Recurrent neural networks for machine translation

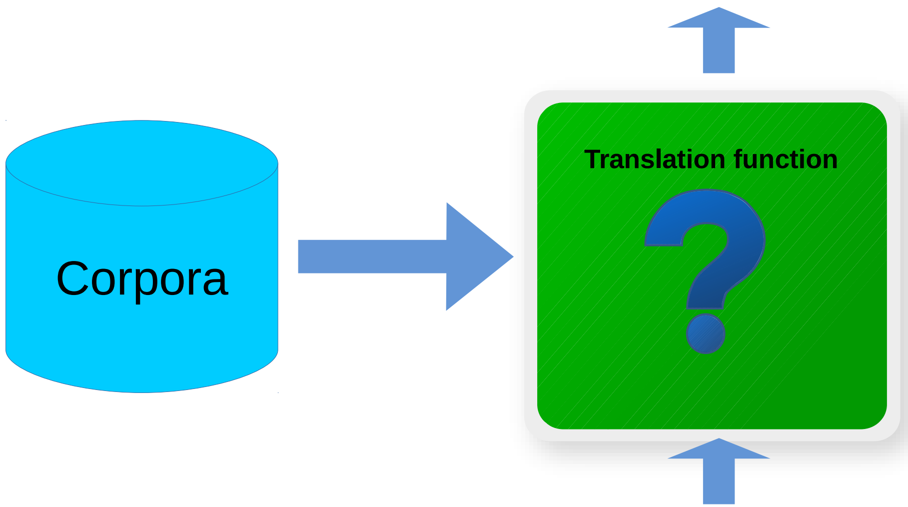
Outline

- Machine translation problem
- Neural machine translation
- Encoder Decoder approach
- Attention-based Neural Machine Translation
- Prominent Neural Machine Translation Models

Statistical Machine Translation

- Translate a source sentence E into a target sentence F
- Set of rules transforming a source sentence into a correct translation
- We don’t even know the set of rules underlying a single language, not to mention the rules underlying a pair of languages.
- Statistical approach where those rules, either implicitly or explicitly, are automatically extracted from a large corpus of text.

f=(La, croissance, économique, s'est, ralenti, ces, dernières, années, .)



e=(economic, growth, has, slowed, down, in, recent, years, .)

Evaluation

- **BLEU** (**B**i**L**ingual **E**valuation **U**nderstudy)
- 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)
- $$BLEU = min\left(1, \frac{output-length}{reference-length}\right)\left(\prod_{i=1}^{4} precision_i\right)^{\frac{1}{4}}$$

Example

SYSTEM A: Israeli officials responsibility of airport safety

2-GRAM MATCH 1-GRAM MATCH

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible

2-GRAM MATCH 4-GRAM MATCH

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%

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 responsibility of airport safety

2-GRAM MATCH   2-GRAM MATCH   1-GRAM

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

Encoder-Decoder Framework for Machine Translation

Encoder-Decoder Framework for Machine Translation

La croissance économique s'est ralenti ces dernières années.

Decode

$$[\, z_1, z_2, \ldots, z_d \,]$$

Encode

economic growth has slowed, down in recent years.

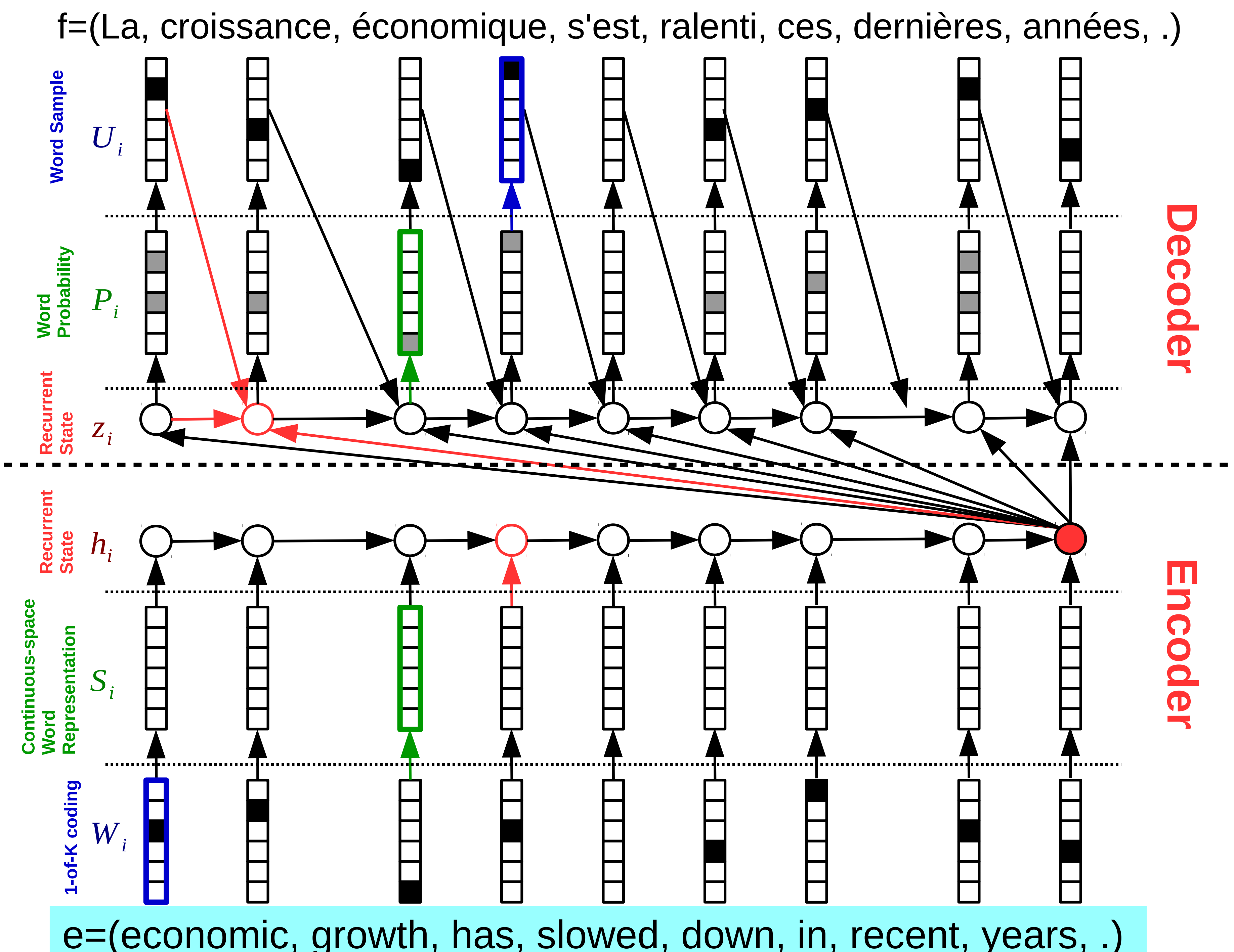
Architectures based on Encoder-Decoder Framework

- **Recurrent Neural Network based**
- Convolution Neural Network based
- Feed Forward Neural Network based

