A Unified Architecture
for
Natural Language Processing

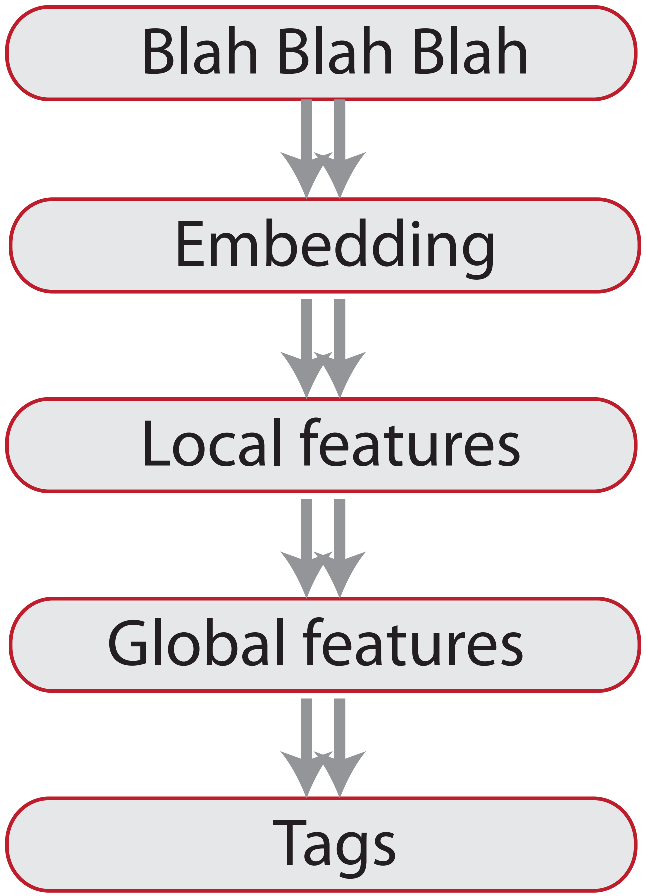
(Deep Neural Networks with Multitask Learning)

Ronan Collobert & **Jason Weston**

ronan@collobert.com jaseweston@gmail.com

NEC Laboratories America

The Big Picture (1/2)

