ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Natural Language Processing with Recurrent Neural Networks

2025–26

Ion Androutsopoulos

http://www.aueb.gr/users/ion/

Contents

- Recurrent neural networks (RNNs), GRUs/LSTMs.
- Bidirectional and stacked RNNs.
- RNNs with self-attention or global max pooling.
- RNNs in text and token classification, RNN language models.
- Obtaining word embeddings from character-based RNNs.
- Hierarchical RNNs.
- Sequence-to-sequence RNN models with attention, and applications in machine translation.
- Variational dropout.
- Universal sentence encoders, LASER.
- Pretraining RNN language models, ELMo.

Extracting contract elements

THIS **AGREEMENT** is made the 15th day of October 2009
(The **"Effective Date"**) BETWEEN:

(1) **Sugar 13 Inc.**, a corporation whose office is at James House, 42-50 Bond Street, London, EW2H TL **("Sugar")**;

(2) **E2 UK Limited**, a limited company whose registered office is at 260 Bathurst Road, Yorkshire, SL3 4SA **("Provider")**.

RECITALS:

A. The Parties wish to enter into a framework agreement which will enable Sugar, from time to time, to [...]
B. [...]

NO THEREFORE IT IS AGREED AS FOLLOWS:

ARTICLE I - DEFINITIONS

"Sugar"	shall mean:	Sugar 13 Inc.
"Provider"	shall mean:	E2 UK Limited
"1933 Act"	shall mean:	Securities Act of 1933

ARTICLE II - TERMINATION

The Service Period will be for five (5) years from the Effective Date (The **"Initial Term"**). The agreement is considered to be terminated in October 16, 2014.

ARTICLE III - PAYMENT - FEES

During the service period monthly payments should occur. The estimated fees for the Initial Term are £154,800.

ARTICLE IV - GOVERNING LAW

This agreement shall be governed and construed in accordance with the Laws of England & Wales. Each party hereby irrevocably submits to the exclusive jurisdiction of the courts sitting in Northern London.

IN WITNESS WHEREOF, the parties have caused their respective duly authorized officers to execute this Agreement.

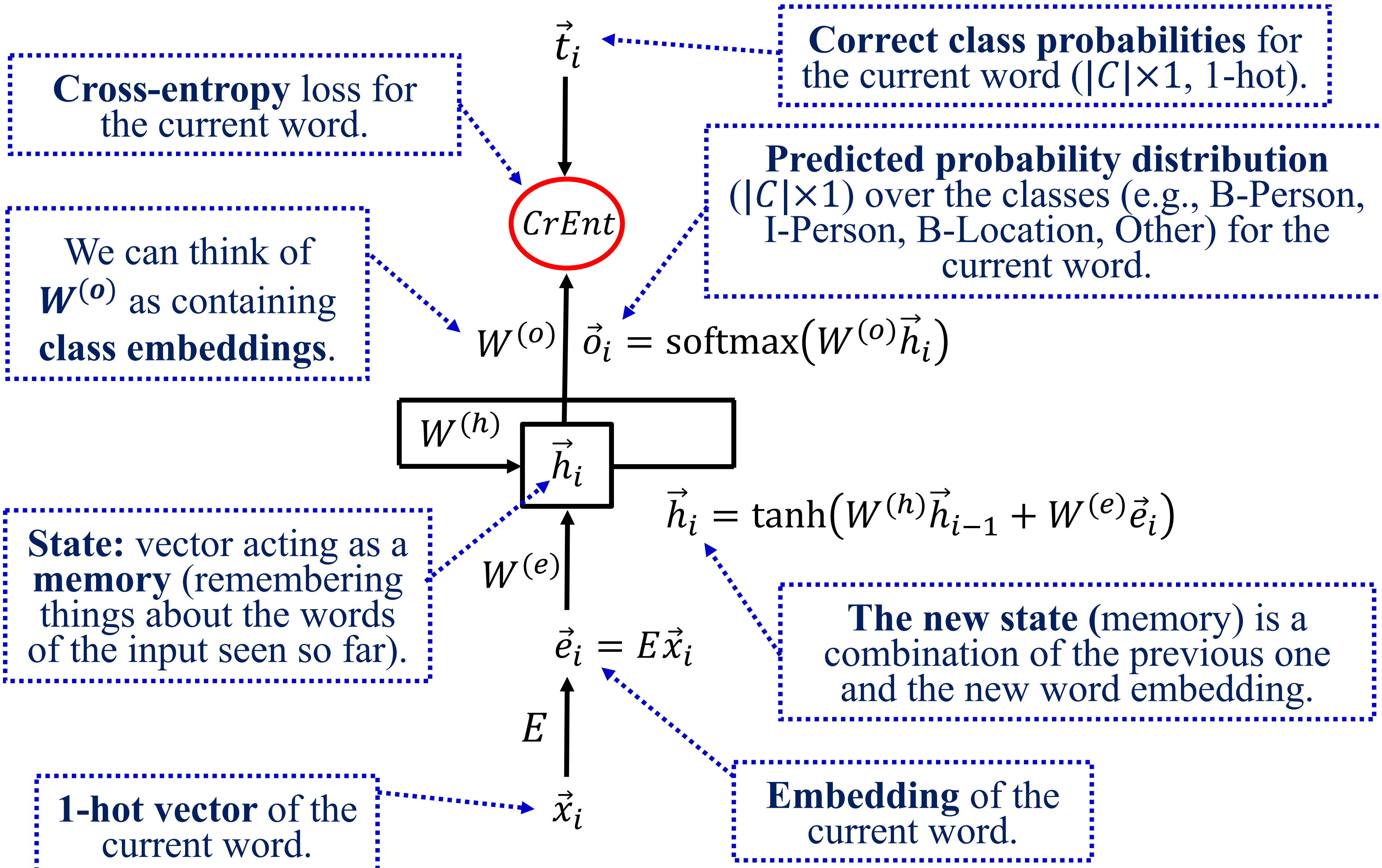
BY: George Fake
Authorized Officer
Sugar 13 Inc.

BY: Olivier Giroux
CEO
E2 UK LIMITED

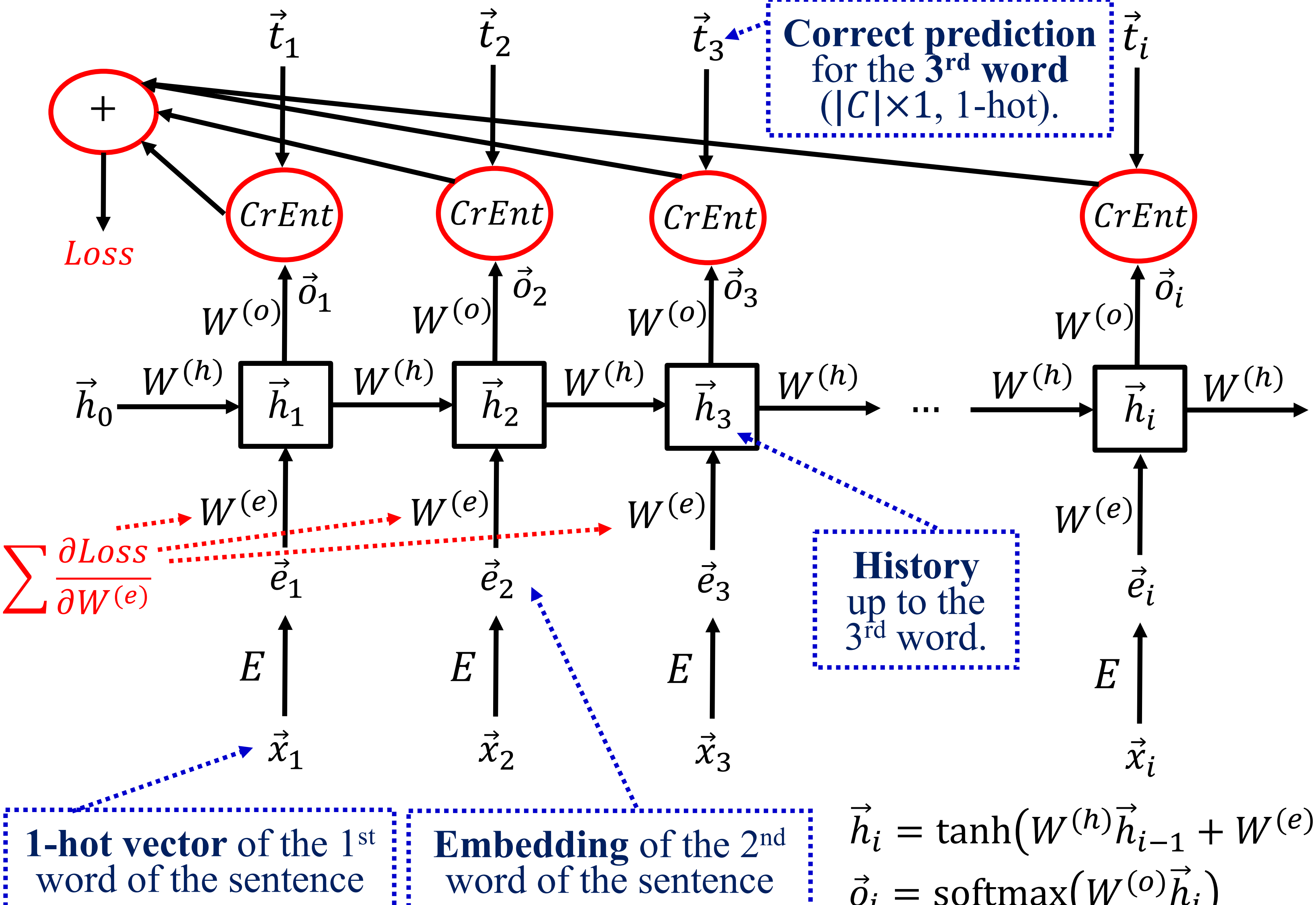
Identify **start/end dates**, **duration**, **contractors**, **amount**, **legislations refs**, **jurisdiction** etc. Similar to **Named Entity Recognition** (NER).

I. Chalkidis, I. Androutsopoulos, A. Michos, "Extracting Contract Elements", ICAIL 2017, http://nlp.cs.aueb.gr/pubs/icail2017.pdf .

RNN-based Named Entity Recognizer



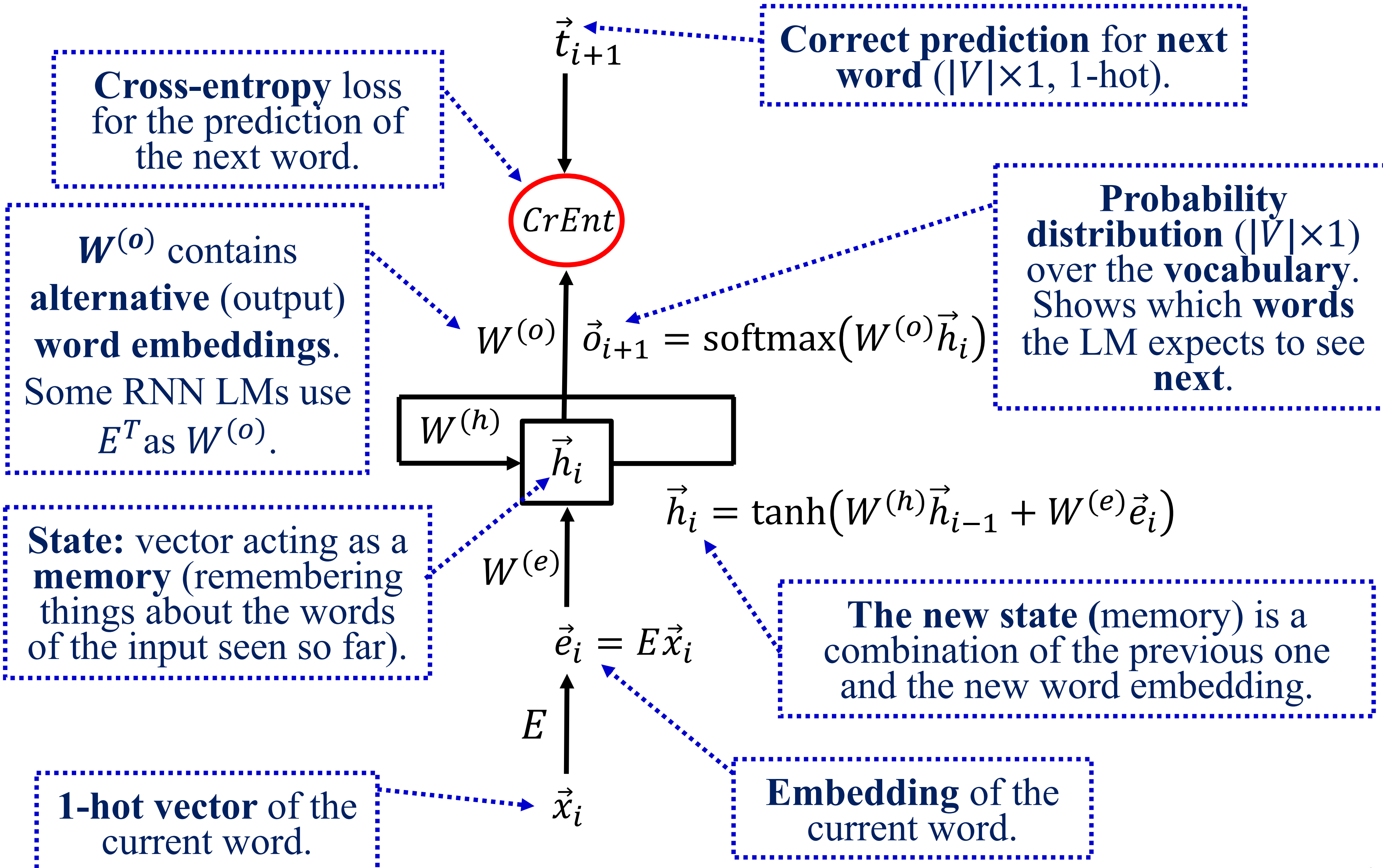
Unrolled RNN



$$\vec{h}_i = \tanh\left(W^{(h)}\vec{h}_{i-1} + W^{(e)}\vec{e}_i\right)$$

$$\vec{o}_i = \text{softmax}\left(W^{(o)}\vec{h}_i\right)$$

RNN language model



Reminder: LMs as next word predictors

- **Sequence probability** using a bigram LM:

