Introduction to Neural Machine Translation

Marine Carpuat

Computer Science

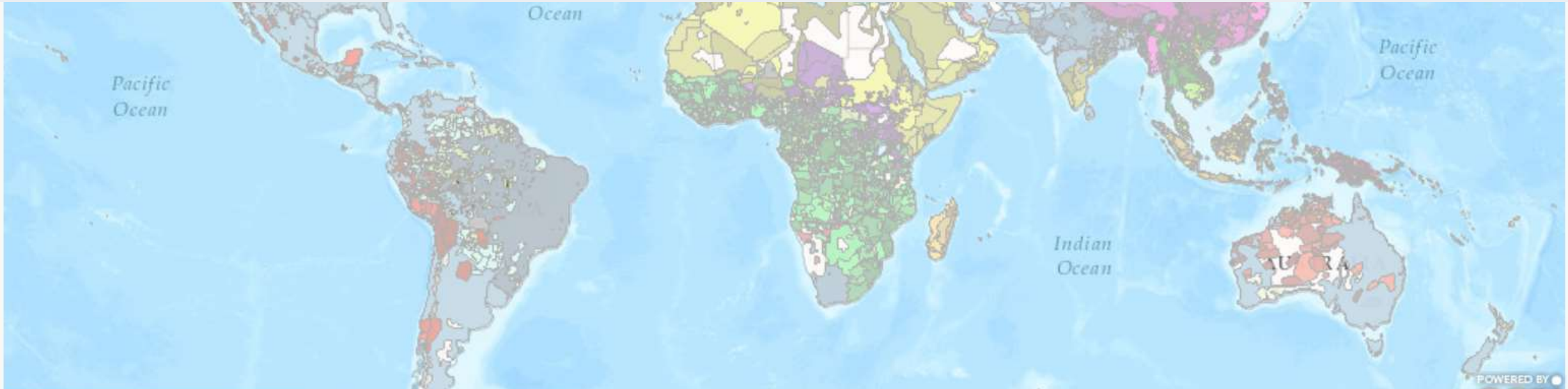
University of Maryland

Maryland Language Science Center

LANGSCAPE DISCOVER ABOUT PARTNERS CONTRIBUTE BLOG

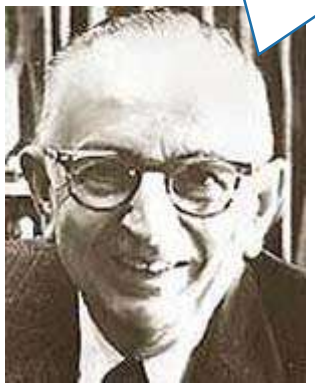
Search for a language, dialect name or major city...

6,800 living languages
600 with written tradition
100 spoken by 95% of population



1947

When I look at an article in Russian, I say to myself: This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.



Warren Weaver

Rule based systems

- Approach
  - Build dictionaries
  - Write transformation rules
  - Refine, refine, refine
- Meteo system for weather forecasts (1976)

```
"have" :=

if
    subject(animate)
    and object(owned-by-subject)
then
    translate to "kade...  aahe"
if
    subject(animate)
    and object(kinship-with-subject)
then
    translate to "laa...  aahe"
if
    subject(inanimate)
then
          translate to "madhye...
aahe"
```

1988

A Statistical Approach to Machine Translation

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin

IBM
Thomas J. Watson Research Center
Yorktown Heights, NY

In this paper, we present a statistical approach to machine translation. We describe the application of our approach to translation from French to English and give preliminary results.

The COLING Paper Review

The validity of statistical (information theoretic) approach to MT has indeed been recognized, as the authors mention, by Weaver as early as 1949. And was universally recognized as mistaken by 1950. (cf. Hutchins, MT: Past, Present, Future, Ellis Horwood, 1986, pp. 30ff. and references therein) The crude force of computers is not science. The paper is simply beyond the scope of COLING.

More about the IBM story: 20 years of bitext workshop

Statistical MT introduced the idea of learning from data

- Use large amounts of existing translations
  - Called parallel corpora (aka bitext)

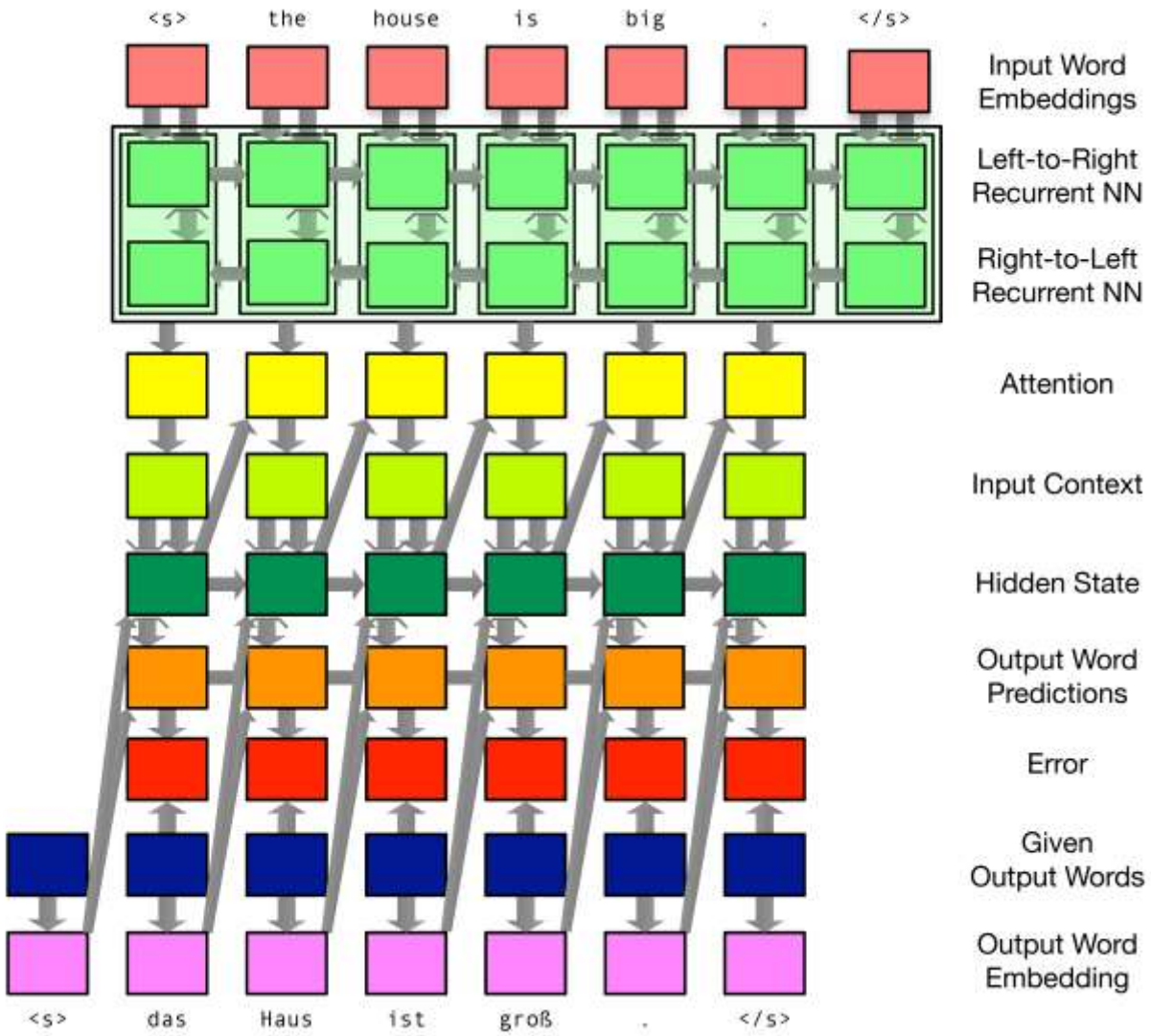
- Use word co-occurrence counts

Sicherheit → security 14,516
Sicherheit → safety 10,015
Sicherheit → certainty 334

Statistical Machine Translation

- 1990s: increased research
- Mid 2000s: phrase-based MT
  - (Moses, Google Translate)
- Around 2010: commercial viability
- Since mid 2010s: neural network models

Neural Machine Translation



Introduced by
Bahdanau, Chio & Bengio 2015

How Good is Machine Translation Today?

