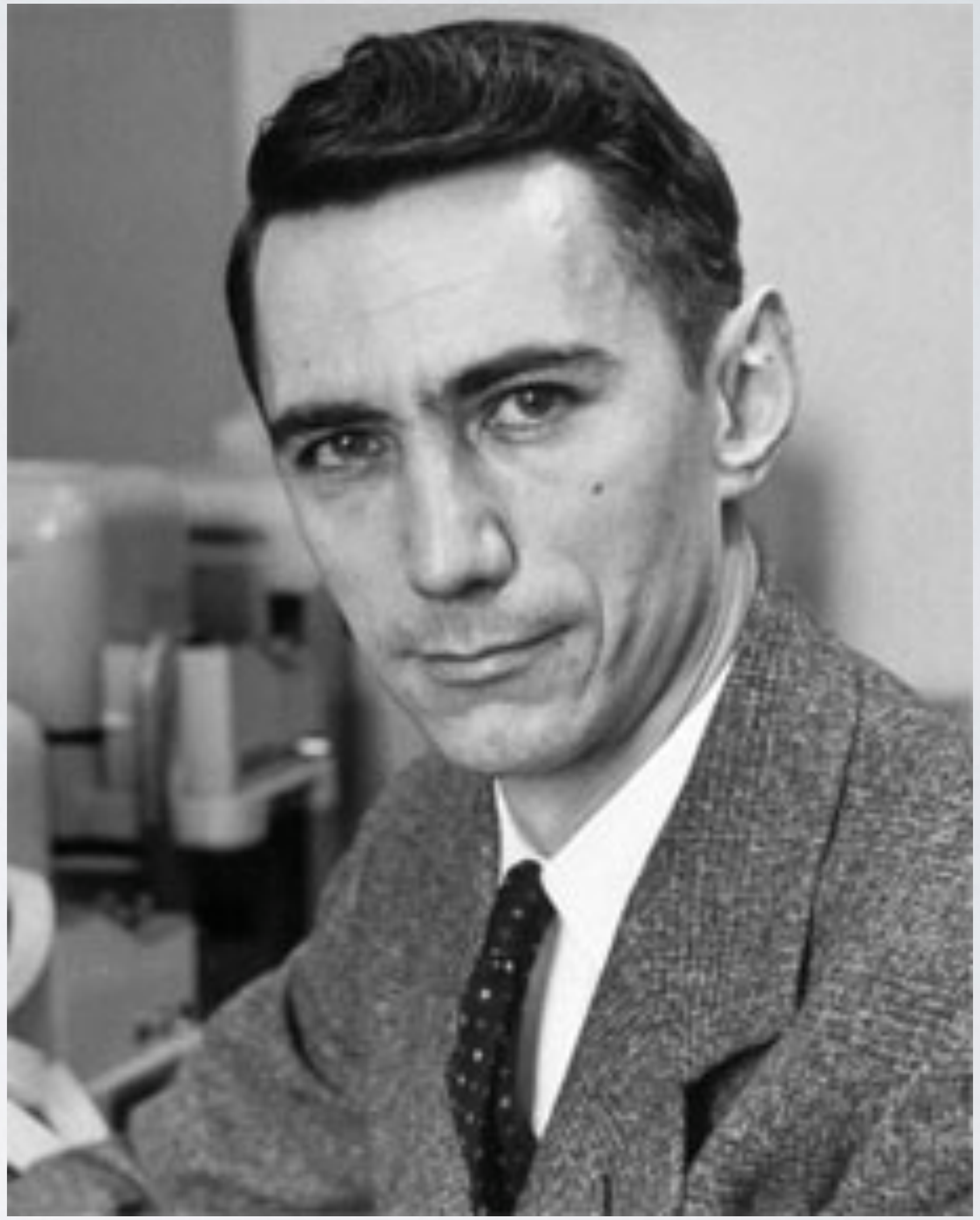
- Shannon (1956)



1956 IRE TRANSACTIONS—INFORMATION THEORY 3

The Bandwagon

CLAUDE E. SHANNON



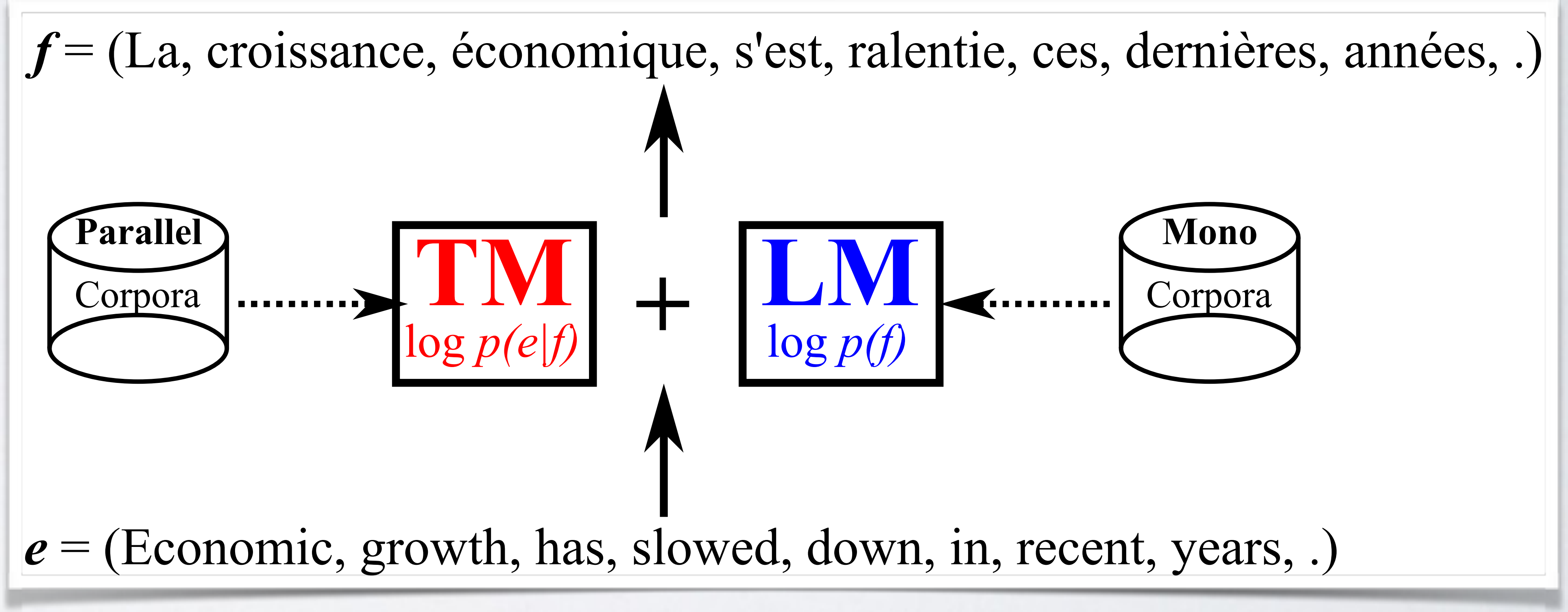
Claude Shannon, 1916-2001

Machine Translation

NEURAL MACHINE TRANSLATION

Topics: Statistical Machine Translation

- $\log p(f|e) = \log p(e|f) + \log p(f)$
  - Translation model: $\log p(e|f)$
    - Fit it with parallel corpora
  - Language model: $\log p(f)$
    - Fit it with monolingual corpora



- The whole task $\log p(f|e)$ is ***conditional* language modelling**.

NEURAL MACHINE TRANSLATION

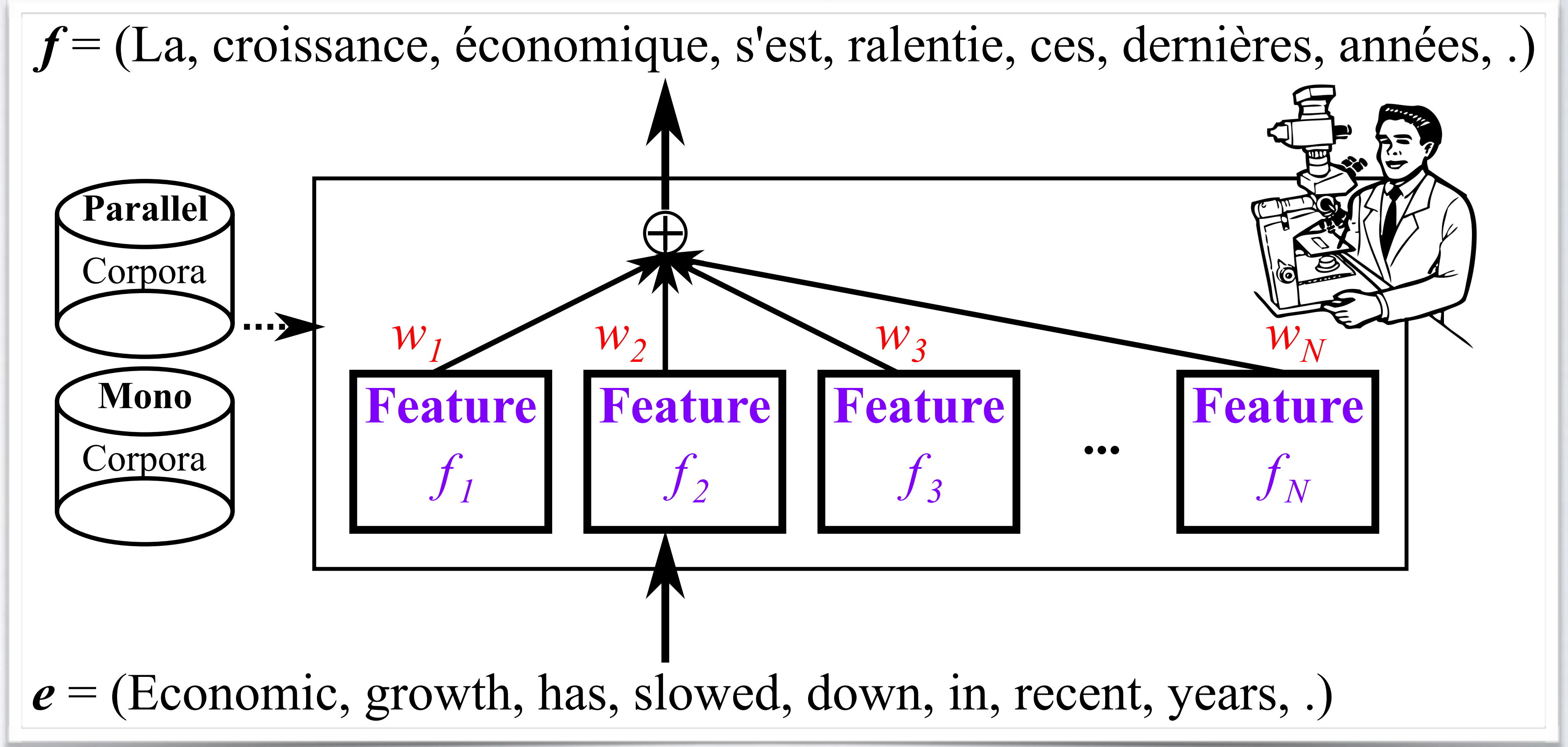
Topics: Statistical Machine Translation - In Reality

- $\log p(f|e) \approx \sum_{n=1}^{N} f_n(e, f) + C$
  - Log-linear model
  - Feature function $f_n(e, f)$
- Steps:

(1) Experts engineer *useful* features

(2) Use a simple log-linear model

(3) Use a strong, external language model



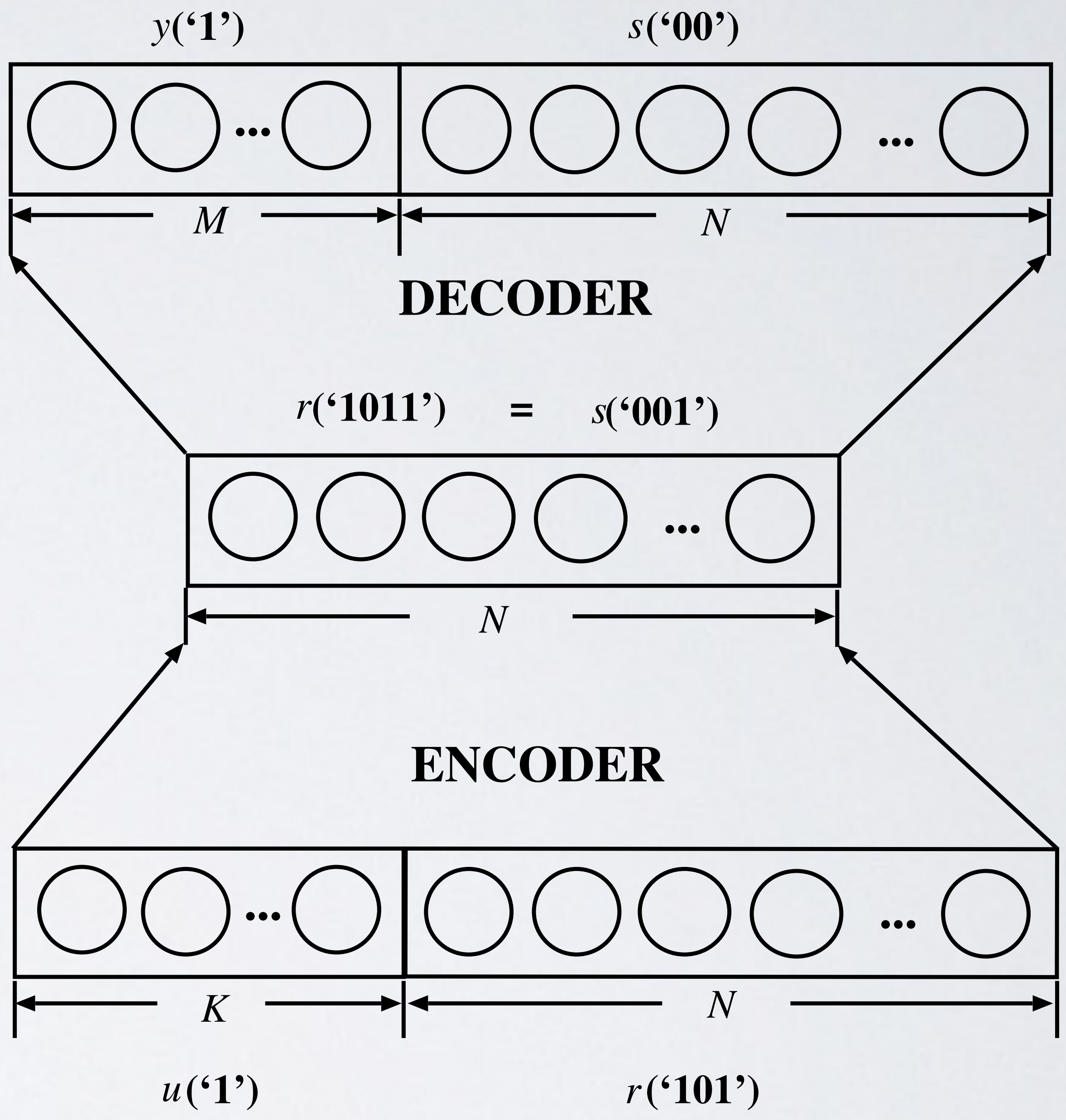
Neural Machine Translation

SPAIN IN 1997

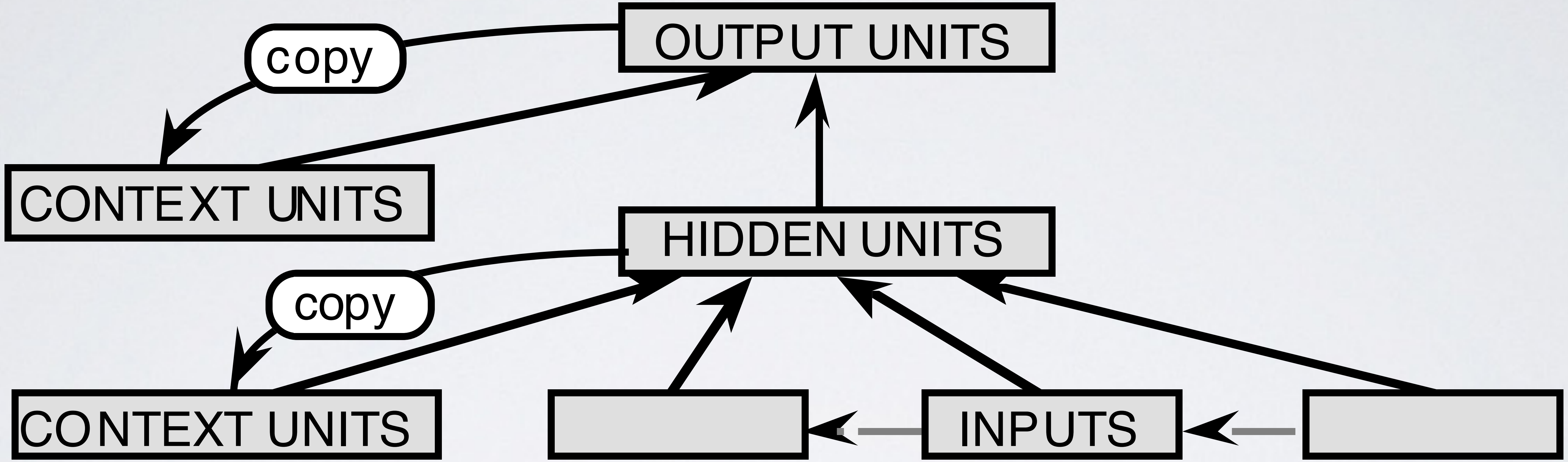
NEURAL MACHINE TRANSLATION

*"We propose .. **Recursive Hetero-Associative Memory** which .. may be applied **to learn general translations from examples** in which different sentences may have the same translation."*