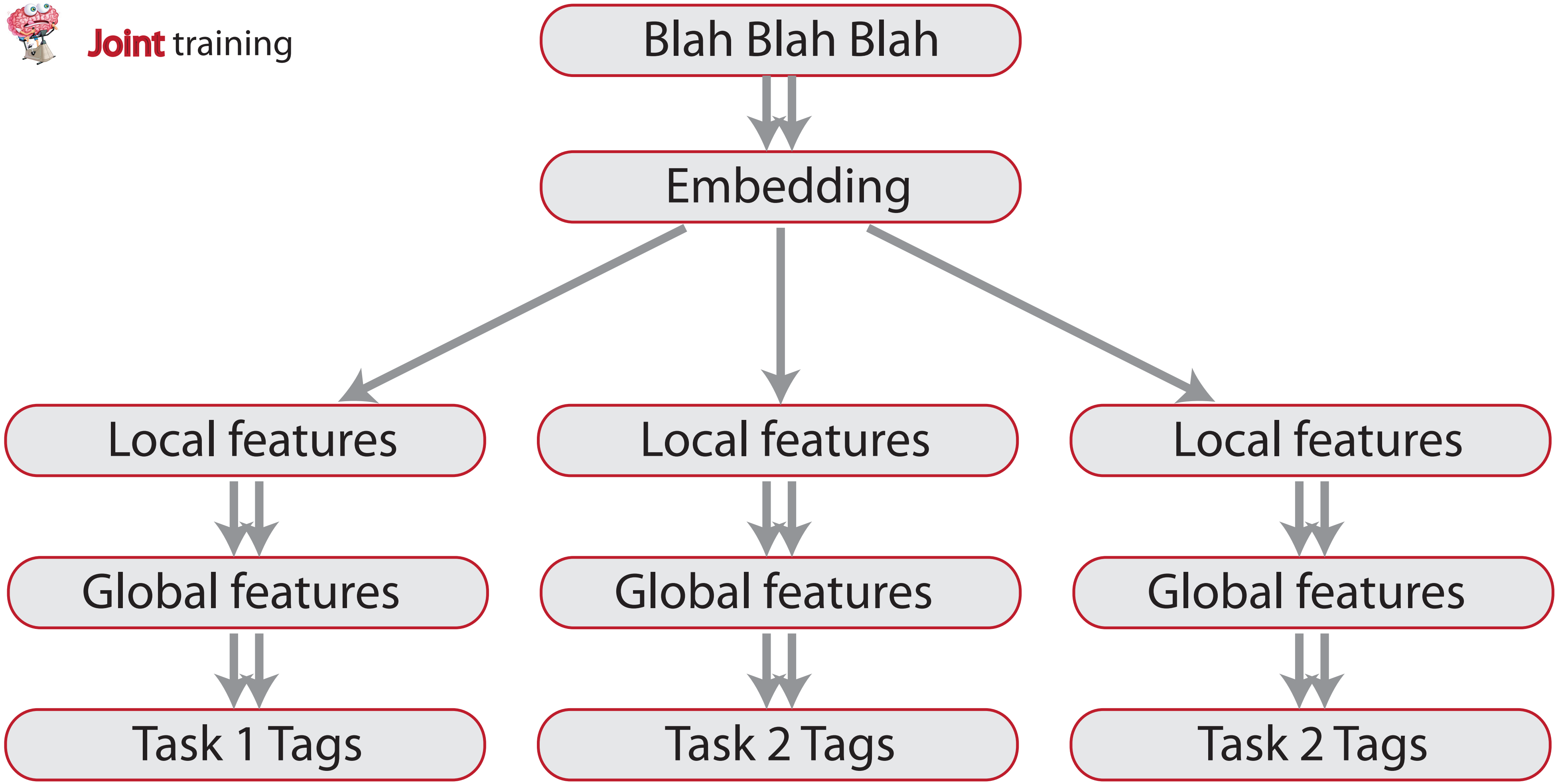


Deep architecture

Unification of NLP tasks

The Big Picture (2/2)

Joint training



NLP Tasks

- Part-Of-Speech Tagging (POS): syntactic roles (noun, adverb...)
- Chunking: syntactic constituents (noun phrase, verb phrase...)
- Name Entity Recognition (NER): person/company/location...
- Semantic Role Labeling (SRL): semantic role

[John]$_{ARG0}$ [ate]$_{REL}$ [the apple]$_{ARG1}$ [in the garden]$_{ARGM-LOC}$

Labeled data: Wall Street Journal ($\sim 1M$ of words)

The Shallow System Way (1/2)