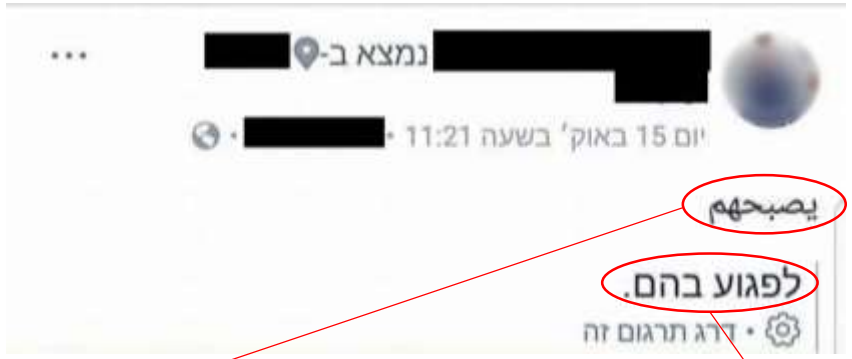
上周，古装剧《**美人私房菜**》临时停播，**意外引**发了关于国产剧收视率造假的热烈讨论。

Last week, the vintage drama "Beauty private dishes" was temporarily suspended, accidentally sparking a heated discussion about the fake ratings of domestic dramas.

民权团体针对密苏里州发出旅**行警告**

Civil rights groups issue travel warnings against Missouri

http://matrix.statmt.org



يصبحهم = Good morning

לפגוע בהם = Hurt them

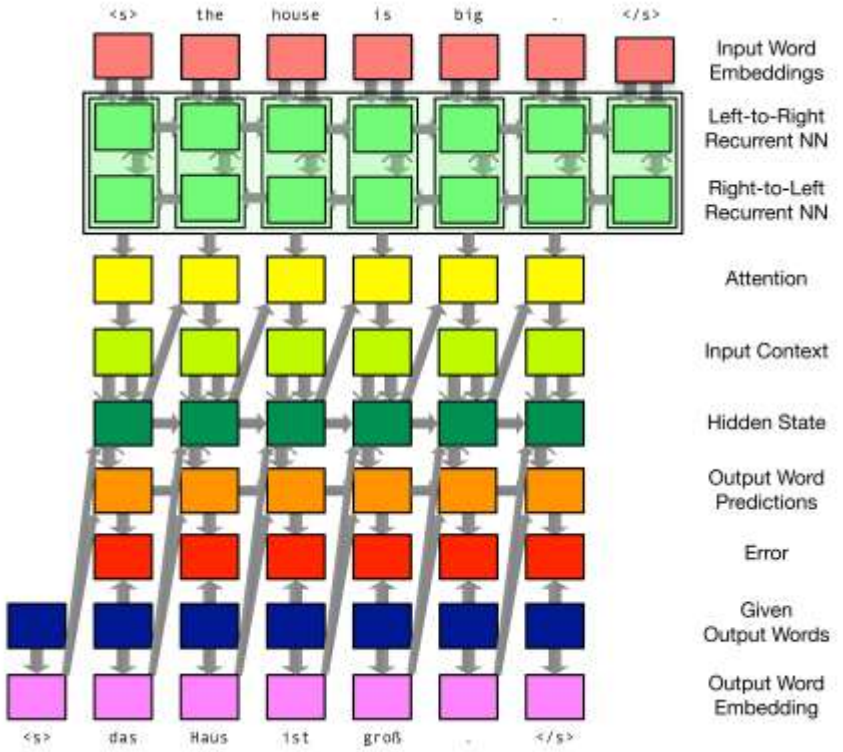
https://www.haaretz.com/israel-news/palestinian-arrested-over-mistranslated-good-morning-facebook-post-1.5459427

hem

https://www.haaretz.com/israel-news/palestinian-arrested-over-mistranslated-good-morning-facebook-post-1.5459427

Roadmap

- Evaluating machine translation
- Introduction to neural networks
- Modeling sequences of words with neural language models
- Translating with encoder-decoder models
- Attention mechanism



Machine Translation Evaluation

With examples and figures by Philipp Koehn (JHU)

Machine Translation Evaluation

- Goal
  - given a source text and its translation
  - measure how good the translation is

- Evaluation is crucial
  - to compare different systems
  - to measure impact of changes in a single system
  - to guide machine learning

How good is a translation?

这个 机场 的 安全 工作 由 以色列 方面 负责 .

Israeli officials are responsible for airport security.

How good is a translation?

这个 机场 的 安全 工作 由 以色列 方面 负责 .

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 o
Israeli side was in charge of the security of this airport.
Israel is responsible for the airport's security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport's security is the responsibility of the Israeli security officials.

No single correct translation, many possible answers!

Machine Translation Evaluation

- Goal
  - given a source text and its translation
  - measure how good the translation is

- 2 types of evaluations
  - Subjective judgments by human evaluators
  - Automatic evaluation metrics

Adequacy and Fluency

- Metrics
  - **Adequacy:** does the output convey the meaning of the input sentence? Is part of the message lost, added, or distorted?
  - **Fluency:** is the output fluent? Involves both grammatical correctness and idiomatic word choices.

- Human judgment
  - Given: machine translation output
  - Given: input and/or reference translation
  - Task: assess quality of MT output

Fluency and Adequacy: Scales

Adequacy	
5	all meaning
4	most meaning
3	much meaning
2	little meaning
1	none

Fluency	
5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible

Let's try:
rate fluency & adequacy on 1-5 scale

- Source:
  N'y aurait-il pas comme une vague hypocrisie de votre part ?
- Reference:
  Is there not an element of hypocrisy on your part?
- System1:
  Would it not as a wave of hypocrisy on yo

Human evaluators often disagree!

Automatic Evaluation Metrics

- Goal: computer program that scores quality of translations
- Advantages: low cost, optimizable, consistent
- Approach
  - Given: MT output
  - Given: human reference translation
  - Task: compute similarity between them

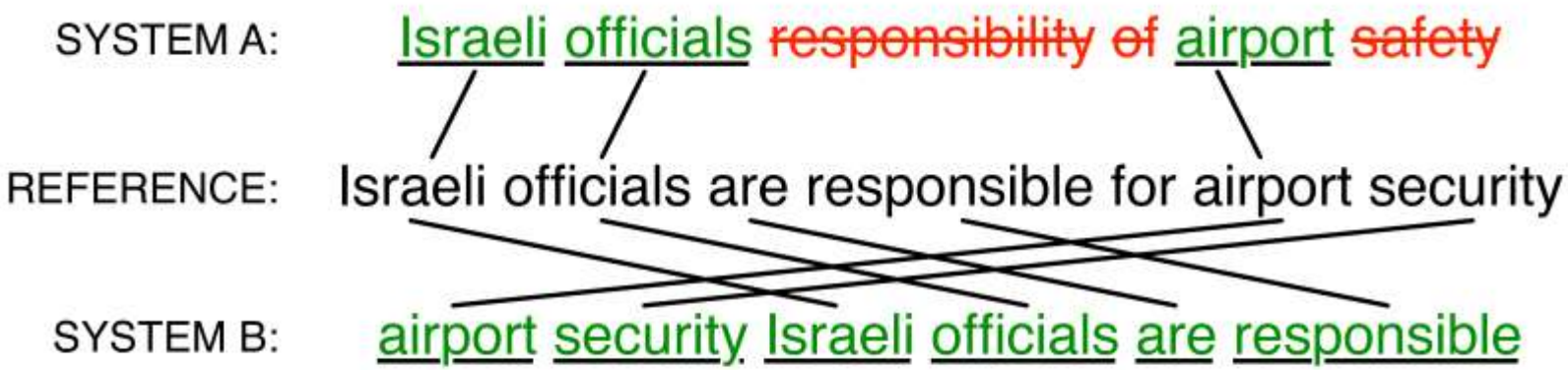
Precision and Recall of Words

Precision $$\frac{correct}{output\text{-}length} = \frac{3}{6} = 50\%$$

Recall $$\frac{correct}{reference\text{-}length} = \frac{3}{7} = 43\%$$

F-measure $$\frac{precision \times recall}{(precision + recall)/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%$$

Precision and Recall of Words



Metric	System A	System B
precision	50%	100%
recall	43%	100%
f-measure	46%	100%

BLEU
Bilingual Evaluation Understudy

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)

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \left(\prod_{i=1}^{4} \textit{precision}_i\right)^{\frac{1}{4}}$$