– Forcada & Ñeco, 1997



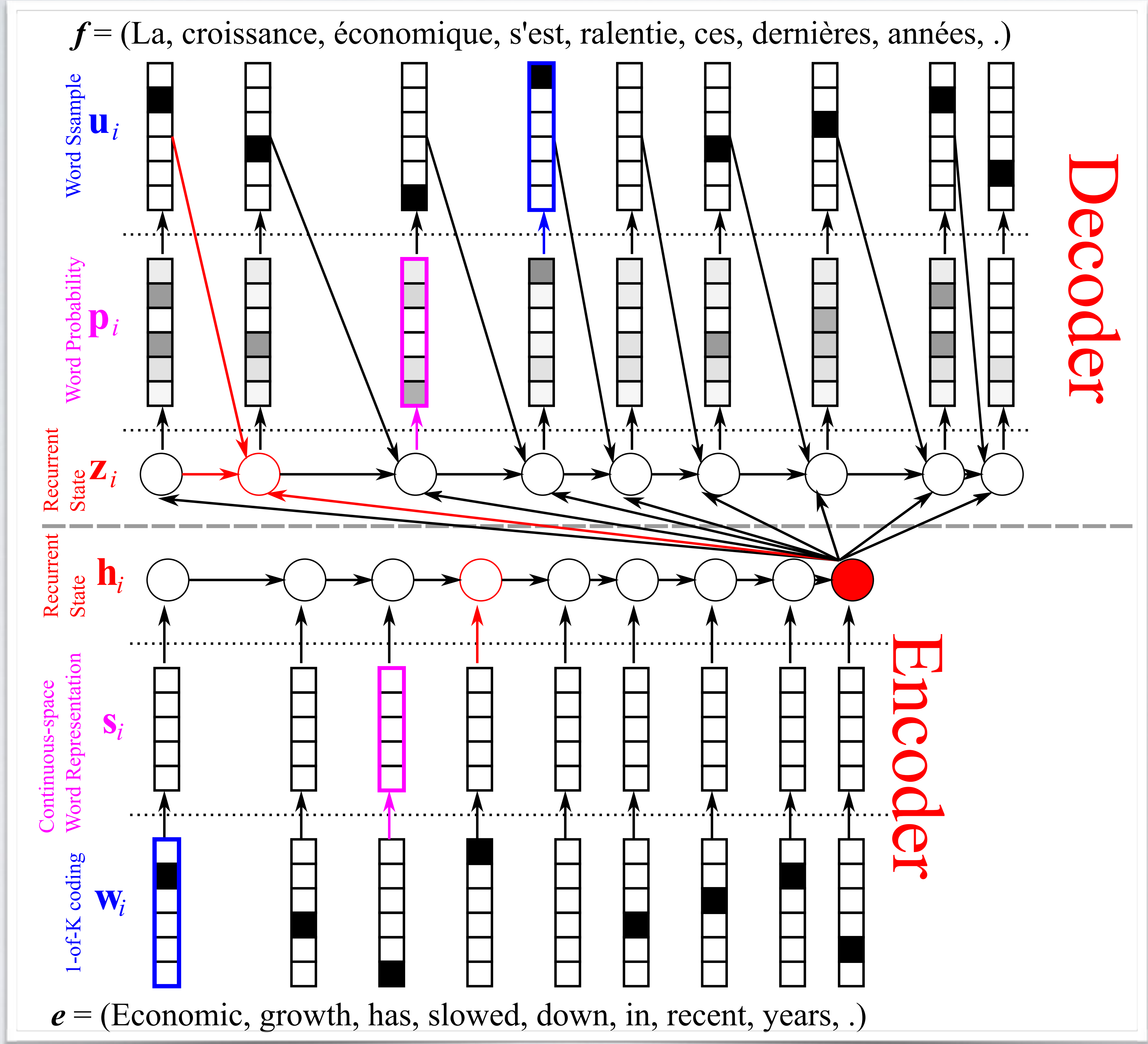
NEURAL MACHINE TRANSLATION



"*Based on these encouraging performances, future work dealing with more complex limited-domain translations seems to be feasible.* ***However, the size of the neural nets required for such applications (and consequently, the learning time) can be prohibitive***"

(Castaño&Casacuberta, 1997)

NEURAL MACHINE TRANSLATION



(Forcada&Ñeco, 1997;
Castaño&Casacuberta, 1997;
Kalchbrenner&Blunsom, 2013;
Sutskever et al., 2014;
Cho et al., 2014)

NEURAL MACHINE TRANSLATION

Topics: Sequence-to-Sequence Learning — Encoder

- Encoder

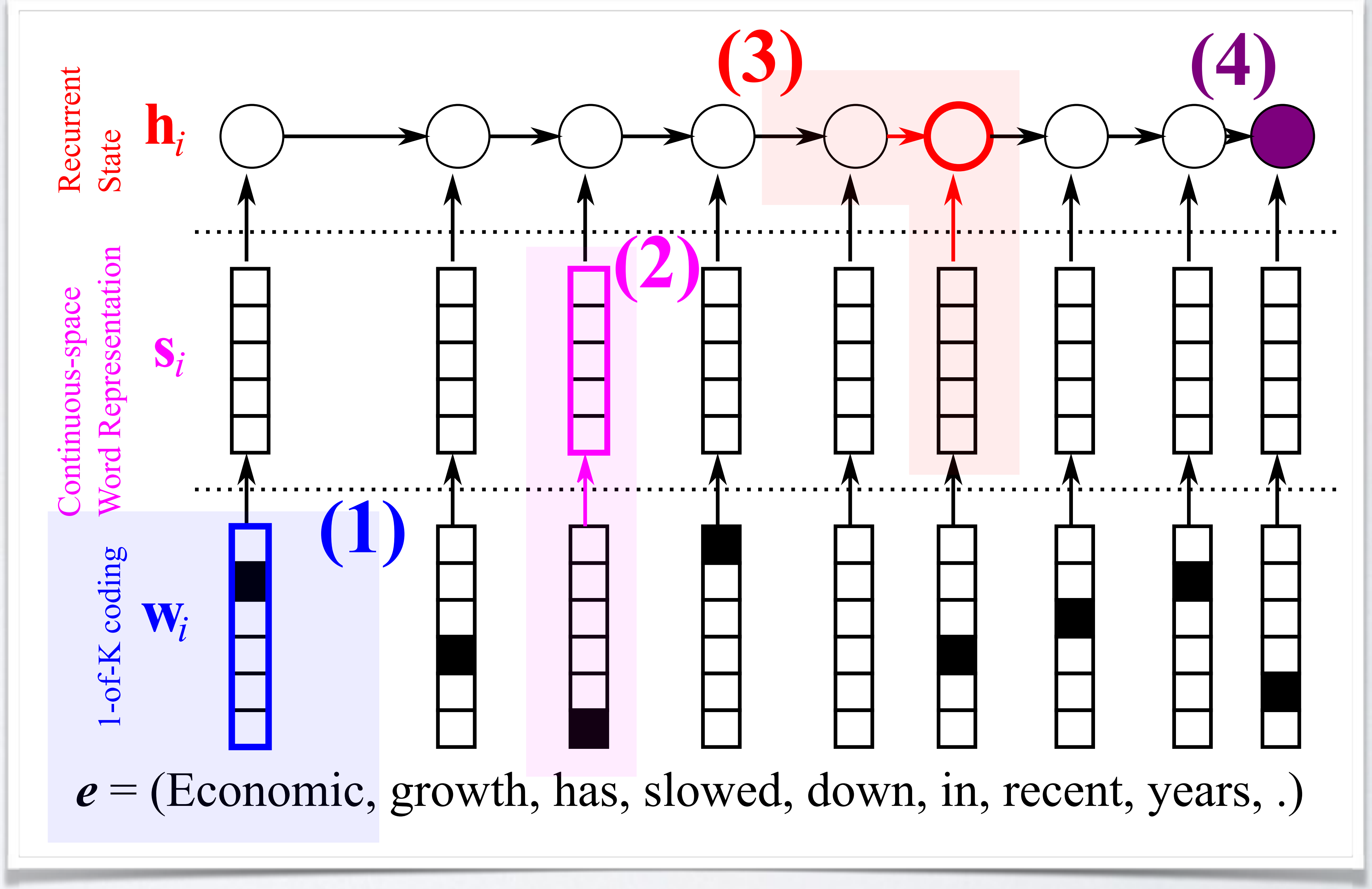
  (1) 1-of-K coding of *source* words

  (2) Continuous-space representation

  $$s_{t'} = W^\top x_{t'}, \text{ where } W \in \mathbb{R}^{|V| \times d}$$

  (3) Recursively read words

  $$h_t = f(h_{t-1}, s_t), \text{ for } t = 1, \ldots, T$$



NEURAL MACHINE TRANSLATION

Topics: Sequence-to-Sequence Learning — Decoder

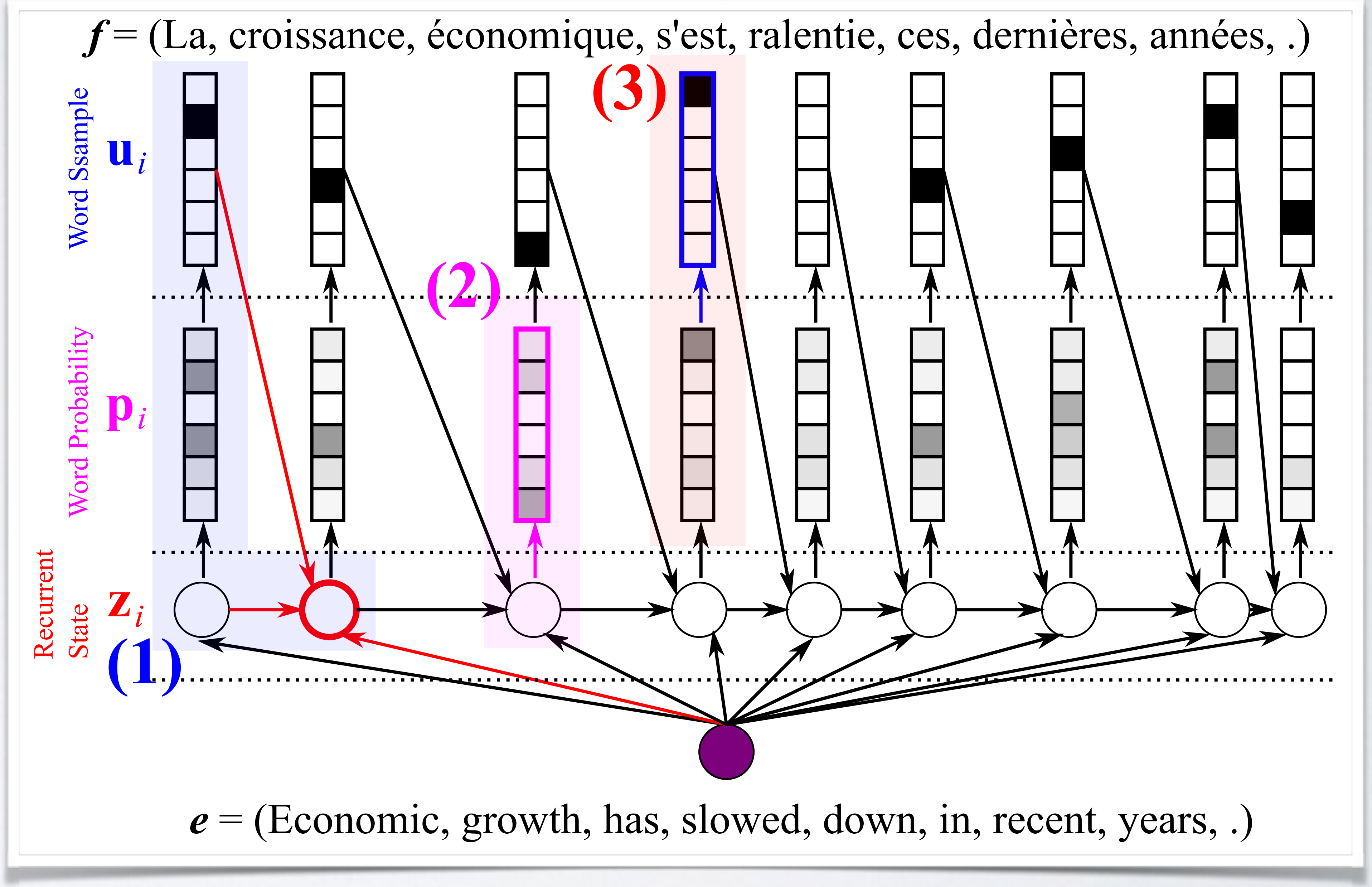
- Decoder
  - (1) Recursively update the memory

    $$z_{t'} = f(z_{t'-1}, u_{t'-1}, h_T)$$

  - (2) Compute the next word prob.