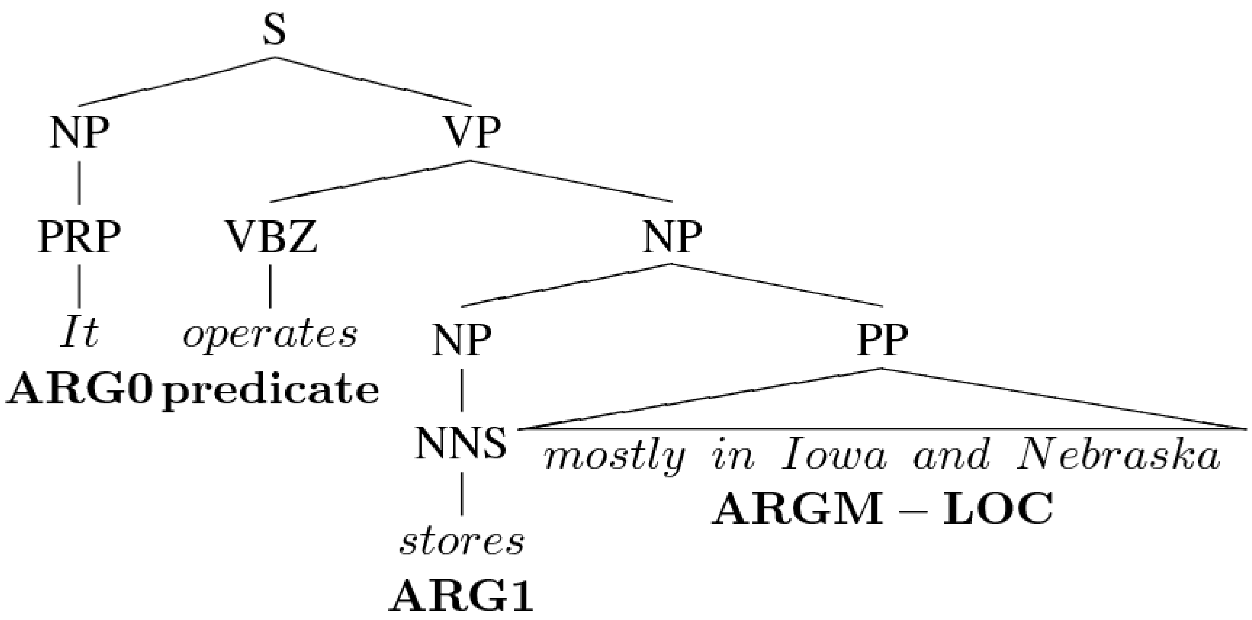
- Choose some good hand designed features

Predicate and POS tag of predicate	Voice: active or passive (hand-built rules)
Phrase type: adverbial phrase, prepositional phrase, . . .	Governing category: Parent node's phrase type(s)
Head word and POS tag of the head word	Position: left or right of verb
Path: traversal from predicate to constituent	Predicted named entity class
Word-sense disambiguation of the verb	Verb clustering
Length of the target constituent (number of words)	NEG feature: whether the verb chunk has a "not"
Partial Path: lowest common ancestor in path	Head word replacement in prepopositional phrases
First and last words and POS in constituents	Ordinal position from predicate + constituent type
Constituent tree distance	Temporal cue words (hand-built rules)
Dynamic class context: previous node labels	Constituent relative features: phrase type
Constituent relative features: head word	Constituent relative features: head word POS
Constituent relative features: siblings	Number of pirates existing in the world. . .

- Feed them to a shallow classifier like SVM

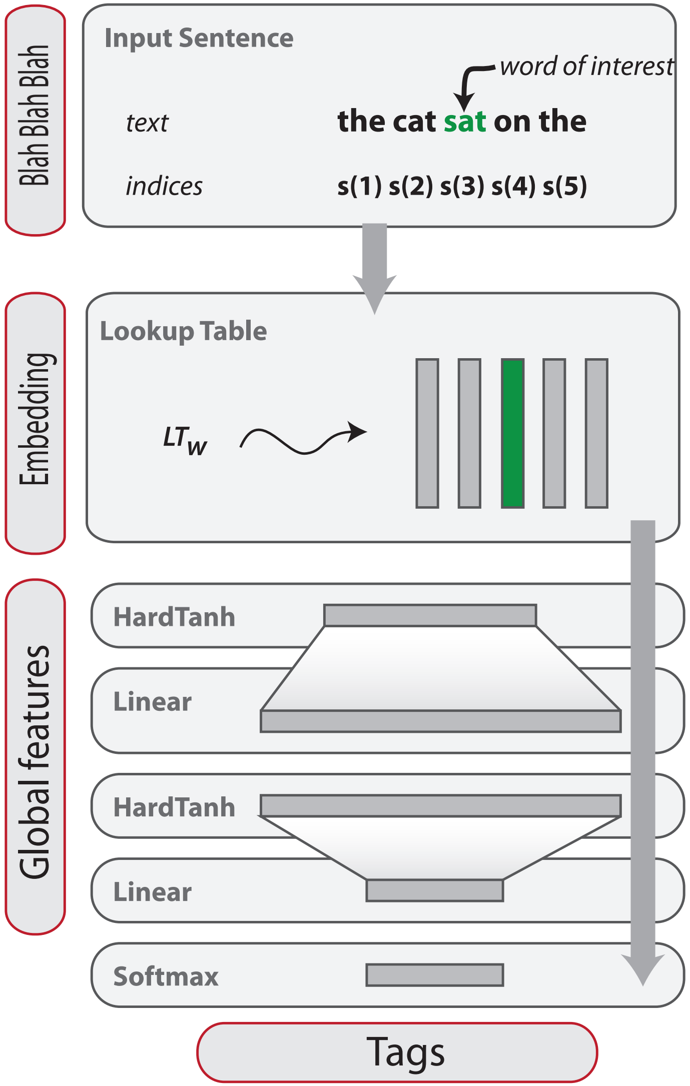
The Shallow System Way (2/2)

