# Sequence-to-Sequence Learning with Neural Networks

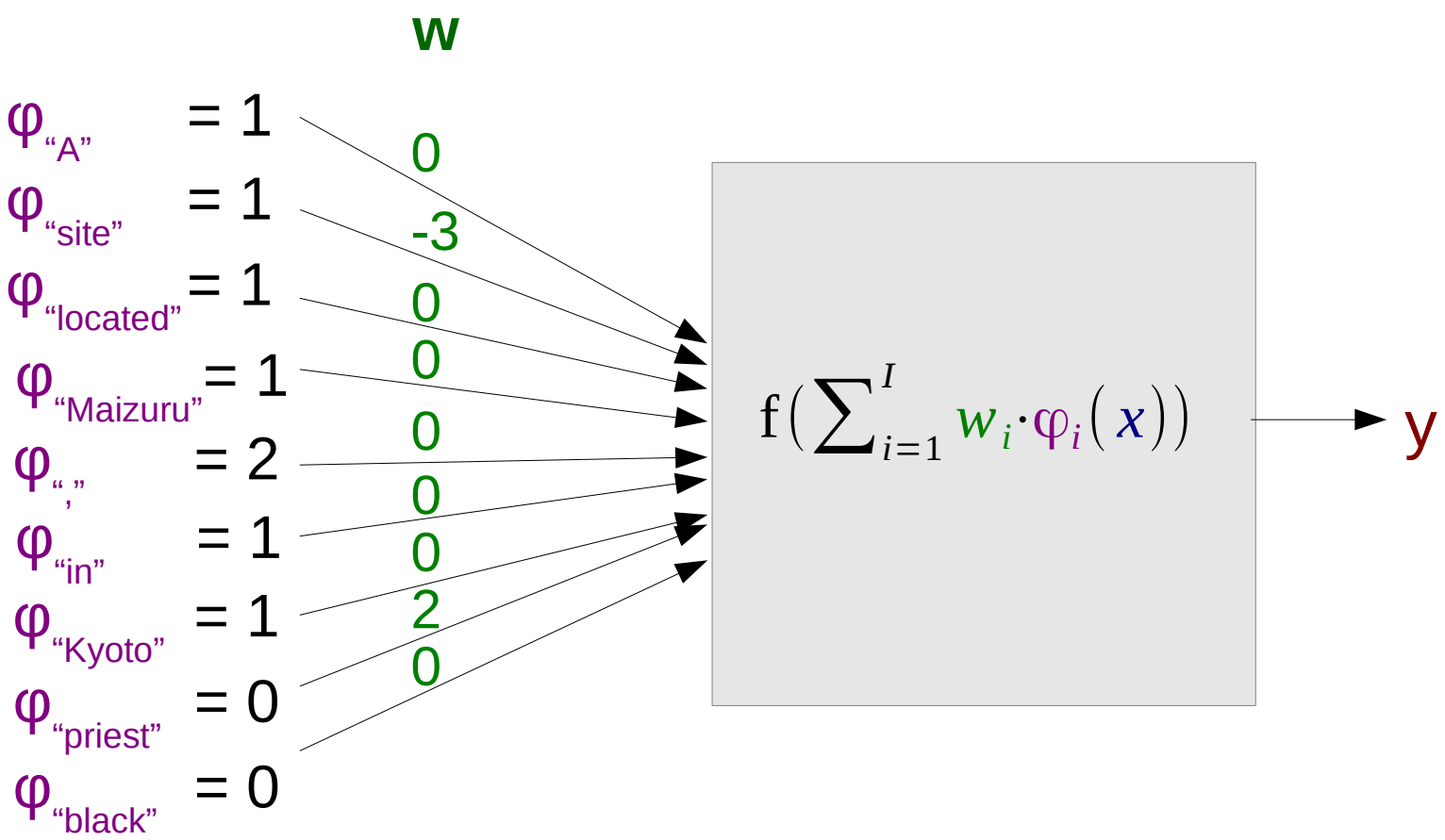
Ilya Sutskever, Oriol Vinyals, Quoc V. Le, NIPS 2014