    $$p(u_{t'} | u_{<t'}) \propto \exp(R_{u_{t'}}^\top z_{t'} + b_{u_{t'}})$$

  - (3) Sample a next word
    - *Beam search is a good idea*

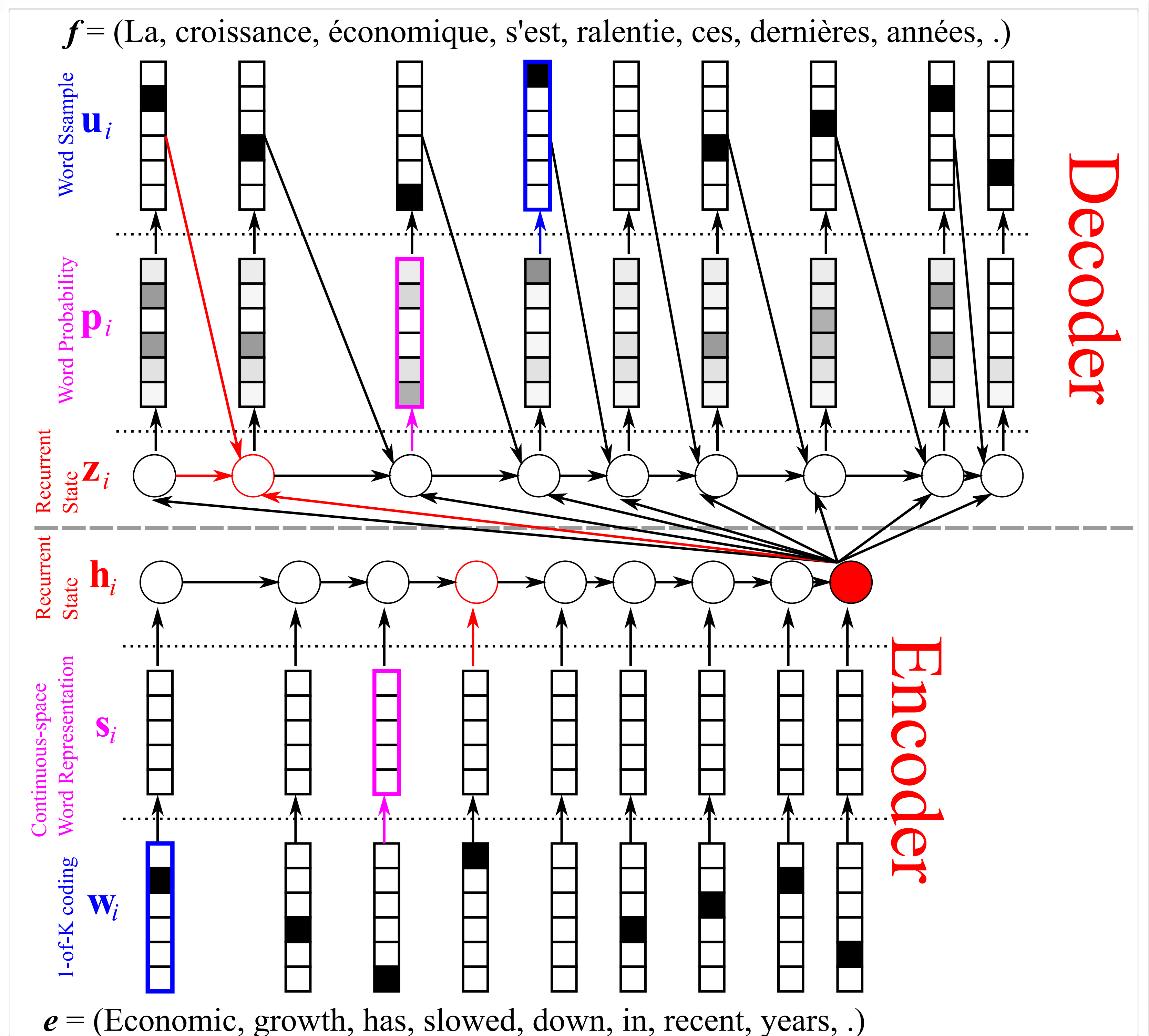


NEURAL MACHINE TRANSLATION

Topics: Sequence-to-Sequence Learning — Issue

- This is quite an unrealistic model.
- *Why?*

"You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector!" Ray Mooney



NEURAL MACHINE TRANSLATION

Topics: Attention-based Model

- Encoder: Bidirectional RNN
  - A set of *annotation* vectors

    $\{h_1, h_2, \ldots, h_T\}$
- Attention-based Decoder

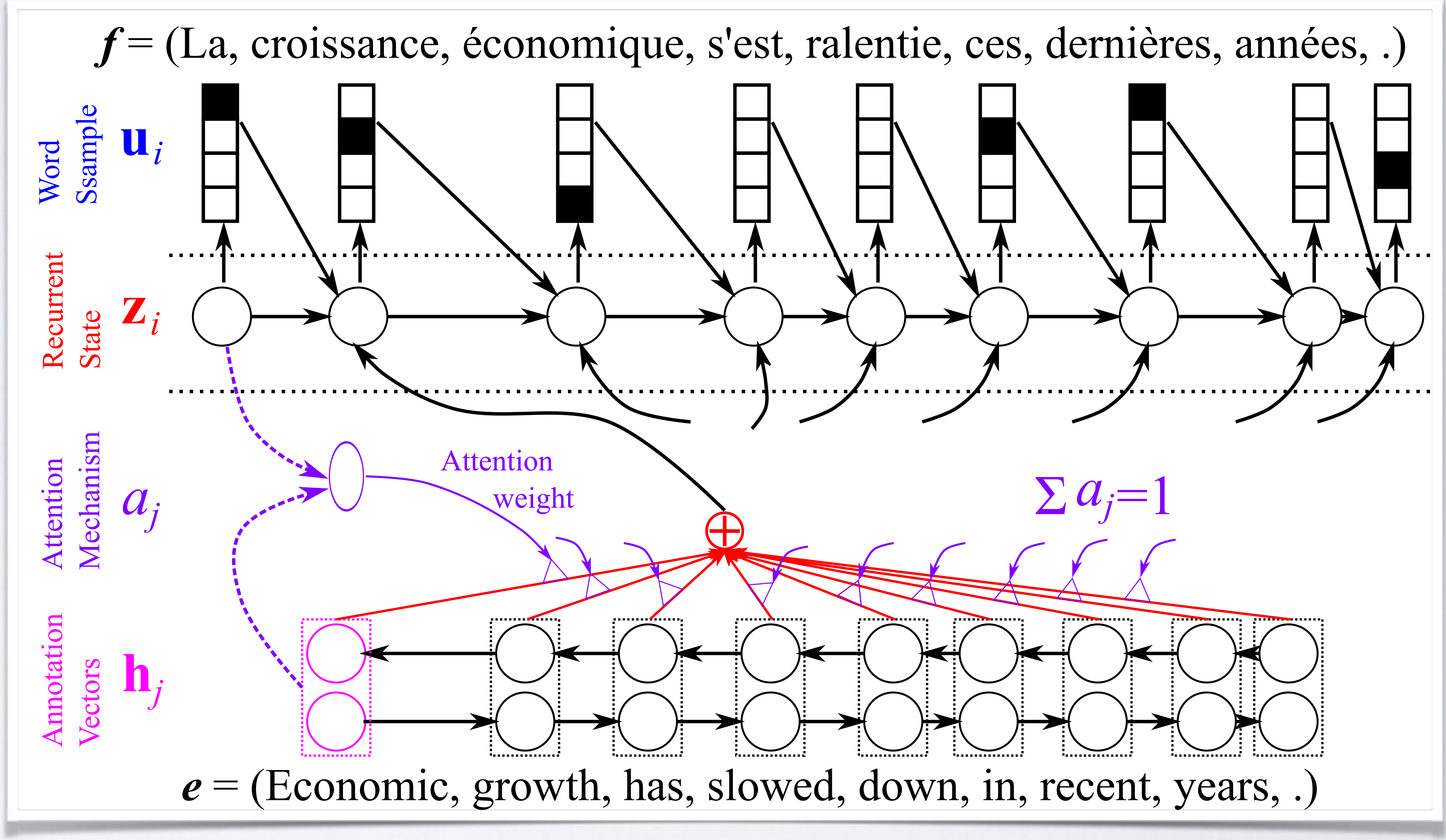
  (1)Compute attention weights

  $\alpha_{t',t} \propto \exp(e(z_{t'-1}, u_{t'-1}, h_t))$

  (2)Weighted-sum of the annotation vectors

  $c_{t'} = \sum_{t=1}^{T} \alpha_{t',t} h_t$

  (3)Use $c_{t'}$ instead of h_T



NEURAL MACHINE TRANSLATION

Topics: Attention-based Model

- Encoder: Bidirectional RNN
  - A set of *annotation* vectors

    $\{h_1, h_2, \ldots, h_T\}$
- Attention-based Decoder

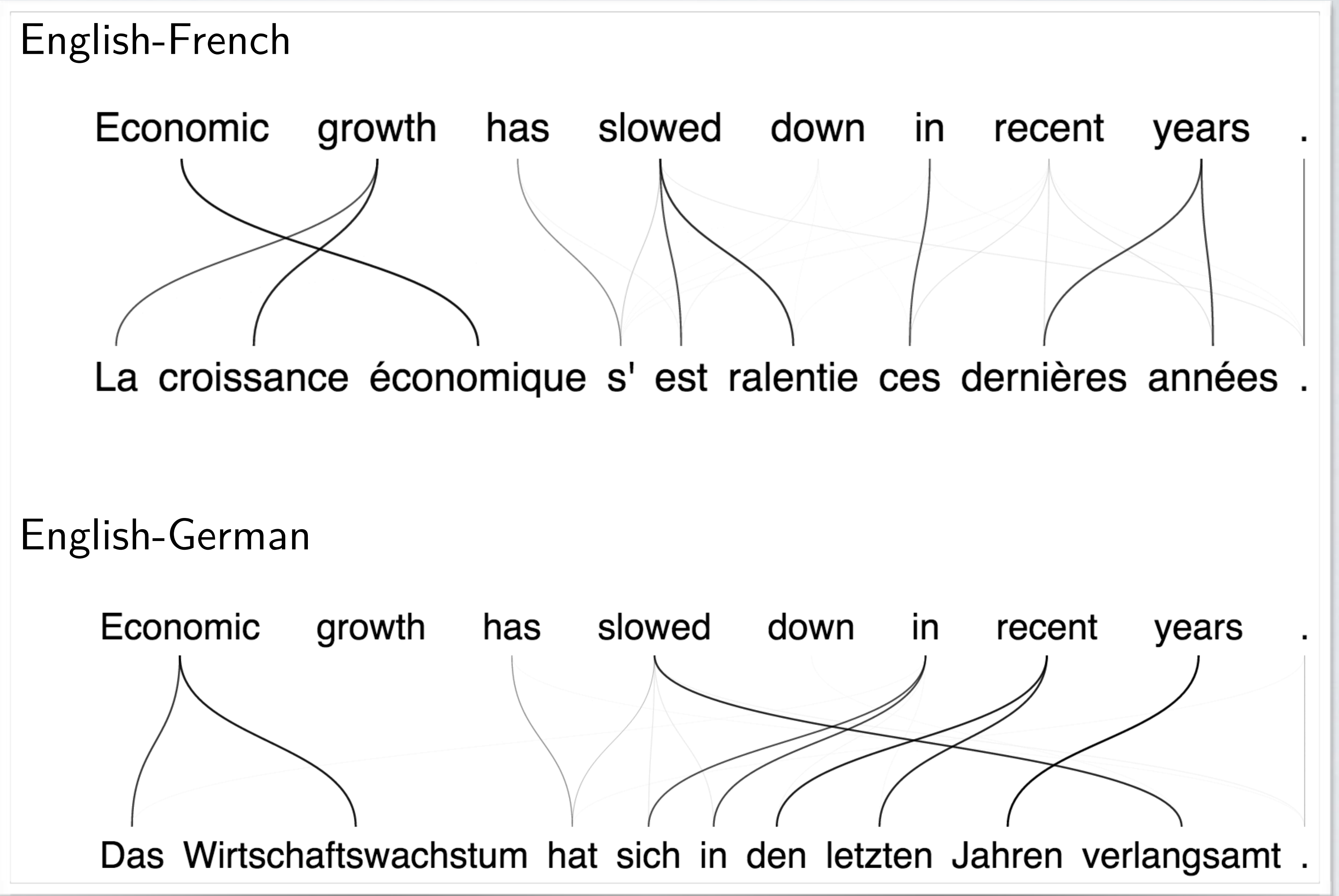
  (1)Compute attention weights

  $\alpha_{t',t} \propto \exp(e(z_{t'-1}, u_{t'-1}, h_t))$

  (2)Weighted-sum of the annotation vectors

  $c_{t'} = \sum_{t=1}^{T} \alpha_{t',t} h_t$

  (3)Use $c_{t'}$ instead of h_T



NEURAL MACHINE TRANSLATION

Topics: Attention-based Model

- How far does the attention mechanism get us?

Model	All	Rel. Improvement
RNNencdec-30	13.93	–
RNNsearch-30	21.50	**54.3%**
RNNencdec-50	17.82	–
RNNsearch-50	26.75	50.1%
RNNsearch-50*	28.45	**59.7%**
Moses	33.30	–

NEURAL MACHINE TRANSLATION

Topics: Very large target vocabulary (Jean et al., 2015)

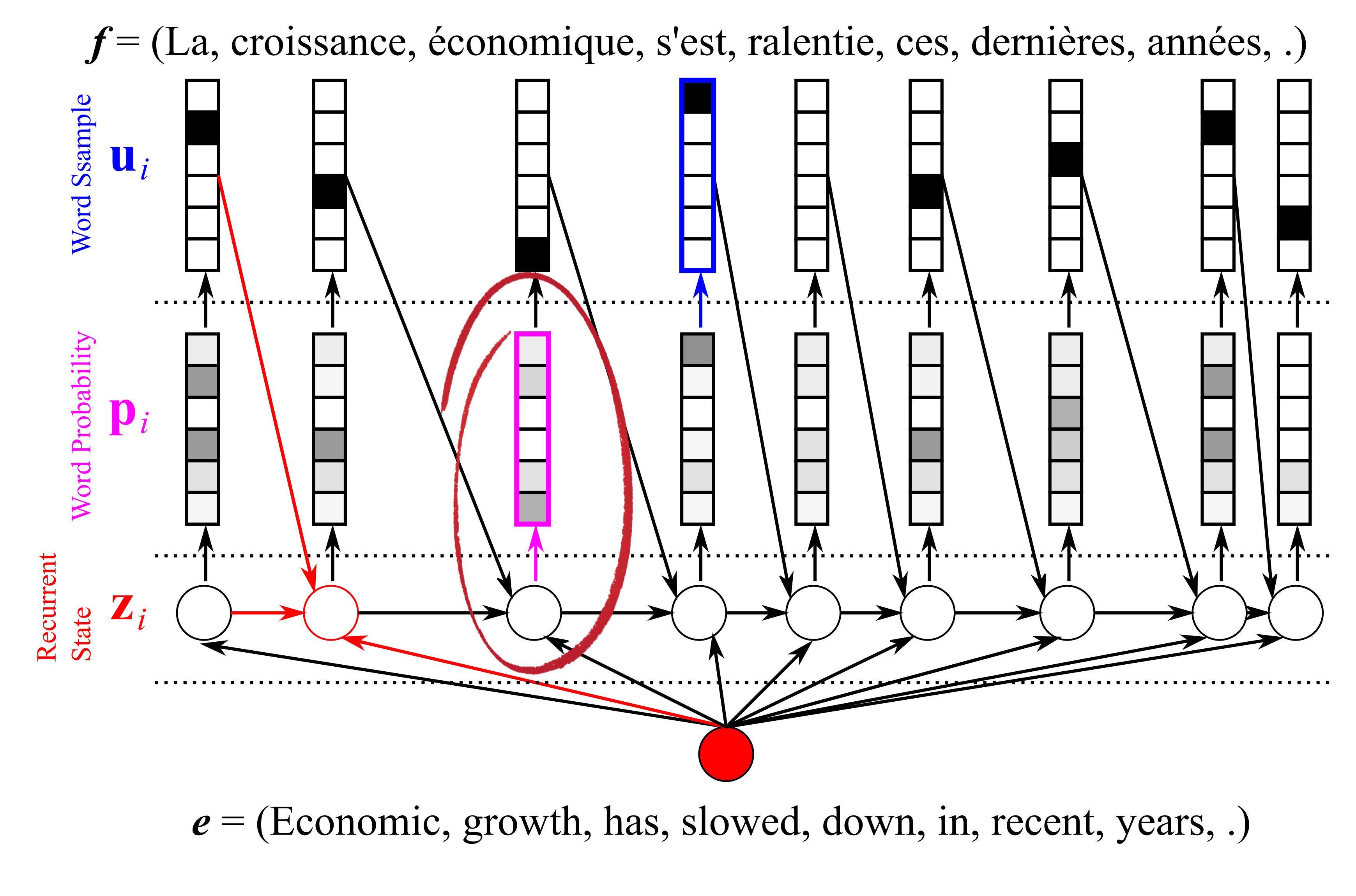
- Where are we spending most time?

  $$p(u_{t'}|u_{<t'}) \propto \exp\left(R_{u_{t'}}^\top z_{t'} + b_{u_{t'}}\right)$$

  - Complexity: $O(|V|d)$
- Where are we spending most memory?

  $$p(u_{t'}|u_{<t'}) \propto \exp\left(R_{u_{t'}}^\top z_{t'} + b_{u_{t'}}\right)$$

  - Complexity: $O(|V|d)$
- $|V|$ is **huge**, and we must compute it more than twenty times per sentence pairs!!