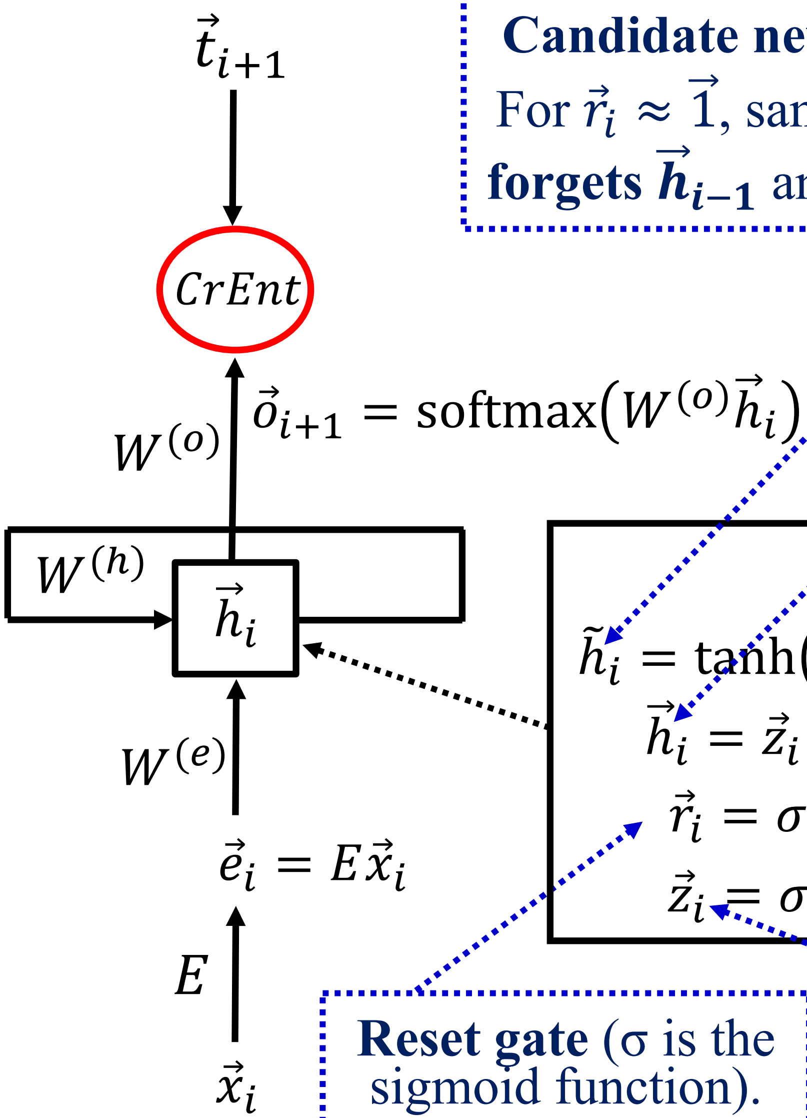
$$P(w_1^k) = P(w_1, \ldots, w_k) = P(w_1) \cdot P(w_2 \mid w_1) \cdot$$

$$P(w_3 \mid w_1, w_2) \cdot P(w_4 \mid w_1^3) \cdots P(w_k \mid w_1^{k-1}) \simeq$$

$$\boxed{P(w_1 \mid start) \cdot P(w_2 \mid w_1) \cdot P(w_3 \mid w_2) \cdots P(w_k \mid w_{k-1})}$$

- We can think of the **LM** as a system that **provides the probabilities $\boldsymbol{P(w_i|w_{i-1})}$**, which we then multiply.
  - Or the probabilities $\boldsymbol{P(w_i|w_{i-2}, w_{i-1})}$ for a **trigram LM**.
  - Or the probabilities $\boldsymbol{P(w_i|h_{i-1})}$ for an LM that considers all the **"history" (previous words) $\boldsymbol{h_{i-1}}$**, e.g., in an **RNN LM**.
- An **LM** typically provides a **distribution $\boldsymbol{P(w|h)}$** showing how probable it is for **every word** $w \in V$ to be the next one.

RNN LM with GRU cells



Candidate new history ($\circ$ denotes pairwise multiplication). For $\vec{r}_i \approx \vec{1}$, same as the $\vec{h}_i$ of a simple RNN cell. For $\vec{\boldsymbol{r}}_{\boldsymbol{i}} \approx \vec{\boldsymbol{0}}$, **forgets** $\vec{\boldsymbol{h}}_{\boldsymbol{i-1}}$ and considers only the current word embedding.

New history. For $\vec{z}_i \approx \vec{0}$, same as $\tilde{h}_i$. For $\vec{\boldsymbol{z}}_{\boldsymbol{i}} \approx \vec{\boldsymbol{1}}$, ignores $\tilde{h}_i$ and **maintains** $\vec{\boldsymbol{h}}_{\boldsymbol{i-1}}$ as $\vec{h}_i$.

GRU cell:

$$\tilde{h}_i = \tanh\left(\vec{r}_i \circ W^{(h)}\vec{h}_{i-1} + W^{(e)}\vec{e}_i\right)$$

$$\vec{h}_i = \vec{z}_i \circ \vec{h}_{i-1} + \left(\vec{1} - \vec{z}_i\right) \circ \tilde{h}_i$$

$$\vec{r}_i = \sigma\left(W^{(r)}\vec{h}_{i-1} + U^{(r)}\vec{e}_i\right)$$

$$\vec{z}_i = \sigma\left(W^{(z)}\vec{h}_{i-1} + U^{(z)}\vec{e}_i\right)$$

Reset gate (σ is the sigmoid function).

Update gate.

LSTM cells are similar, but with **more gates**. See http://colah.github.io/posts/2015-08-Understanding-LSTMs/

More about RNNs

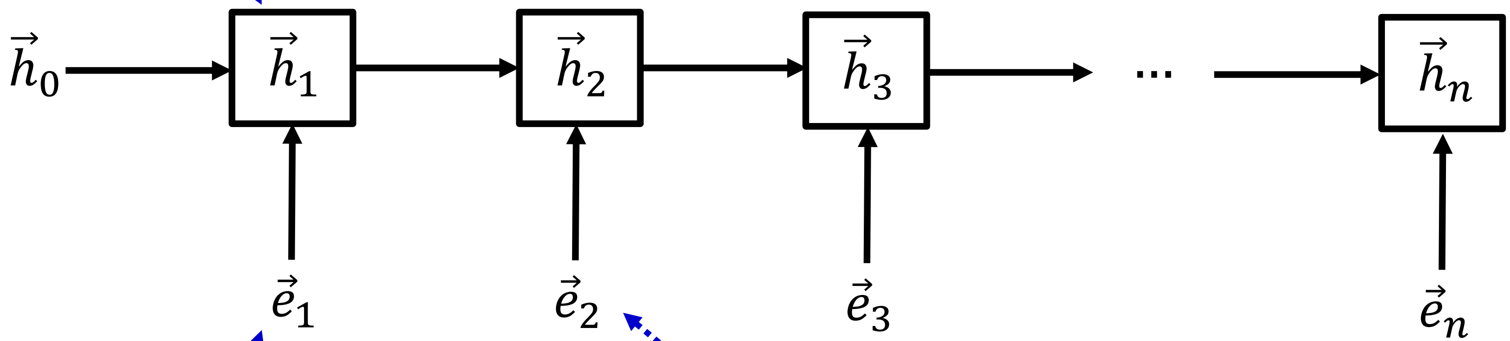
- Trained by **backpropagation** (with **unrolled** view).
  - For **each sentence** (**or window**), **feed** it to the **unrolled RNN**, compute the **loss** and **backpropagate**, **adding gradients** obtained for the **same matrix** (e.g., same $W^{(h)}$ at each cell).
  - **GRU** or **LSTM** cells help avoid **vanishing gradients**.
  - The norms of the **gradients** can be **clipped** (when larger than a max value) to avoid **exploding gradients**.
  - Use **layer normalization**, not batch normalization in RNNs.
- We can also **learn** the **word embeddings** (E) with an RNN LM. Billions of **free training examples**!
  - We can **re-use the word embeddings** in **other NLP tasks**.
  - With a **large vocabulary**, **softmax** is too **slow** (alternatives: small vocabulary, hierarchical softmax, negative sampling).

What about the right-context of each token?

Revised embedding of the 1st word. Knows we are at the beginning of a sentence.

Revised embedding of the 2nd word. Knows the left-context.

We can also treat the $\vec{h}_i$ vectors as the **memory** of the RNN, but in recent NLP work, it's easier to think of them as **revised word embeddings**.



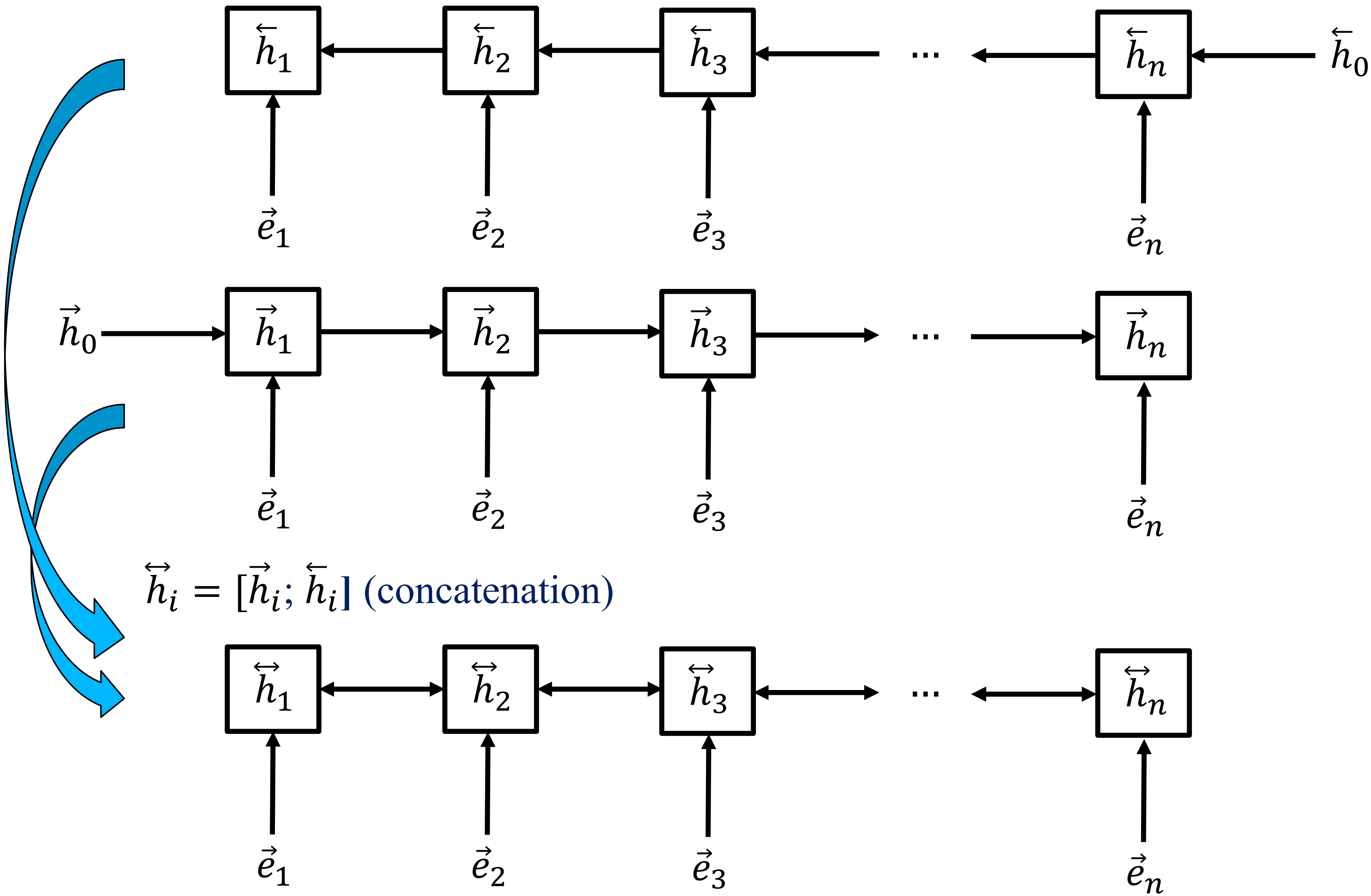
Embedding of the 1st word of the sentence