Introduced by Graham Neubig, NAIST

2014-11-01

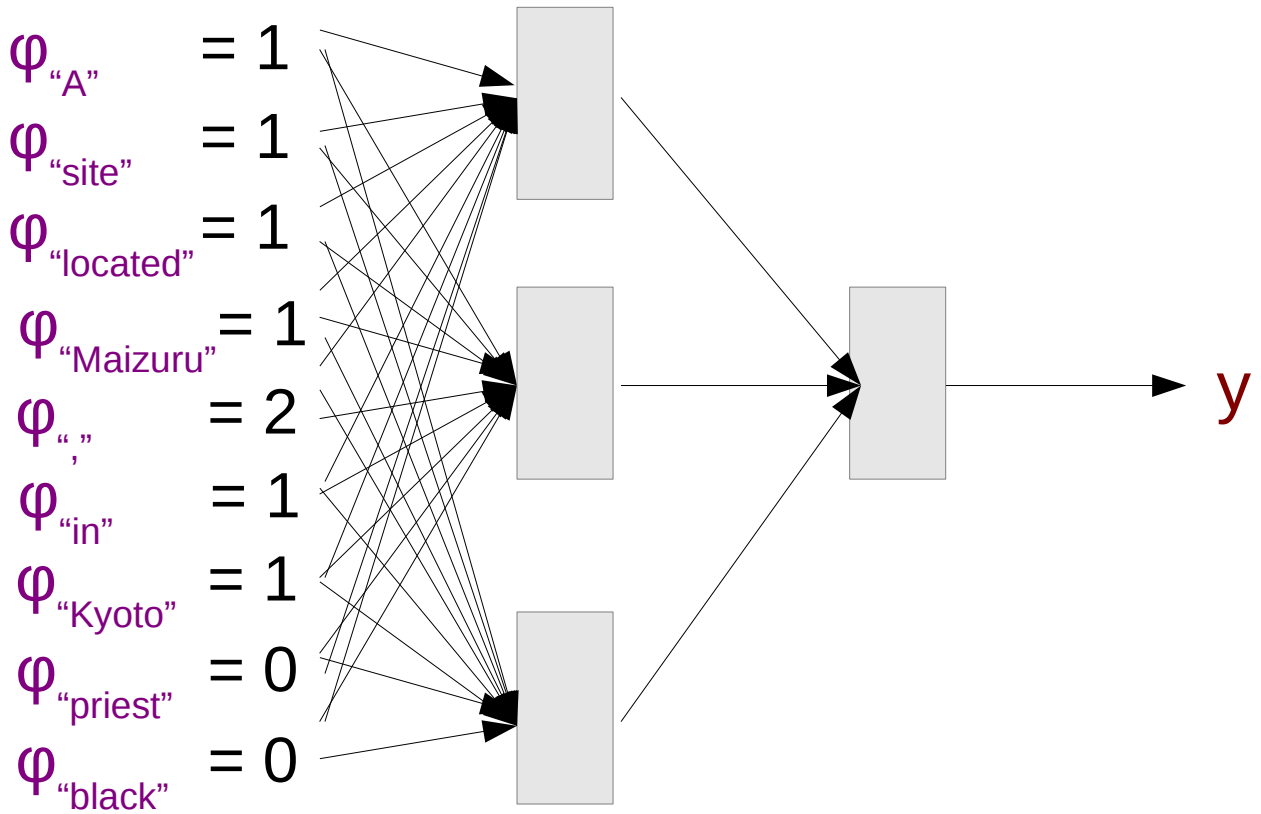
# Review: Recurrent Neural Networks

# Perceptron


w
φ“A” = 1
φ“site” = 1
φ“located” = 1
φ“Maizuru” = 1
φ“,” = 2
φ“in” = 1
φ“Kyoto” = 1
φ“priest” = 0
φ“black” = 0
0
-3
0
0
0
0
0
2
0
$f\left(\sum_{i=1}^{I} w_i \cdot \varphi_i(x)\right)$
y