Typically computed over the entire corpus, not single sentences

BLEU examples

SYSTEM A: Israeli officials (2-GRAM MATCH) responsibility of airport (1-GRAM MATCH) safety

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security (2-GRAM MATCH) Israeli officials are responsible (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

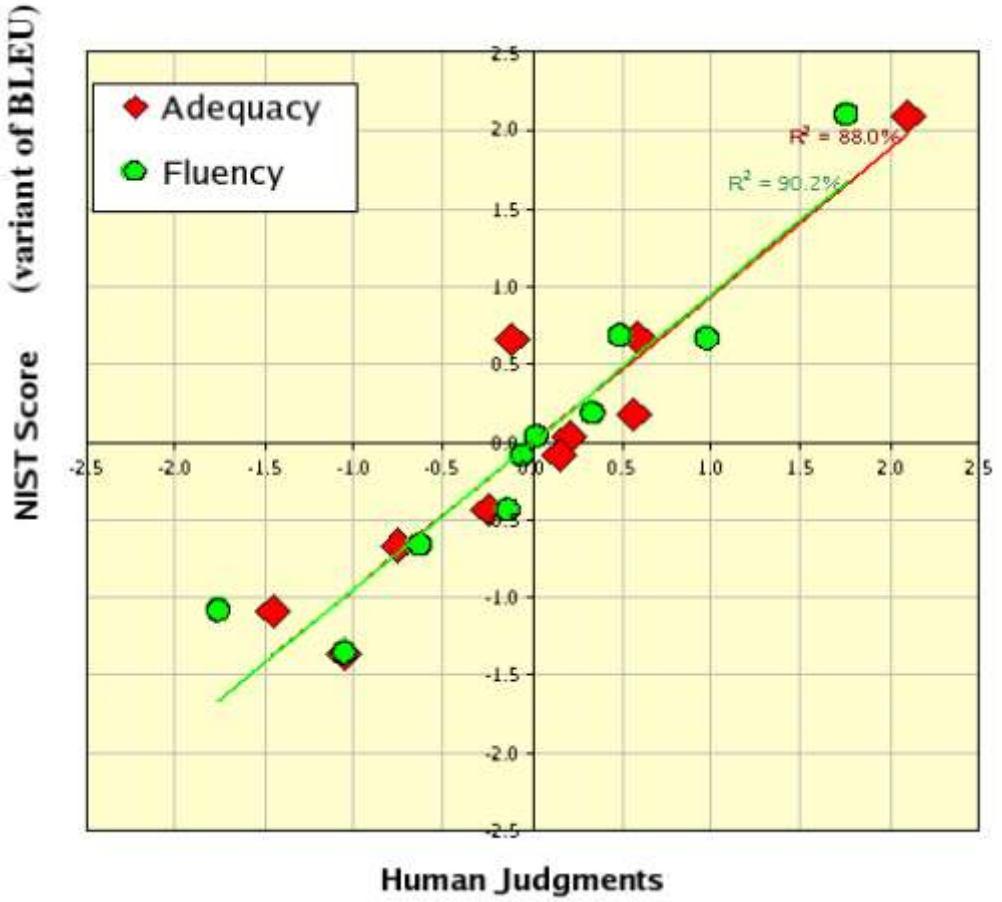
SYSTEM: Israeli officials (2-GRAM MATCH) responsibility of (2-GRAM MATCH) airport (1-GRAM) safety

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

Drawbacks of Automatic Metrics

- Ground truth is defined by reference translation(s)
- All words are treated as equally relevant
- Operate on local level
- Absolute value of scores is not informative
- Human translators score low on BLEU

Yet automatic metrics such as BLEU correlate with human judgement*



Automatic Evaluation Metrics

- Provide cheap but imperfect estimates of translation quality
- Essential for system development

- Still many open research questions
  - Do we need new metrics as quality improves?
  - How can we evaluate document translation?
  - Which errors matter most to users?

Introduction to Neural Networks

With examples and figures by Graham Neubig (CMU) & Philipp Koehn (JHU)

Let's Consider a Binary Classification Problem

Given an introductory sentence in Wikipedia

predict whether the article is about a person

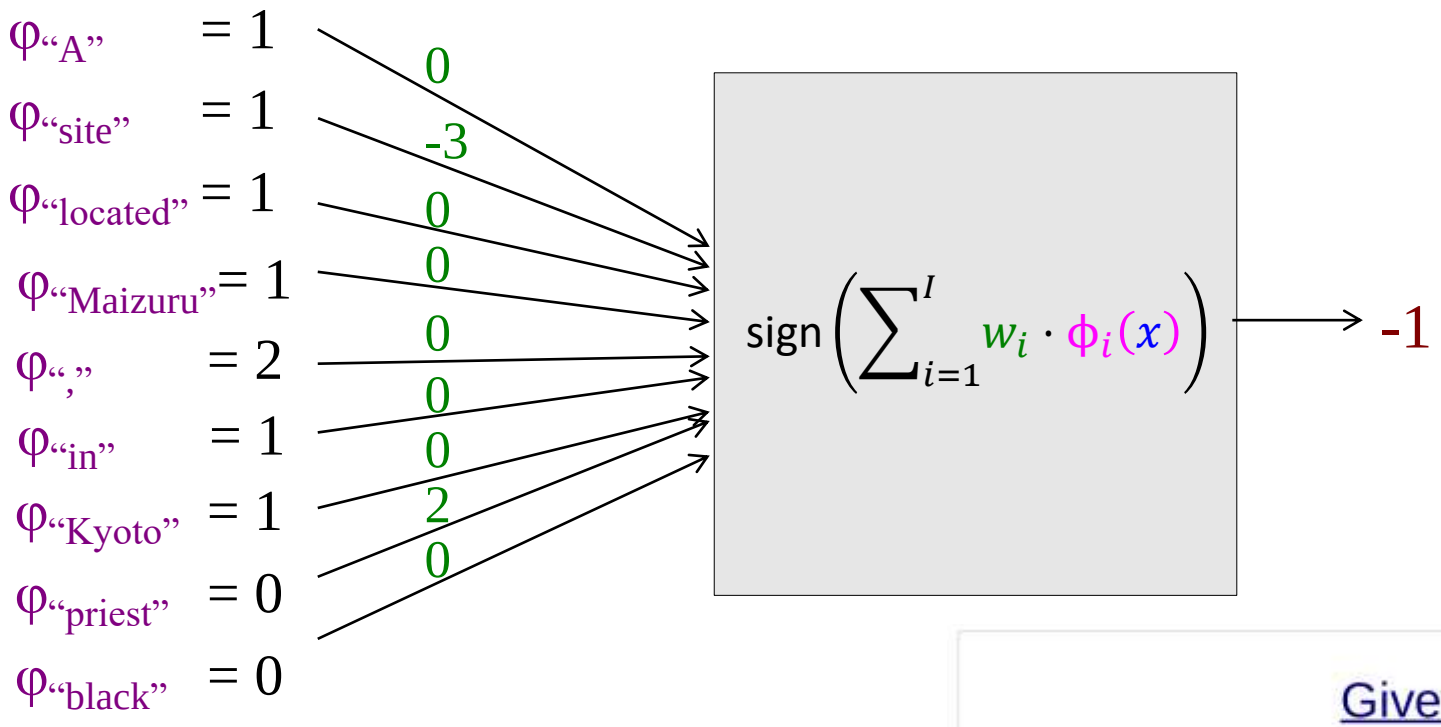
Example & figures by Graham Neubig

Binary Classification with the Perceptron

$$
\begin{aligned}
y &= \operatorname{sign}\left(\boldsymbol{w} \cdot \boldsymbol{\varphi}(x)\right) \\
&= \operatorname{sign}\left(\sum_{i=1}^{I} w_i \cdot \varphi_i(x)\right)
\end{aligned}
$$