Embedding of the 2nd word of the sentence

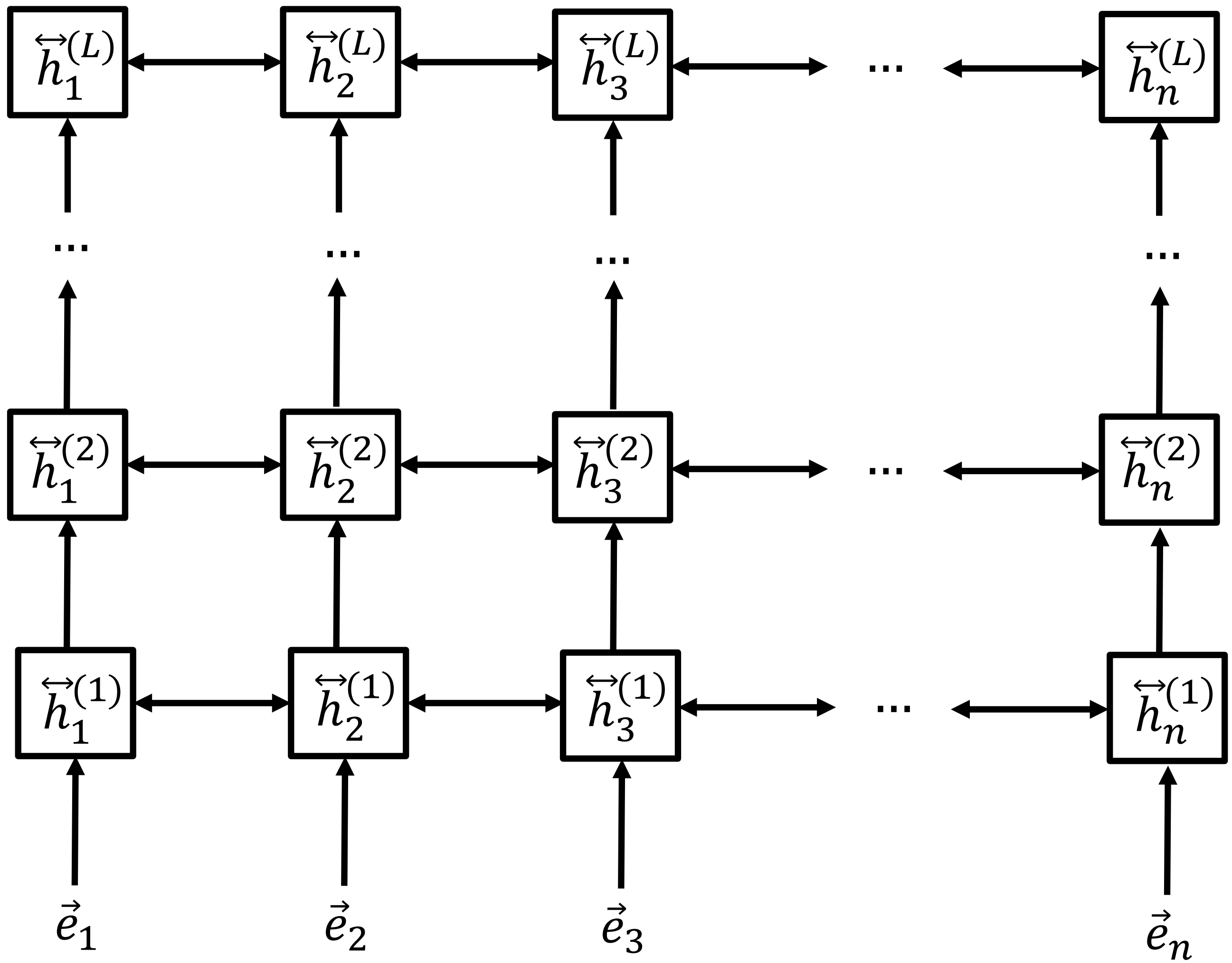
g is an **activation function** (e.g., sigmoid). More complex update mechanisms in practice: **LSTM** or **GRU** cells.

$$\vec{h}_i = g\left(W^{(h)}\vec{h}_{i-1} + W^{(e)}\vec{e}_i + \vec{b}^{(h)}\right)$$

Bidirectional RNN (biRNN)



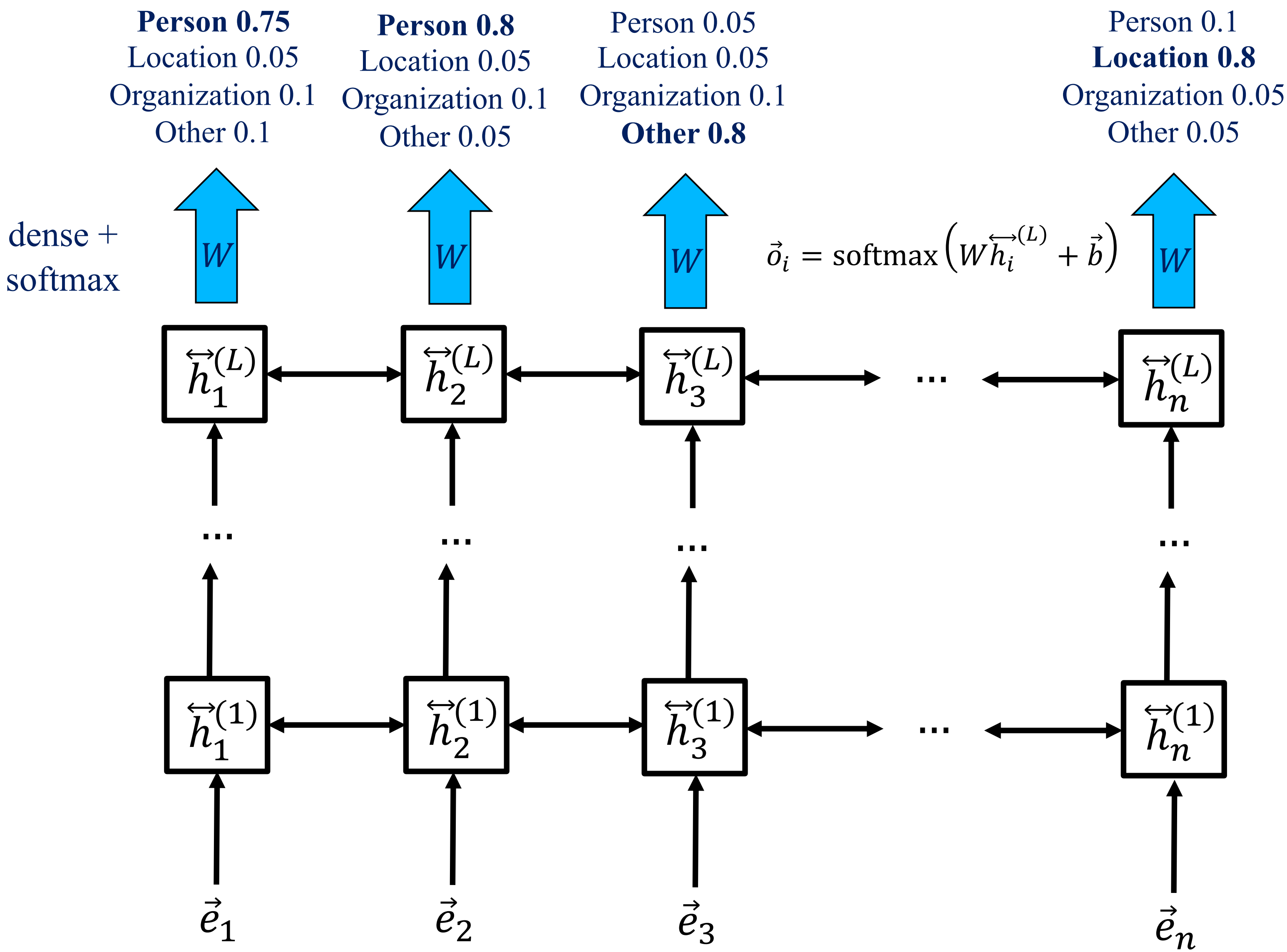
Stacked bidirectional RNN



Each layer revises the word embeddings of the previous (lower) layer. The **embeddings** become **increasingly more context-aware** and also **increasingly more appropriate** for the **particular task** we address…

Token classification with a stacked biRNN

Compare to the correct predictions (**sum** the **cross-entropy loss** for **all token positions**) and **backpropagate** to **adjust all the weights**, including the weights of the stacked biRNN.



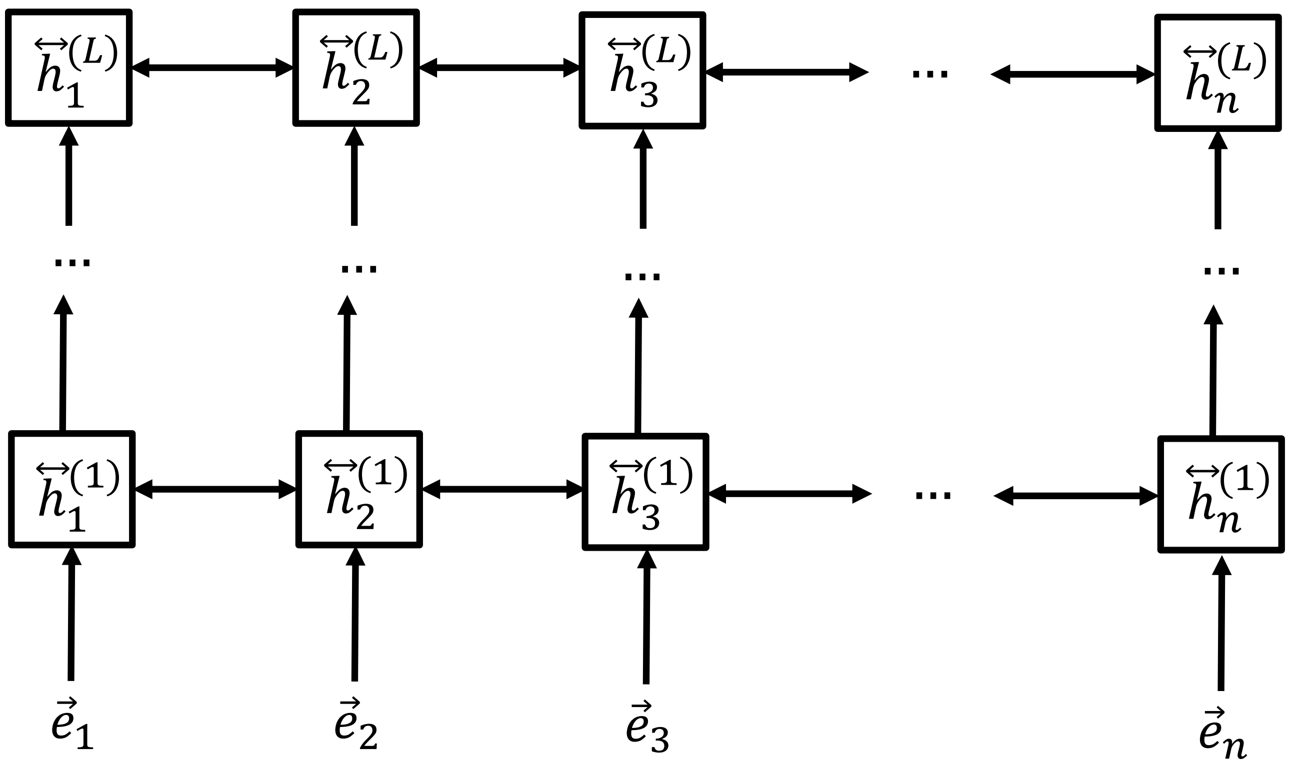
Text classification with stacked biRNN

Compare (via **categorical cross entropy**) the **predicted $\vec{o}$** to the **correct 1-hot distribution** and **backpropagate** to adjust all the weights, including the weights of the stacked biRNN.

$$\vec{o} = \mathrm{softmax}\left(W \vec{h}_{max} + \vec{b}\right)$$

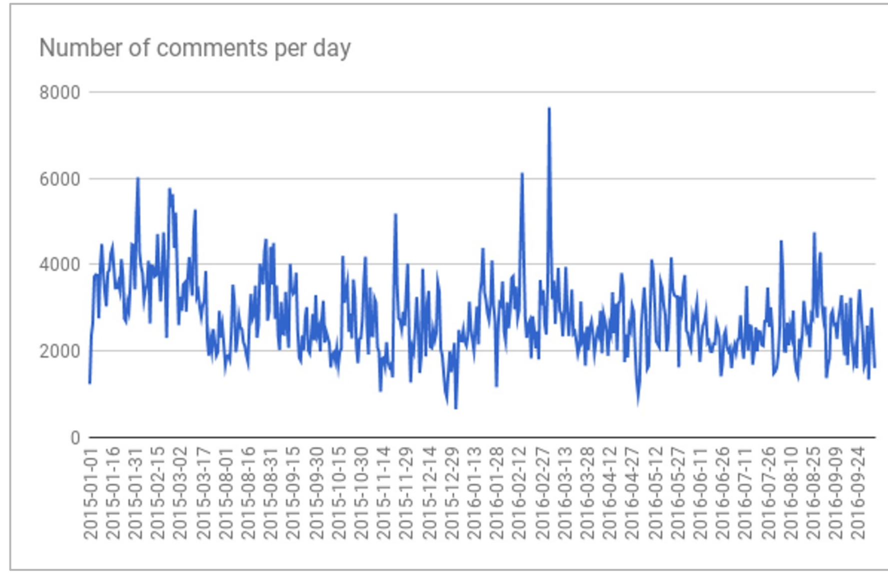
$$\overleftrightarrow{h}_{max} = \left\langle \max\left(\overleftrightarrow{h}^{(L)}_{*,1}\right), \max\left(\overleftrightarrow{h}^{(L)}_{*,2}\right), \ldots, \max\left(\overleftrightarrow{h}^{(L)}_{*,n}\right) \right\rangle^{\mathrm{T}}$$

Global max-pooling creates a **single vector** containing the **max per dimension** of all the $\overleftrightarrow{h}^{(L)}_i$. We pass it through a **dense layer and softmax (or MLP)** to obtain a **probability per class**.

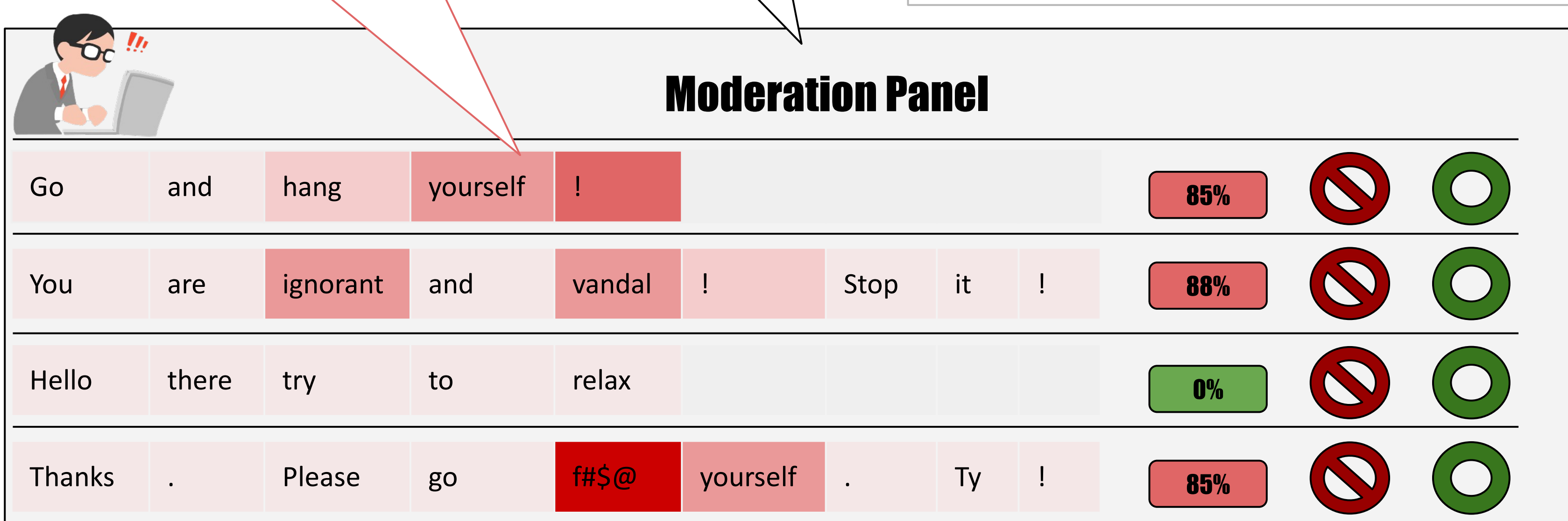


User comment moderation

A **moderation panel assists the moderators** to **detect abusive comments**, and leads to **quicker publication** of non-abusive comments.