- Cascade features: e.g. extract POS, construct a parse tree

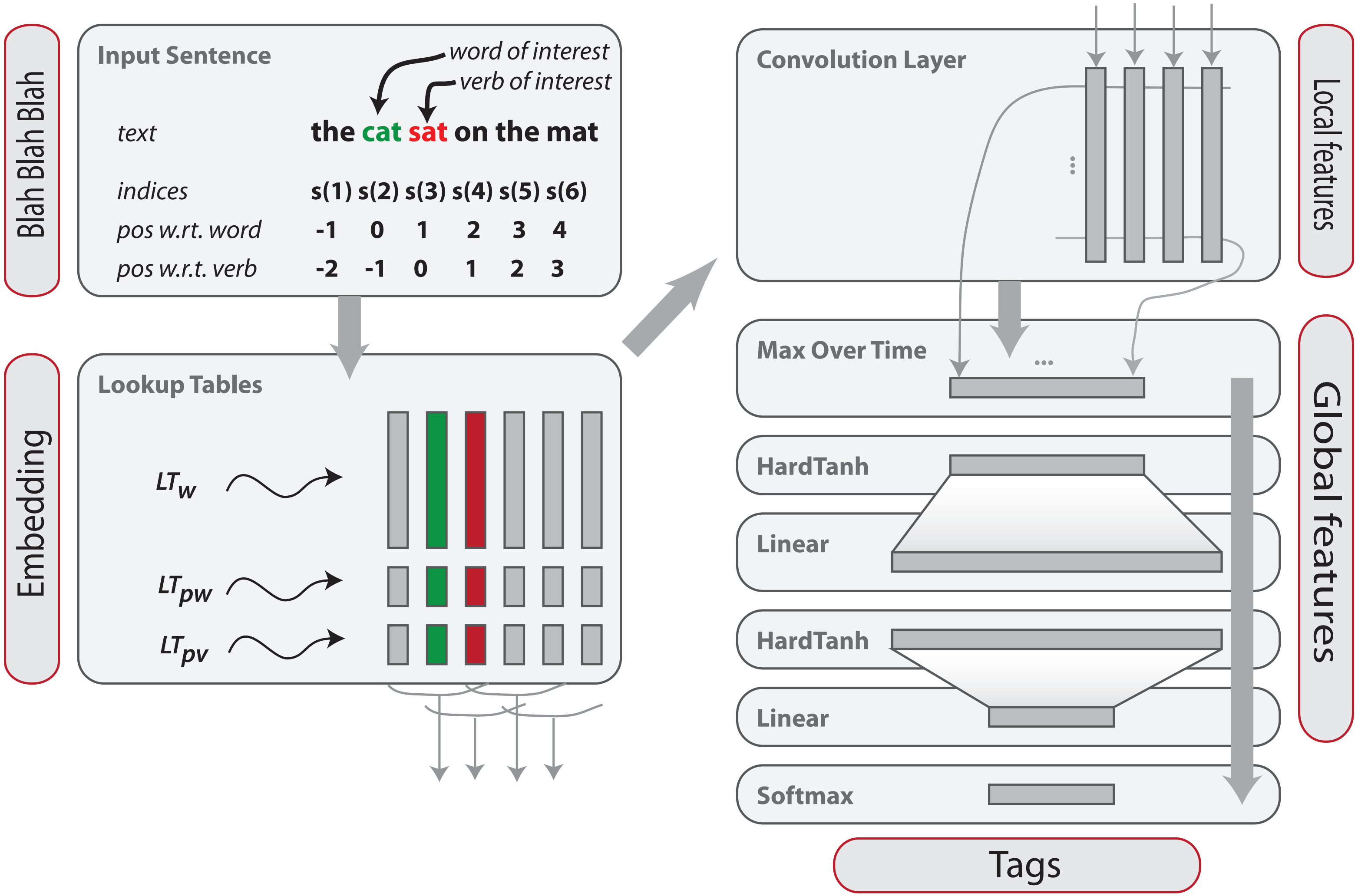


- Extract hand-made features from the parse tree