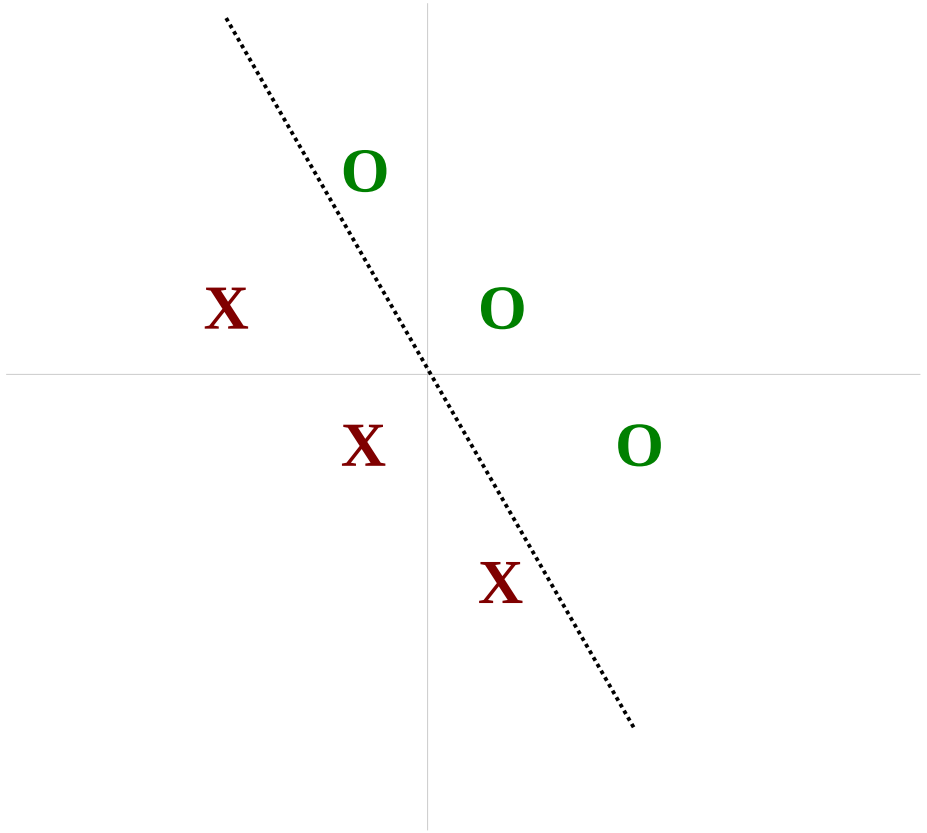
- x: the input
- **φ(x)**: vector of feature functions $\{\varphi_1(x), \varphi_2(x), \ldots, \varphi_I(x)\}$
- **w**: the weight vector $\{w_1, w_2, \ldots, w_I\}$
- y: the prediction, +1 if “yes”, -1 if “no”
  - (sign(v) is +1 if v >= 0, -1 otherwise)

Making Predictions with the Perceptron



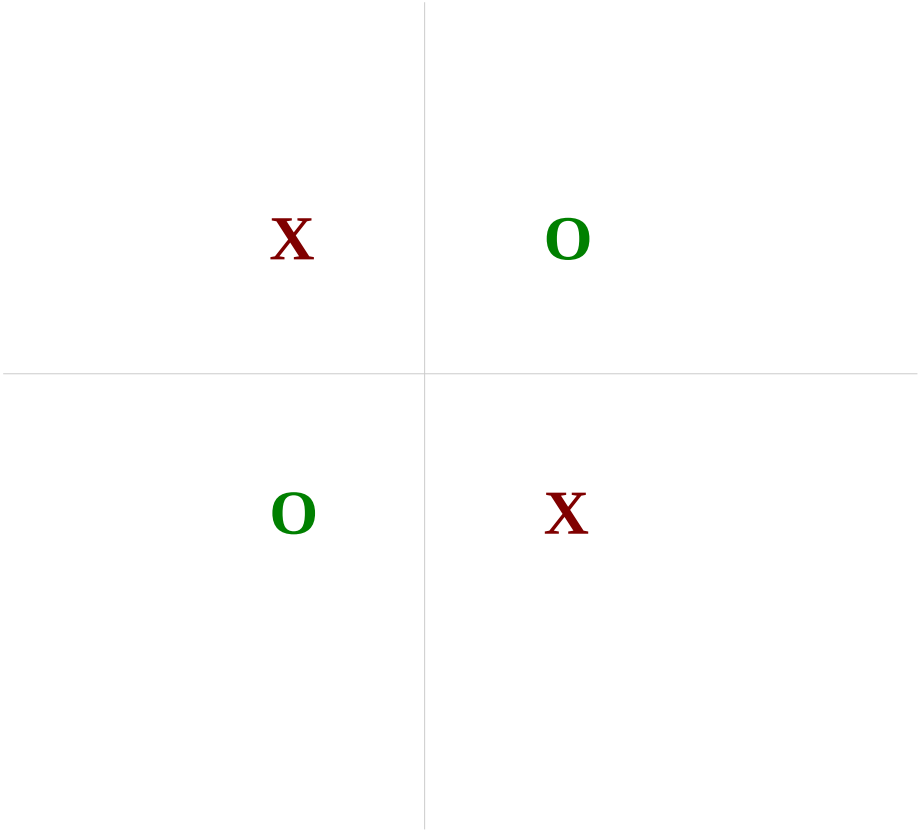
Given	Predict
Gonso was a Sanron sect priest (754-827) in the late Nara and early Heian periods.	→
Shichikuzan Chigogataki Fudomyoo is a historical site located at Magura, Maizuru City, Kyoto Prefecture.	→