Key steps

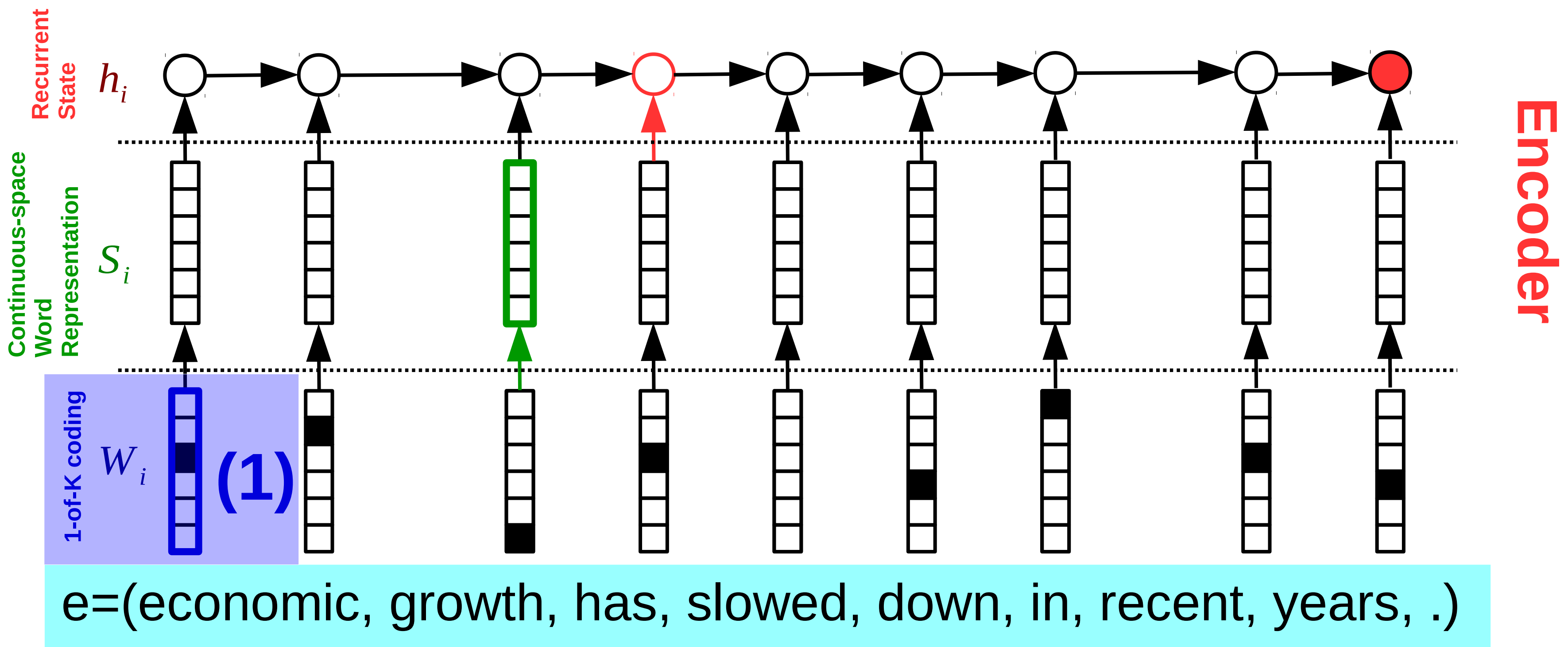
- Embed
- Encode
- Attend (Only in attention based architectures)
- Predict

Neural machine translation system



The Encoder

Step 1: Word to one-hot vector

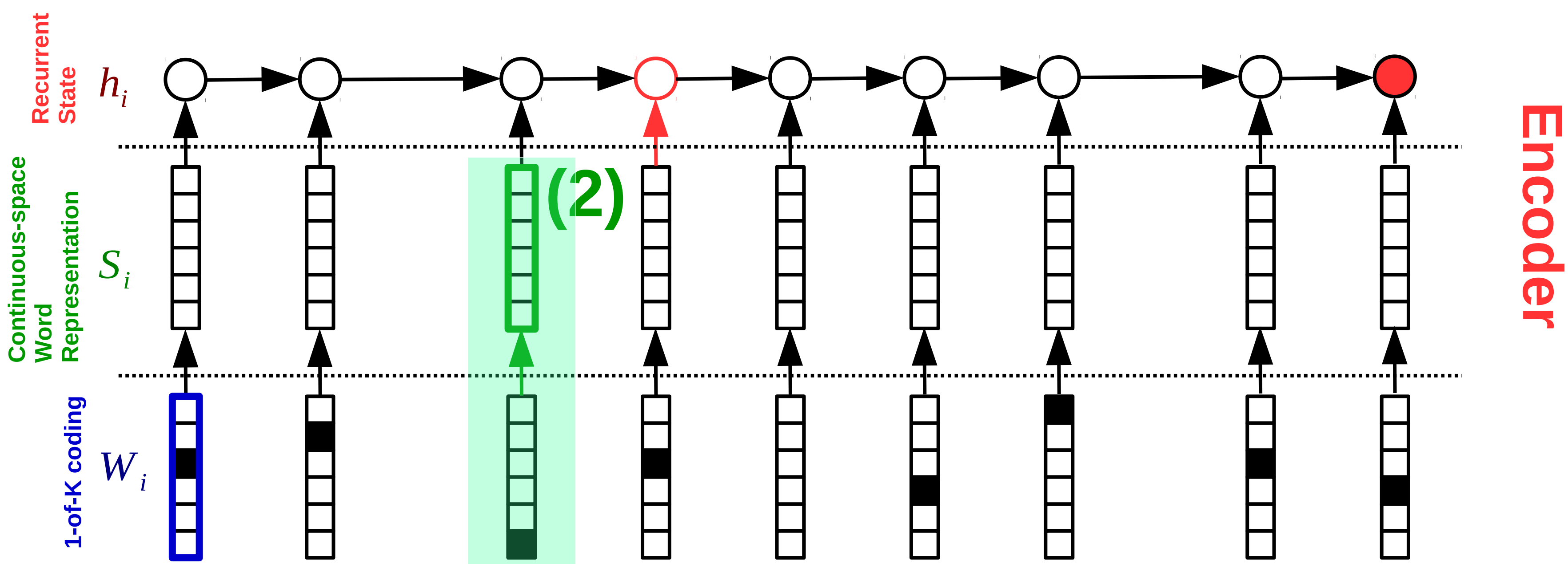


Step 2: One-hot vector to continuous-space representation

Projects the 1-of-K coded vector with a matrix E to d-dimensional (typically 100 – 500) continuous word representation

$$s_i = E_{d\times K} x_i$$

s_i updated to maximize the translation performance

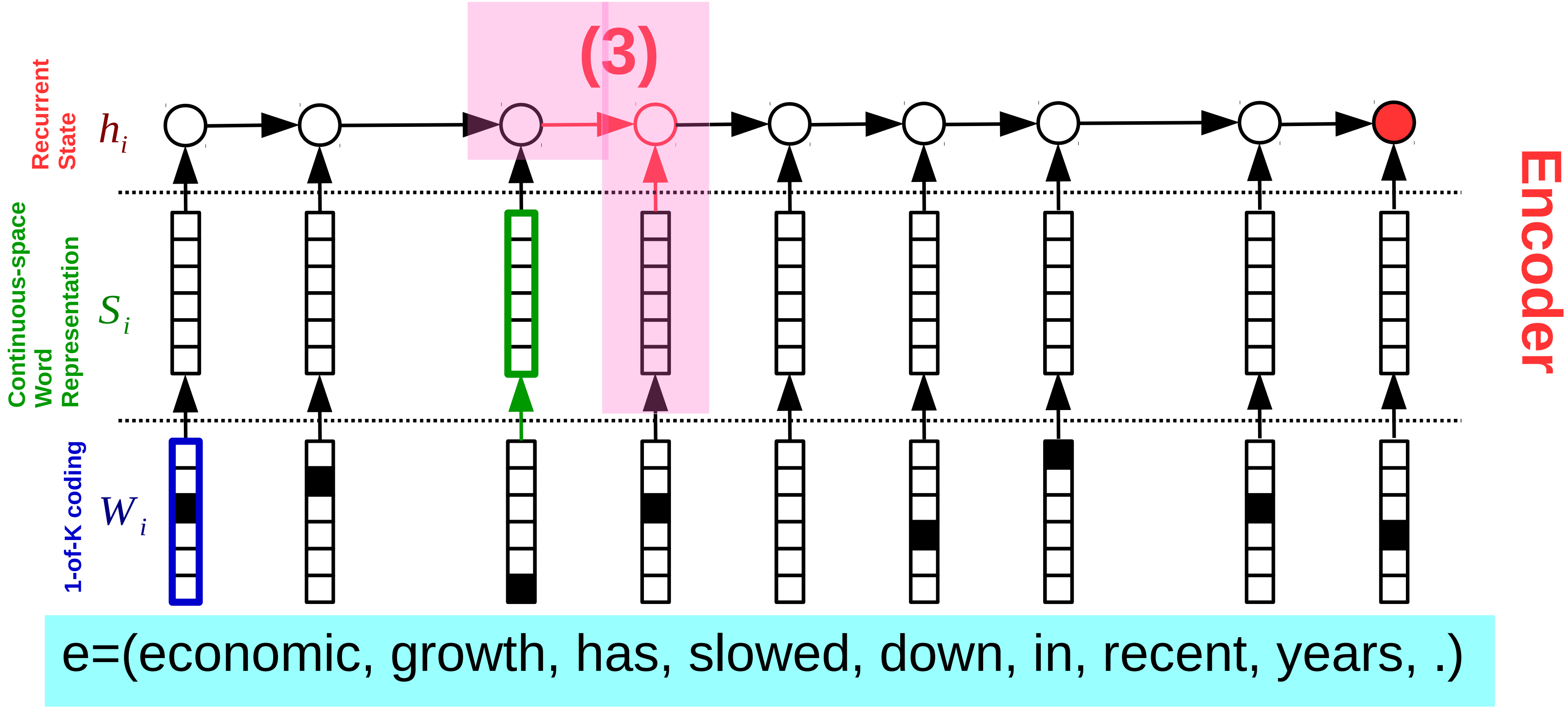


e=(economic, growth, has, slowed, down, in, recent, years, .)