- Feed these features to a shallow classifier like SVM

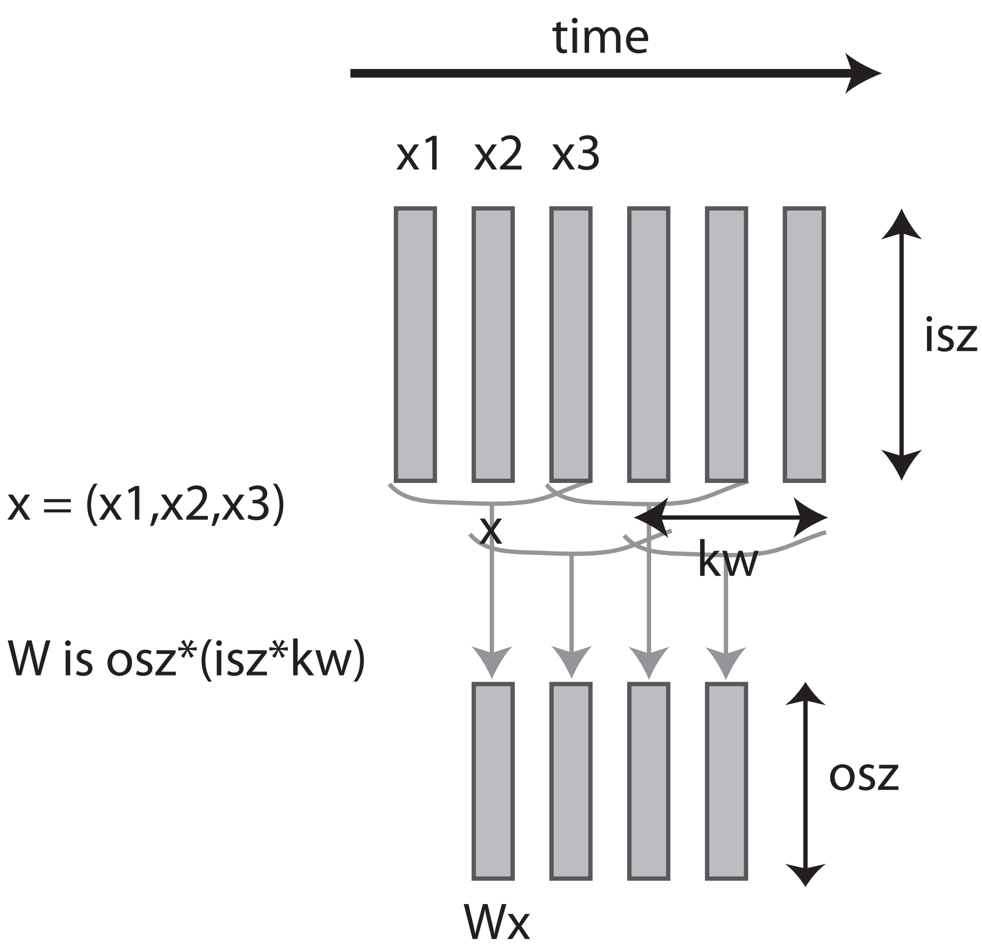
The Deep Learning Way (1/2)



The Deep Learning Way (2/2)