The Perceptron: Geometric interpretation

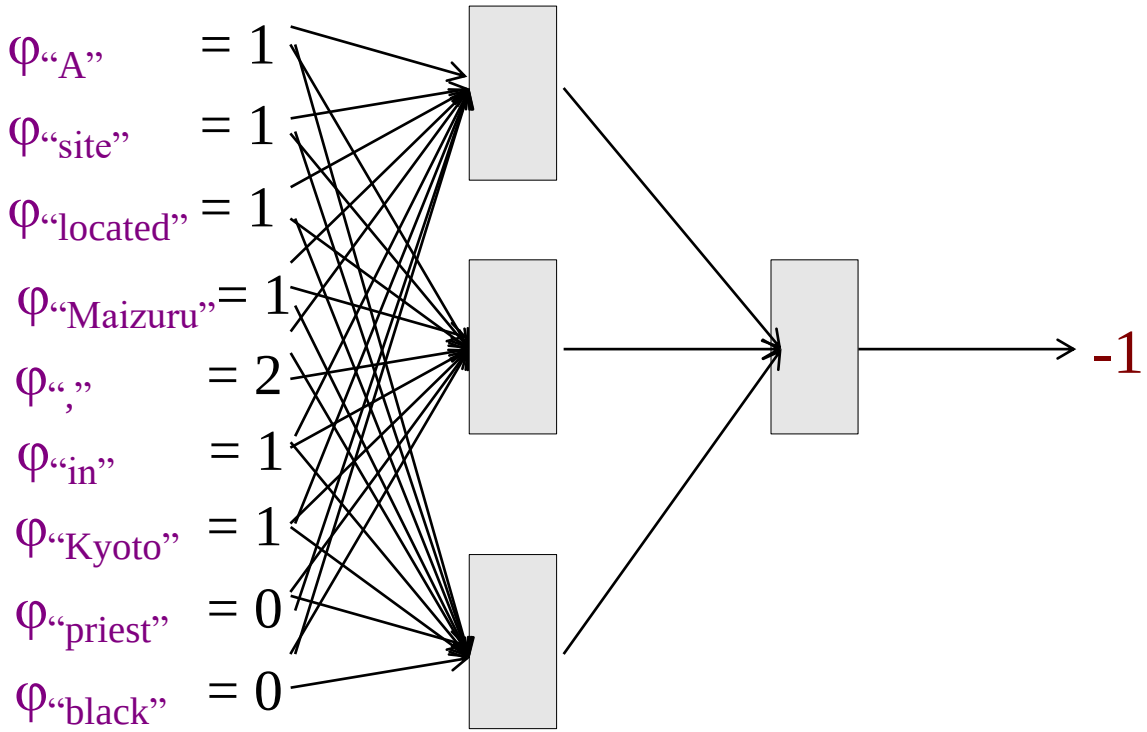


Limitation of perceptron

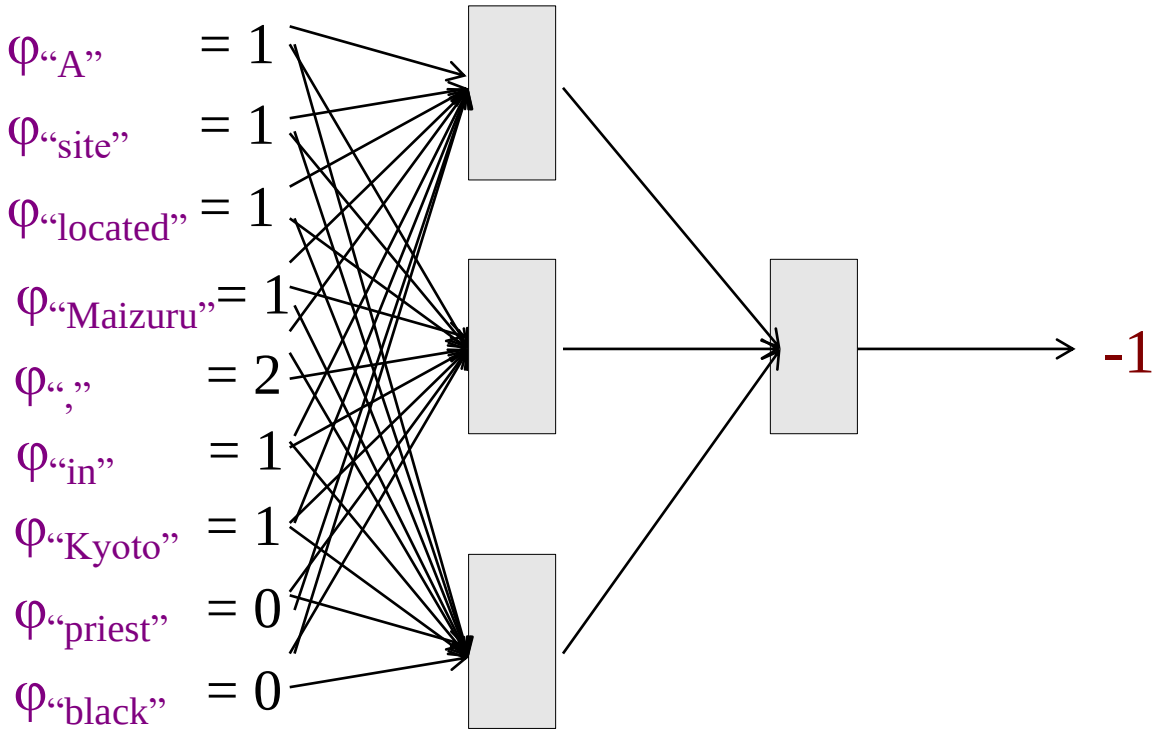
- can only find **linear separations** between positive and negative examples



Binary Classification with a Multi-layer Perceptron



Multi-layer Perceptrons are a kind of “Neural Network” (NN)

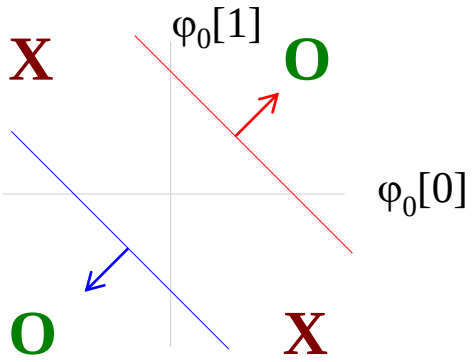


- Input (aka features)
- Output
- Nodes
- Layers
- Hidden layers
- Activation function (non-linear)

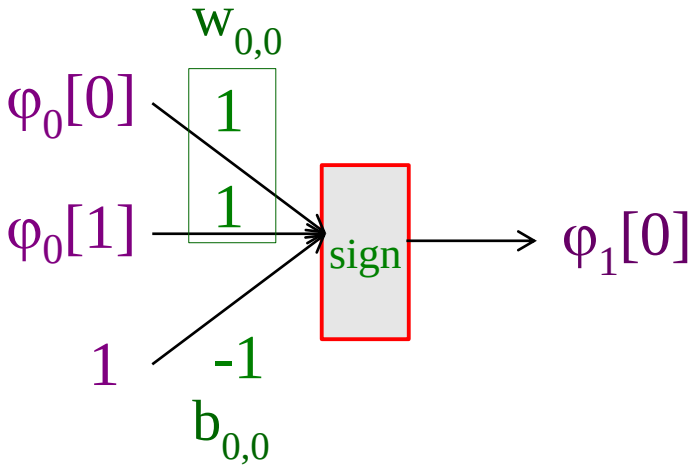
Example: binary classification with a NN