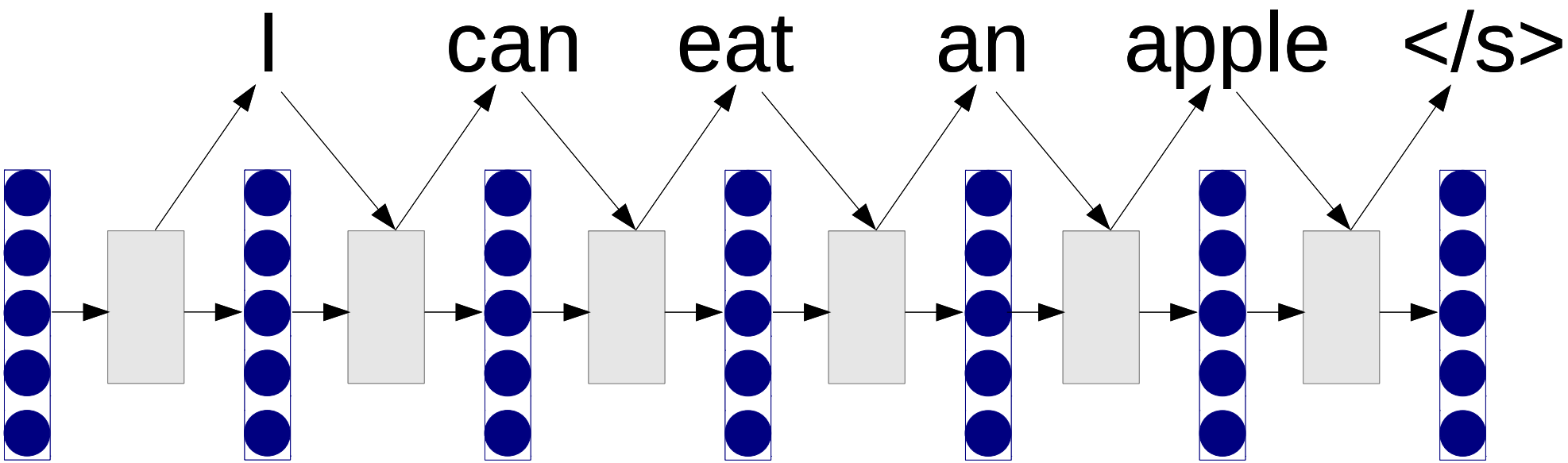
# Neural Net

- Combine multiple perceptrons



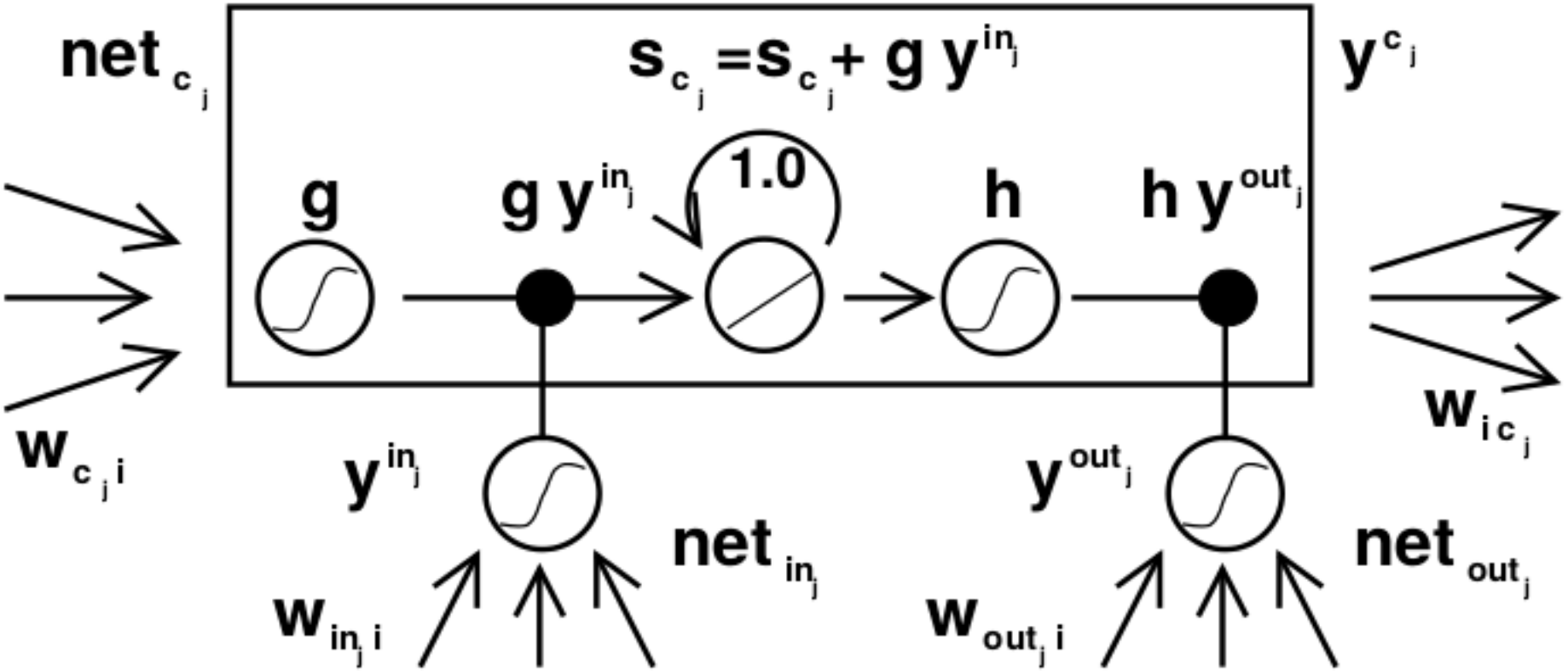

- Learning of complex functions possible

# Recurrent Neural Nets

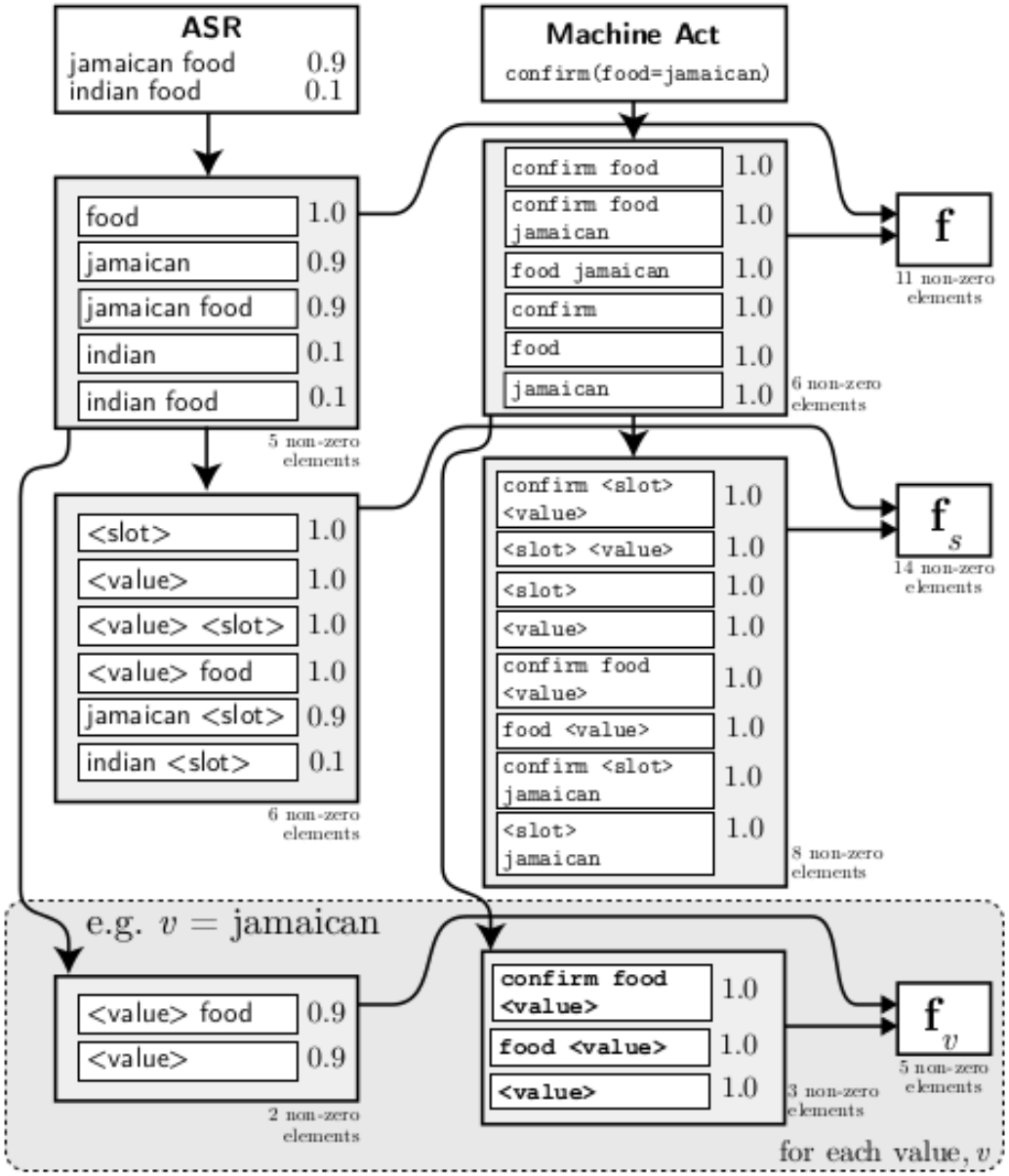