Highlighting suspicious words using an **RNN** with **self-attention**.



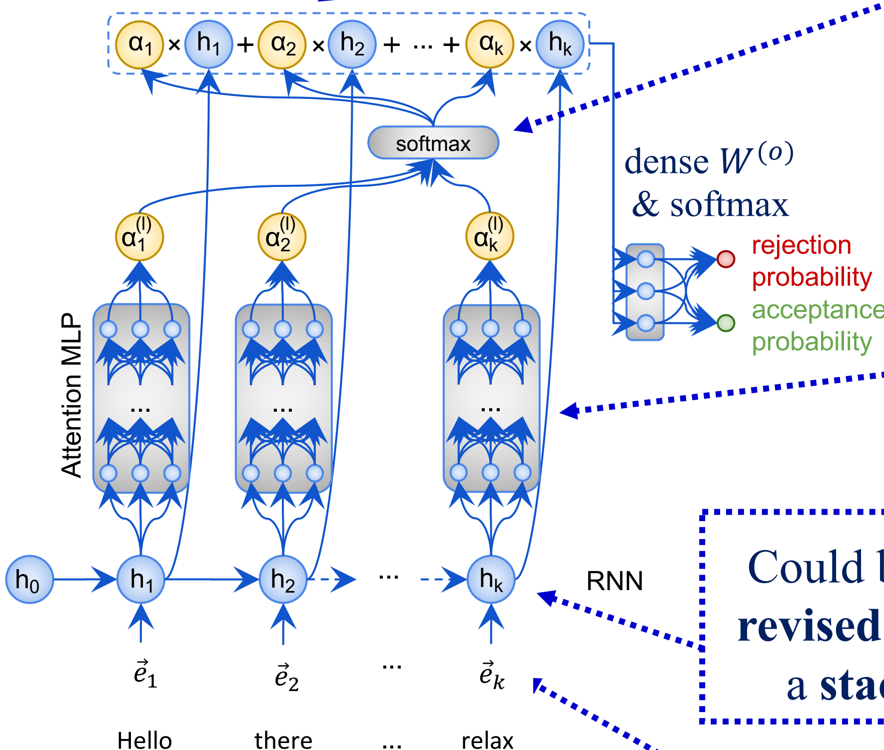
J. Pavlopoulos, P. Malakasiotis and I. Androutsopoulos, "Deeper Attention to Abusive User Content Moderation", EMNLP 2017, http://nlp.cs.aueb.gr/pubs/emnlp2017.pdf.

RNN with deep self-attention

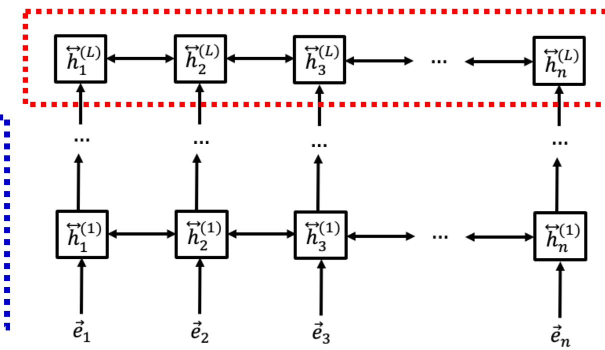
The **entire input text** is now represented by the **weighted (by a_i scores) sum** of the **revised embeddings** of its words.

The **softmax** ensures all the a_i scores are between 0 and 1, and that they sum to 1.

We use an **MLP** (the **same at all time-steps**) to obtain an **attention score** (importance) a_i **for each word from its revised embedding** h_i. We could also use a **single dense layer**: $a_i = W^{(a)} h_i$.



Could be the **top-level revised embeddings** of a **stacked biRNN**.



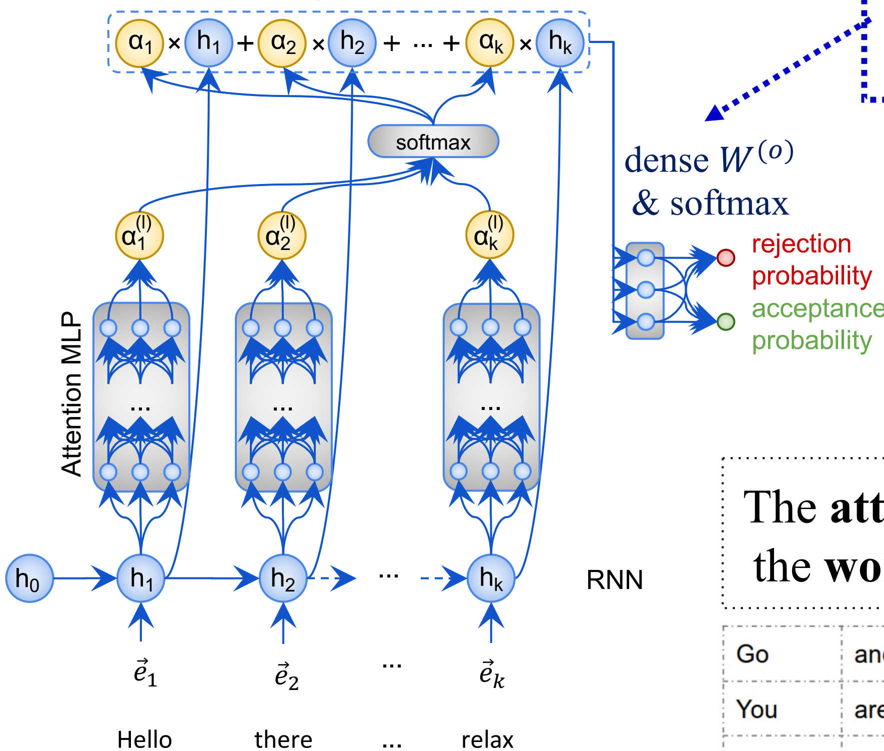
Initial word embeddings (e.g., via **Word2Vec**).

J. Pavlopoulos, P. Malakasiotis and I. Androutsopoulos, "Deeper Attention to Abusive User Content Moderation", EMNLP 2017, http://nlp.cs.aueb.gr/pubs/emnlp2017.pdf.

RNN with deep self-attention

The **entire input text** is now represented by the **weighted (by a_i scores) sum** of the **revised embeddings** of its words.

We pass the **weighted sum vector** (point) through another **dense layer and softmax** to obtain a **probability** score for **each class** (here accept, reject).



Compare to the correct predictions with a **cross-entropy loss** and **backpropagate** to **adjust the weights** of the **entire neural net**, including the MLP and RNN(s).

The **attention scores** a_i can also be used to **highlight** the **words** that influence the system's decision most.

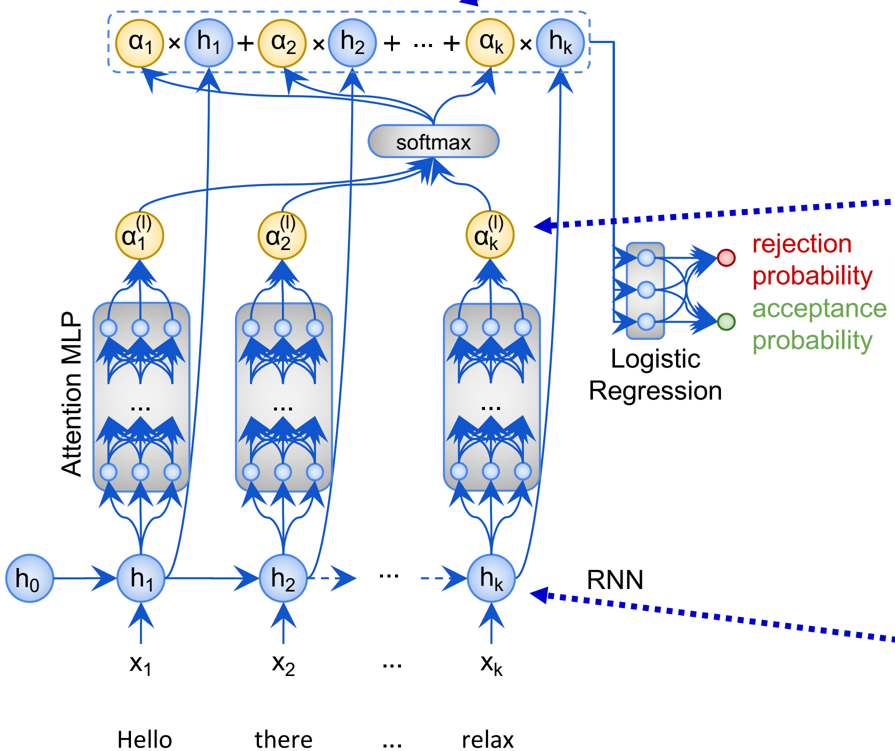
Go	and	hang	yourself	!				
You	are	ignorant	and	vandal	!	Stop	it	!
Thanks	.	Please	go	[redacted]	yourself	.	ty	!

J. Pavlopoulos, P. Malakasiotis and I. Androutsopoulos, "Deeper Attention to Abusive User Content Moderation", EMNLP 2017, http://nlp.cs.aueb.gr/pubs/emnlp2017.pdf.

RNN with deep self-attention

$$h_{sum} = \sum_{t=1}^{k} a_t h_t$$

$$P_{a-\text{RNN}}(reject|c) = \sigma(W_p h_{sum} + b_p)$$



$$a_t^{(1)} = \text{RELU}(W^{(1)} h_t + b^{(1)})$$

$$\cdots$$

$$a_t^{(l-1)} = \text{RELU}(W^{(l-1)} a_t^{(l-2)} + b^{(l-1)})$$

$$a_t^{(l)} = W^{(l)} a_t^{(l-1)} + b^{(l)}$$

$$a_t = \mathtt{softmax}(a_t^{(l)}; a_1^{(l)}, \dots, a_k^{(l)})$$

$$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h)$$

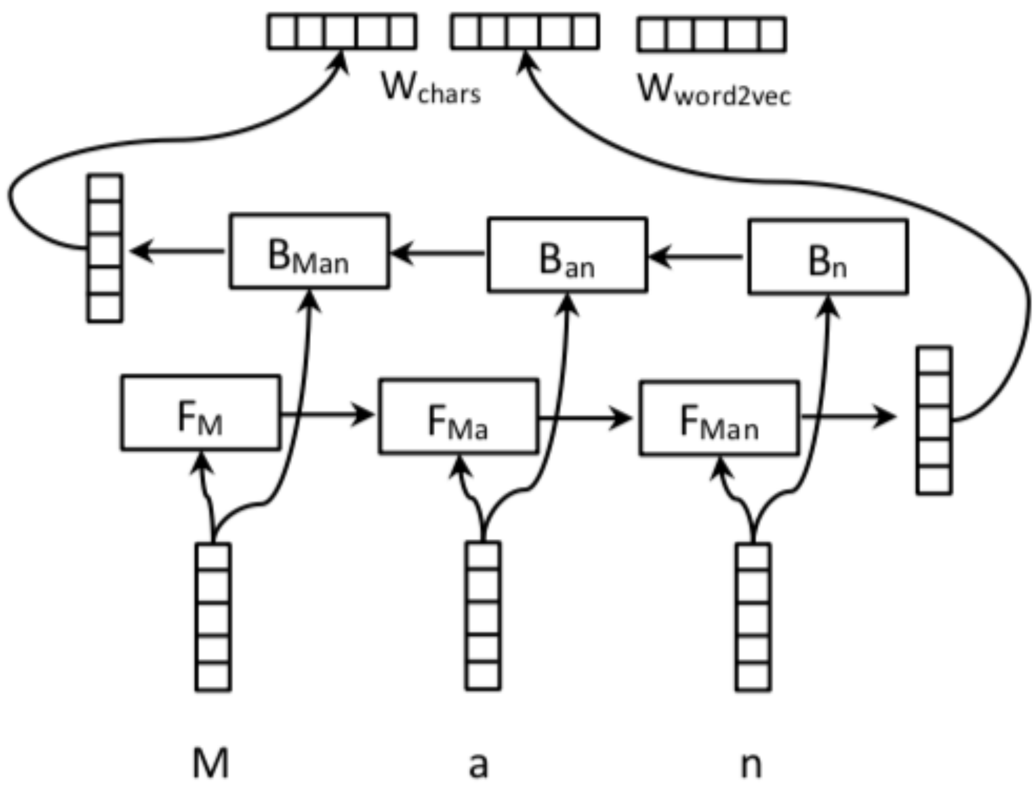
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

J. Pavlopoulos, P. Malakasiotis and I. Androutsopoulos, "Deeper Attention to Abusive User Content Moderation", EMNLP 2017, http://nlp.cs.aueb.gr/pubs/emnlp2017.pdf.