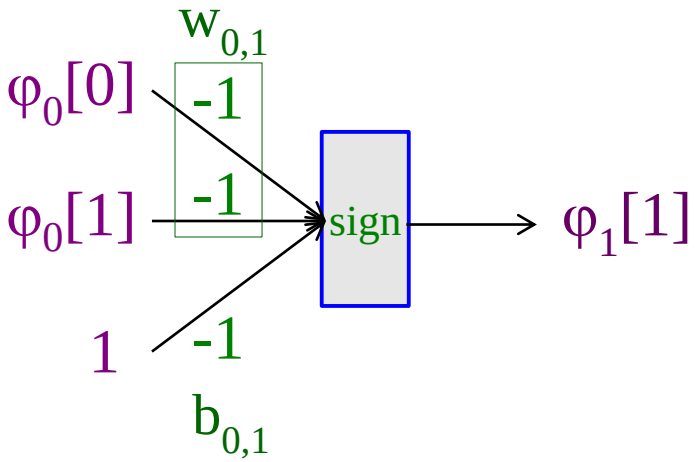
- Create two classifiers

$\varphi_0(x_1) = \{-1, 1\}$ $\varphi_0(x_2) = \{1, 1\}$



$\varphi_0(x_3) = \{-1, -1\}$ $\varphi_0(x_4) = \{1, -1\}$

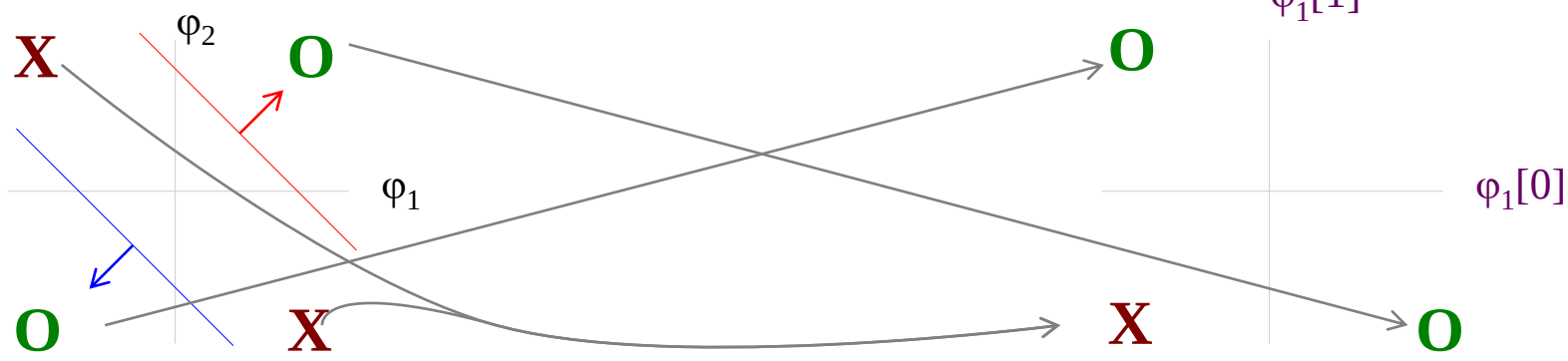




Example: binary classification with a NN

- These classifiers map to a new space

$\varphi_0(x_1) = \{-1, 1\}$ $\varphi_0(x_2) = \{1, 1\}$



$\varphi_0(x_3) = \{-1, -1\}$ $\varphi_0(x_4) = \{1, -1\}$

$\varphi_1(x_3) = \{-1, 1\}$

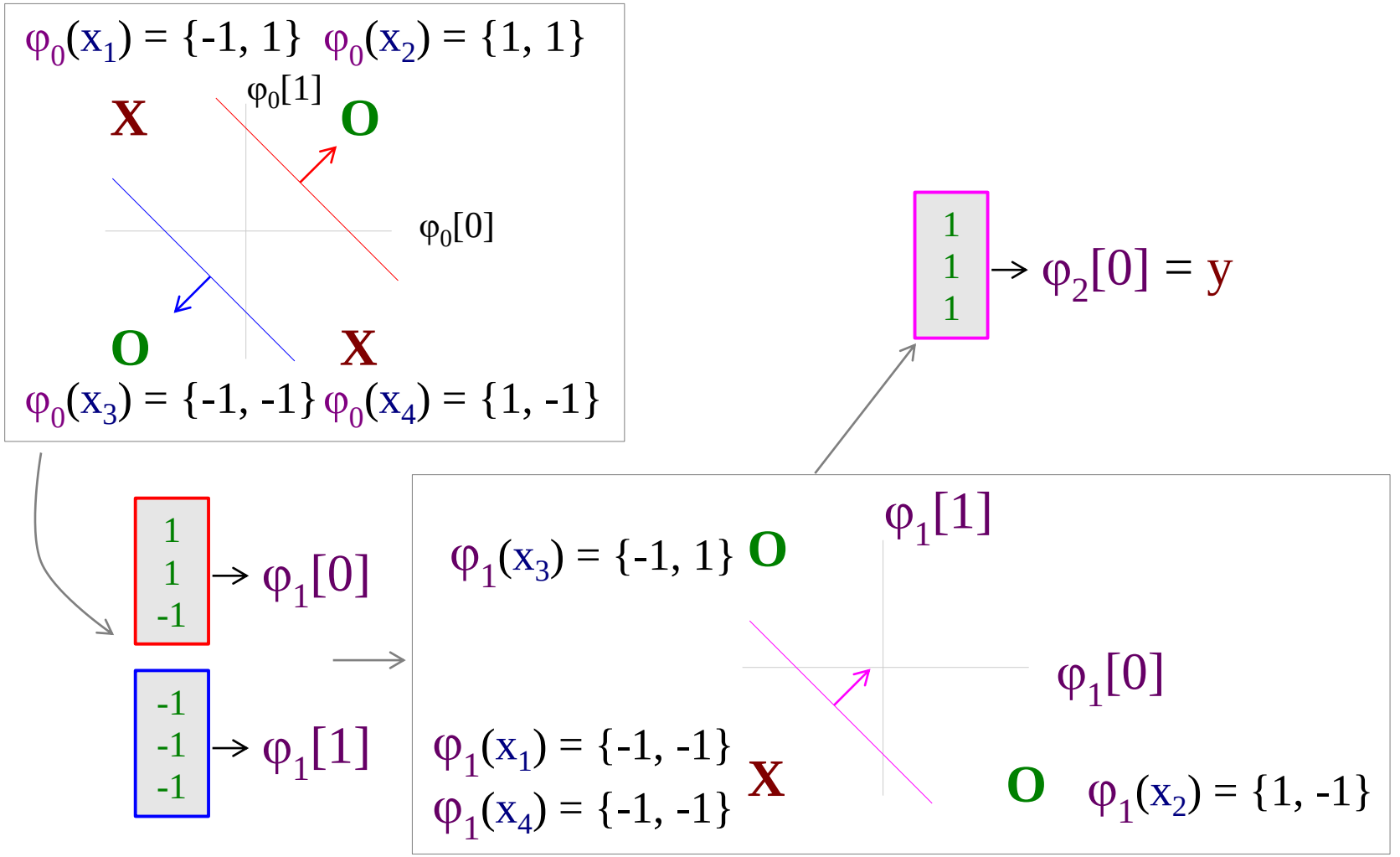
$\varphi_1(x_1) = \{-1, -1\}$ $\varphi_1(x_2) = \{1, -1\}$