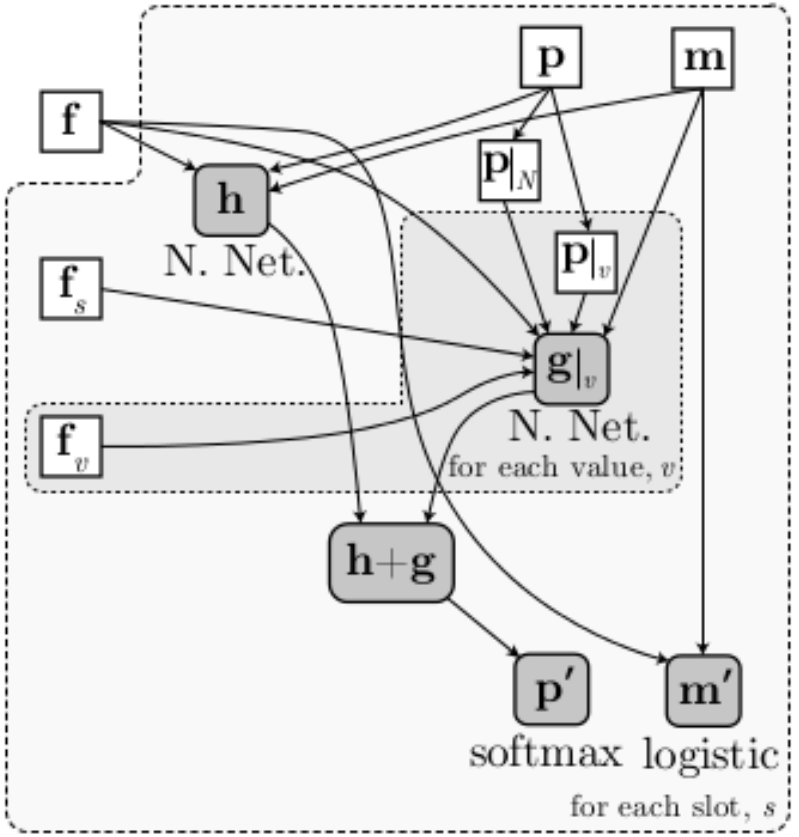
# Long Short Term Memory [Hochreiter+ 97]

- Problem: RNNs suffer from vanishing gradient
- Solution: Create units that decide when to activate