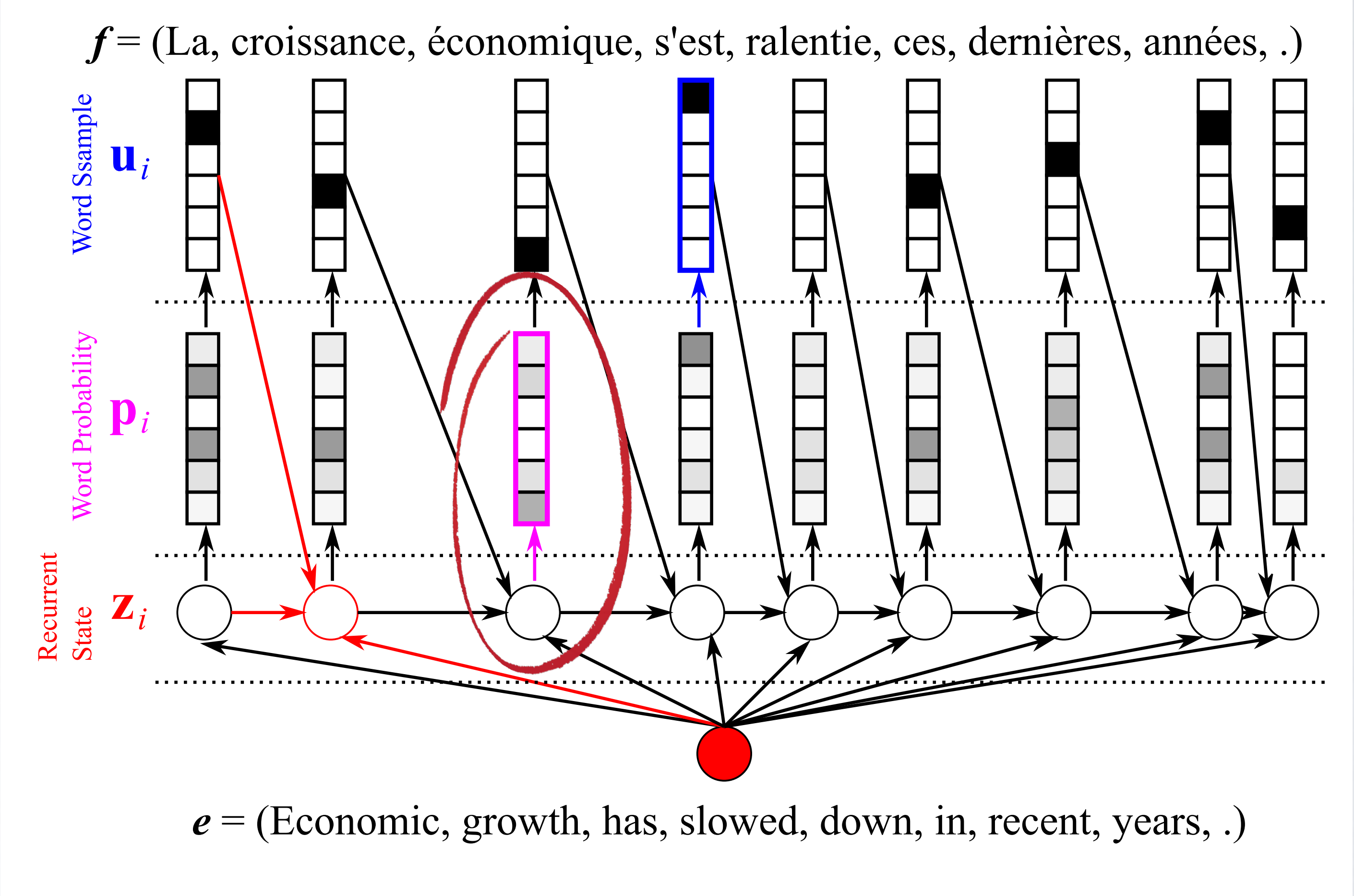


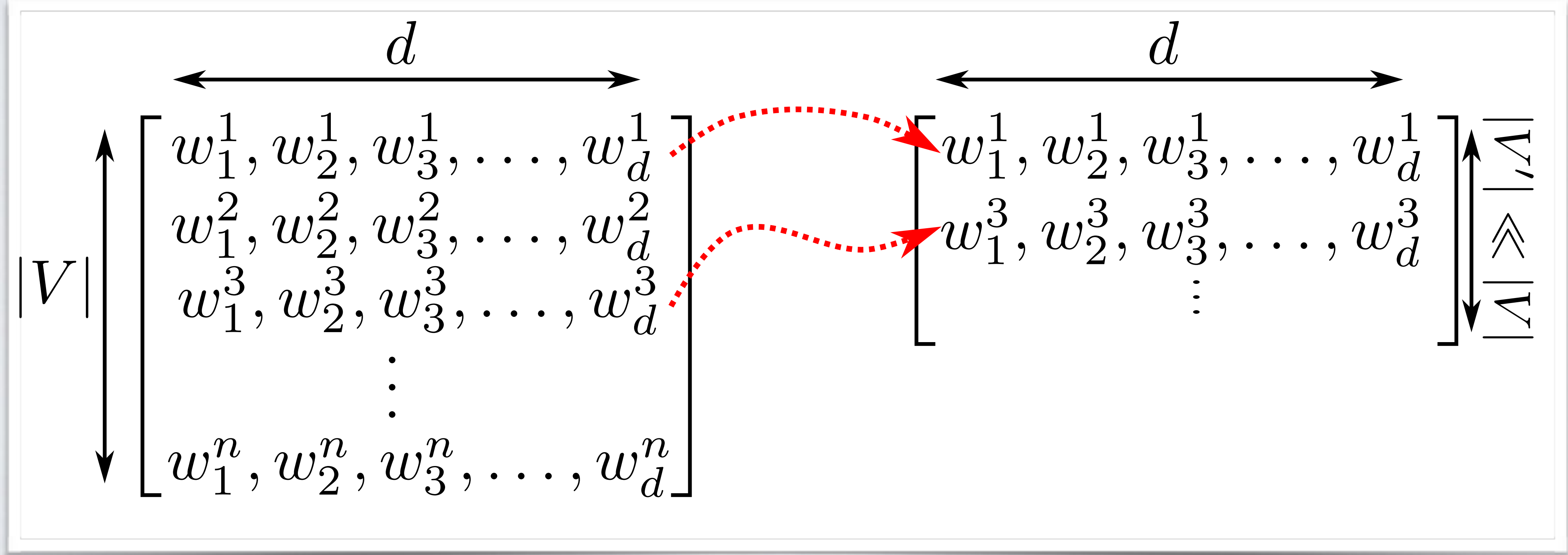
NEURAL MACHINE TRANSLATION

Topics: Very large target vocabulary (Jean et al., 2015)

- (Biased) Importance Sampling *without Sampling*

$$p(y_t \mid y_{<t}, x) = \frac{\exp\left\{w_t^\top \phi\left(y_{t-1}, z_t, c_t\right)\right\}}{\sum_{k:\mathbf{y_k}\in\mathbf{V}} \exp\left\{w_k^\top \phi\left(y_{t-1}, z_t, c_t\right)\right\}}$$
$$\approx \frac{\exp\left\{w_t^\top \phi\left(y_{t-1}, z_t, c_t\right)\right\}}{\sum_{k:\mathbf{y_k}\in\mathbf{V'}} \exp\left\{w_k^\top \phi\left(y_{t-1}, z_t, c_t\right)\right\}}$$

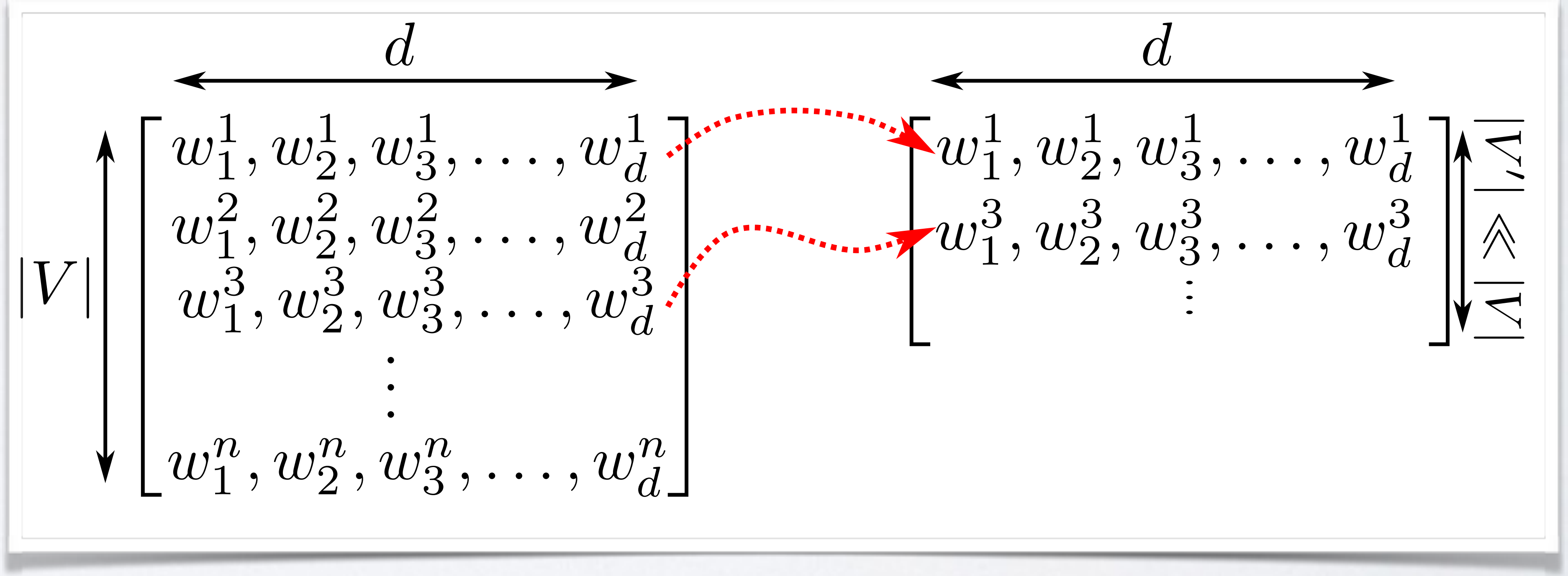




NEURAL MACHINE TRANSLATION

Topics: Very large target vocabulary (Jean et al., 2015)

- How we do choose V'?
  - Training Time:
    - Divide a training corpus into D subsets
    - Build a vocabulary V' for each subset separately
  - Test Time:
    - K-most frequent words
    - K' words that are aligned to source words



NEURAL MACHINE TRANSLATION

Topics: Very large target vocabulary (Jean et al., 2015)

(a) **English→French (WMT-14)**

	NMT(A)	Google	P-SMT
NMT	32.68	30.6*	**37.03•**
+Cand	33.28	–	
+UNK	33.99	32.7°	
+Ens	**36.71**	**36.9°**	

(b) **English→German (WMT-15)**

Model	Note
24.8	Neural MT
24.0	U.Edinburgh, Syntactic SMT
23.6	LIMSI/KIT
22.8	U.Edinburgh, Phrase SMT
22.7	KIT, Phrase SMT

(c) **English→Czech (WMT-15)**

Model	Note
18.3	Neural MT
18.2	JHU, SMT+LM+OSM+Sparse
17.6	CU, Phrase SMT
17.4	U.Edinburgh, Phrase SMT
16.1	U.Edinburgh, Syntactic SMT

NEURAL MACHINE TRANSLATION

Topics: Subword-level Machine Translation (Sennrich et al., 2015)

- Character n-grams (byte pair encoding) [+ Frequent words]

system	sentence
source	health research institutes
reference	Gesundheitsforschungsinstitute
WDict	Forschungsinstitute
C2-50k	Fo|rs|ch|un|gs|in|st|it|ut|io|ne|n
BPE-60k	Gesundheits|forsch|ungsinstitu|ten
BPE-J90k	Gesundheits|forsch|ungsin|stitute
source	asinine situation
reference	dumme Situation
WDict	asinine situation → UNK → asinine
C2-50k	as|in|in|e situation → As|in|en|si|tu|at|io|n
BPE-60k	as|in|line situation → A|in|line-|Situation
BPE-J90K	as|in|line situation → As|in|in-|Situation

Table 6: English→German translation examples. "\|" marks subword boundaries.

system	sentence
source	Mirzayeva
reference	Мирзаева (Mirzaeva)
WDict	Mirzayeva → UNK → Mirzayeva
C2-50k	Mi|rz|ay|ev|a → Ми|рз|ае|ва (Mi|rz|ae|va)
BPE-60k	Mirz|ayeva → Мир|за|ева (Mir|za|eva)
BPE-J90k	Mir|za|yeva → Мир|за|ева (Mir|za|eva)
source	rakfisk
reference	ракфиска (rakfiska)
WDict	rakfisk → UNK → rakfisk
C2-50k	ra|kf|is|k → ра|кф|ис|к (ra|kf|is|k)
BPE-60k	rak|f|isk → пра|ф|иск (pra|f|isk)
BPE-J90k	rak|f|isk → рак|ф|иска (rak|f|iska)

Table 7: English→Russian translation examples. "\|" marks subword boundaries.

NEURAL MACHINE TRANSLATION

Topics: Subword-level Machine Translation (Sennrich et al., 2015)

- Character n-grams (byte pair encoding) [+ Frequent words]

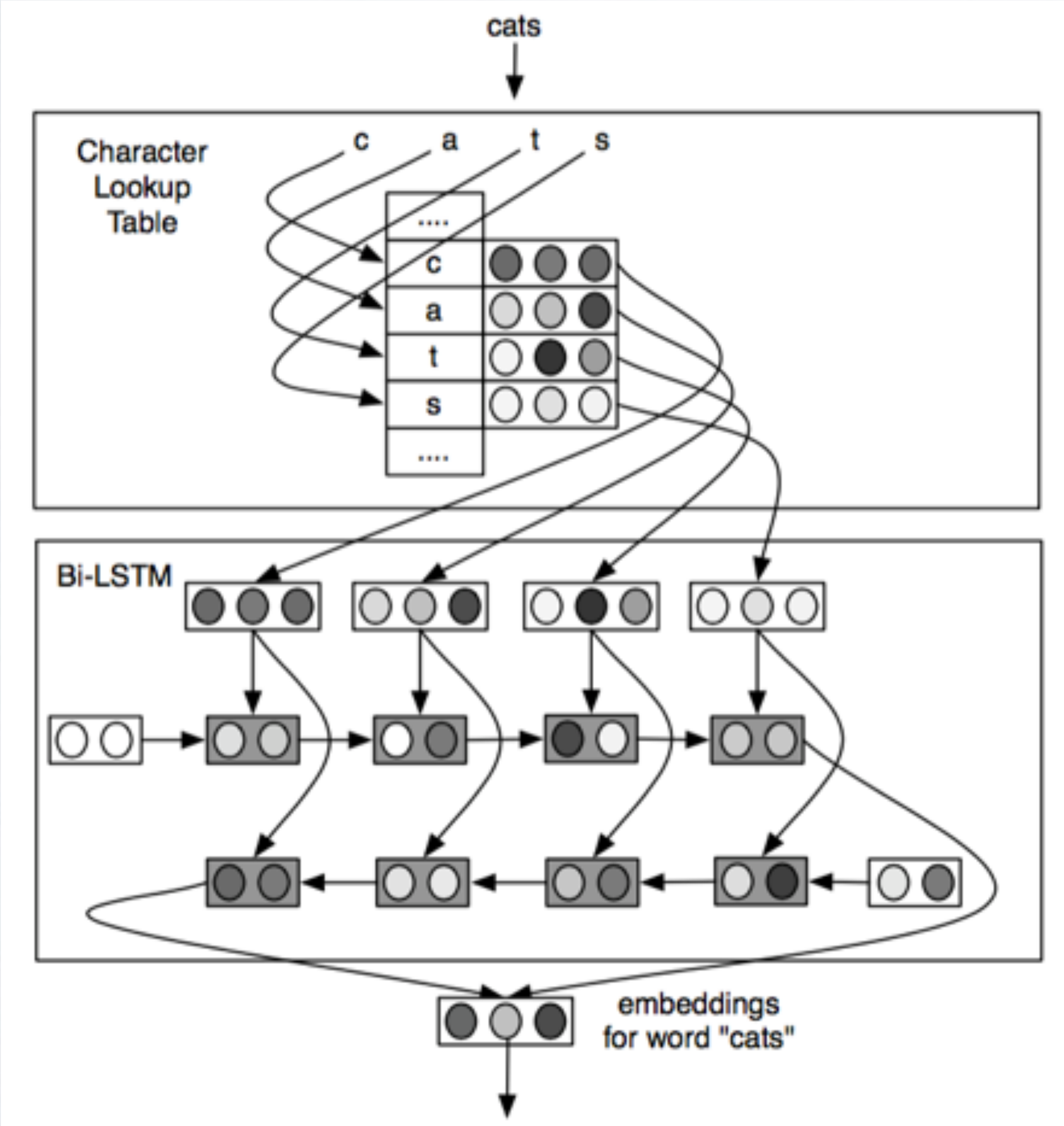
			vocabulary		BLEU			
					newstest2014		newstest2015	
name	segmentation	shortlist	source	target	single	ens-4	single	ens-4
syntax-based (Sennrich and Haddow, 2015)					22.6	-	24.4	-
WUnk	-	-	300 000	500 000	17.1	18.8	19.9	21.7
WDict	-	-	300 000	500 000	18.1	19.9	21.1	23.1
MDict	morfessor	-	300 000	500 000	18.1	20.0	20.5	22.7
C2-3/500k	char-bigrams	3/500 000	310 000	510 000	18.4	20.3	21.8	23.0
C2-50k	char-bigrams	50 000	60 000	60 000	18.7	20.7	21.9	**23.9**
C3-50k	char-trigrams	50 000	100 000	100 000	18.9	20.5	21.5	**23.9**
BPE-60k	BPE	-	60 000	60 000	18.6	**20.8**	21.1	23.6
BPE-J90k	BPE (joint)	-	90 000	90 000	**19.4**	**20.8**	**22.2**	23.7