Convolutions



- Extract local features – share weights through time/space

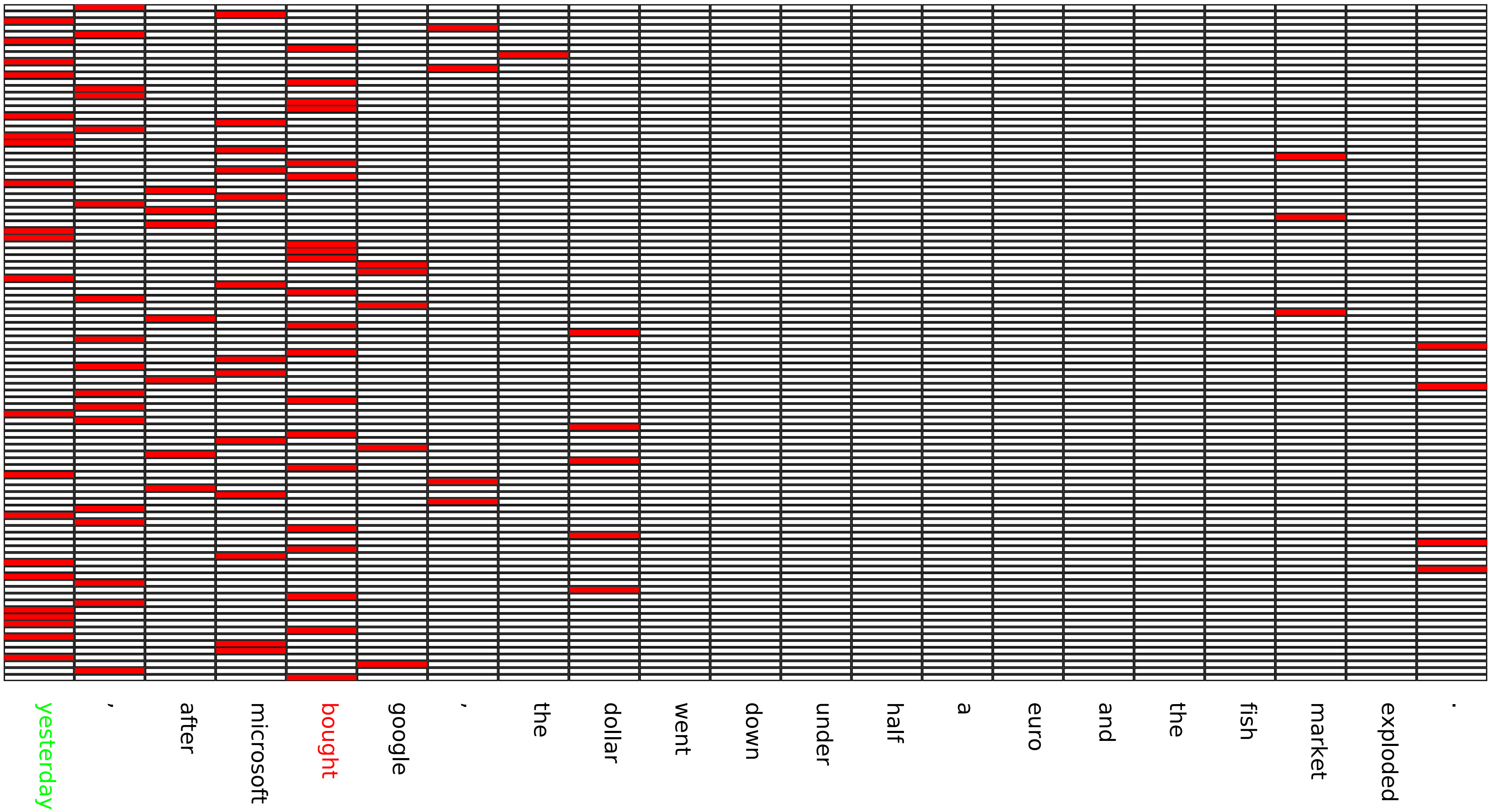
- Used with success in image (Le Cun, 1989) and speech (Bottou & Haffner, 1989)

- Lookup-table is a special case: convolution with kernel size of 1 and input i^{th} word