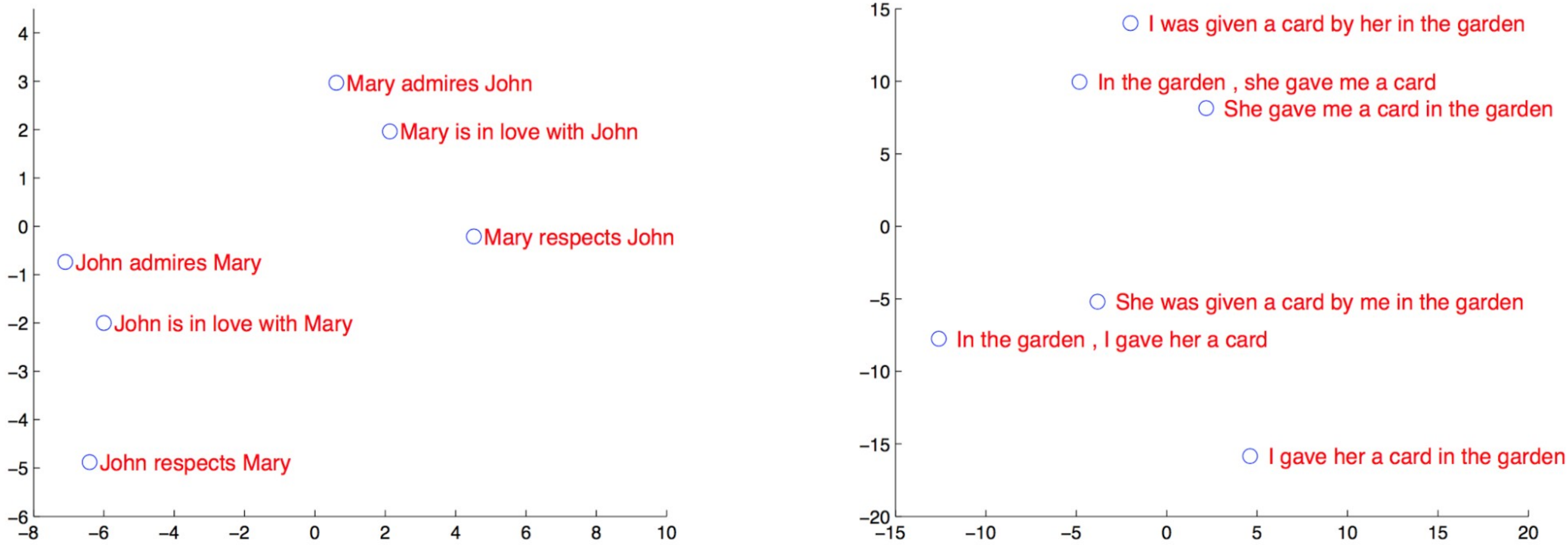
Step 3: Sequence summarization by RNN

Sequence of continuous vectors s_i summarized by RNN

$$h_i = \phi_\theta(h_{i-1}, s_i)$$



Summary sentence representation vectors



Sentence Representations from [Sutskever et al., 2014]. Similar sentences are close together

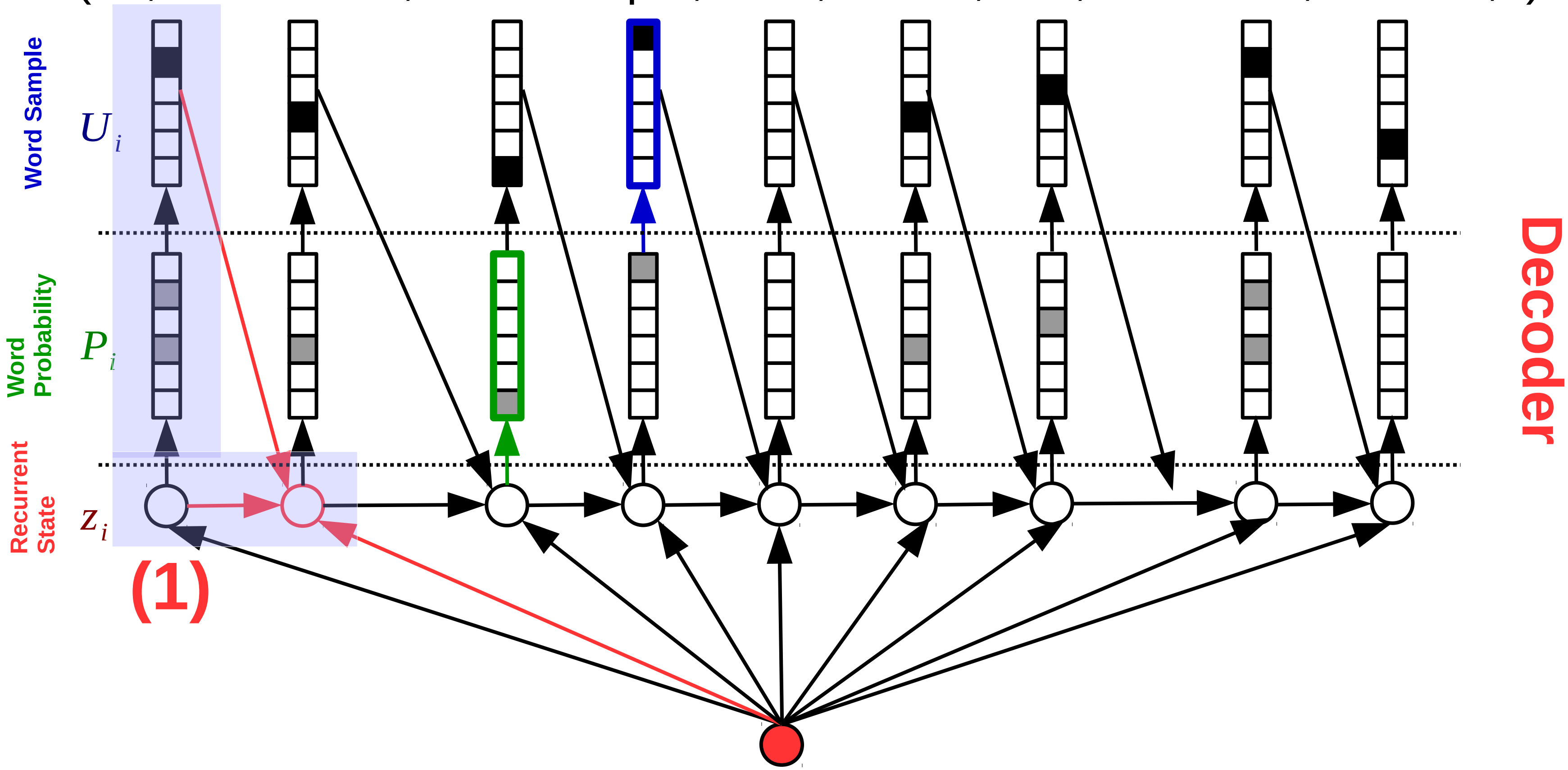
t-distributed stochastic neighbor embedding (t-sne)

The Decoder

Step 1: Compute internal hidden state of decoder

$$z_i = \phi_{\theta'}(h_T, u_{i-1}, z_{i-1})$$

f=(La, croissance, économique, s'est, ralenti, ces, dernières, années, .)