RNNs that produce word embeddings from character embeddings

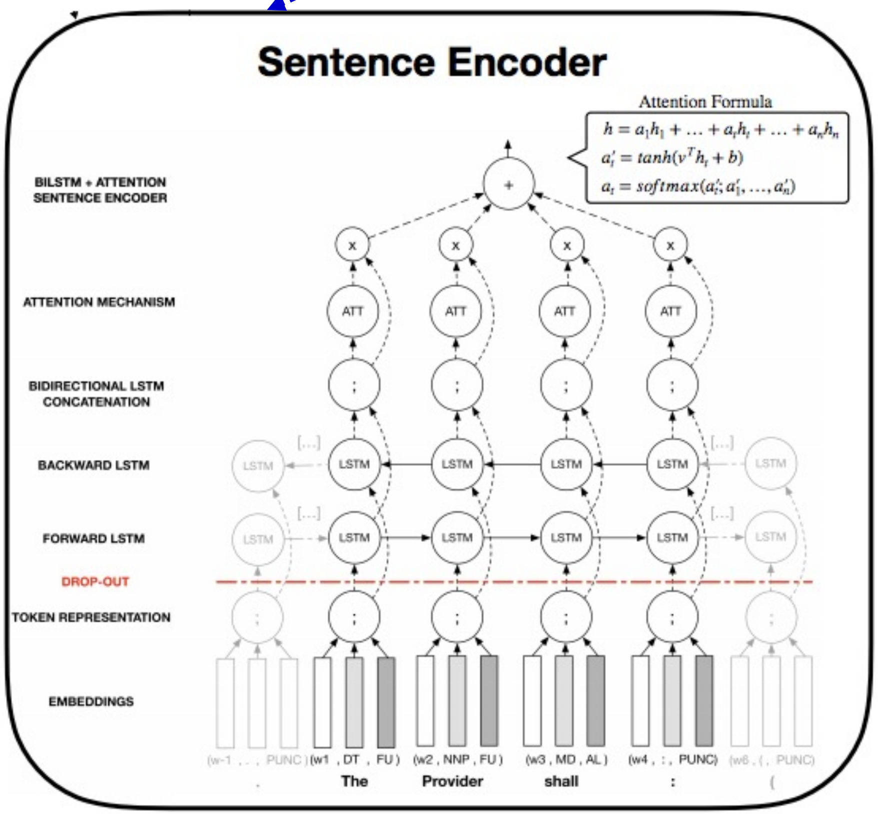


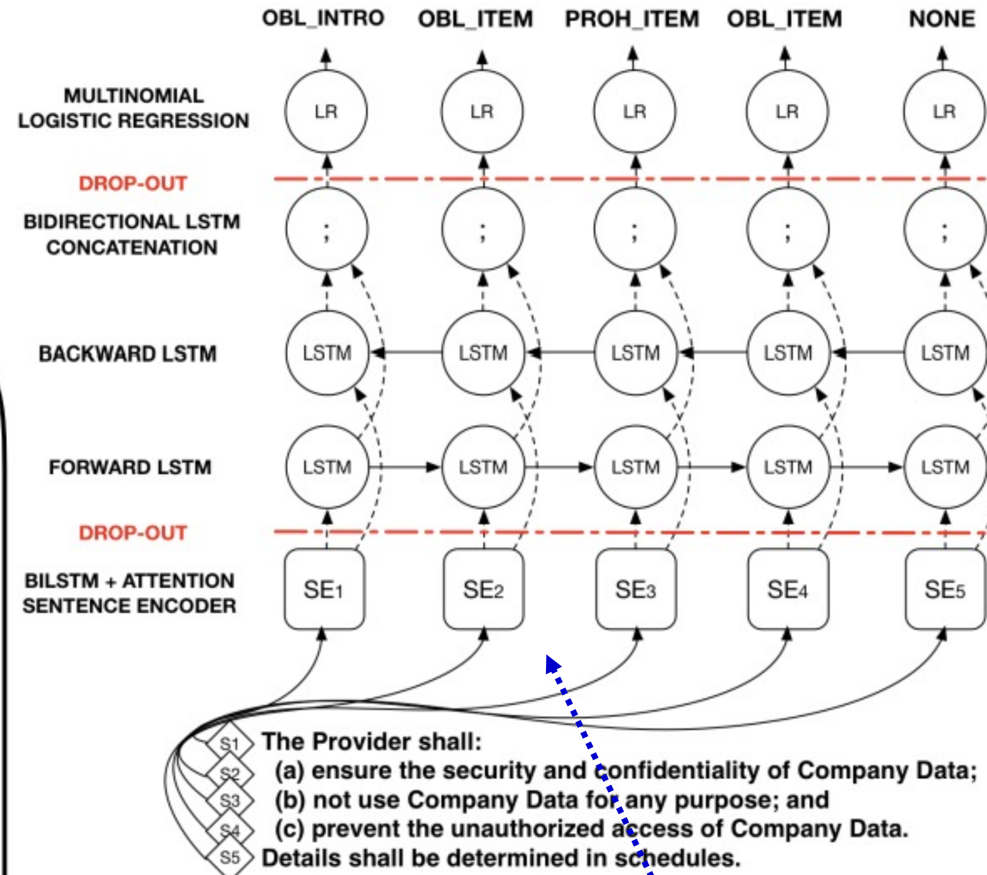
Word embedding layer, part of a larger network. We **concatenate** the **word embedding** we get from the **character-level biLSTM** with the **Word2Vec** embedding. The **character embeddings** are **learned** during back-propagation.

G, Bekoulis, J, Deleu, T, Demeester, C. Develder, "Joint entity recognition and relation extraction as a multi-head selection problem", Expert Systems with Applications, Vol, 114, pp. 34-45, 2018. Figure from the pre-print https://arxiv.org/abs/1804.07847.

Sequence labeling with a Hierarchical RNN

The **lower RNN** reads the words of each sentence and converts the sentence to a **sentence embedding**.





The **upper RNN** reads a sequence of sentence embeddings and **classifies each sentence**.

I. Chalkidis, I. Androutsopoulos, A. Michos, "Obligation and Prohibition Extraction Using Hierarchical RNNs", ACL 2018. http://www.aclweb.org/anthology/P18-2041

Legal judgment prediction for ECHR cases

Case ID: 001-148227 **Violated Articles:** Article 3 **Predicted Violation:** YES (0.97%)

1. The applicant was born in 1955 and lives in Kharkiv .

2. On 5 May 2004 the applicant was arrested by four police officers on suspicion of bribe - taking . The police officers took him to the Kharkiv Dzerzhynskyy District Police Station , where he was held overnight . According to the applicant , the police officers beat him for several hours , forcing him to confess .

3. On 6 May 2004 the applicant was taken to the Kharkiv City Prosecutor 's Office . He complained of ill-treatment to a senior prosecutor from the above office . The prosecutor referred the applicant for a forensic medical examination .

4. On 7 May 2004 the applicant was diagnosed with concussion and admitted to hospital .

5. On 8 May 2004 the applicant underwent a forensic medical examination , which established that he had numerous bruises on his face , chest , legs and arms , as well as a damaged tooth .

6. On 11 May 2004 criminal proceedings were instituted against the applicant on charges of bribe-taking . They were eventually terminated on 27 April 2007 for lack of corpus delicti .

7. On 2 June 2004 the applicant lodged another complaint of ill - treatment with the Kharkiv City Prosecutor 's Office .

Figure 1: Attention over words (colored words) and facts (vertical heat bars) as produced by HAN.

Words with **high attention** scores.

Sentences with **high attention** scores.

Biased against **particular locations**?

I. Chalkidis, I. Androutsopoulos and N. Aletras, "Neural Legal Judgment Prediction in English", ACL 2019. https://www.aclweb.org/anthology/P19-1424/

RNNs for Machine Translation

From the slides of R. Socher's course "Deep Learning for NLP", 2015. http://cs224d.stanford.edu/

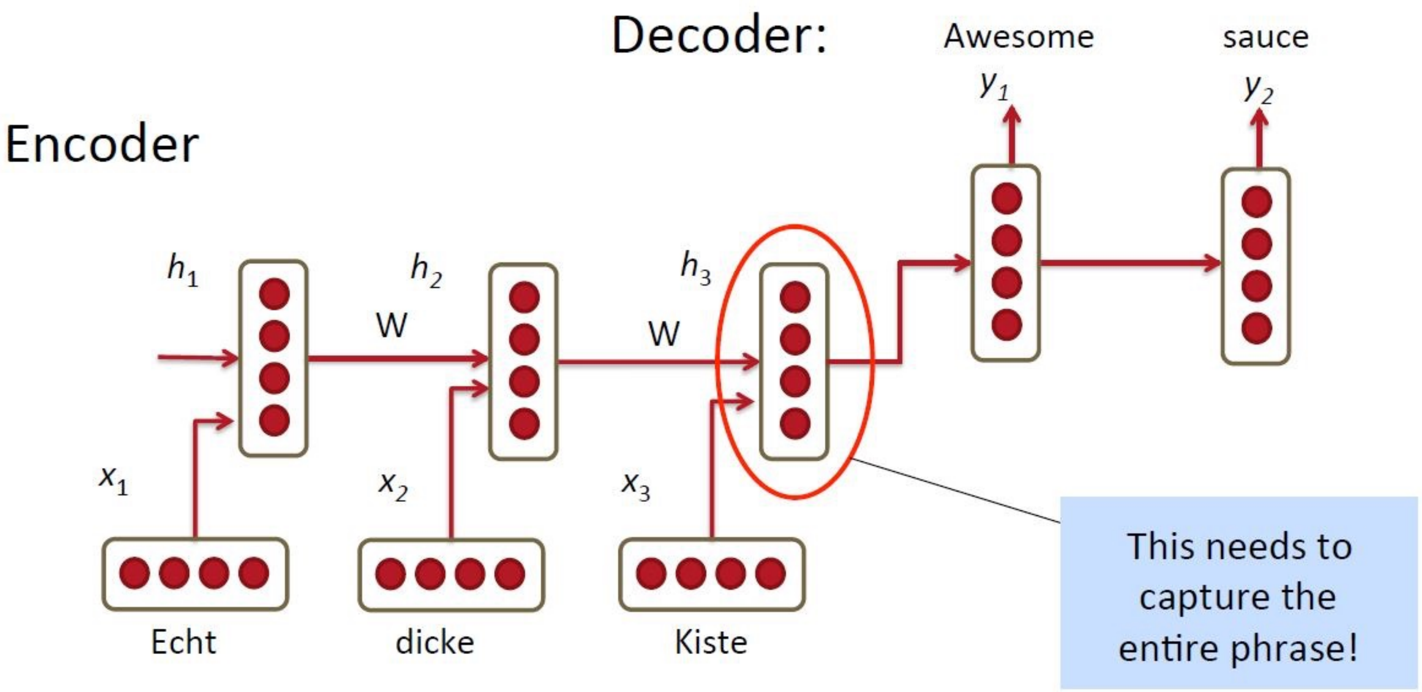
Encoder: $h_t = \phi(h_{t-1}, x_t) = f\left(W^{(hh)} h_{t-1} + W^{(hx)} x_t\right)$

Decoder: $h_t = \phi(h_{t-1}) = f\left(W^{(hh)} h_{t-1}\right)$

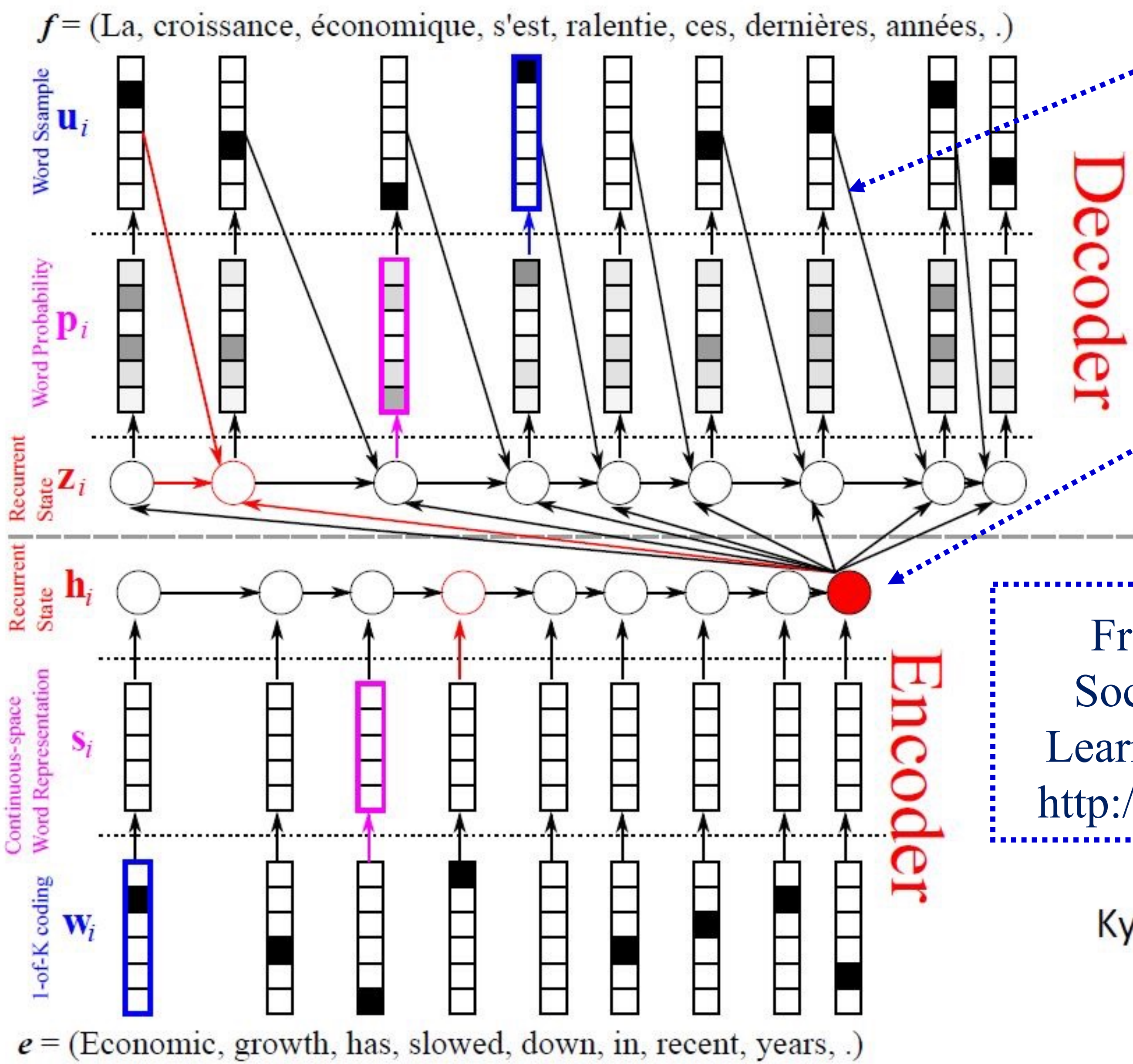
$$y_t = softmax\left(W^{(S)} h_t\right)$$