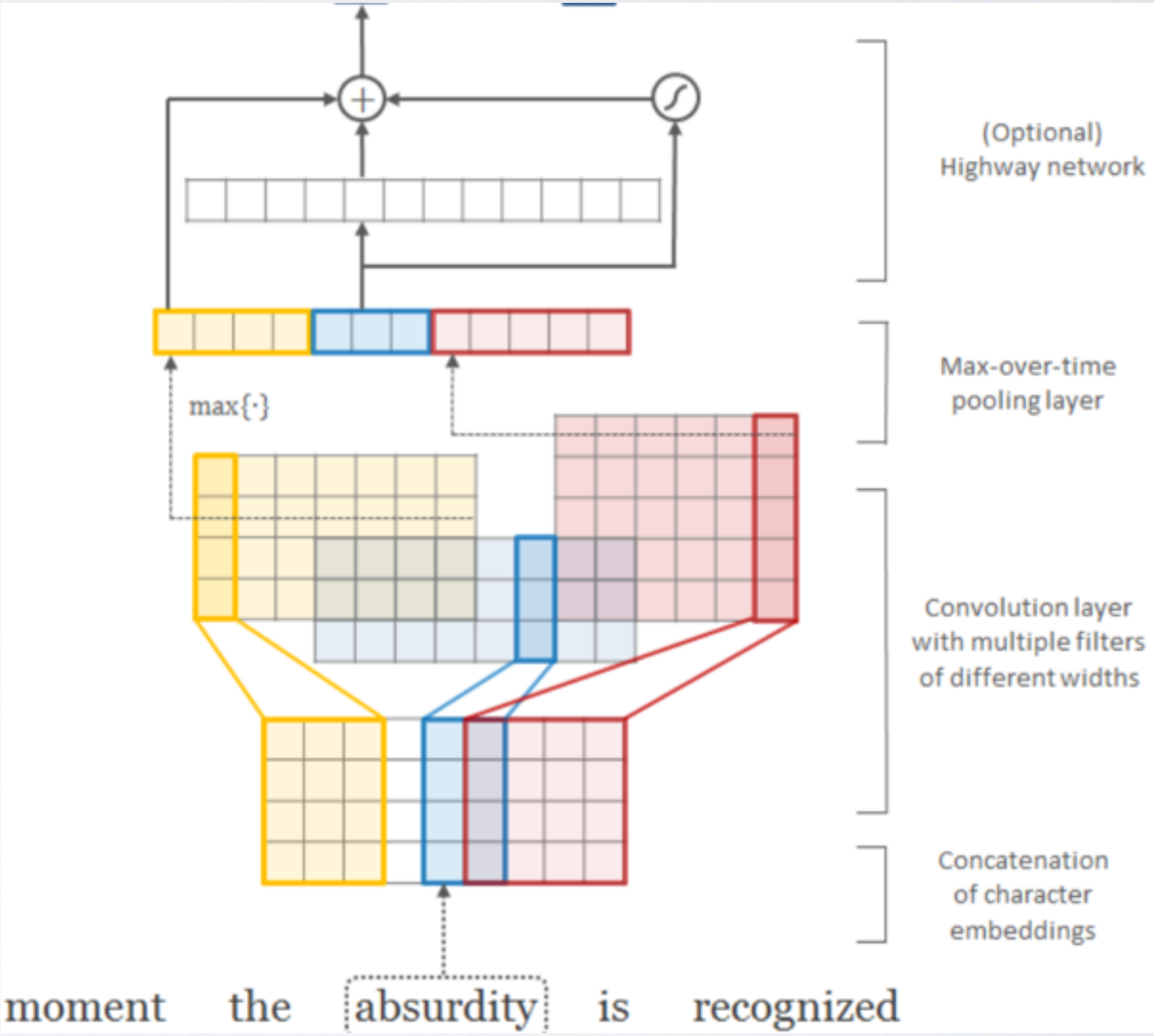
Table 2: English→German translation performance (BLEU) on newstest2014 and newstest2015 test sets. Ens-4: ensemble of 4 models. Best NMT system in bold.

NEURAL MACHINE TRANSLATION

Topics: Subword-level Language Modelling (Kim et al., 2015; Ling et al., 2015)

- Directly processing characters





NEURAL MACHINE TRANSLATION

Topics: Very large target vocabulary (Jean et al., 2015)

(d) German→English (WMT-15)

Model	Note
29.3	U.Edinburgh, Phrase SMT
29.1	KIT, Phrase SMT
28.9	Aachen, Phrase SMT
28.7	JHU, Phrase SMT
28.7	U.Edinburgh, Syntax SMT
27.9	Neural MT

(e) Czech→English (WMT-15)

Model	Note
26.2	JHU, Phrase SMT
24.5	U.Edinburgh, Syntax SMT
23.8	Neural MT
20.4	Dublin, Phrase SMT
14.5	Illinois

(f) Finnish→English (WMT-15)

Model	Note
17.9	U.Edinburgh, Syntax SMT
17.6	Dublin, Rule-based SMT
16.4	Uppsala, Phrase SMT
15.9	Prompsit Language Engineering
15.7	Illinois
13.6	Neural MT

Is neural MT particularly weak when translating to English?

NEURAL MACHINE TRANSLATION

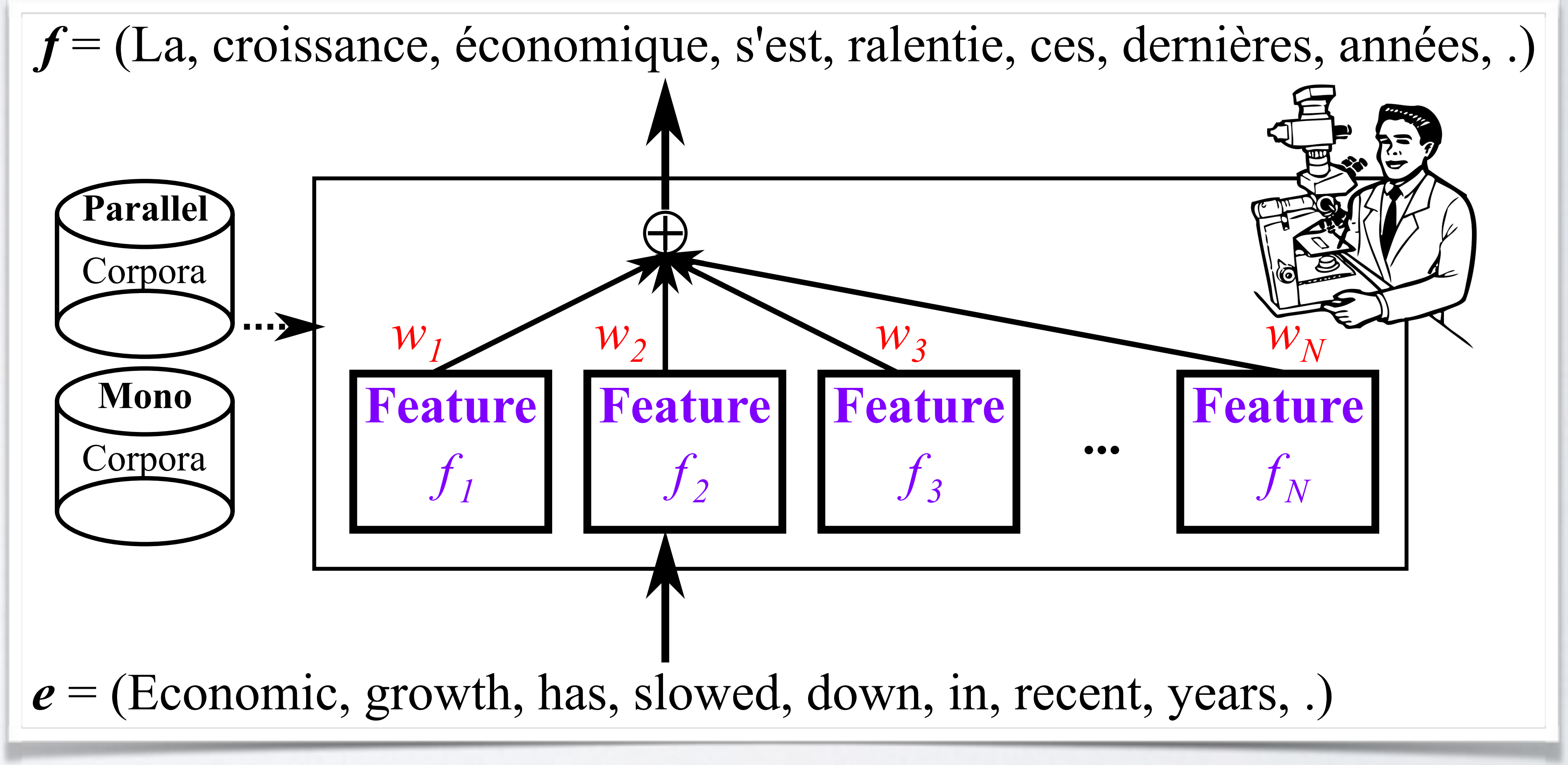
Topics: Statistical Machine Translation - Recap

- $\log p(f|e) \approx \sum_{n=1}^{N} f_n(e,f) + C$
  - Log-linear model
  - Feature function $f_n(e,f)$
- Steps:

  (1) Experts engineer *useful* features

  (2) Use a simple log-linear model

  (3) Use a strong, external language model

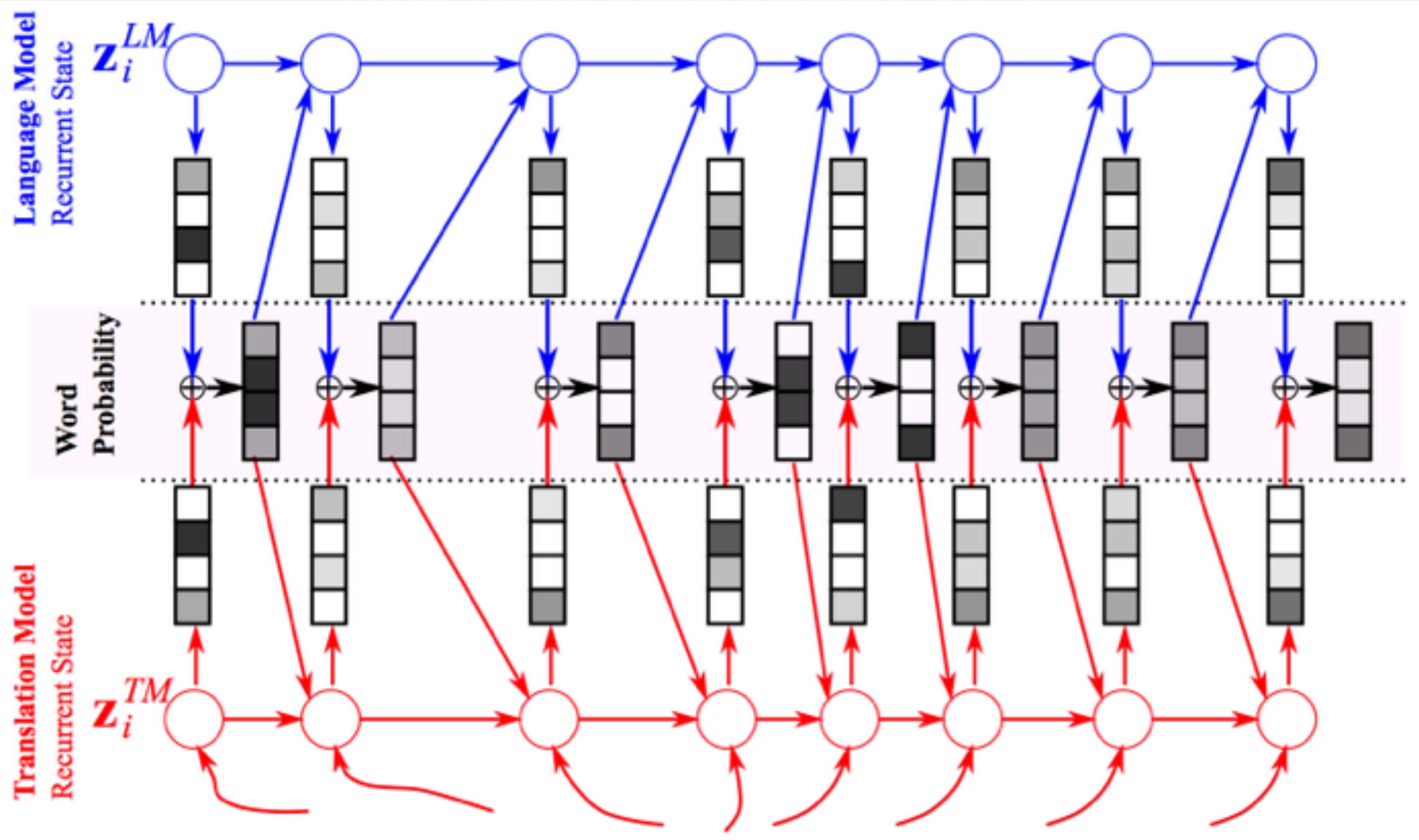


NEURAL MACHINE TRANSLATION

Topics: Incorporating Target Language Model (Gulcehre&Firat et al., 2015)

- Shallow Fusion: Log-Linear Interpolation between TM and LM

$$\log p(y_t|y_{<t},x) = \log p^{\mathrm{TM}}(y_t|y_{<t},x) + \beta \log p^{\mathrm{LM}}(y_t|y_{<t})$$



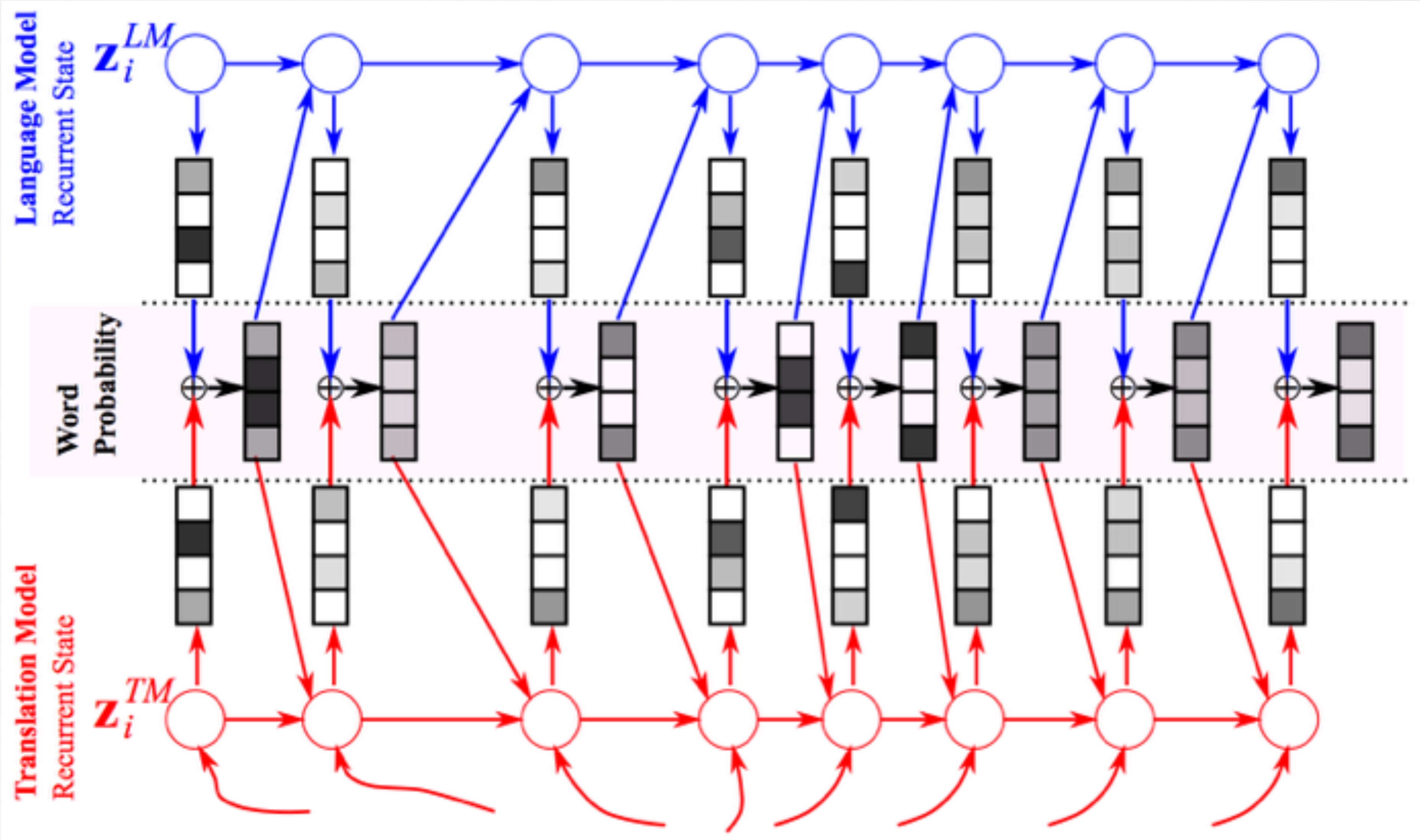
NEURAL MACHINE TRANSLATION

Topics: Incorporating Target Language Model (Gulcehre&Firat et al., 2015)

- Shallow Fusion: Log-Linear Interpolation between TM and LM

$$\log p(y_t|y_{<t}, x) = \log p^{\mathrm{TM}}(y_t|y_{<t}, x) + \beta \log p^{\mathrm{LM}}(y_t|y_{<t})$$

- Advantages:
  - Single tunable parameter β
- Disadvantages:
  - Is is really ***linear?***