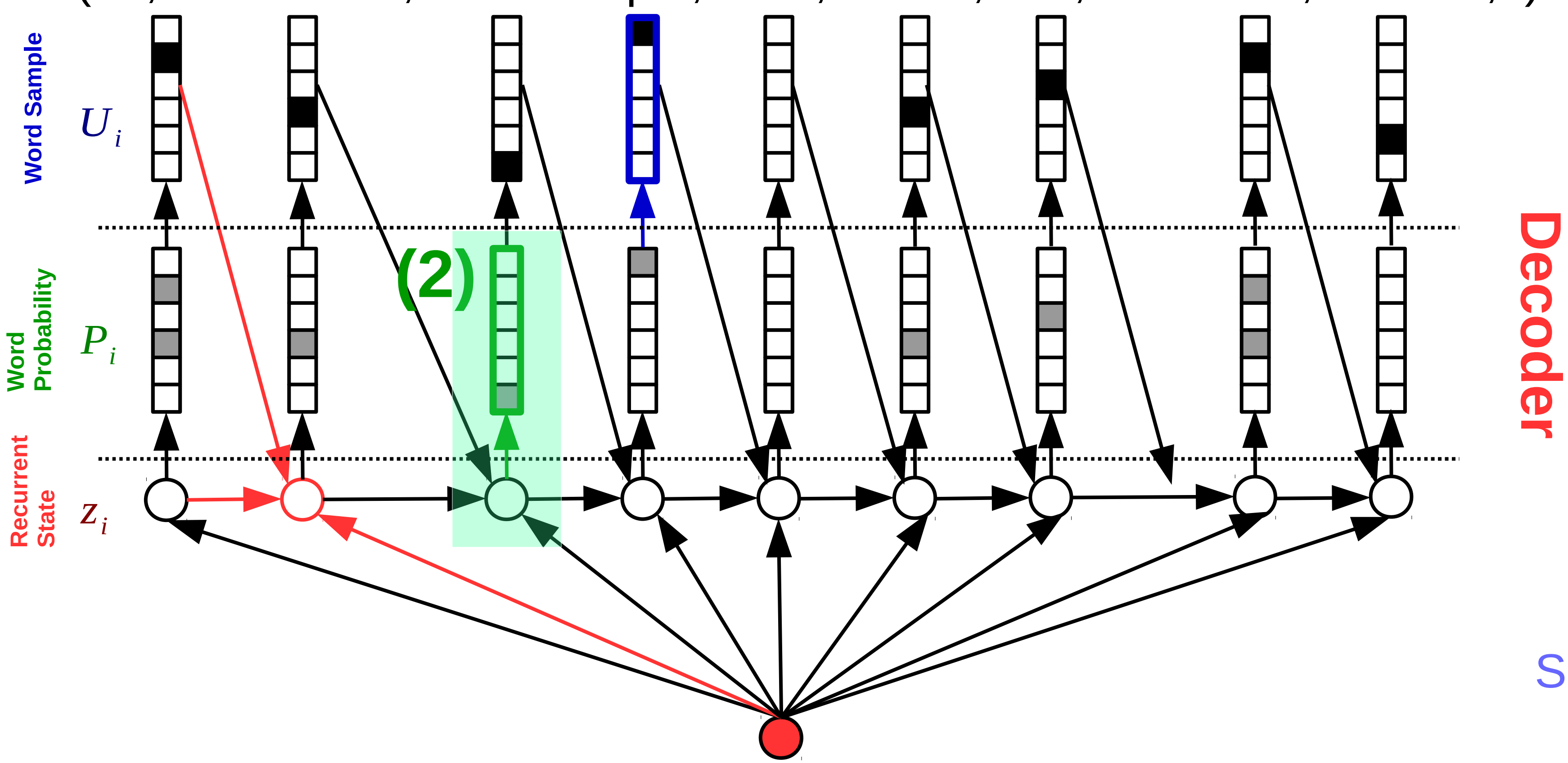
Step 2: Next word probability

Score and normalize target words

$$e(k) = w_k^T z_i + b_k$$

$$p(w_i = k \mid w_1, w_2, \ldots, w_{i-1}, h_T) = \frac{\exp(e(k))}{\sum_j \exp(e(j))}$$

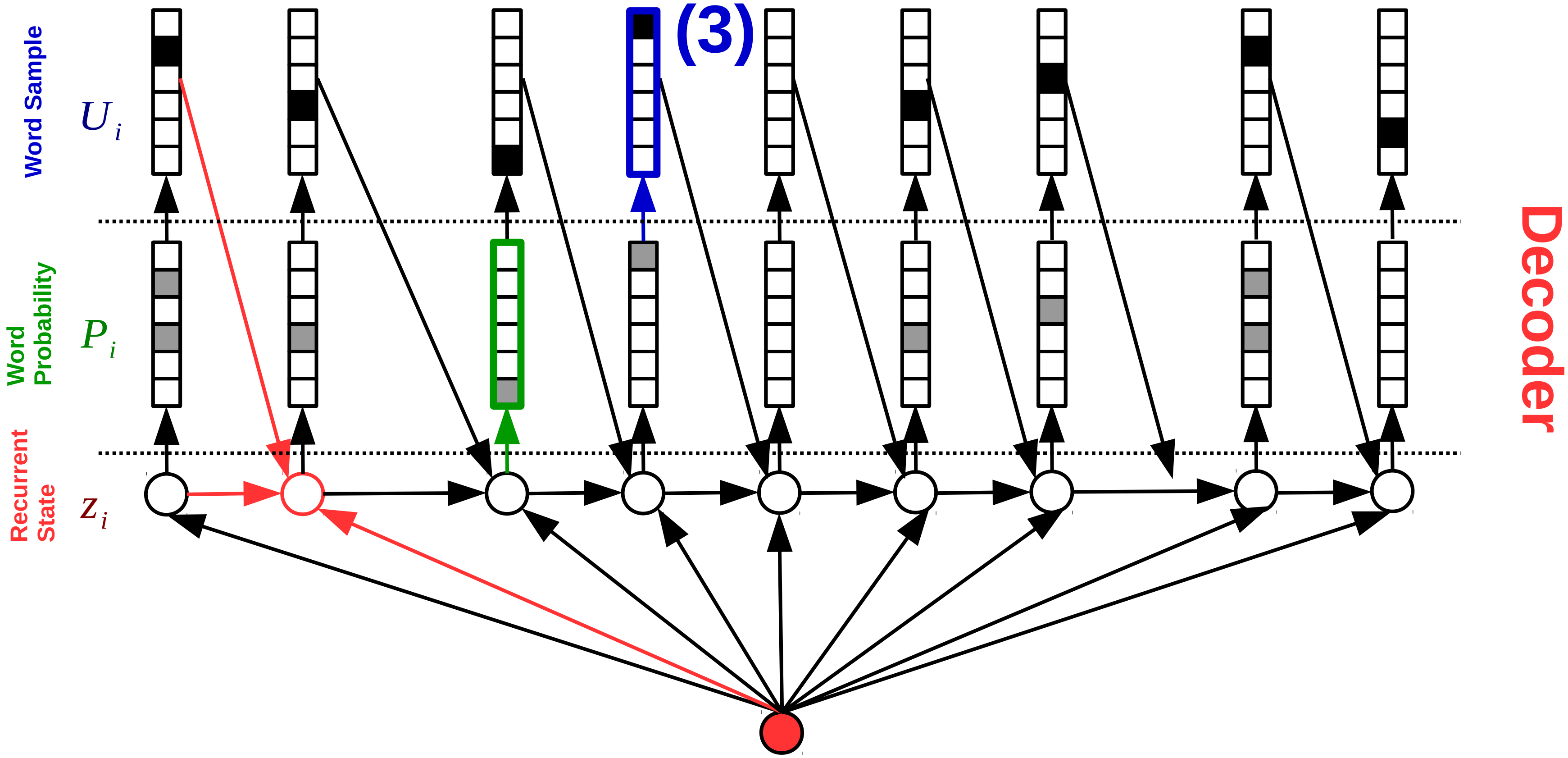
f=(La, croissance, économique, s'est, ralenti, ces, dernières, années, .)



Softmax

Step 3: Sample next word

f=(La, croissance, économique, s'est, ralenti, ces, dernières, années, .)