Minimize cross entropy error for all target words conditioned on source words

$$\max_\theta \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(y^{(n)} | x^{(n)})$$



Different picture, same idea



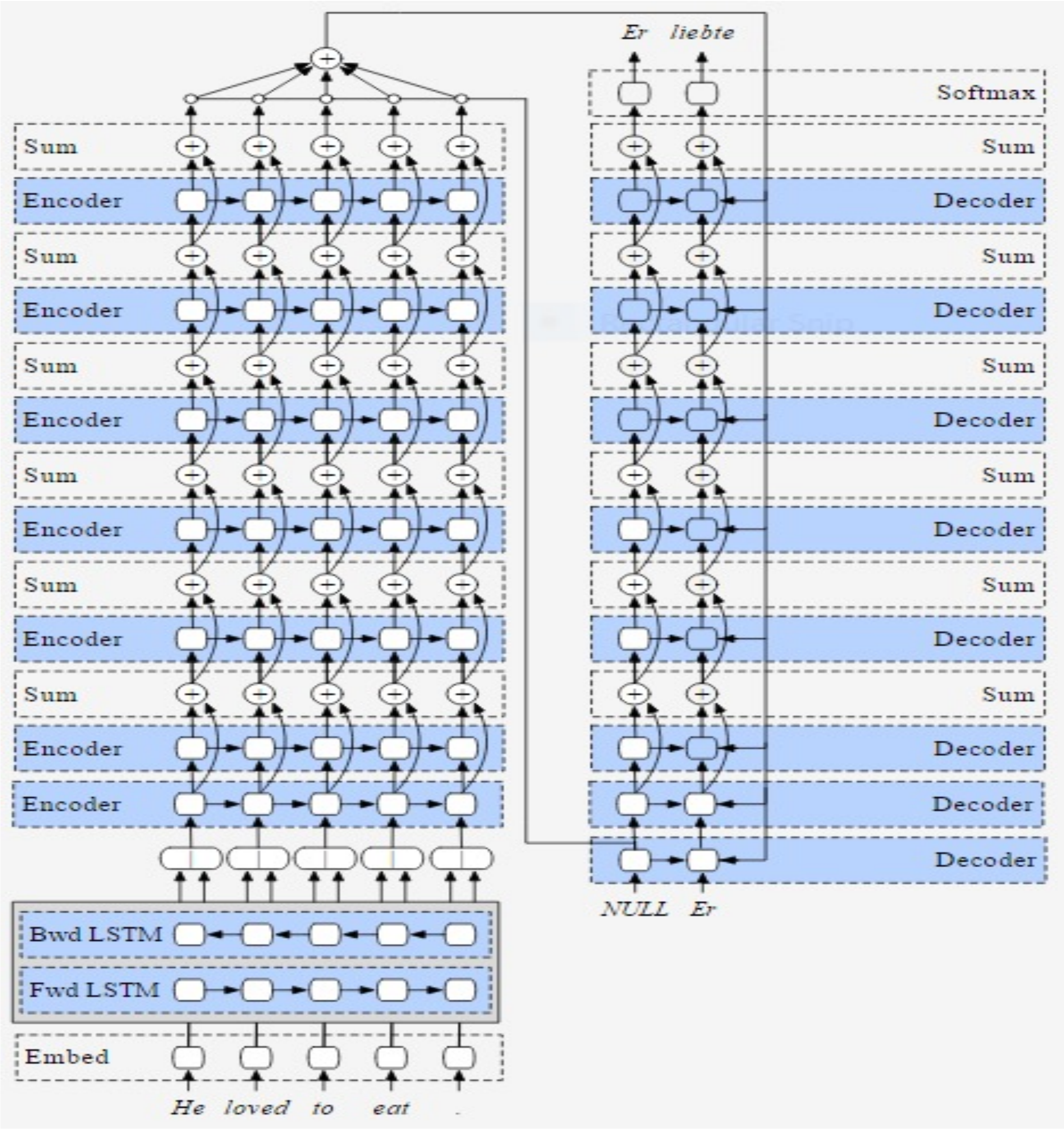
Embedding of the **previously generated word**.

Last $\vec{h}_i$ of the **encoder RNN**. Treated as **embedding** of the entire **input sentence**.

From the slides of R. Socher's course "Deep Learning for NLP", 2015. http://cs224d.stanford.edu/

Kyunghyun Cho et al. 2014

RNN-based Machine Translation

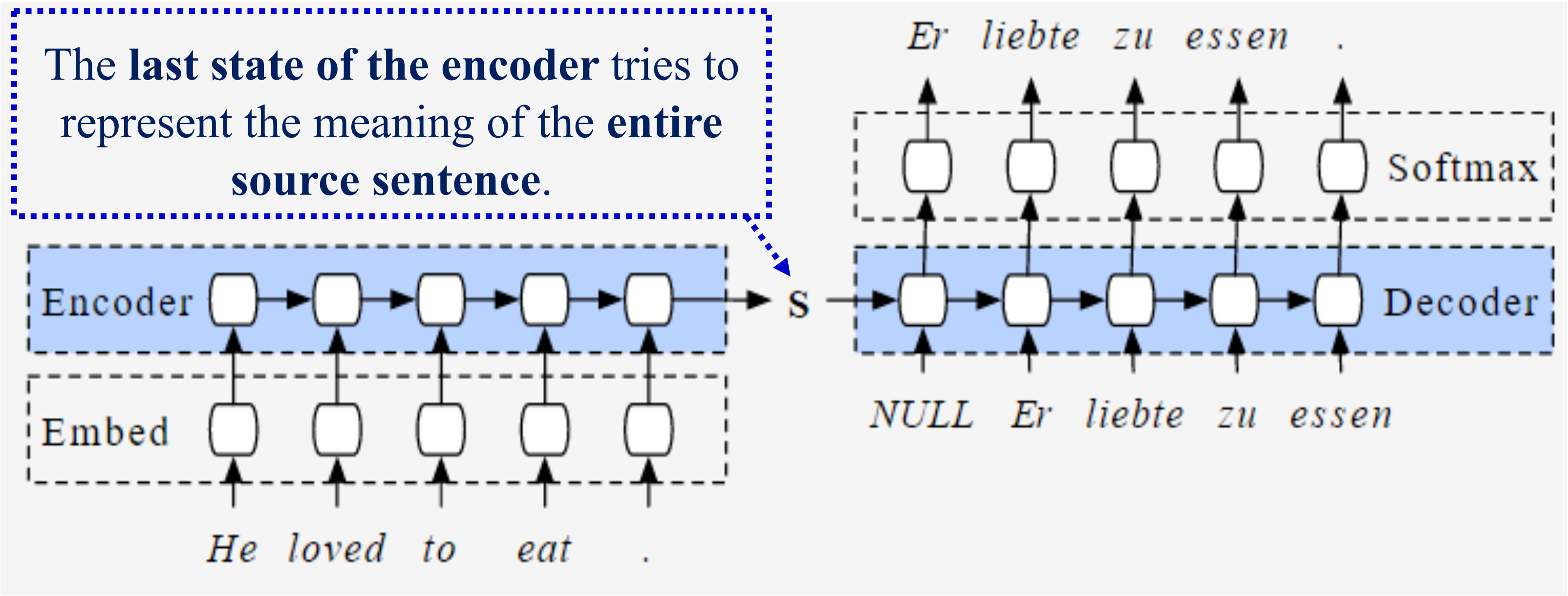


Google's paper: https://arxiv.org/abs/1609.08144

Images from Stephen Merity's http://smerity.com/articles/2016/google_nmt_arch.html

Easier to explain step by step…

Basic Encoder-Decoder NMT



Images from Stephen Merity's http://smerity.com/articles/2016/google_nmt_arch.html

During training, at each decoding time-step, we can always use the **correct previous word** ("**teacher forcing**"); or we can **randomly use the correct or** (increasingly more often) the **predicted** previous word (**scheduled sampling**).

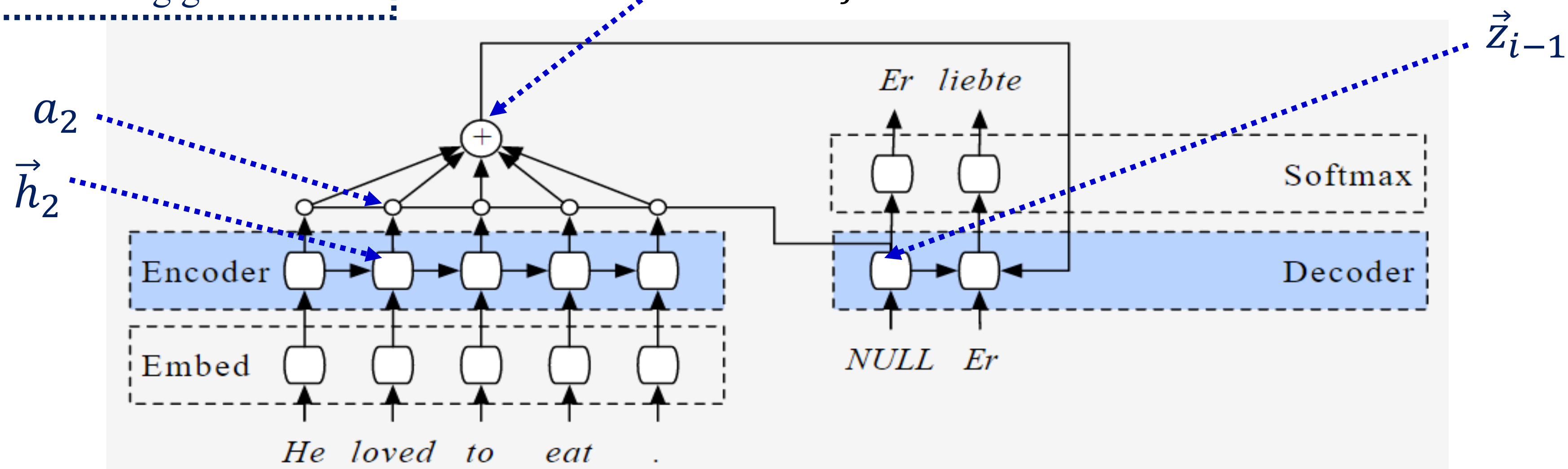
During testing (inference), we always use the **predicted previous word**; and we **greedily select the most probable next word,** or we **sample from the distribution of the next word**, or we use **beam search** to find the translation $y_1, \ldots y_m$ with the highest probability: $p(y_1|z_1)\; p(y_2|y_1, z_2)\; p(y_3|y_2, z_3) \ldots\; p(y_m|y_{m-1}, z_m)$ where z_i are the states of the decoder.

Encoder-Decoder with attention

The **source sentence** is now represented by the **weighted sum** of the **encoder states**:

Attention also mitigates vanishing gradients.

$$\vec{h}_{sum} = \sum_j a_j \vec{h}_j$$



For each German word, the **attention scores** over the English words **change**!

Each "**attention**" **weight $\boldsymbol{a_j}$** is a **function** of the **corresponding encoder state $\vec{\boldsymbol{h}}_{\boldsymbol{j}}$** and the **previous state $\vec{\boldsymbol{z}}_{\boldsymbol{i-1}}$ of the decoder** (memory of translation so far), e.g.:

$$\tilde{a}_j = v^T \cdot f\left(W^{(h)}\vec{h}_j + W^{(z)}\vec{z}_{i-1}\right) = v^T \cdot f\left(W\left[\vec{h}_j; \vec{z}_{i-1}\right]\right), \qquad a_j = softmax(\tilde{a}_j)$$

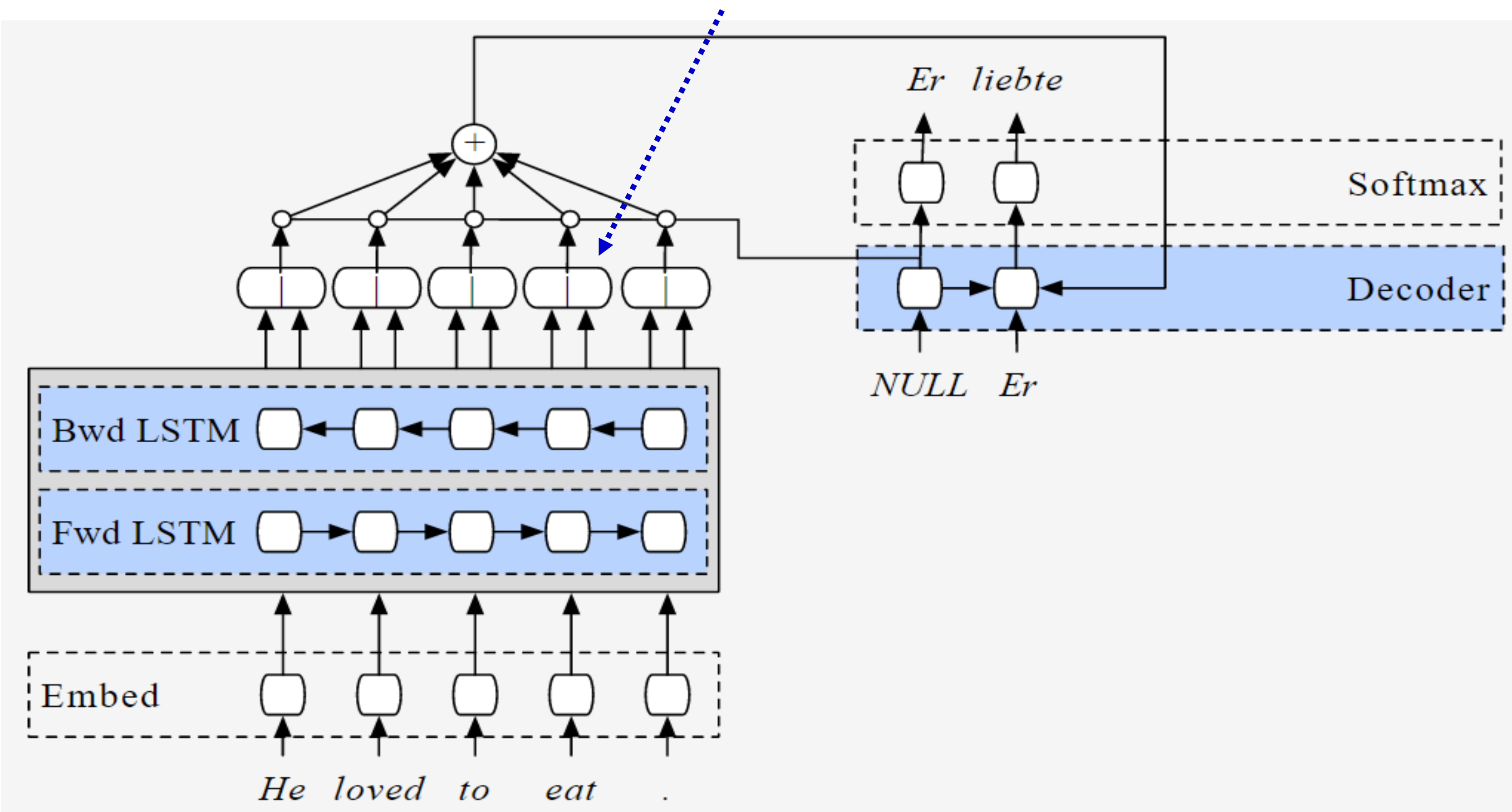
with a **softmax** to make the $\boldsymbol{a_j}$ weights sum to 1.