# Dialogue State Tracking with RNNs [Henderson+ 14]



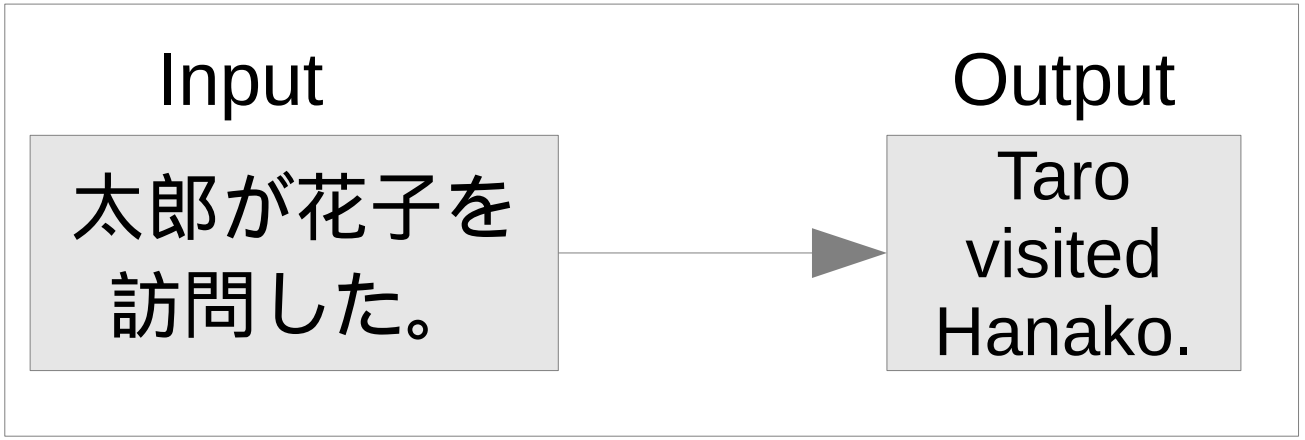




f: features
s: slot
m: memory
p: probability over goals

# Sequence-to-Sequence Learning with Neural Networks

# Task: Machine Translation

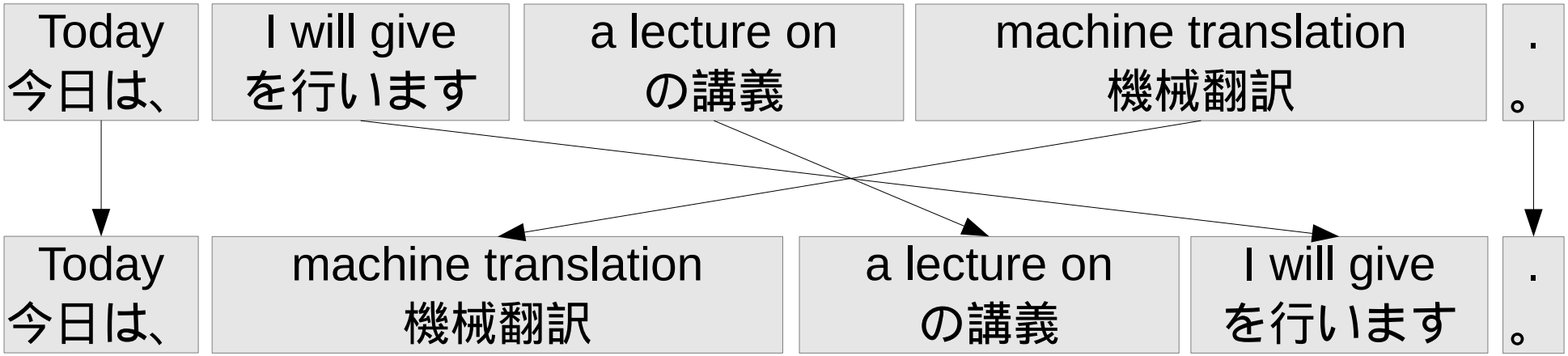
- Mapping from input to output sentence



# Traditional Method: Phrase-based MT

- Translate phrases, reorder

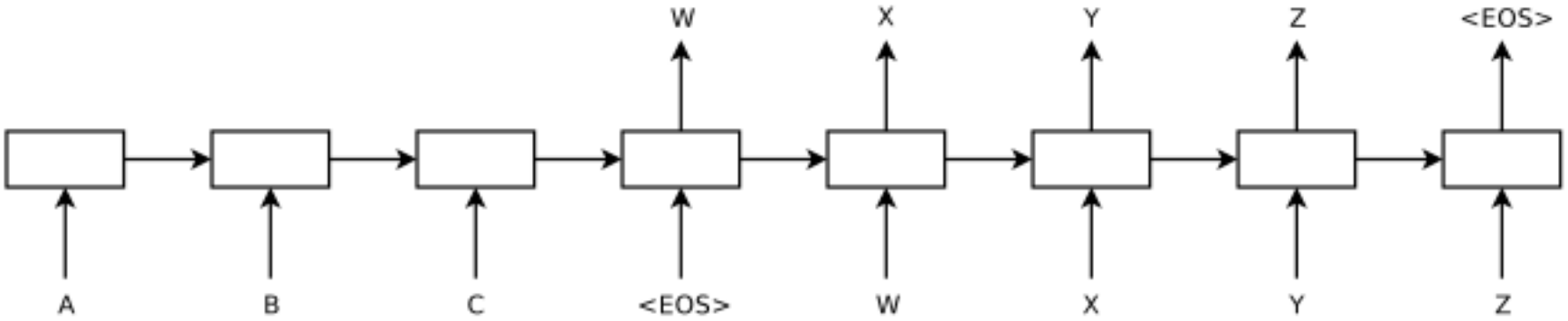
Today I will give a lecture on machine translation .