Training

Training: Maximum Likelihood Estimation

- Given parallel corpus D of training examples

$$D = \{ (x^{1,} y^{1}), (x^{2,} y^{2}), \ldots, (x^{n}, y^{n}) \}$$

- NMT can compute

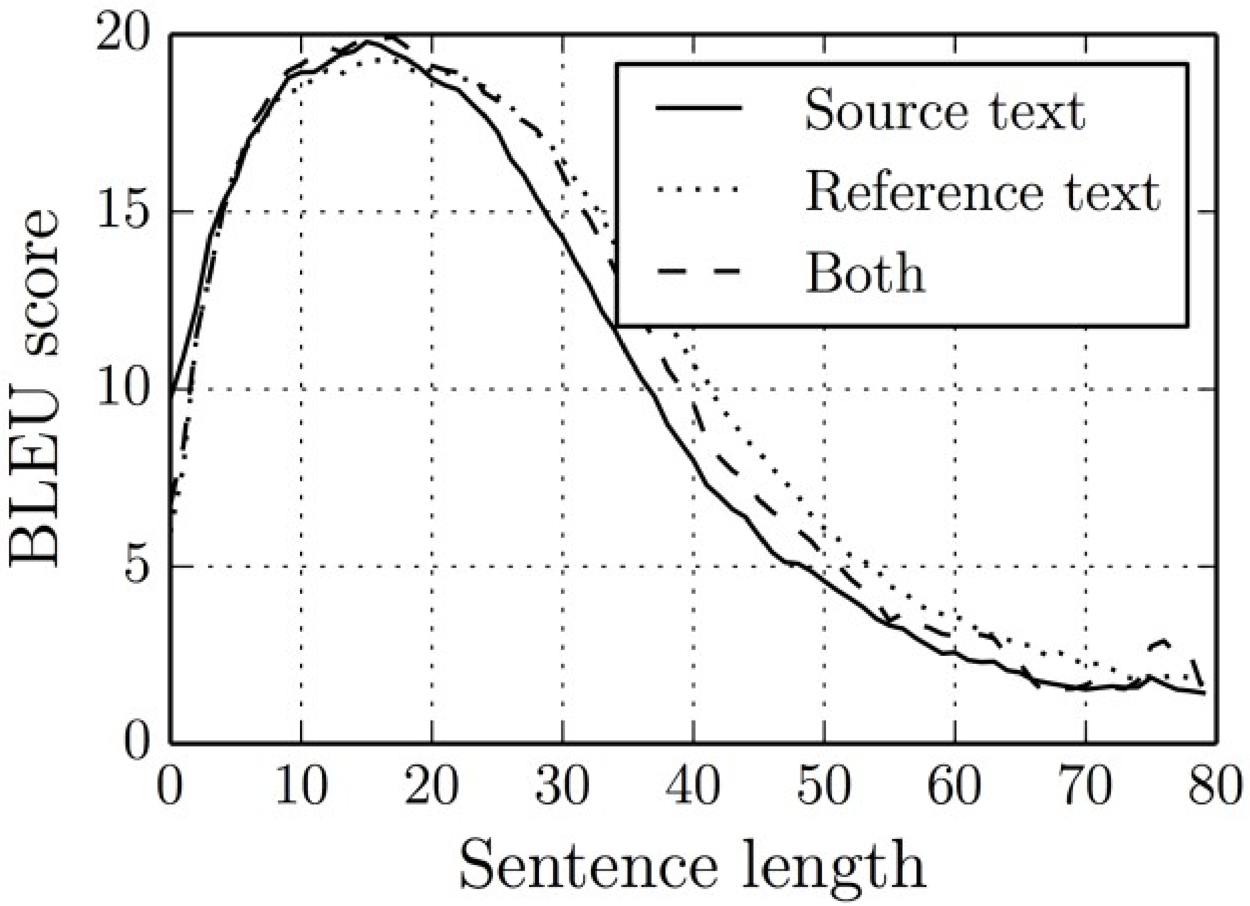
$$\log P(y^{n} | x^{n}, \theta)$$

- Log-likelihood of training corpus

$$L(D, \theta) = \frac{1}{N} \sum_{n=1}^{N} \log P(y^{n} \vee x^{n}, \theta),$$

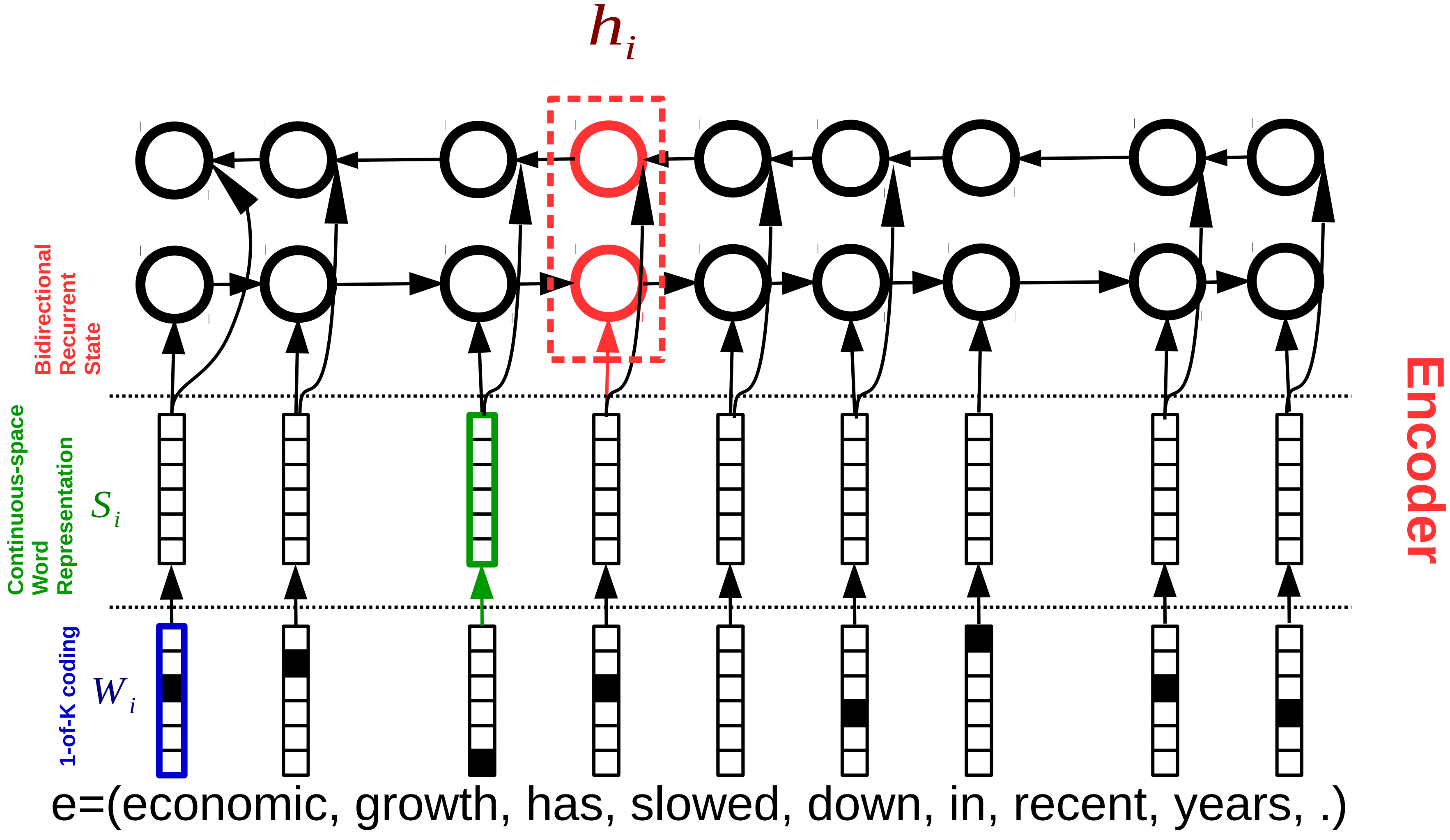
- Maximize log-likelihood using stochastic gradient descent (SGD)