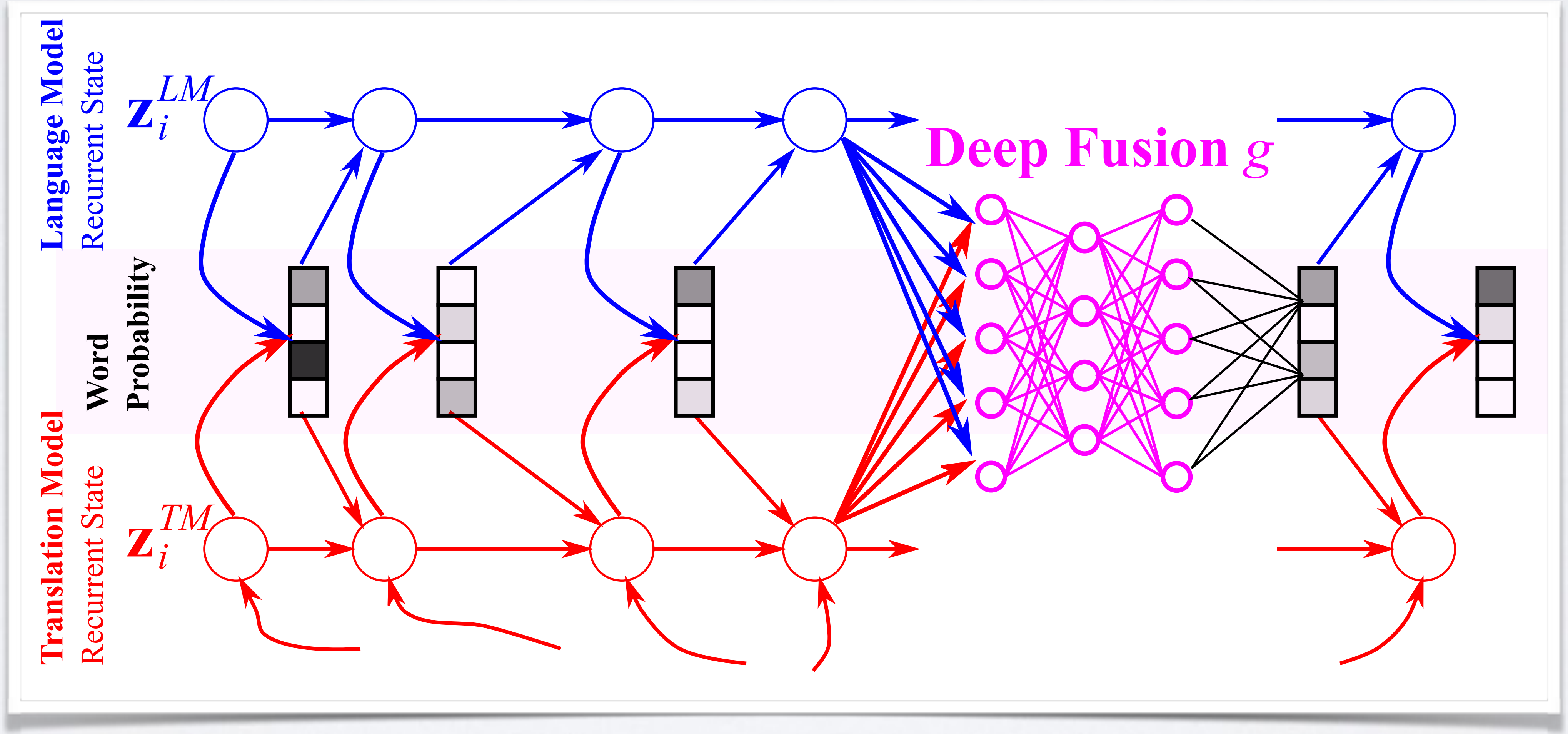


NEURAL MACHINE TRANSLATION

Topics: Incorporating Target Language Model (Gulcehre&Firat et al., 2015)

- Deep Fusion: Nonlinear interpolation between LM and TM

$$p(y_t|y_{<t},x) \propto \exp(y_t^\top(W_o f_{o,\theta}(z_t^{\text{LM}}, g_t \cdot z_t^{\text{TM}}, y_{t-1}, c_t) + b_o))$$



NEURAL MACHINE TRANSLATION

Topics: Incorporating Target Language Model (Gulcehre&Firat et al., 2015)

- Deep Fusion: Nonlinear interpolation between LM and TM

$$p(y_t|y_{<t}, x) \propto \exp(y_t^\top (W_o f_{o,\theta}(z_t^{\text{LM}}, g_t \cdot z_t^{\text{TM}}, y_{t-1}, c_t) + b_o))$$

- Advantages
  - No linearity assumed: the core philosophy of deep learning
  - Context-Dependent Fusion
- Disadvantages
  - Works only with a continuous-space LM: NLM or RNN-LM
  - Computationally demanding (comparatively to shallow fusion)

NEURAL MACHINE TRANSLATION

Topics: Deep Fusion of Target Language Model (Gulcehre&Firat et al., 2015)

(a) **German/Czech→English (WMT-15)**

	De-En	Cs-En
Neural MT	23.61	21.89
+Shallow	23.69	22.18
+Deep	**24.00**	**22.36**

(b) **Chinese*→English (OpenMT-15)**

	SMS/Chat	CTS
Phrase SMT	14.73	21.68
Hiero SMT	14.71	21.43
Neural MT	17.36	**23.59**
+Shallow	16.42	22.83
+Deep	**17.64**	23.5

(b) **Turkish→English (IWSLT-2014)**

	tst2011	tst2012	tst2013	Test 2014
Previous Best	18.83	18.93	18.70	-
Neural MT	18.40	18.77	19.86	18.64
+Shallow	18.48	18.80	19.87	18.66
+Deep	**20.17**	**20.23**	**21.34**	**20.56**

Neural MT is comparable to, or better than, phrase-based MT

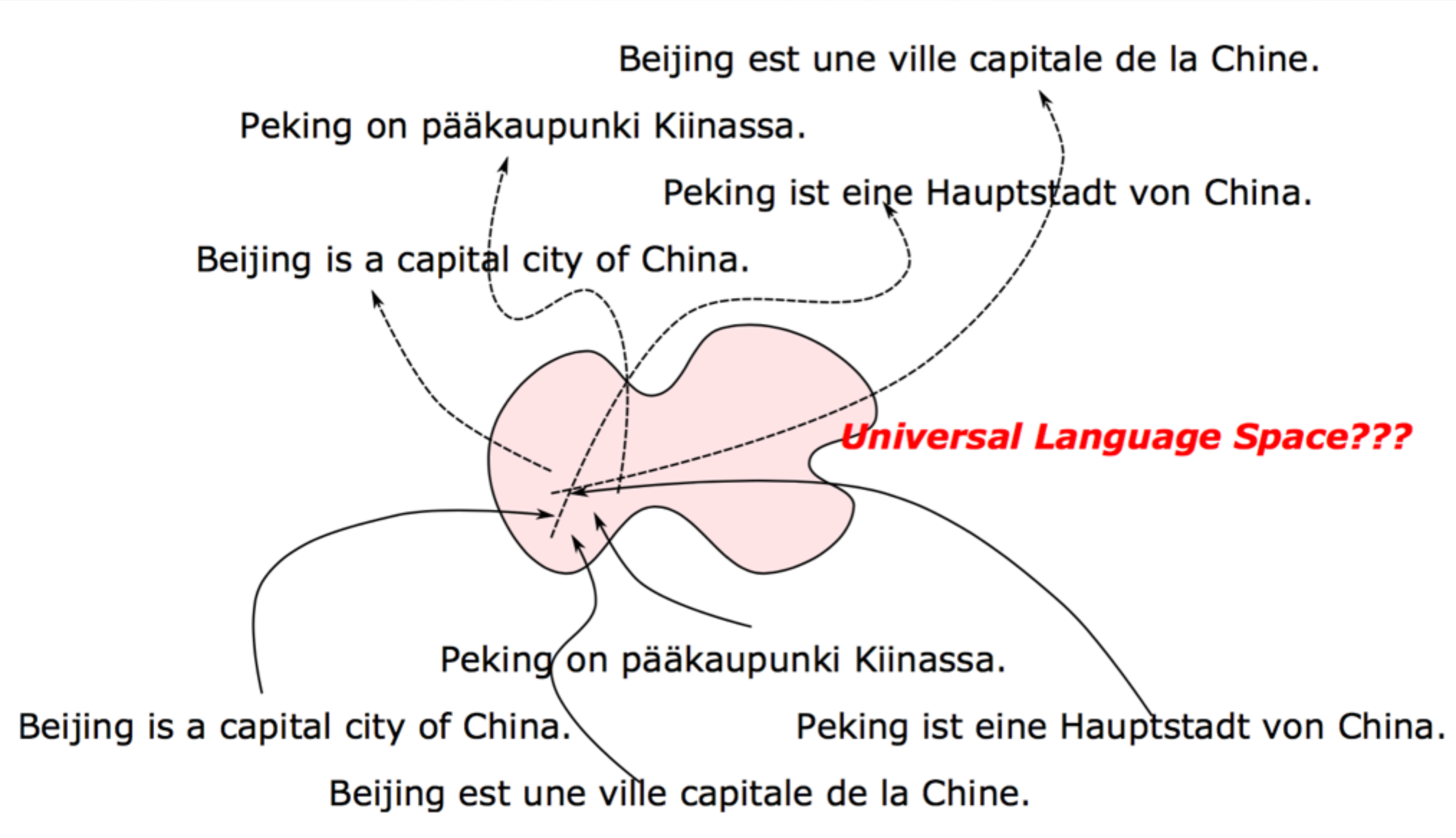
- Multi-task learning for multiple language translation (Dong et al., 2015)
- Neural Machine Translation of Rare Words with Subword Units (Sennrich et al., 2015)
- Variable-Length Word Encodings for Neural Translation Models (Chitnis&DeNero, 2015)
- Addressing the rare word problem in neural machine translation (Luong et al., 2015)
- Effective Approaches to Attention-based Neural Machine Translation (Luong et al., 2015)
- *and the list continues…*

.. an extremely promising approach to MT through .. deep learning ..

Advances in natural language processing
by Hirschberg & Manning (2015)

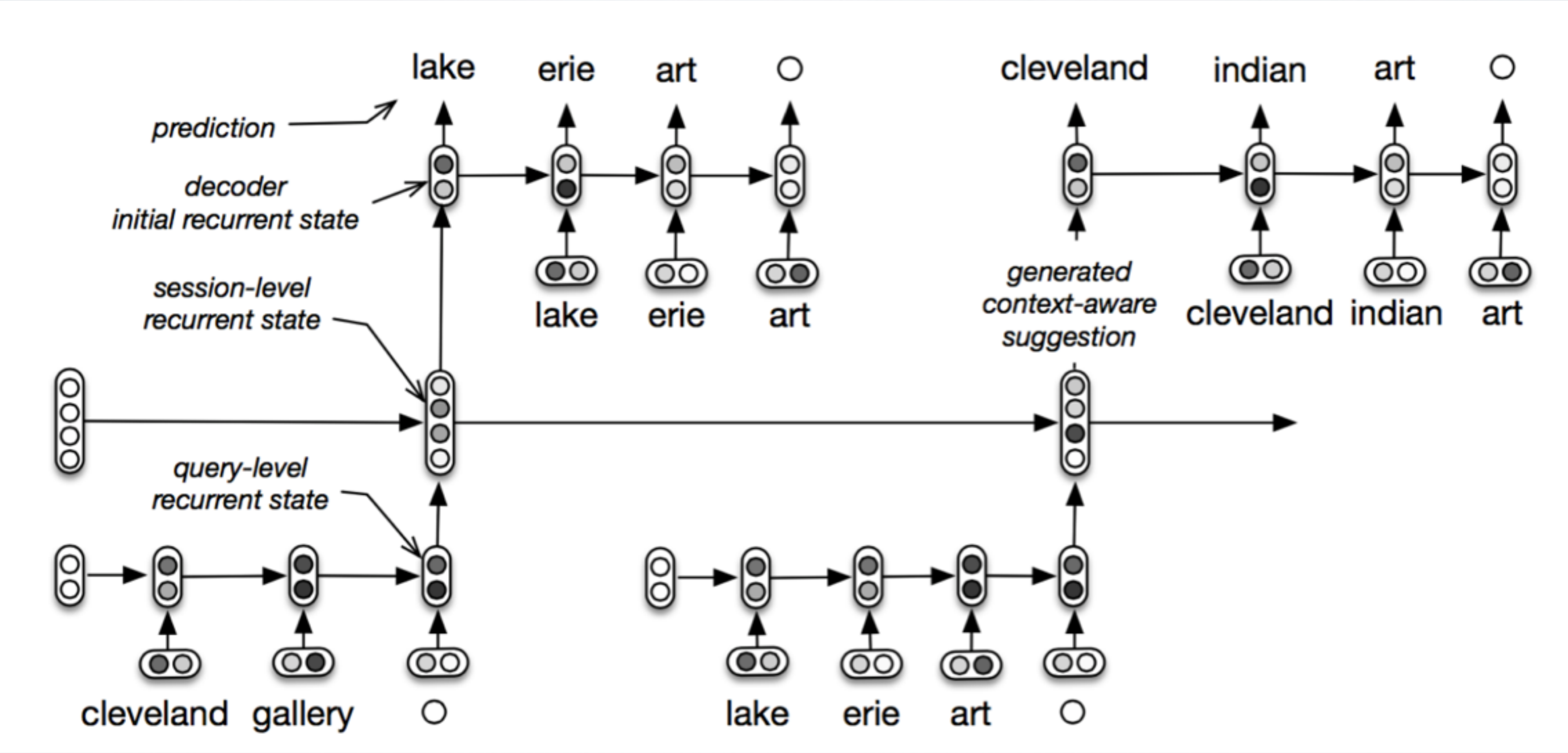
What next?

MULTILINGUAL TRANSLATION



Dong et al. (2015)

TOWARD DISCOURSE-LEVEL MT



Hierarchical Recurrent Encoder–Decoder (HRED) by Sordoni et al. (2015)

Neural MT beyond MT

- Memory Networks (Weston et al., 2014)
- Neural Turing Machines (Graves et al., 2014)
- Pointer Networks (Vinyals et al., 2015)
- Grammar as a Foreign Languages (Vinyals et al., 2014)
- Teaching machines to read and comprehend (Hermann et al., 2015)
- Reasoning about Entailment with Neural Attention (Rocktaschel et al., 2015)
- *and the list continues…*

Any supervised learning task is a translation task

Going beyond Natural Languages

Is a human language special?