今日は、機械翻訳の講義を行います。

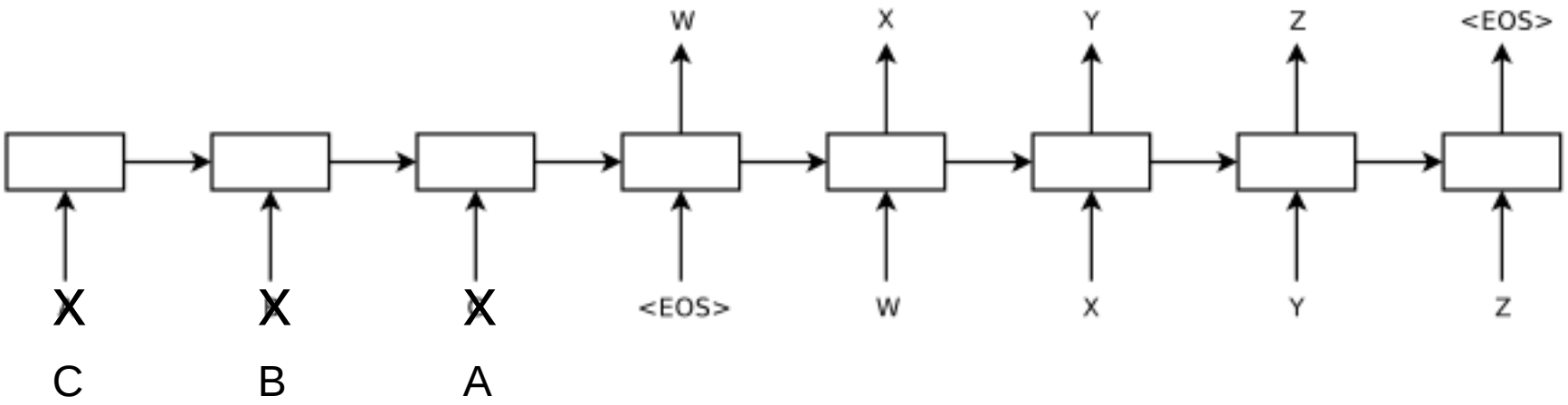
- Requires alignment, phrase extraction, scoring (phrase, reordering), NP-hard decoding, tuning

# Proposed Method: Memorize Sequence, Generate Sequence




- Left-to-right beam search (size 2 was largely sufficient)
- Also can use for reranking

# Proposed Method: Reversal Trick



# Experimental Setup

- Network details
  - 160,000/80,000 word input/output (all other UNK)
  - 4 hidden LSTM layers of 1,000 cells
  - 1,000 dimensional word representations
- Training
  - Stochastic gradient descent
  - 8 GPUs (1 for each hidden layer, 4 for output)
  - 6,300 words per second, 10 days total
- Data details
  - ~340M words of English-French data from WMT14

# Results

| Method | test BLEU score (ntst14) |
|---|---|
| Bahdanau et al. [2] | 28.45 |
| Baseline System [29] | 33.30 |
| Single forward LSTM, beam size 12 | 26.17 |
| Single reversed LSTM, beam size 12 | 30.59 |
| Ensemble of 5 reversed LSTMs, beam size 1 | 33.00 |
| Ensemble of 2 reversed LSTMs, beam size 12 | 33.27 |
| Ensemble of 5 reversed LSTMs, beam size 2 | 34.50 |
| Ensemble of 5 reversed LSTMs, beam size 12 | **34.81** |

Table 1: The performance of the LSTM on WMT'14 English to French test set (ntst14). Note that an ensemble of 5 LSTMs with a beam of size 2 is cheaper than of a single LSTM with a beam of size 12.

| Method | test BLEU score (ntst14) |
|---|---|
| Baseline System [29] | 33.30 |
| Cho et al. [5] | 34.54 |
| State of the art [9] | **37.0** |
| Rescoring the baseline 1000-best with a single forward LSTM | 35.61 |
| Rescoring the baseline 1000-best with a single reversed LSTM | 35.85 |
| Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs | **36.5** |
| Oracle Rescoring of the Baseline 1000-best lists | ~45 |

Table 2: Methods that use neural networks together with an SMT system on the WMT'14 English to French test set (ntst14).