Problem with simple encoder-decoder architectures



Soft Attention Mechanism for Neural Machine Translation

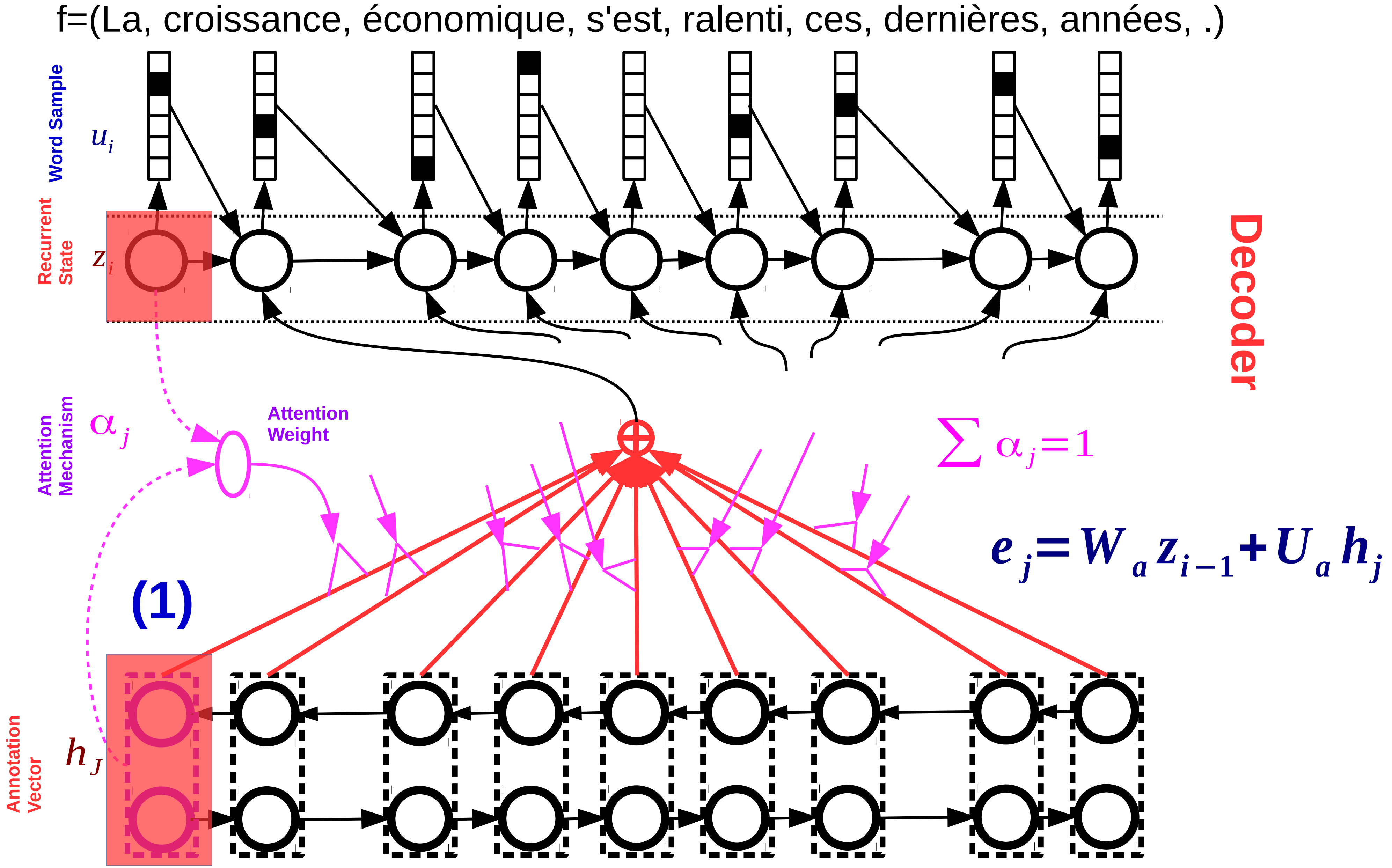
Bidirectional recurrent neural networks for encoding a source sentence



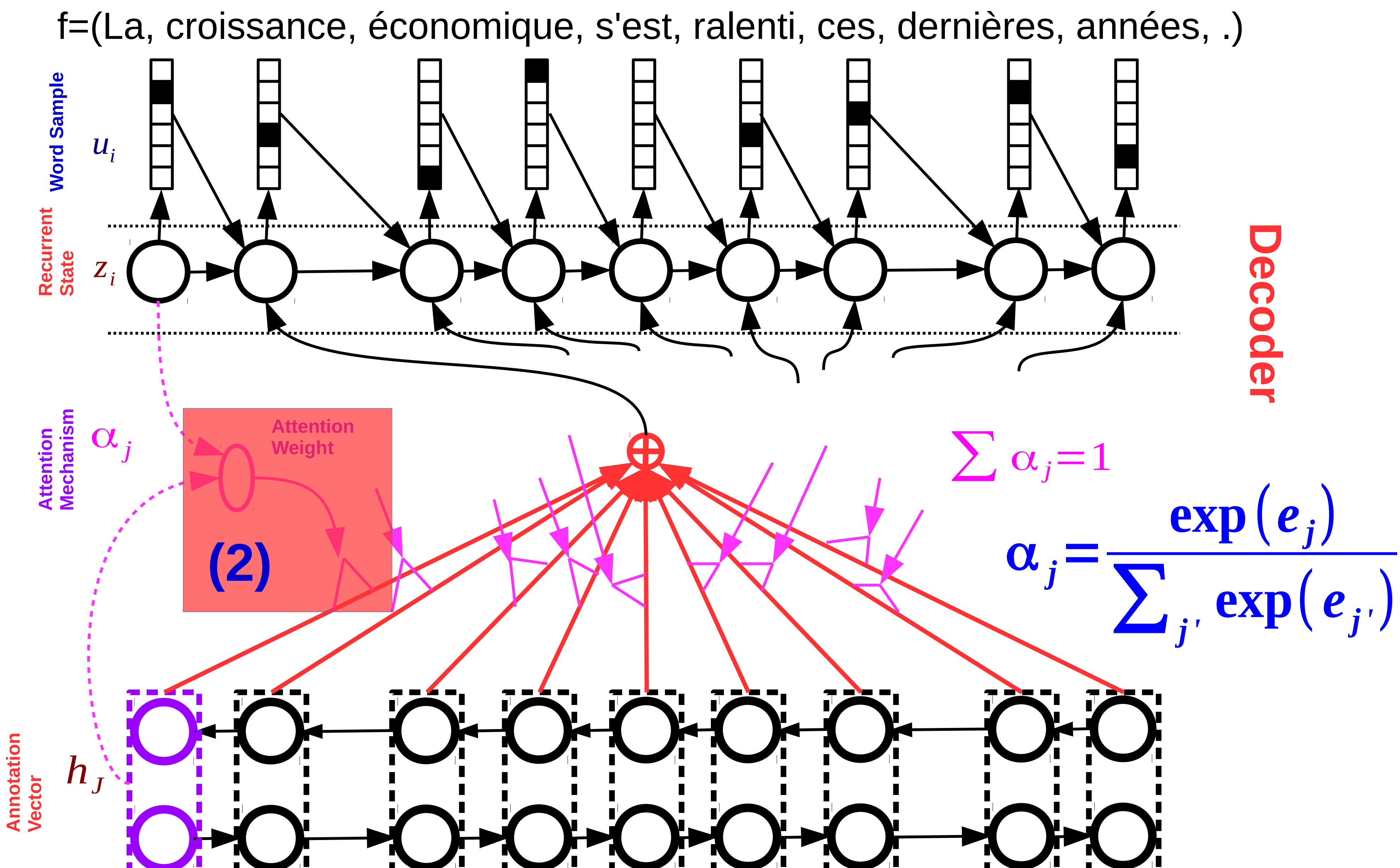
Attention Mechanism

- To decide i-th target word, calculate relevance score between i-th target word and every source word
- Attention mechanism is implemented by a neural network
- Input to NN (z_{i-1}, h_j)

Attention Mechanism (step 1)