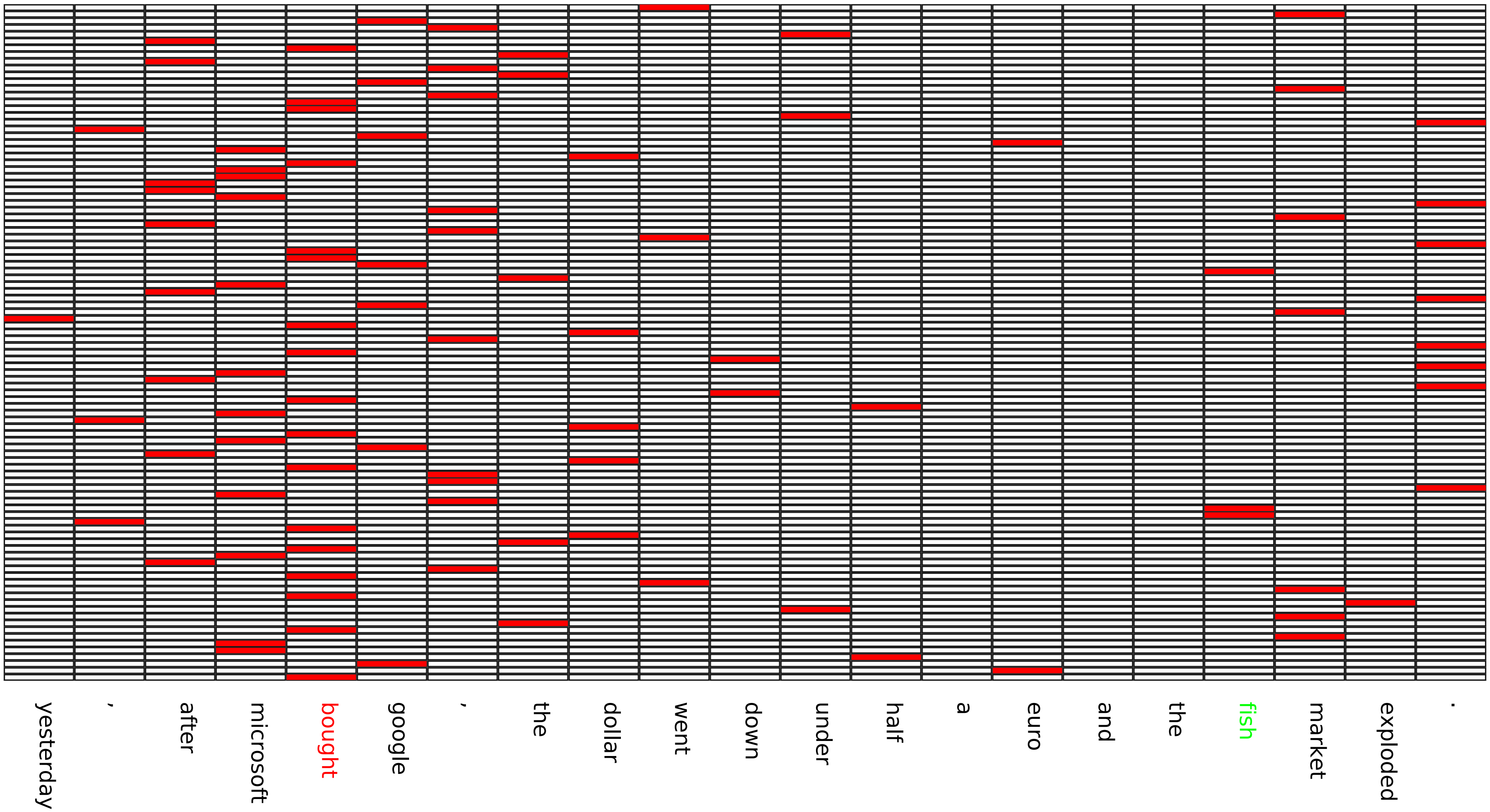
  $(0, 0, \ldots, 1, 0, \ldots, 0)$ 1 at position i

  Bengio et al (2001)