BEYOND NATURAL LANGUAGES

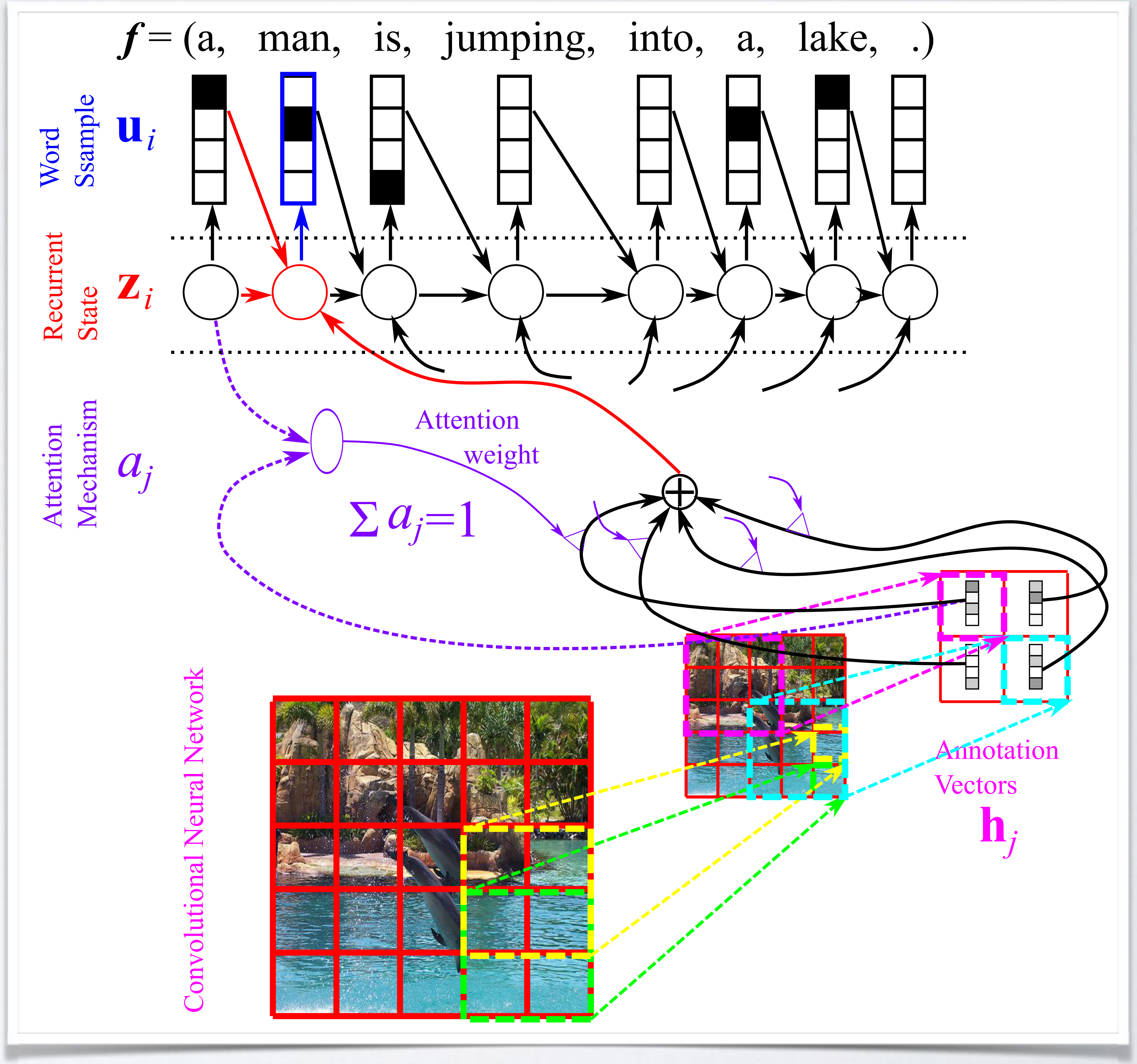
Topics: Beyond Natural Languages

— Image Caption Generation

- Task: *conditional* language modelling

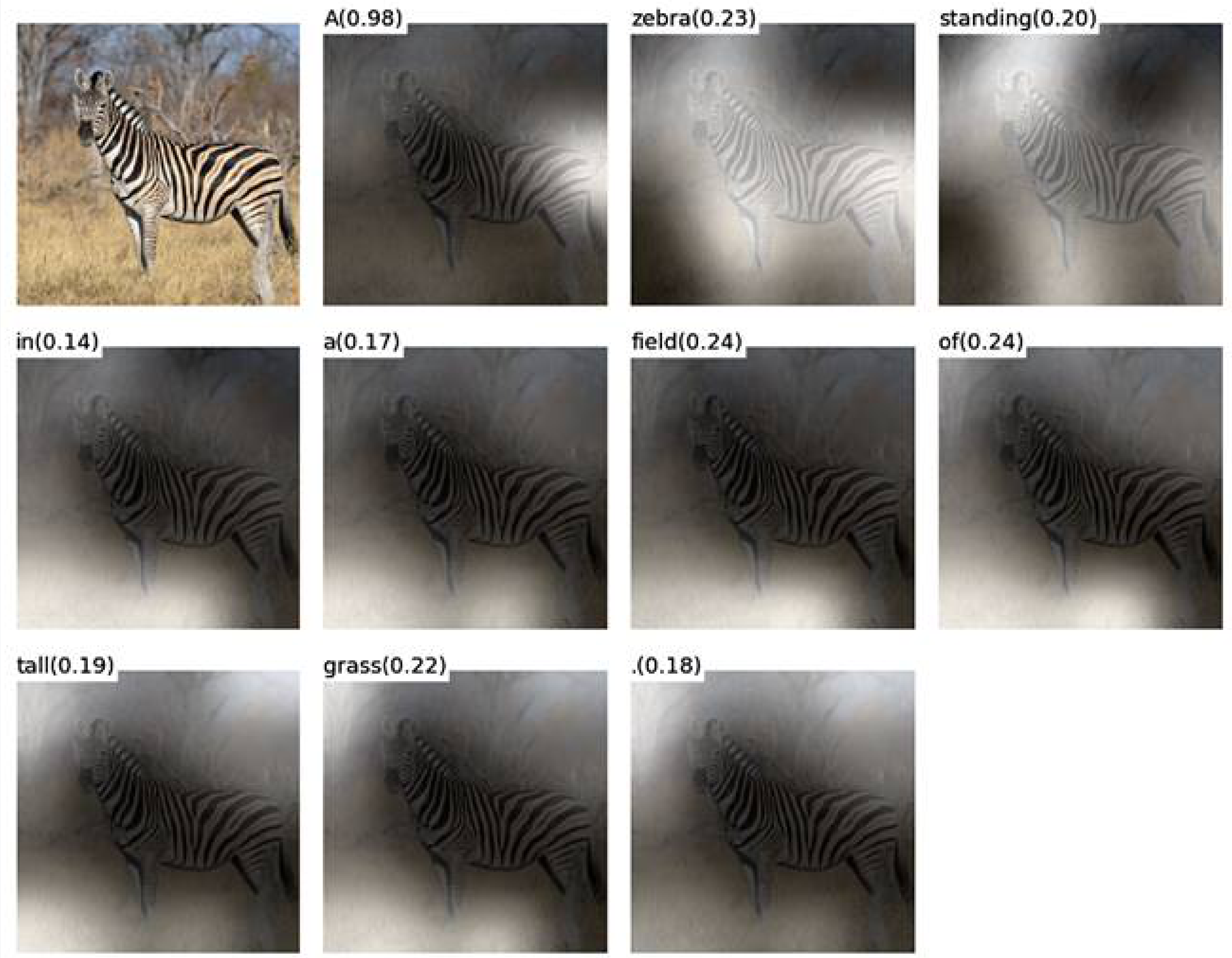
$p(\text{Two}, \text{dolphins}, \text{are}, \text{diving}|$  $) = ?$

- Encoder: convolutional network
  - Pretrained as a classifier or autoencoder
- Decoder: recurrent neural network
  - RNN Language model
  - With attention mechanism (Xu et al., 2015)



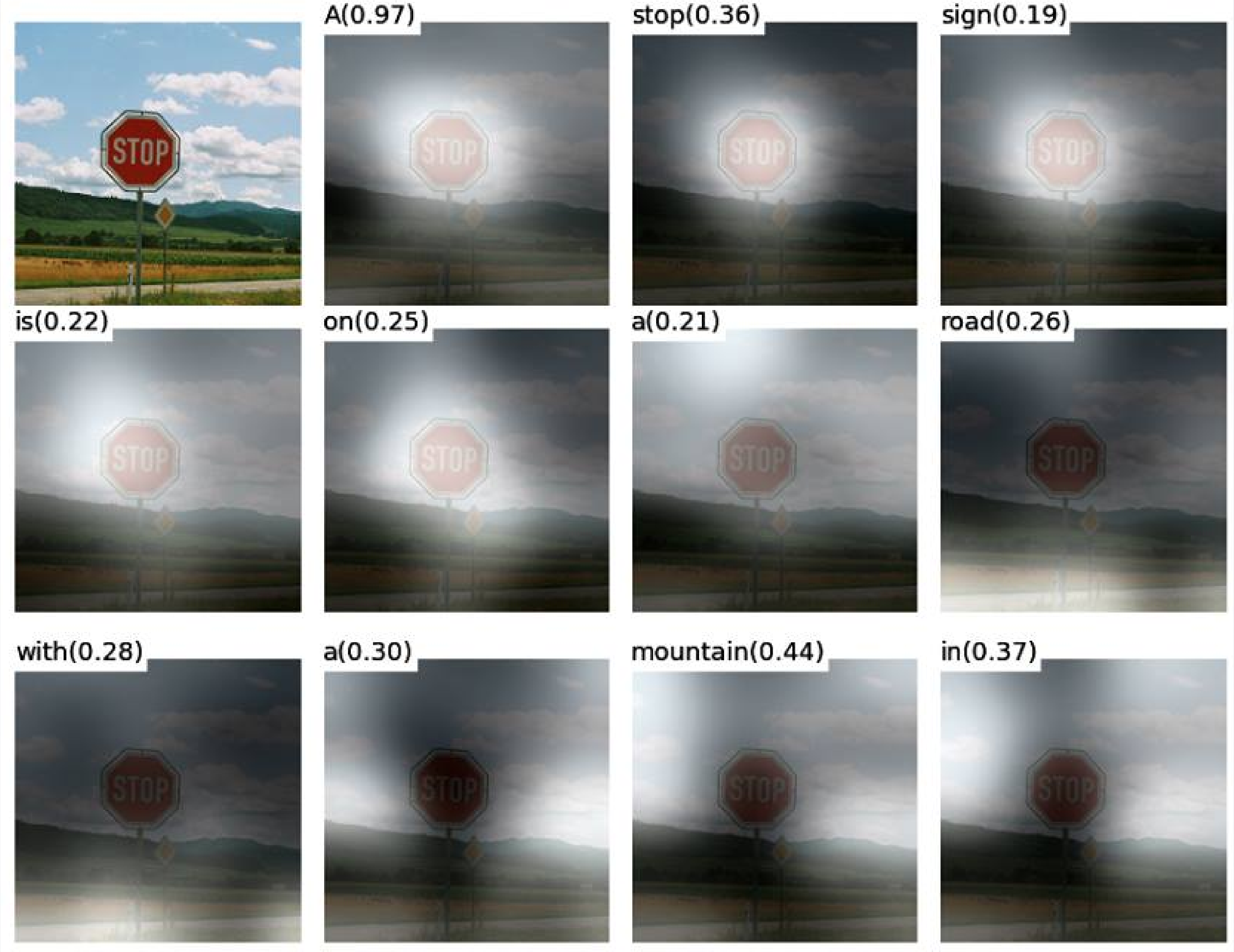
BEYOND NATURAL LANGUAGES

Topics: Beyond Natural Languages — Image Caption Generation (Examples)



BEYOND NATURAL LANGUAGES

Topics: Beyond Natural Languages — Image Caption Generation (Examples)

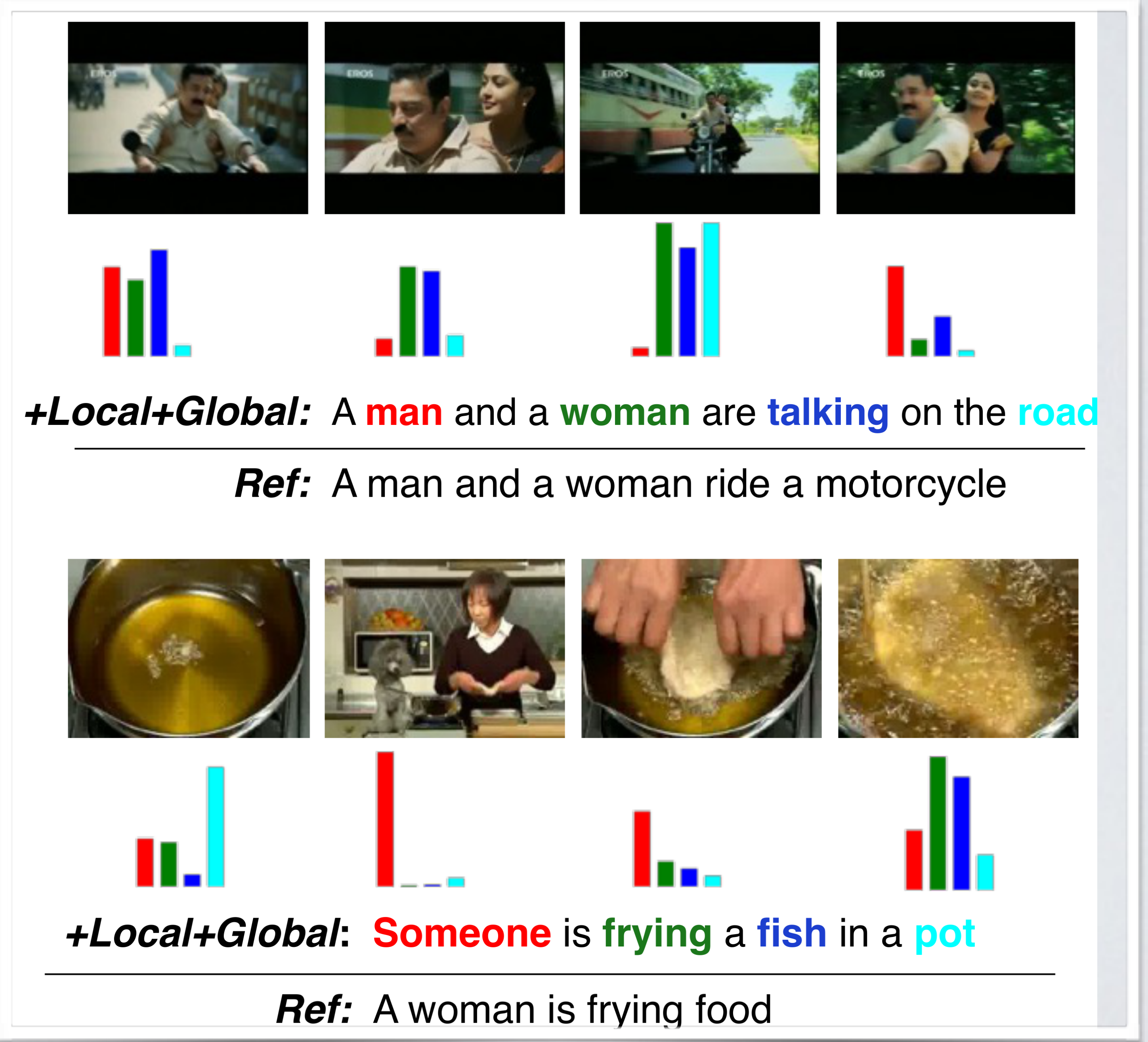


BEYOND NATURAL LANGUAGES

Topics: Beyond Natural Languages — Attention Models

- End-to-End Speech Recognition (Chorowski et al., 2015; Chan et al., 2015)
- Video Description Generation (Yao et al., 2015)
- Discrete Optimization (Vinyals et al., 2015)
- and many more...

  (Cho et al., 2015) and references therein



NEW YORK UNIVERSITY

NYU | GRADUATE SCHOOL OF ARTS & SCIENCE

NYU | COURANT INSTITUTE OF MATHEMATICAL SCIENCES

- **Department of Computer Science**
  - Ph.D. Programme: Application dl. 12th December
- **Center for Data Science**
  - M.Sc. Programme in Data Science: Application dl. 4th Februrary

Teaching Machines to Read, Comprehend and Answer

Based on (Hermann et al., 2015; Blunsom, 2015)

READING COMPREHENSION

Topics: Teaching machines to read and comprehend

CNN article:

Document	The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." ...
Query	Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.
Answer	Oisin Tymon