Google's paper: https://arxiv.org/abs/1609.08144

Images from Stephen Merity's http://smerity.com/articles/2016/google_nmt_arch.html

Bidirectional LSTM encoder

The encoder is now a **bidirectional LSTM**. The **encoder state** for the ***j*-th word of the source** sentence is the **concatenation** of the **corresponding states** of the **forward** and **backward** LSTM.



Google's paper: https://arxiv.org/abs/1609.08144
Images from Stephen Merity's http://smerity.com/articles/2016/google_nmt_arch.html

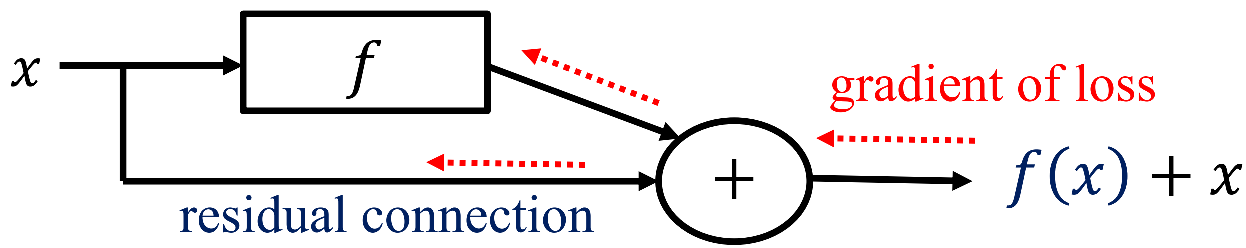
Residual connections

- Given a **block** (part of a neural network) that would **normally compute**:

$$y = f(x)$$

where x, y are vectors, we **add its input to its output**.

$$y = f(x) + x$$



- During **backpropagation**, the **addition** gate just **copies the gradient** from its output to its inputs (lecture 13).
- If the **gradient vanishes** in the ***f* block**, it will reach the previous blocks (that produced x) via the residual.

Residual connections – continued

- Residual connections **allow stacking** more layers/blocks **without vanishing gradients**.
- They also give the network the **option not to use *f*** (by learning weights that always produce $f(x) = 0$), but still **pass on to following blocks** the **information *x*** that a previous block computed.
- **Residuals** were first used (and are still used) in **CNNs**, but they are now also used in **RNNs** and **Transformers**.

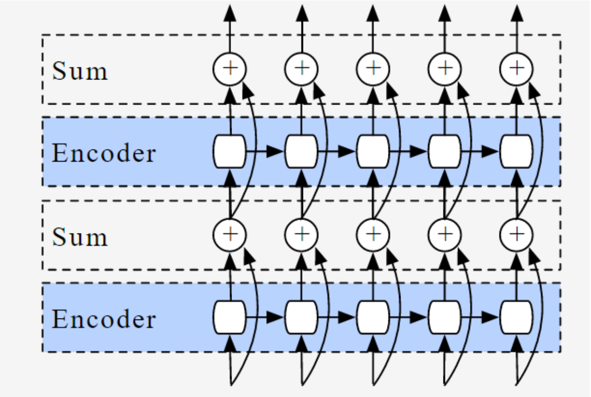
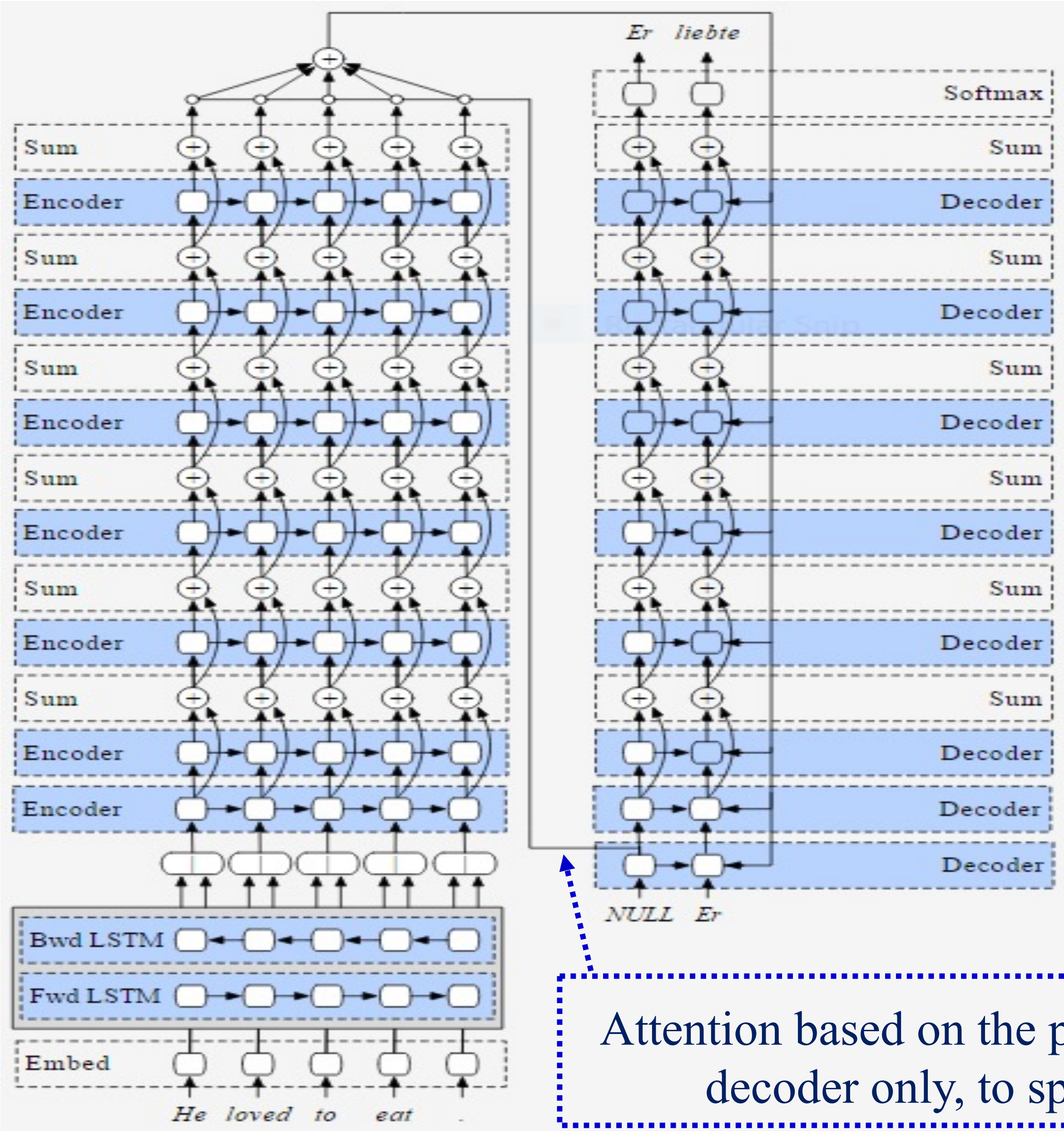


Image from Stephen Merity's http://smerity.com/articles/2016/google_nmt_arch.html

RNN-based Machine Translation



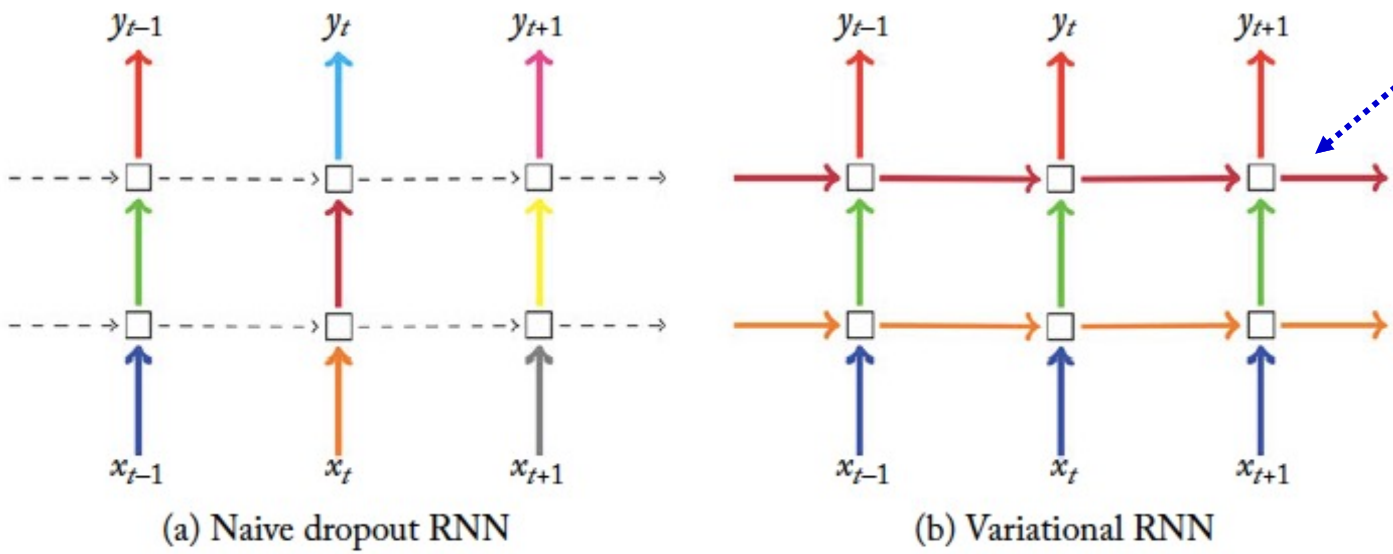
Google's paper:
https://arxiv.org/abs/1609.08144

Images from Stephen Merity's
http://smerity.com/articles/2016/google_nmt_arch.html

Attention based on the previous state of the bottom decoder only, to speed up computations.

Additional optional reading slides.

Dropout vs. Variational Dropout

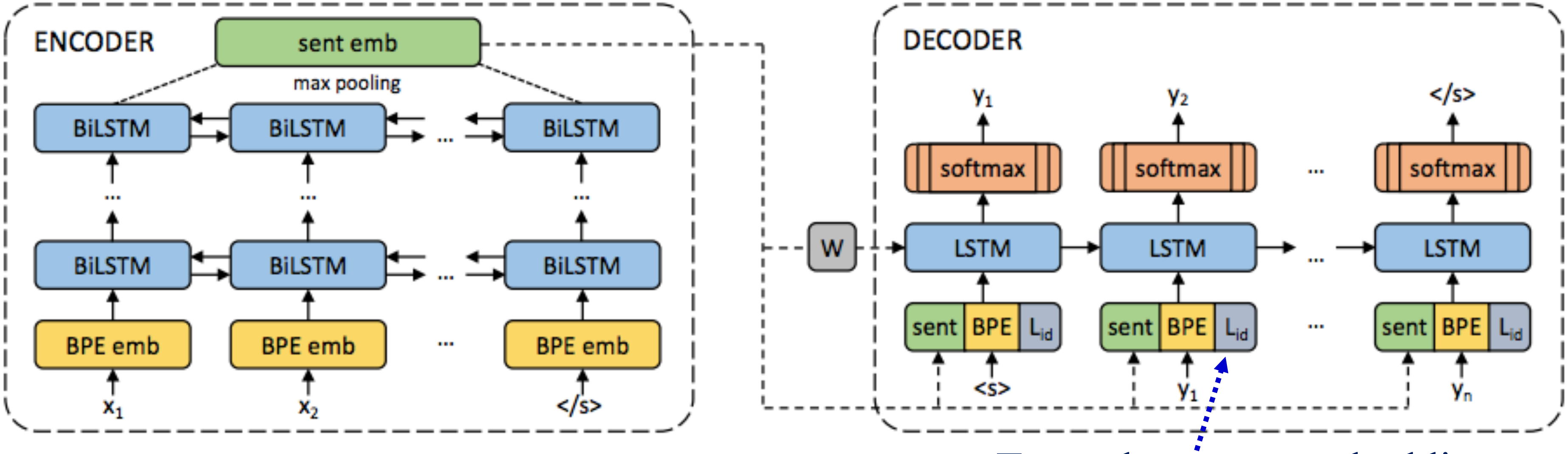


Dropout **masks** sampled once for each sequence, then kept the **same across all time-steps**. Better for RNNs.