# Learned Phrase Representations

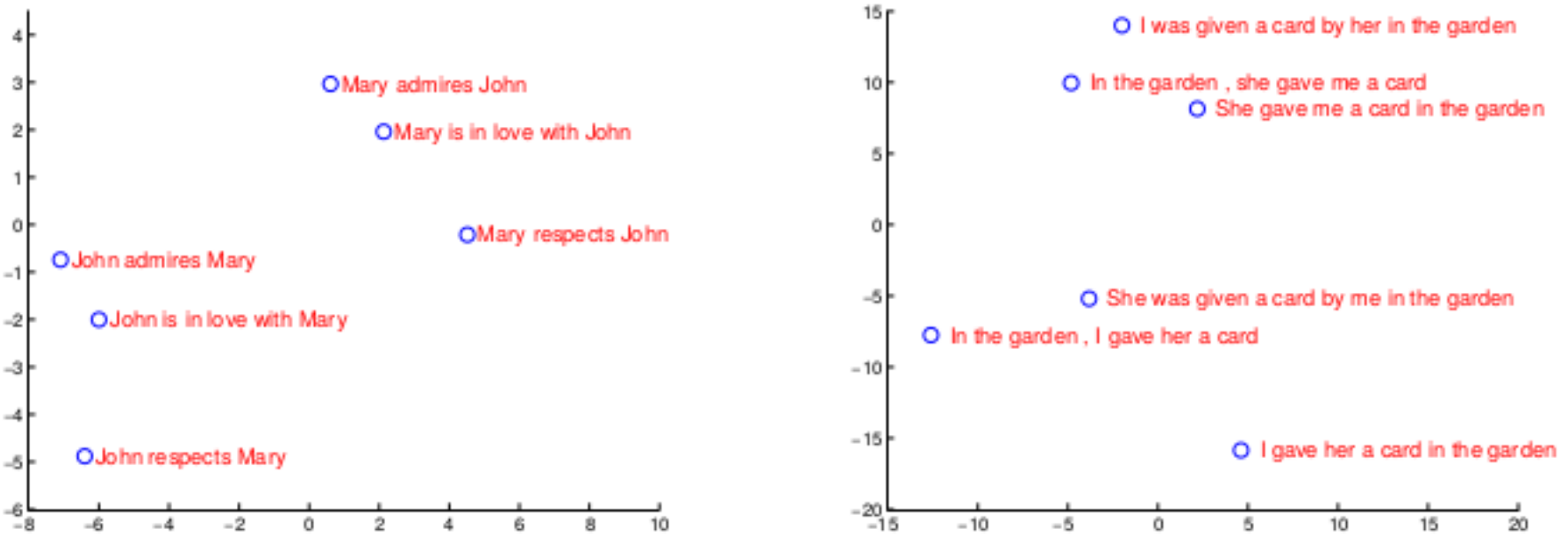



Figure 2: The figure shows a 2-dimensional PCA projection of the LSTM hidden states that are obtained after processing the phrases in the figures. The phrases are clustered by meaning, which in these examples is primarily a function of word order, which would be difficult to capture with a bag-of-words model. Notice that both clusters have similar internal structure.