$\varphi_1(x_4) = \{-1, -1\}$

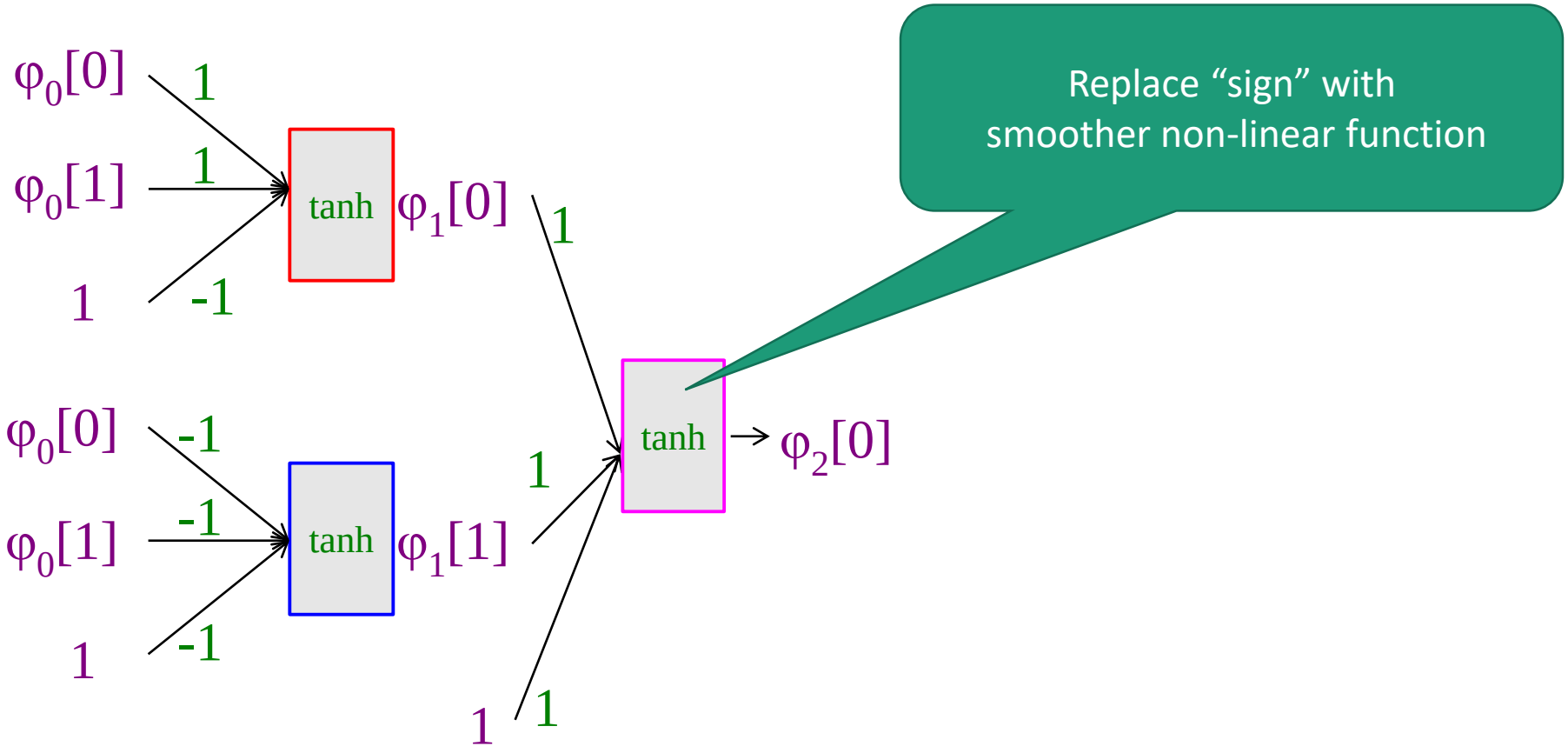
1
1
-1
→ $\varphi_1[0]$

-1
-1
-1
→ $\varphi_1[1]$

Example: binary classification with a NN

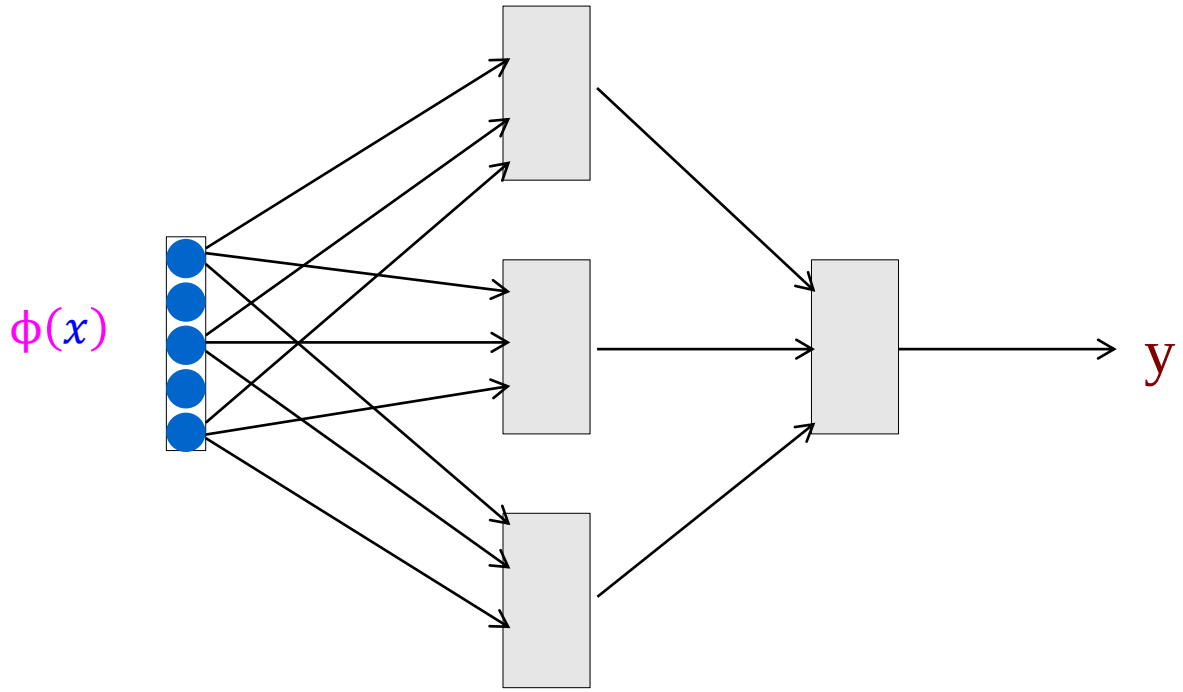


Example: the Final Net



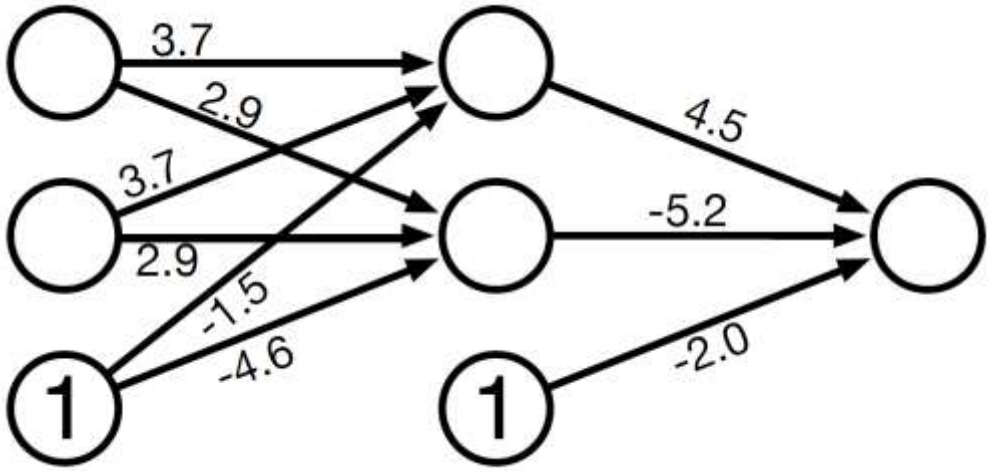
Feed Forward Neural Networks

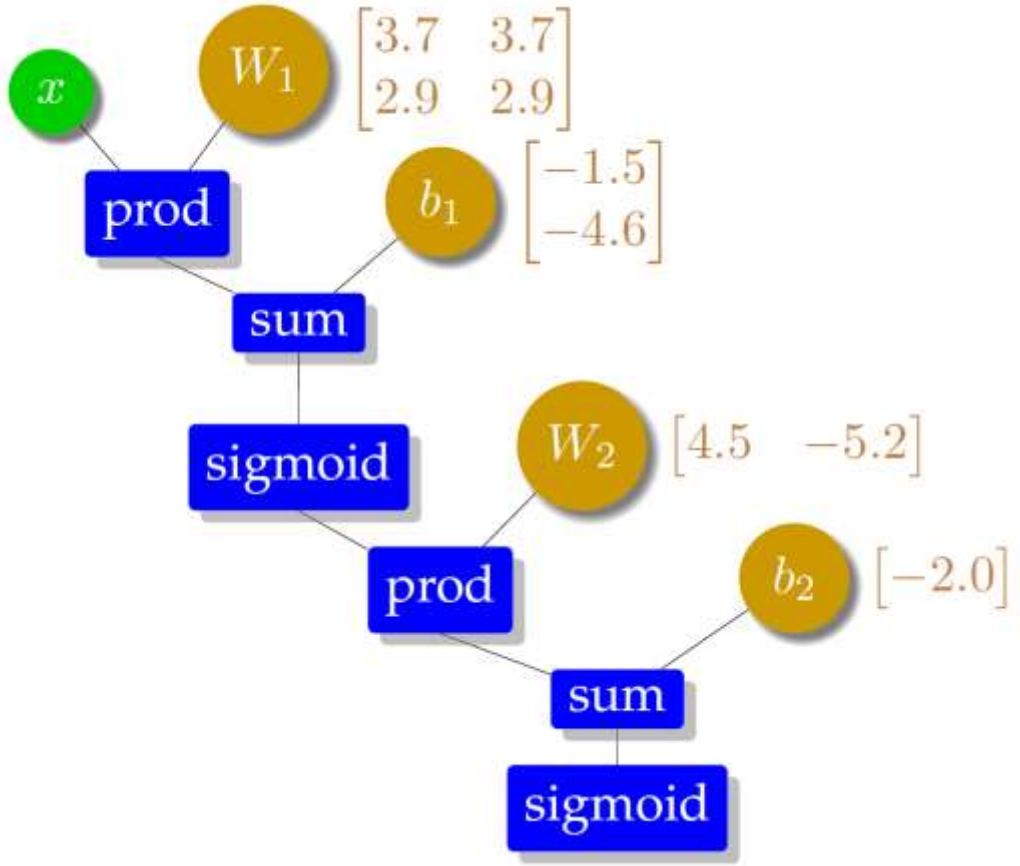
All connections point **forward**



It is a directed acyclic graph (DAG)