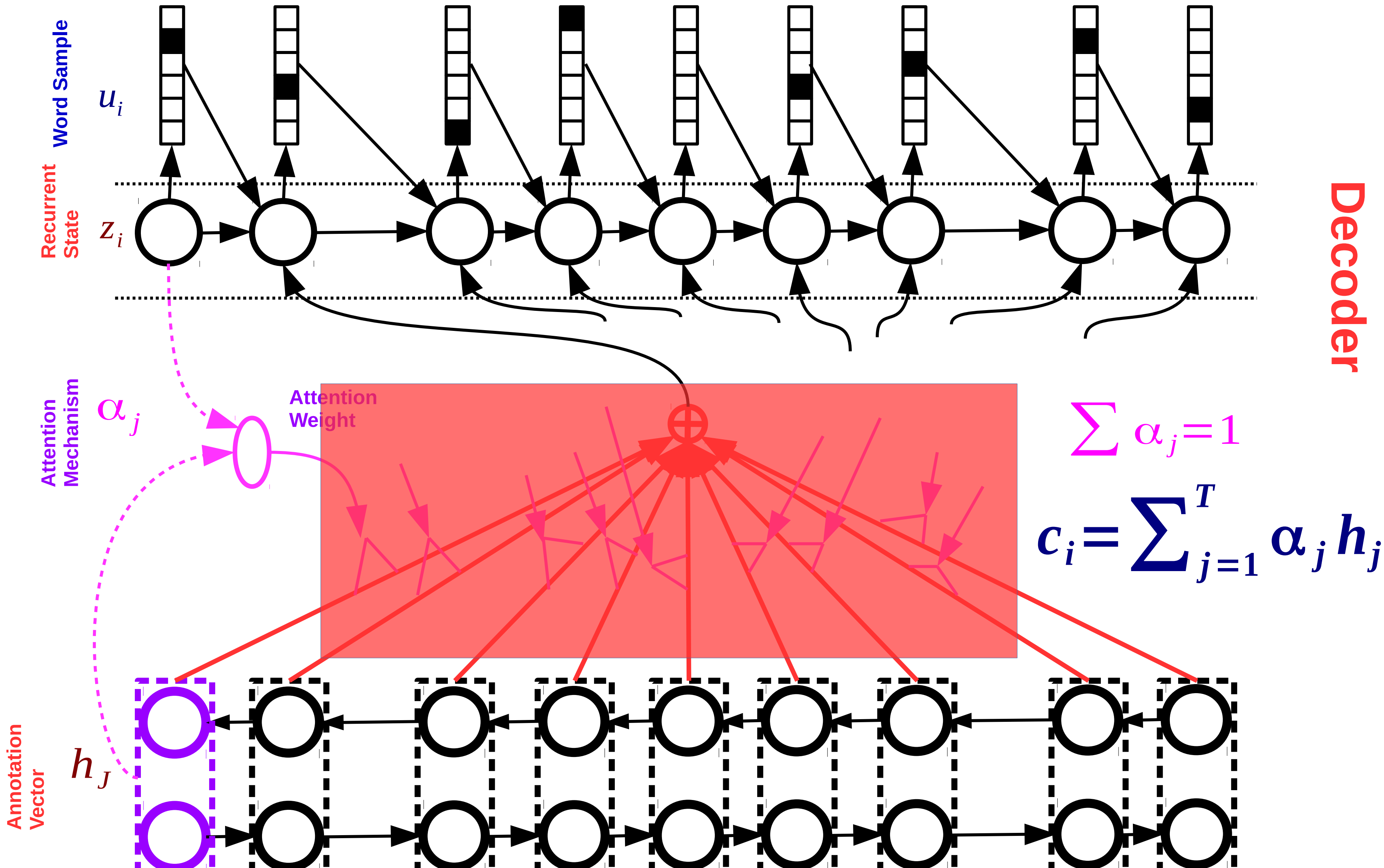


Attention Mechanism (step 2)

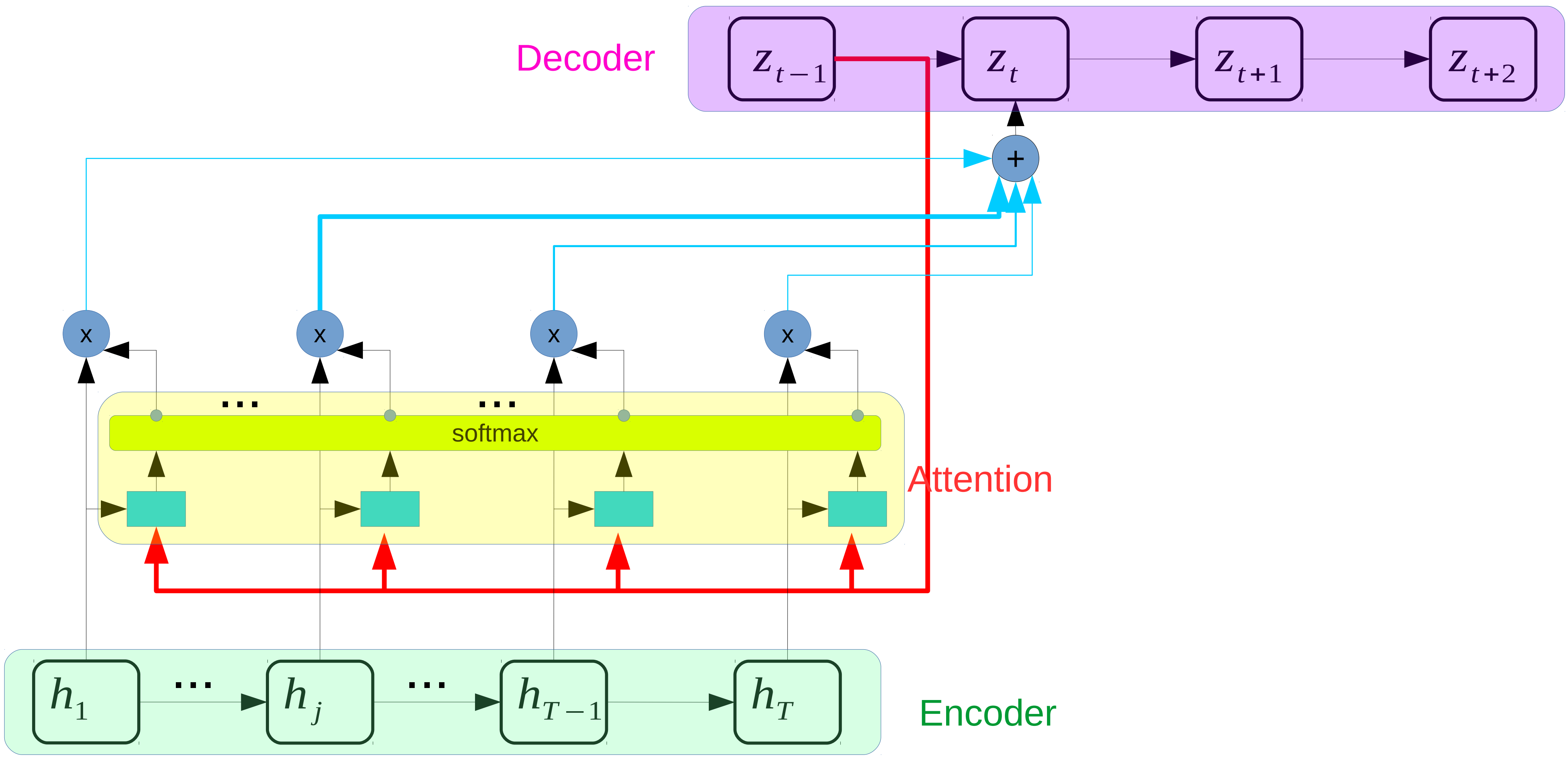


Attention Mechanism (step 3)

f=(La, croissance, économique, s'est, ralenti, ces, dernières, années, .)