# Effect of Length

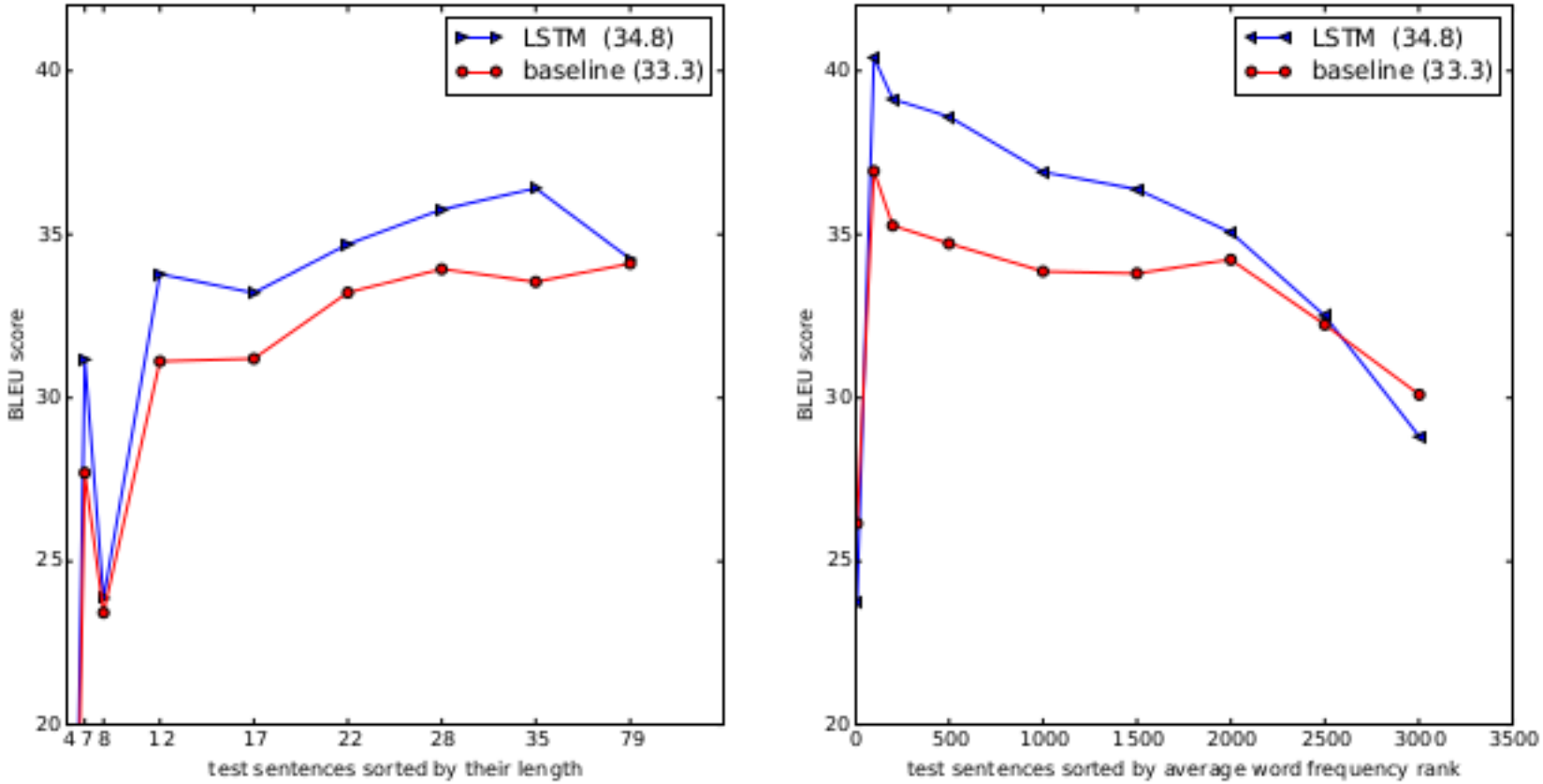


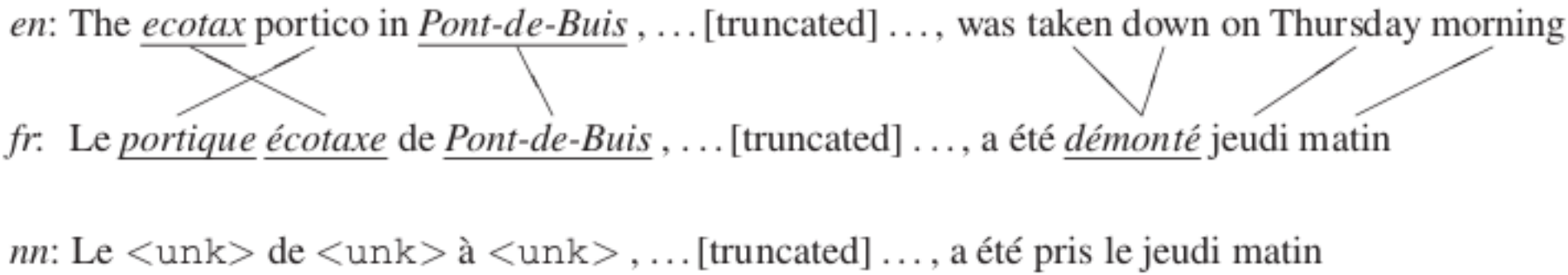

Figure 3: The left plot shows the performance of our system as a function of sentence length, where the x-axis corresponds to the test sentences sorted by their length and is marked by the actual sequence lengths. There is no degradation on sentences with less than 35 words, there is only a minor degradation on the longest sentences. The right plot shows the LSTM's performance on sentences with progressively more rare words, where the x-axis corresponds to the test sentences sorted by their "average word frequency rank".

# Examples/Problems with UNK

| Type | Sentence |
|---|---|
| **Our model** | Ulrich UNK , membre du conseil d' administration du constructeur automobile Audi , affirme qu' il s' agit d' une pratique courante depuis des années pour que les téléphones portables puissent être collectés avant les réunions du conseil d' administration afin qu' ils ne soient pas utilisés comme appareils d' écoute à distance . |
| **Truth** | Ulrich Hackenberg , membre du conseil d' administration du constructeur automobile Audi , déclare que la collecte des téléphones portables avant les réunions du conseil , afin qu' ils ne puissent pas être utilisés comme appareils d' écoute à distance , est une pratique courante depuis des années . |
| **Our model** | " Les téléphones cellulaires , qui sont vraiment une question , non seulement parce qu' ils pourraient potentiellement causer des interférences avec les appareils de navigation , mais nous savons , selon la FCC , qu' ils pourraient interférer avec les tours de téléphone cellulaire lorsqu' ils sont dans l' air " , dit UNK . |
| **Truth** | " Les téléphones portables sont véritablement un problème , non seulement parce qu' ils pourraient éventuellement créer des interférences avec les instruments de navigation , mais parce que nous savons , d' après la FCC , qu' ils pourraient perturber les antennes-relais de téléphonie mobile s' ils sont utilisés à bord " , a déclaré Rosenker . |
| **Our model** | Avec la crémation , il y a un " sentiment de violence contre le corps d' un être cher " , qui sera " réduit à une pile de cendres " en très peu de temps au lieu d' un processus de décomposition " qui accompagnera les étapes du deuil " . |
| **Truth** | Il y a , avec la crémation , " une violence faite au corps aimé " , qui va être " réduit à un tas de cendres " en très peu de temps , et non après un processus de décomposition , qui " accompagnerait les phases du deuil " . |

Table 3: A few examples of long translations produced by the LSTM alongside the ground truth translations. The reader can verify that the translations are sensible using Google translate.

# Addressing the Rare Word Problem in Neural Machine Translation [Luong+ 14]




- **Copyable model:** label unk words

  **en:** The $unk_1$ portico in $unk_2$ …

  **fr:** Le $unk_n$ $unk_1$ de $unk_2$ ...

- **Positional all model:** label word positions (i=j-d, $e_i$ $f_j$)

  **en:** The <unk> portico in <unk> ...
  **fr:** Le $pos_0$ <unk> $pos_{-1}$ <unk> $pos_1$ de $pos_n$ <unk> $pos_{-1}$ ...

- **Positional unk:** label unk positions

  **en:** The <unk> portico in <unk> ...
  **fr:** Le $unkpos_1$ $unkpos_{-1}$ de $unkpos_1$ ...

# Results with PosUnk

| System | BLEU |
|---|---|
| State of the art [7] | **37.0** |
| *Standard MT + neural components* | |
| LIUM [19] – neural language model | 33.3 |
| Cho et al. [5] – phrase table neural features | 34.5 |
| Sutskever et al. [22] – ensemble 5 LSTMs, reranking | 36.5 |
| *Purely neural machine translation systems* | |
| Bahdanau et al. [2] – bi-directional gated single RNN | 28.5 |
| Sutskever et al. [22] – single LSTM | 30.6 |
| Sutskever et al. [22] – ensemble of 5 LSTMs | 34.8 |
| *Our purely neural machine translation systems* | |
| Single depth-4 LSTM | 29.5 |
| Single depth-4 LSTM + PosUnk | 31.8 (+2.3) |
| Single depth-6 LSTM | 30.4 |
| Single depth-6 LSTM + PosUnk | 32.7 (+2.3) |
| Ensemble of 8 LSTMs | 34.1 |
| Ensemble of 8 LSTMs + PosUnk | **36.9 (+2.8)** |