Figure 15.2: Gal's proposal for RNN dropout (b), vs. the previous suggestion by Pham et al. [2013], Zaremba et al. [2014] (a). Figure from Gal [2015], used with permission. Each square represents an RNN unit, with horizontal arrows representing time dependence (recurrent connections). Vertical arrows represent the input and output to each RNN unit. Colored connections represent dropped-out inputs, with different colors corresponding to different dropout masks. Dashed lines correspond to standard connections with no dropout. Previous techniques (naive dropout, left) use different masks at different time steps, with no dropout on the recurrent layers. Gal's proposed technique (Variational RNN, right) uses the same dropout mask at each time step, including the recurrent layers.

Figure from: Y. Goldberg, *Neural Network Models for Natural Language Processing*, Morgan & Claypool Publishers, 2017. See also https://adriangcoder.medium.com/a-review-of-dropout-as-applied-to-rnns-72e79ecd5b7b

Universal sentence encoders

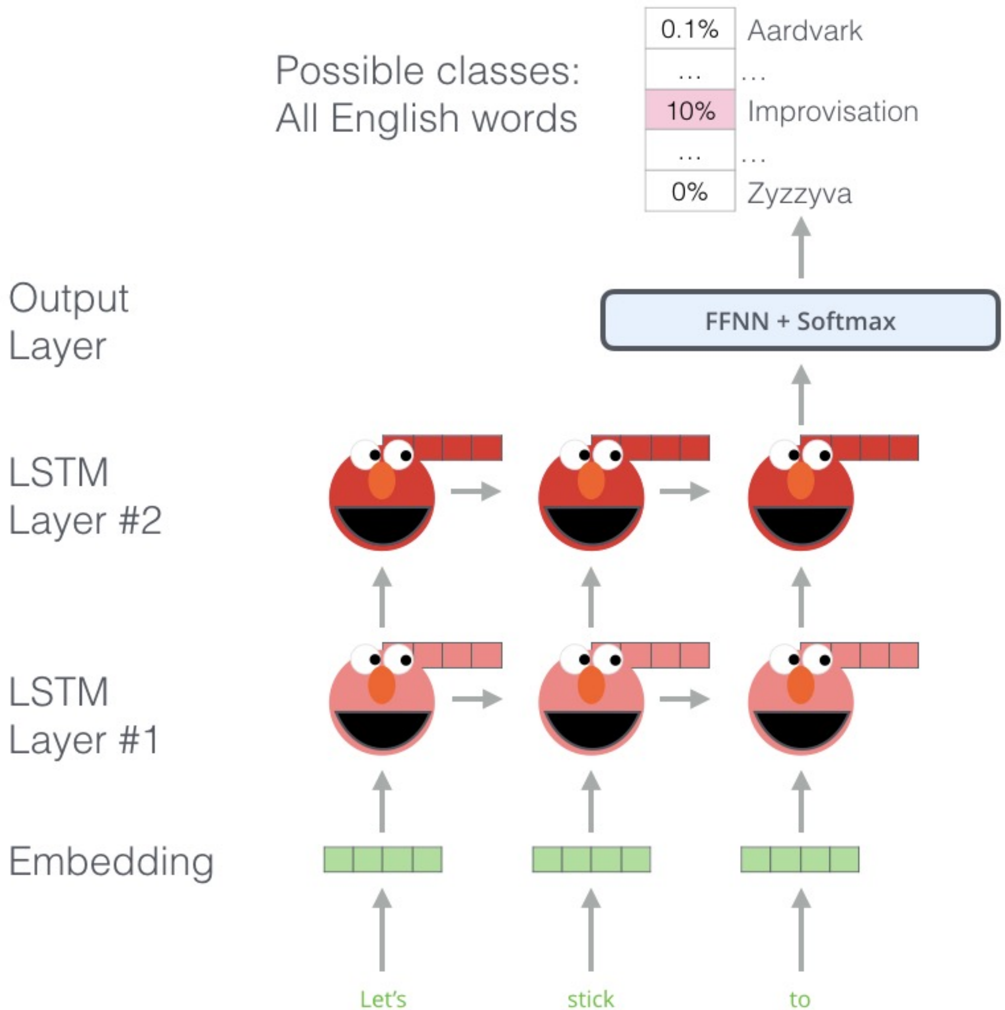


Target language embedding

- **Laser: Trained** on **parallel corpora** of 93 languages.
  - Using the **same encoder** and **decoder** for all languages.
  - **Shared vocabulary** of **sub-word units** (BPEs).
- E.g., we can **train a classifier** on **English tweets**, and use the **same trained classifier** on **Greek tweets**.
  - In **both languages**, we use the **same encoder** to convert each **tweet** to a **feature vector**.

M. Artetxe and H. Schwenk, "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond". https://arxiv.org/abs/1812.10464
https://code.fb.com/ai-research/laser-multilingual-sentence-embeddings/

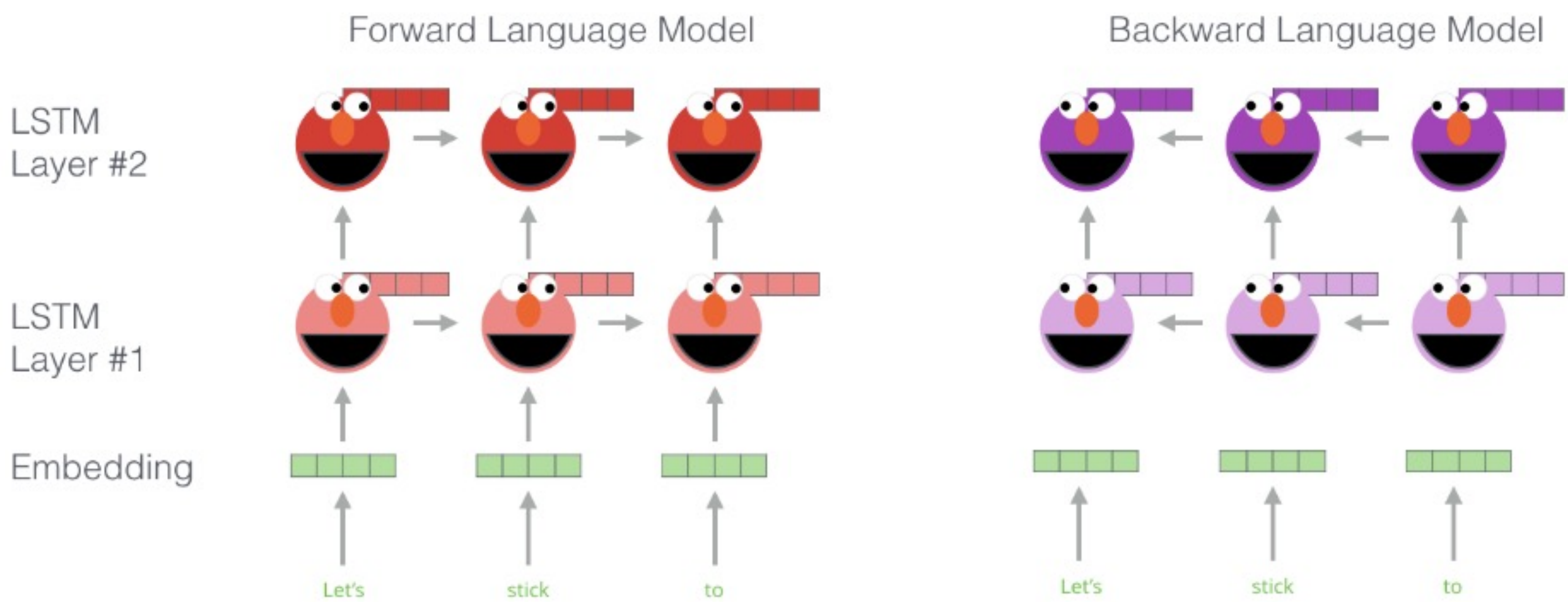
ELMo – Pretraining LMs to obtain context aware word embeddings



Bottom **word embeddings** generated by **CNNs operating on characters**.

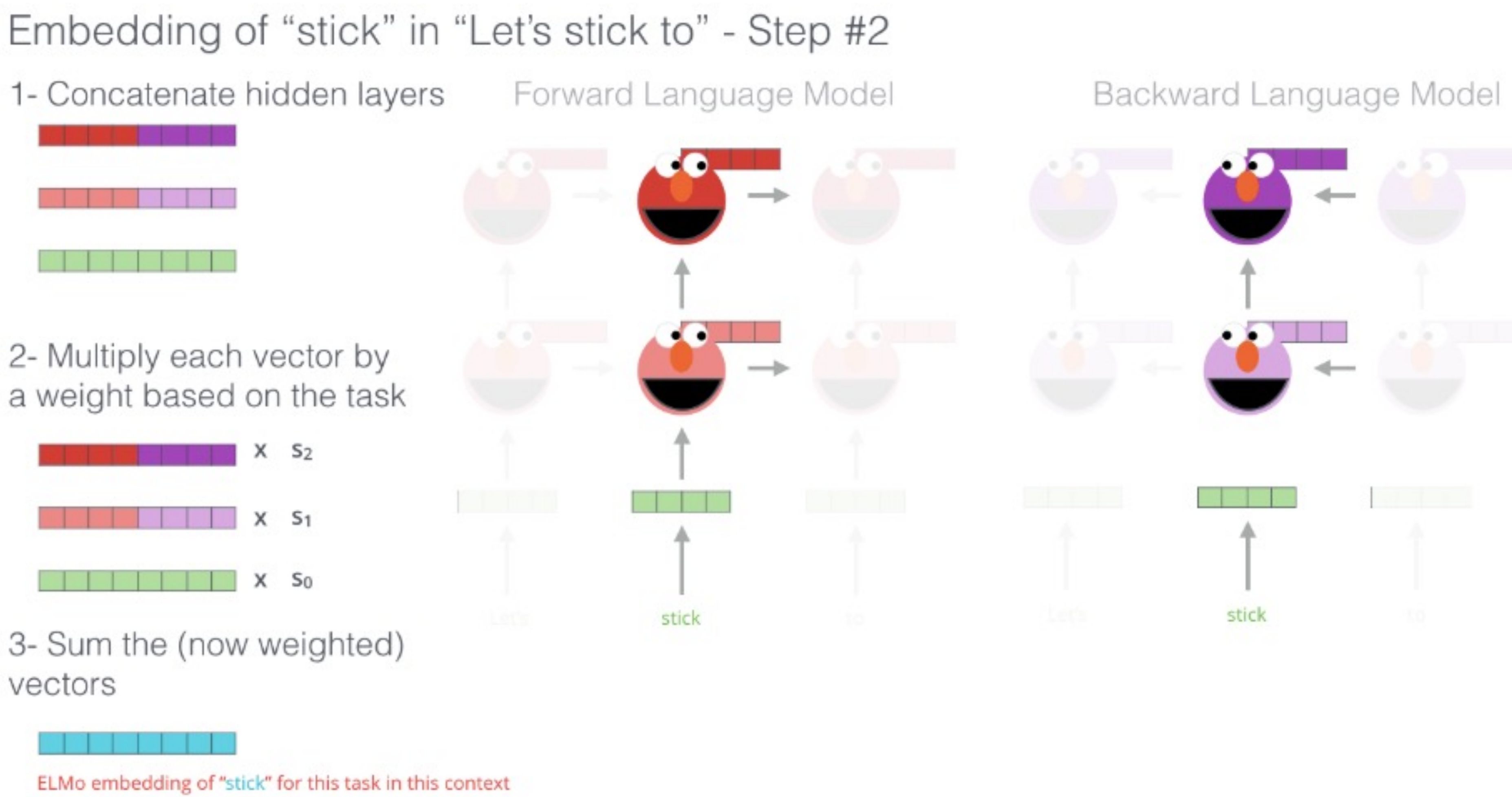
Figures from J. Alammar's "The Illustrated BERT, ELMo, and co." http://jalammar.github.io/illustrated-bert/. ELMo paper: Peters et al. "Deep Contextualized Word Representations", NAACL-HLT 2018. http://aclweb.org/anthology/N18-1202

ELMo – Pretraining LMs to obtain context aware word embeddings



Figures from J. Alammar's "The Illustrated BERT, ELMo, and co." http://jalammar.github.io/illustrated-bert/. ELMo paper: Peters et al. "Deep Contextualized Word Representations", NAACL-HLT 2018. http://aclweb.org/anthology/N18-1202

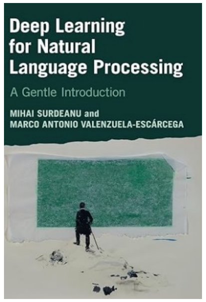
ELMo – Pretraining LMs to obtain context aware word embeddings



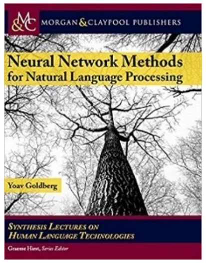
Figures from J. Alammar’s “The Illustrated BERT, ELMo, and co.” http://jalammar.github.io/illustrated-bert/. ELMo paper: Peters et al. “Deep Contextualized Word Representations”, NAACL-HLT 2018. http://aclweb.org/anthology/N18-1202

Recommended reading

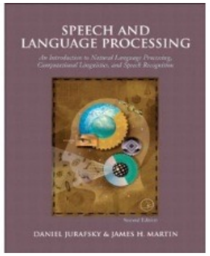
- M. Surdeanu and M.A. Valenzuela-Escarcega, *Deep Learning for Natural Language Processing: A Gentle Introduction,* Cambridge Univ. Press, 2024.
  - Chapters 11, 12, 14. See https://clulab.org/gentlenlp/text.html
  - Also available at AUEB's library.



- Y. Goldberg, *Neural Network Models for Natural Language Processing*, Morgan & Claypool Publishers, 2017.
  - Mostly chapters 14–17.



- Jurafsky and Martin's, *Speech and Language Processing* is being revised (3rd edition) to include DL methods.
  - http://web.stanford.edu/~jurafsky/slp3/



Recommended reading

- A. Zhang et al., *Dive into Deep Learning*.
  - Freely available at: https://d2l.ai/
  - See Chapters 9 and 10 for RNNs.