Alternate view of attention mechanism



Summary of attention based NMT

- Annotation vectors $\left(h_1, \ldots, h_T\right)$ Where $h_j^T = \left[\overleftarrow{h}_j^T ; \overrightarrow{h}_j^T\right]$
- Relevance weight or an alignment weight of j-th annotation vector for t-th target word (f is FF NN)

$$\alpha_{tj} = \frac{\exp\left(f\left(z_{(t-1)}, h_j, y_{t-1}\right)\right)}{\sum_{k=1}^{T} \exp\left(f\left(z_{(t-1)}, h_k, y_{t-1}\right)\right)}$$

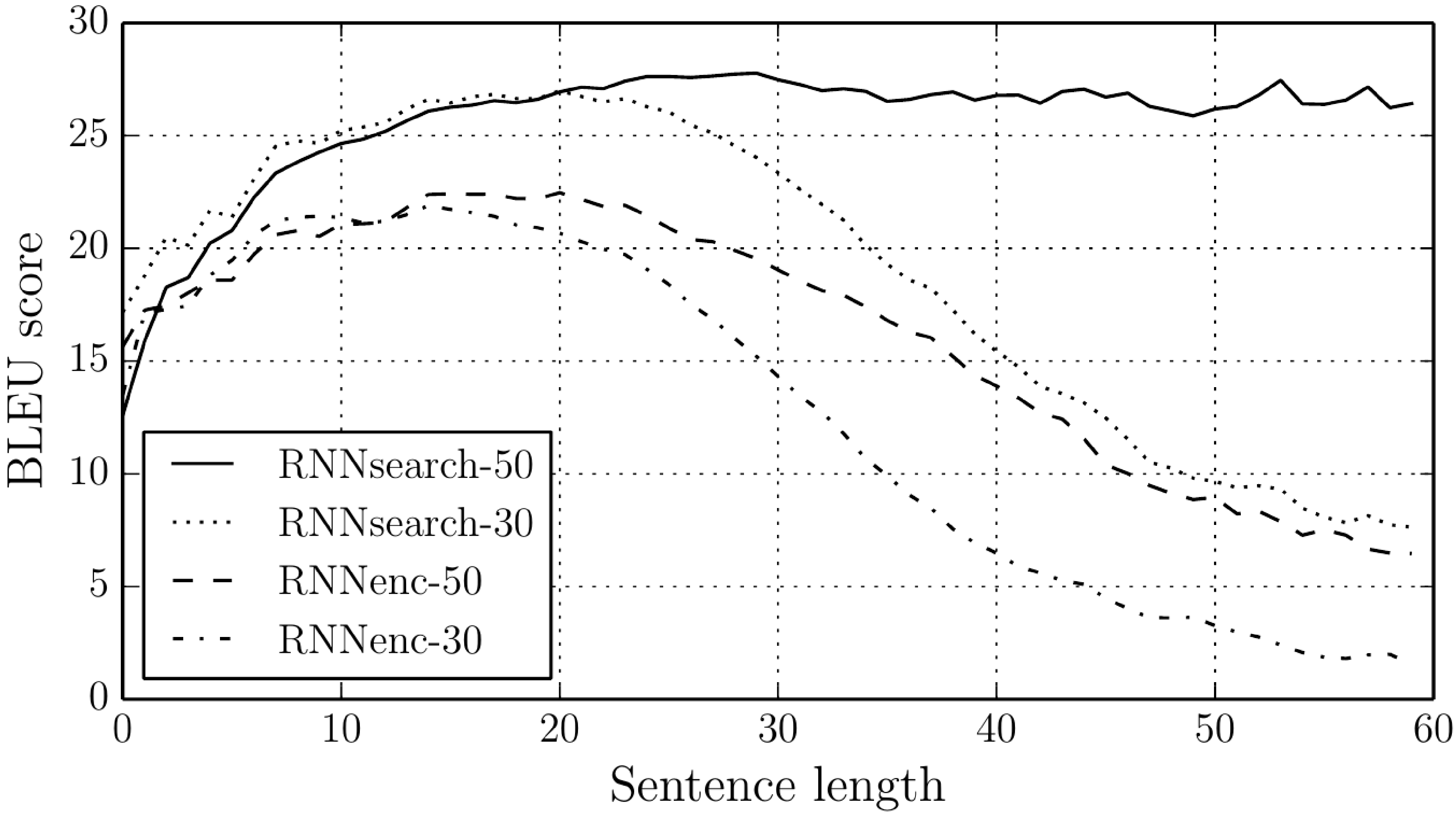
- Context vector of t-th word

$$c_t = \sum_{j=1}^{T} \alpha_{tj} h_j$$

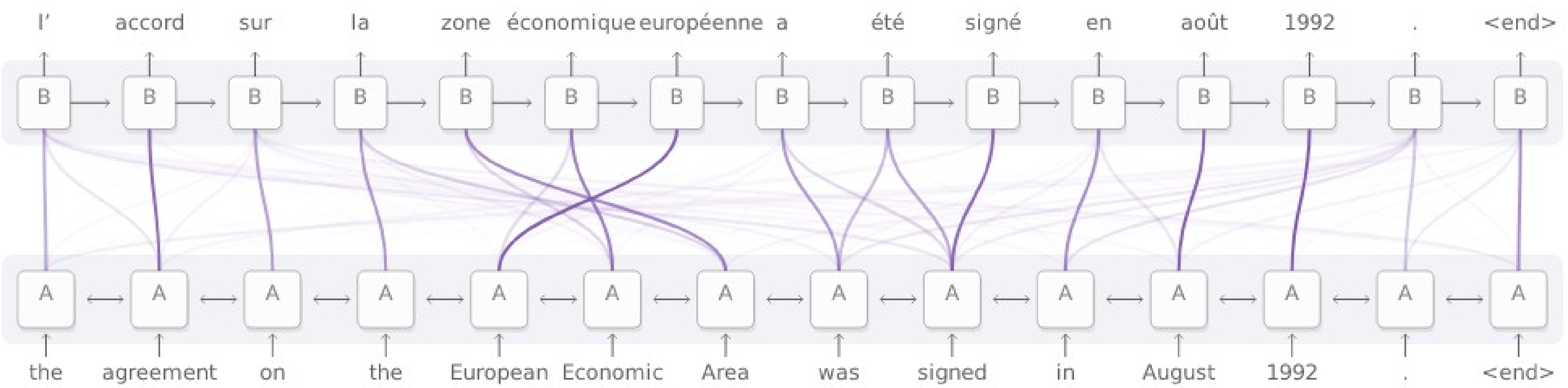
- Decoder's hidden state

$$z_t = f_r\left(z_{t-1}, y_{t-1}, c_t\right)$$

Performance



Visualization of attention



Prominent approaches for neural machine translation

Translation Model	Training time	BLEU (difference from baseline)
Transformer (T2T)	3 days on 8 GPU	**28.4 (+7.8)**
SliceNet (T2T)	6 days on 32 GPUs	26.1 (+5.5)
GNMT + Mixture of Experts	1 day on 64 GPUs	26.0 (+5.4)
ConvS2S	18 days on 1 GPU	25.1 (+4.5)
GNMT	1 day on 96 GPUs	24.6 (+4.0)
ByteNet	8 days on 32 GPUs	23.8 (+3.2)
MOSES (phrase-based baseline)	N/A	20.6 (+0.0)

BLEU scores (higher is better) on the standard WMT English-German translation task

References

- Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate.” arXiv preprint arXiv:1409.0473 (2014).
- Cho, Kyunghyun et al. “Learning phrase representations using RNN encoder-decoder for statistical machine translation.” arXiv preprint arXiv:1406.1078 (2014).
- Cho, Kyunghyun, Aaron Courville, and Yoshua Bengio. “Describing Multimedia Content using Attention-based Encoder–Decoder Networks.” arXiv preprint arXiv:1507.01053 (2015).
- Graves, Alex, Greg Wayne, and Ivo Danihelka. “Neural Turing Machines.” arXiv preprint arXiv:1410.5401 (2014).
- Gulcehre, Caglar et al. “On Using Monolingual Corpora in Neural Machine Translation.” arXiv preprint arXiv:1503.03535 (2015).
- Kalchbrenner, Nal, and Phil Blunsom. “Recurrent Continuous Translation Models.” EMNLP 2013: 1700-1709.
- Koehn, Philipp. Statistical machine translation. Cambridge University Press, 2009.
- Pascanu, Razvan et al. “How to construct deep recurrent neural networks.” arXiv preprint arXiv:1312.6026 (2013).
- Schwenk, Holger. “Continuous space language models.” Computer Speech & Language 21.3 (2007): 492-518.
- Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. “Sequence to sequence learning with neural networks.” Advances in Neural Information Processing Systems 2014: 3104-3112.
- http://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-with-gpus/
- Minh Thang Luong, Hieu Pham, Christopher D. Manning“Effective Approaches to Attention-based Neural Machine Translation.” arXiv:1508.04025 (2015).
- Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin. “Convolutional Sequence to Sequence Learning”, arXiv:1705.03122 (2017)
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi. “ Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation” arXiv:1609.08144
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones. “ Attention Is All You Need”, arXiv:1706.03762