READING COMPREHENSION

Topics: Teaching machines to read and comprehend

— *Deep LSTM Reader*

- Document Reader

  $h_t = f(h_{t-1}, w_t)$, for all $t = 1, \ldots, T$

  - Summary of the document: h_T

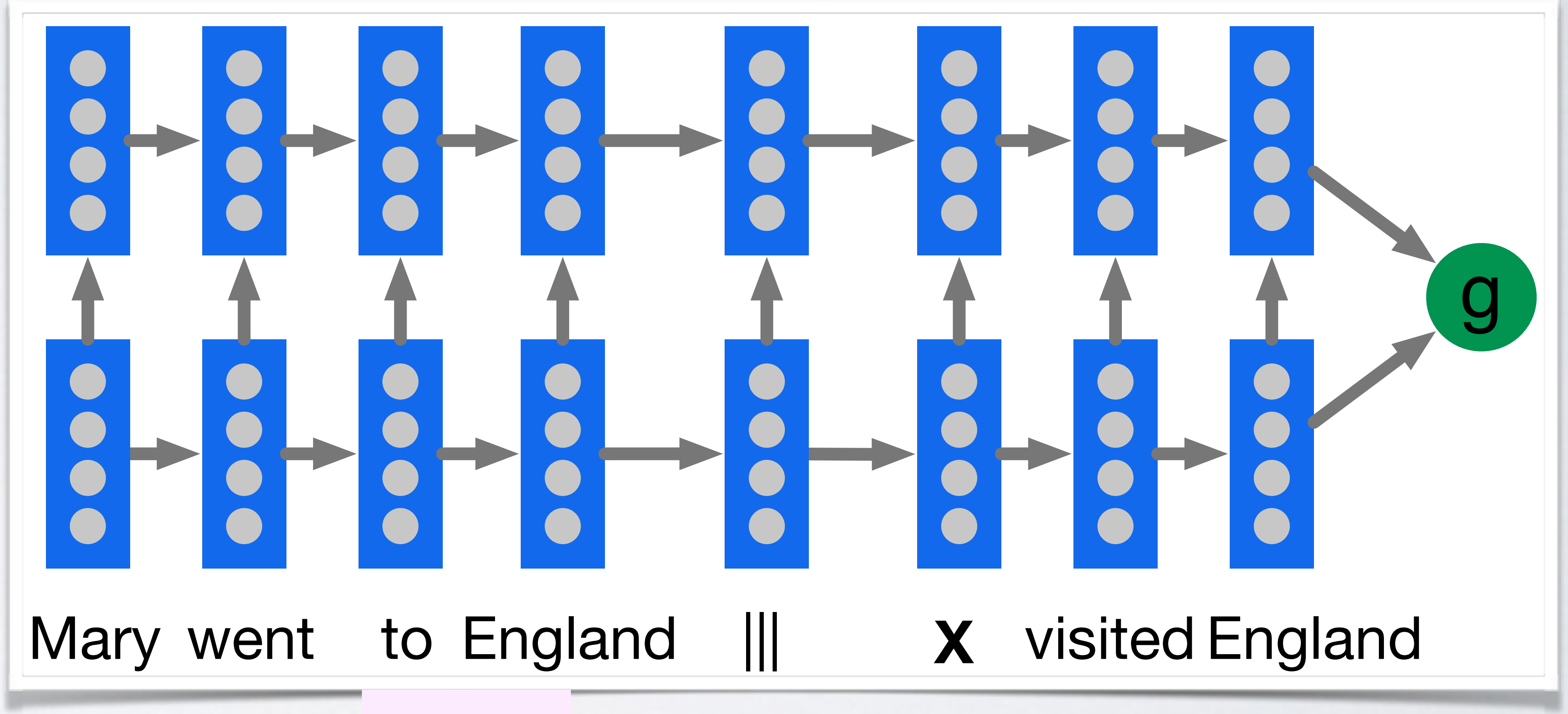
- Query Reader

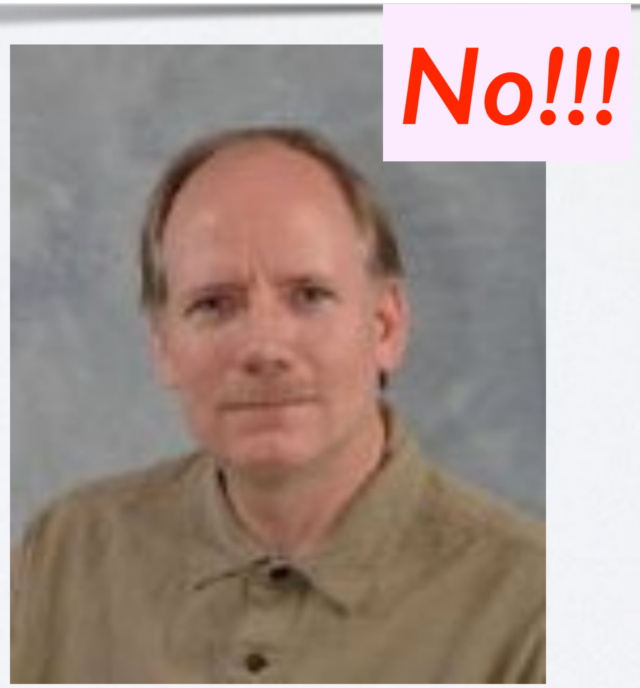
  $z_t = f(z_{t-1}, w'_t)$, for all $t = 1, \ldots, T'$

  - Summary of the query: $z_{T'}$

- Answer selection

  $p(a|\{w_t\}_{t=1}^{T}, \{w_{t'}\}_{t'=1}^{T'}) = g_a(h_T, z_T)$

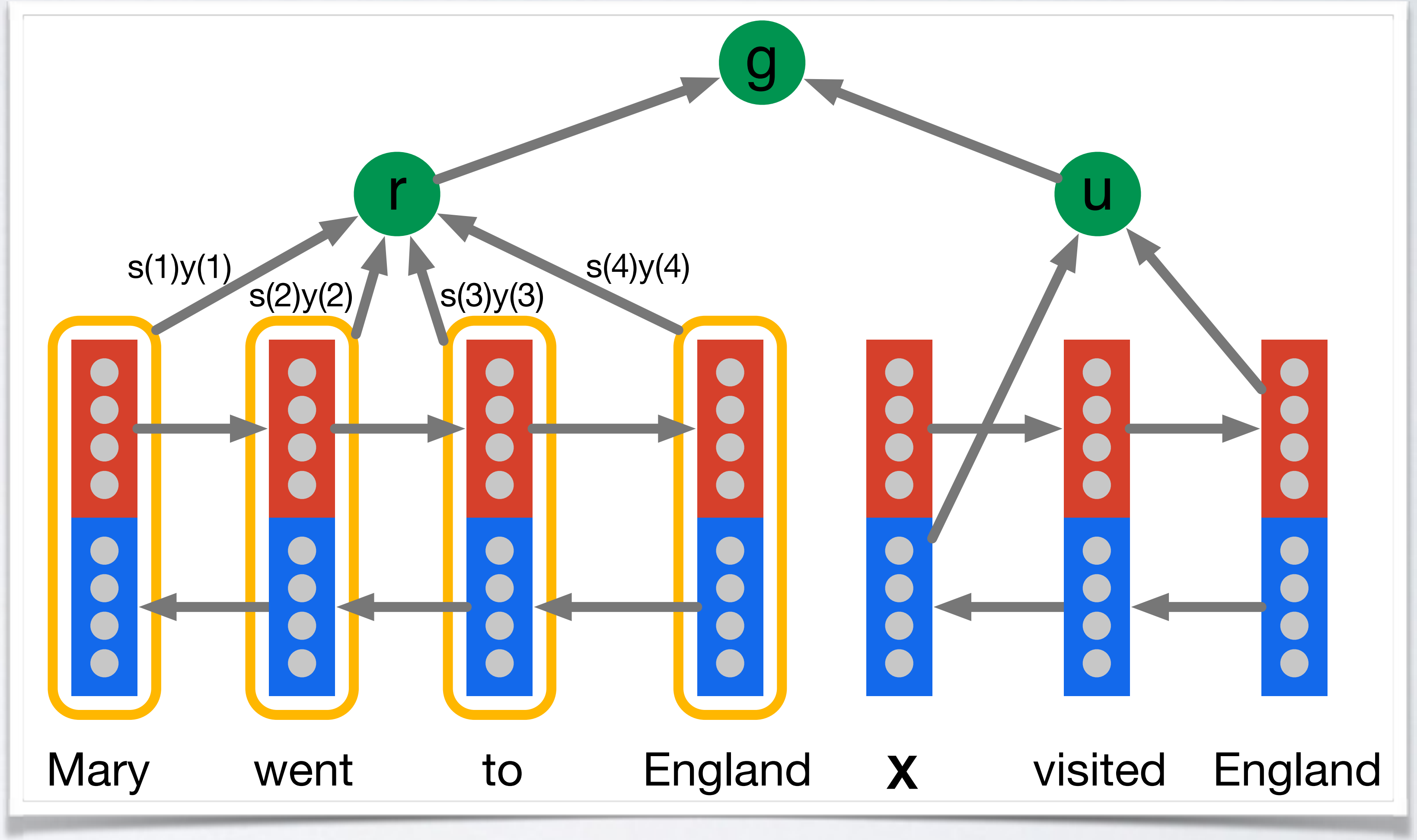




READING COMPREHENSION

Topics: Teaching machines to read and comprehend

— *Attentive Reader*



- Document Reader: BiRNN
  - Annotation vectors: $\{h_1, h_2, \ldots, h_T\}$
- Query Reader: $z_{T'}$
- Answer selection
  - Attention mechanism $\alpha_t \propto e(h_t, z_{T'})$
  - Query-dependent document summary $c = \sum_{t=1}^{T} \alpha_t h_t$
  - Answer selection: $p(a|\{w_t\}_{t=1}^{T}, \{w_{t'}\}_{t'=1}^{T'}) = g_a(z_{T'}, c)$

READING COMPREHENSION

Topics: Teaching machines to read and comprehend

— *Attentive Reader* (Examples)

- Visualize the attention

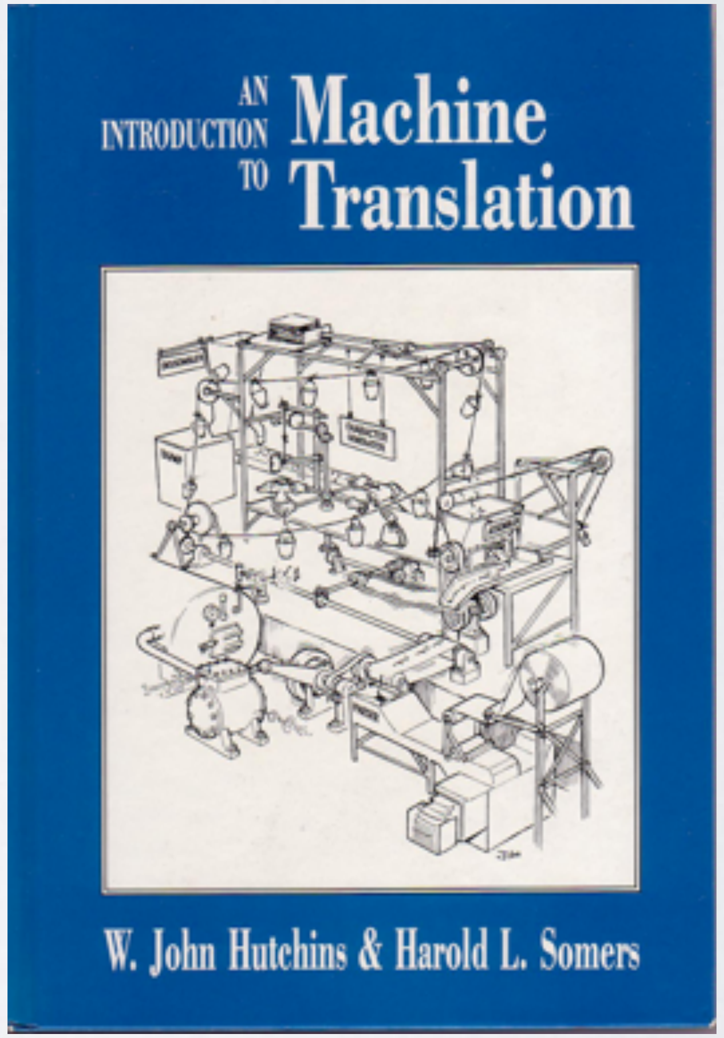
by *ent40* ,*ent62* correspondent updated 9:49 pm et ,thu march 19 ,2015 (*ent62*) a *ent88* was killed in a parachute accident in *ent87* ,*ent28* ,near *ent66* ,a *ent47* official told *ent62* on wednesday .he was identified thursday as special warfare operator 3rd class *ent49* ,29 ,of *ent44* ,*ent13* .`` *ent49* distinguished himself consistently throughout his career .he was the epitome of the quiet professional in all facets of his life ,and he leaves an inspiring legacy of natural tenacity and focused commitment for posterity ,'' the *ent47* said in a news release .*ent49* joined the seals in september after enlisting in the *ent47* two years earlier .he was married ,the *ent47* said .initial indications are the parachute failed to open during a jump as part of a training exercise .*ent49* was part of a *ent57* - based *ent88* team .

ent47 identifies deceased sailor as **X** ,who leaves behind a wife

Connectionist Approach to Natural Language Understanding

The relevance of the ***connectionist model to natural language processing is clear enough****. The traditional stratificational approach to parsing and generation (morphology, syntax, semantics) .. is not seriously accepted .. as a psychologically real model of how humans understand and communicate.*

Hutchins and Somers (1992)



*With **a neural network**, we don't encode any hard principles. The model **infers the important structures, properties and relationships directly from raw data**, in a way that allows it to best describe achieve its objective.*

Hill (2015)

https://medium.com/@felixhill/deep-consequences-fa823a588e97

CONNECTIONIST NLP

Topics: No such thing as (universal) word embeddings

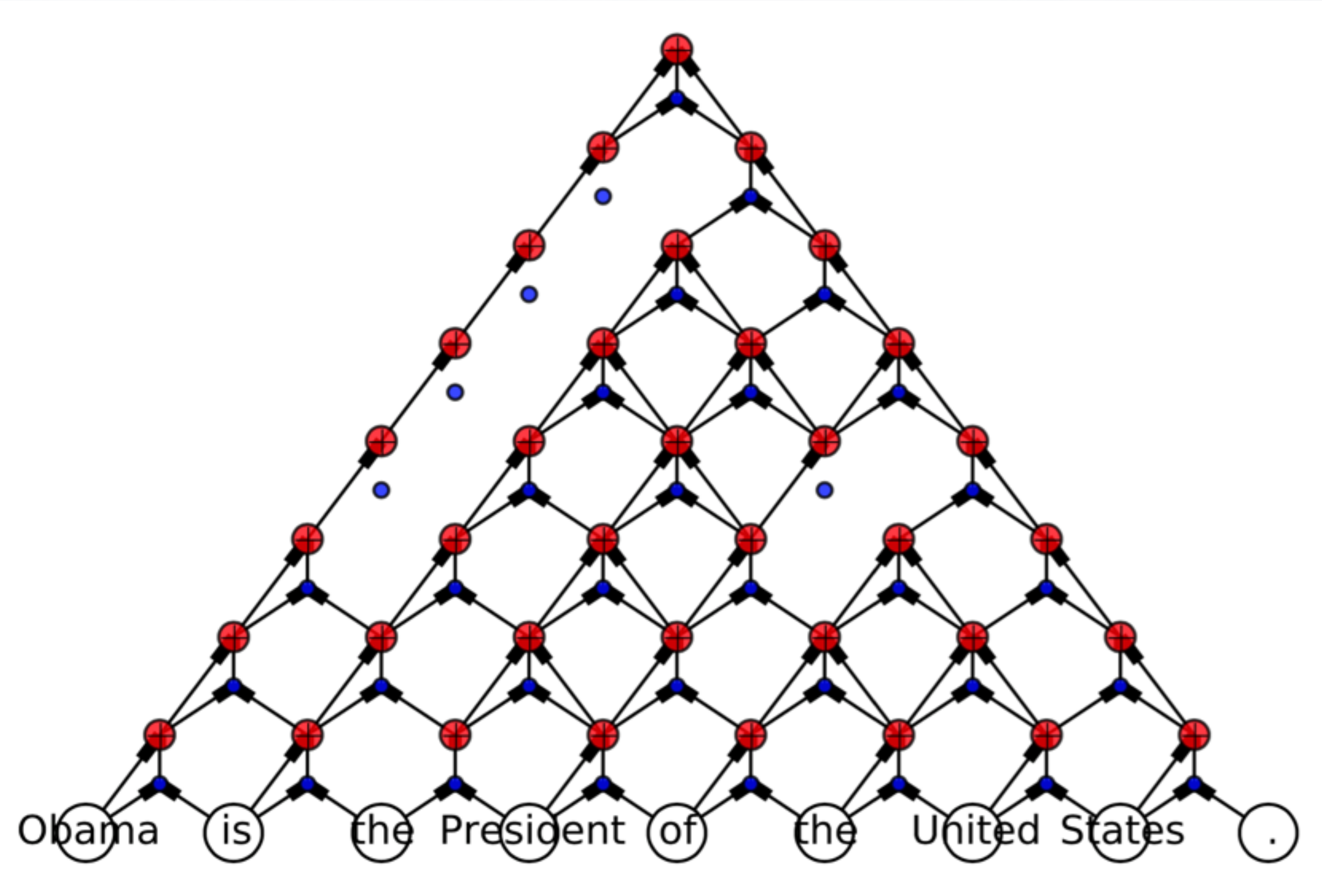
- Word embeddings are nothing but the *first layer weight matrix*
- Objective functions matter a lot (Hill et al., 2014; Hill et al., 2015)

	Skipgram	Glove	CW	FD	Bi[illegible]	RNNenc	RNNsearch
teacher	*vocational*	*student*	*student*	*elementa[illegible]*	*[illegible]ulty*	*professor*	*instructor*
	in-service	*pupil*	*tutor*	[illegible]	*professors*	*instructor*	*professor*
	college	*university*	*ment[illegible]*	*[illegible]oom*	*teach*	*trainer*	*educator*
eaten	*spoiled*	*cooked*	[illegible]	*ate*	*eating*	*ate*	*ate*
	squeezed	*e[illegible]*	*[illegible]eeled*	*meal*	*eat*	*consumed*	*consumed*
	cooked	[illegible]	*cooked*	*salads*	*baking*	*tasted*	*eat*
Britain	*North[illegible]*	*[illegible]land*	*Luxembourg*	*UK*	*UK*	*UK*	*England*
	[illegible]	*Kingdom*	*Belgium*	*British*	*British*	*British*	*UK*
	[illegible]reland	*Great*	*Madrid*	*London*	*England*	*America*	*Syria*

Don't hammer everything with monolingual word embeddings!!!

CONNECTIONIST NLP

Topics: Compositionality naturally arises



Cho et al. (2014)

CONNECTIONIST NLP

Topics: Neural net will capture underlying structures

- As long as the structures are needed to achieve the goal