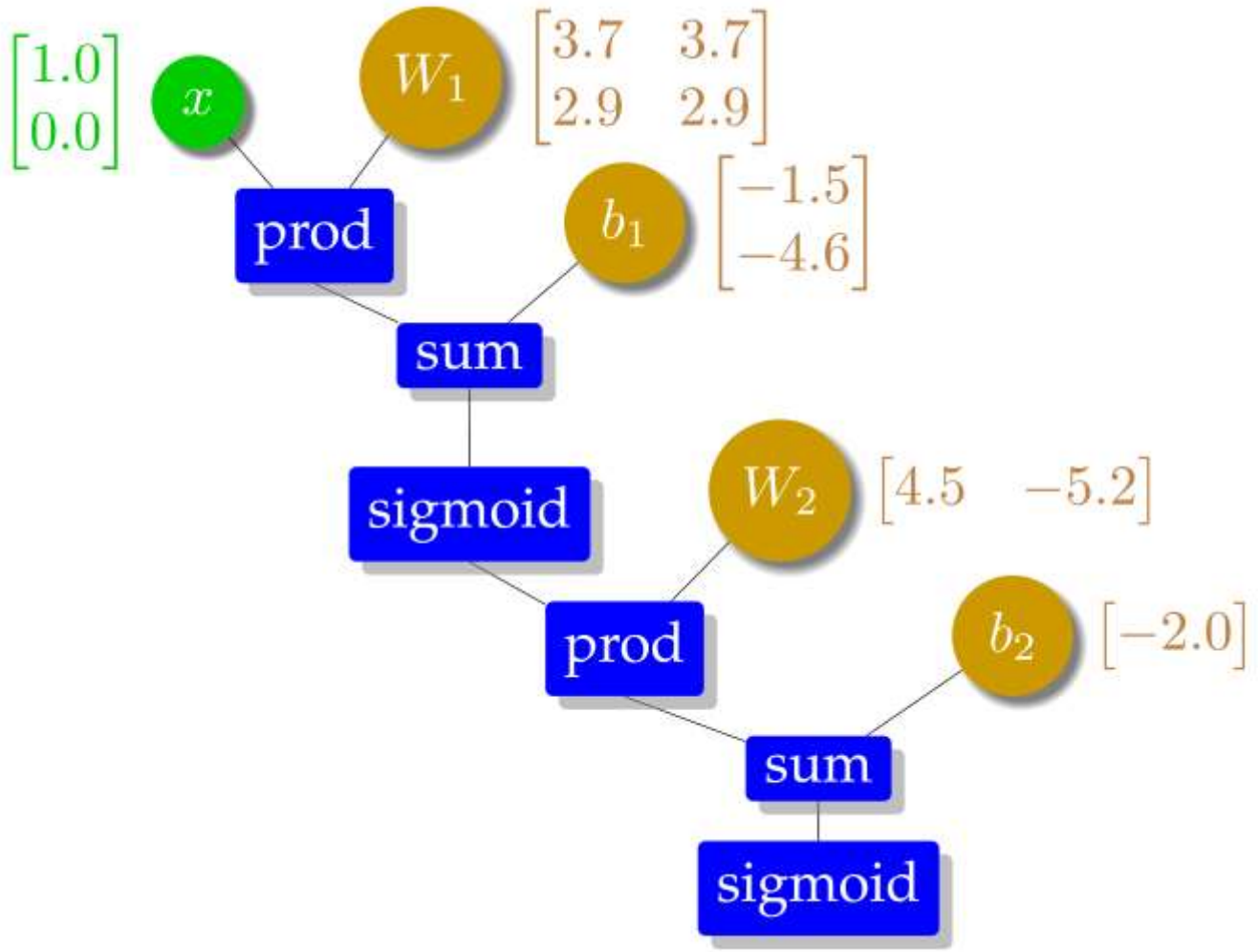
Neural Networks as Computation Graphs



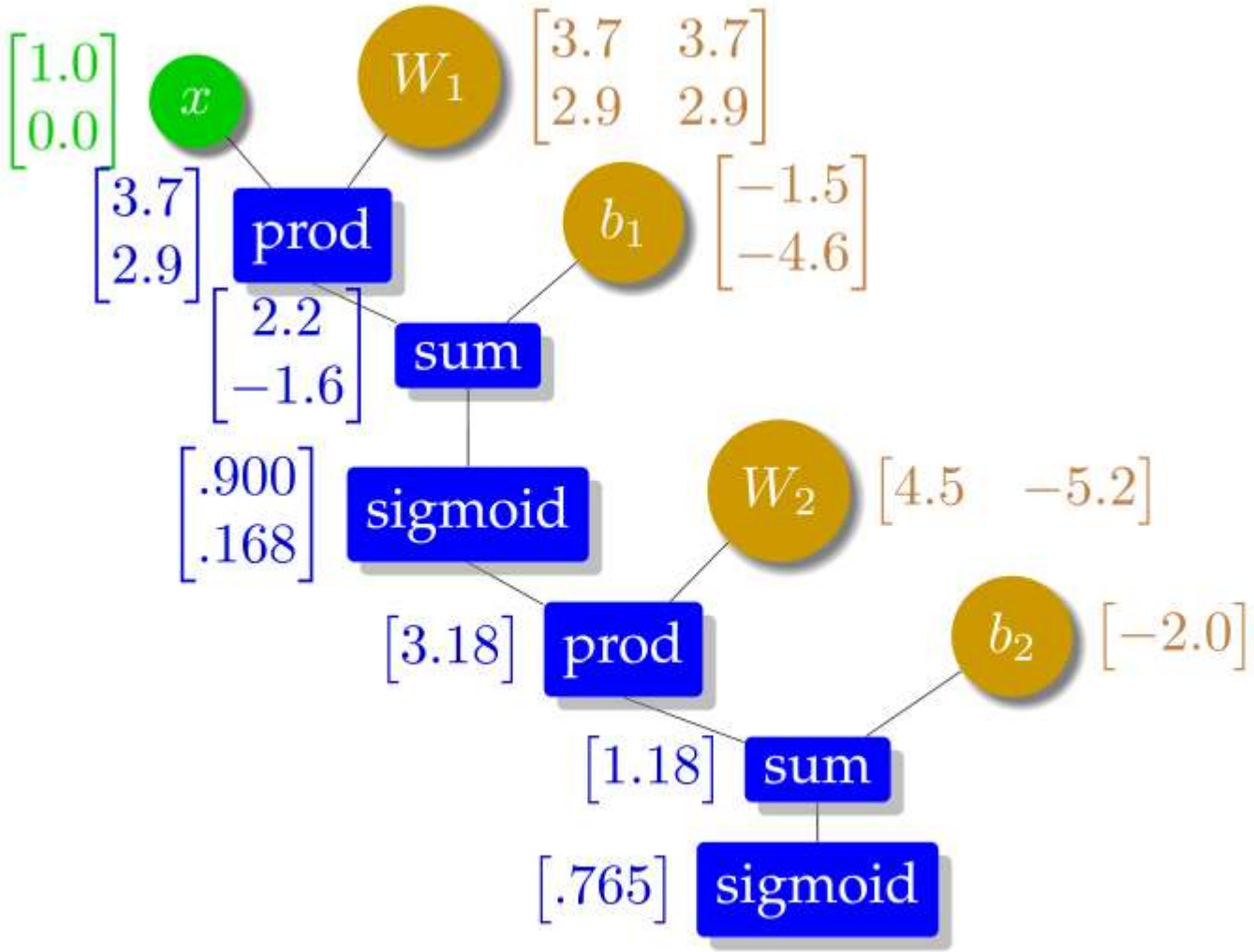


Example & figures by Philipp Koehn

Computation Graphs Make Prediction Easy: Forward Propagation



Computation Graphs Make Prediction Easy: Forward Propagation



How do we estimate the parameters (aka "train") a neural net?

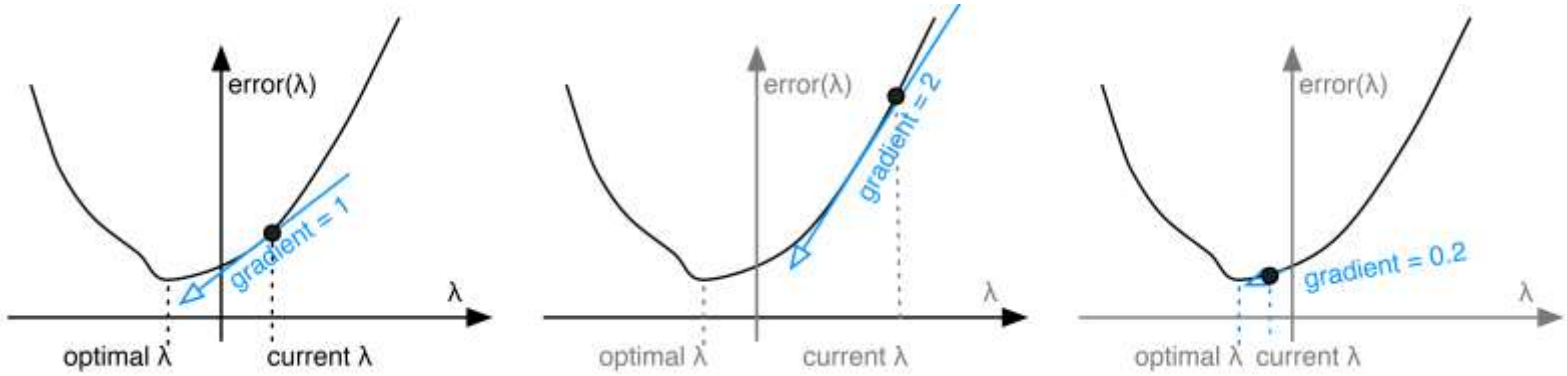
For training, we need:

- Data: (a large number of) examples paired with their correct class (x,y)
- error function: quantify how bad our prediction y is compared to the truth t
  - Let's use squared error:

$$\text{error} = \frac{1}{2}(t - y)^2$$

Stochastic Gradient Descent

- We view the error as a function of the trainable parameters, on a given dataset
- We want to find parameters that minimize the error



Stochastic Gradient Descent

Start with some initial parameter values

Go through the training data one example at a time

$w = 0$

for *I* iterations

  for each labeled pair *x*, *y* in the data

$$w = w - \mu \frac{d\text{error}(w, x, y)}{dw}$$

Take a step down the gradient

Computation Graph: A Powerful Abstraction

- To build a system, we only need to:
  - Define network structure
  - Define loss
  - Provide data
  - (and set a few more hyperparameters to control training)
- Given network structure
  - Prediction is done by forward pass through graph (forward propagation)
  - Training is done by backward pass through graph (back propagation)
  - Based on simple matrix vector operations
- Forms the basis of neural network libraries
  - Tensorflow, Pytorch, mxnet, etc.

Roadmap

- Evaluating machine translation
- Introduction to neural networks
- Modeling sequences of words with neural language models
- Translating with encoder-decoder models
- Attention mechanism