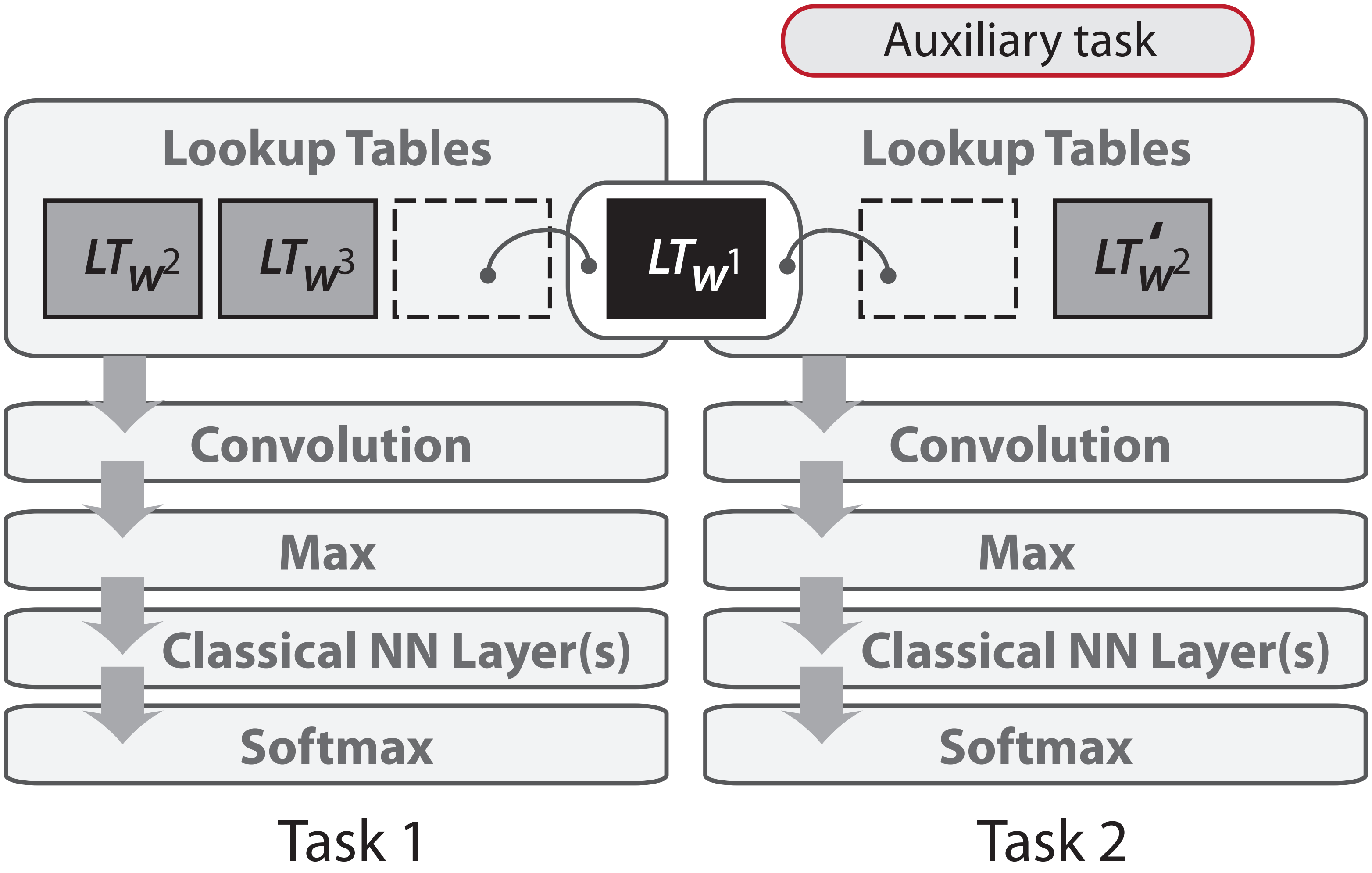
Removing The Time Dimension (1/2)



Removing The Time Dimension (2/2)



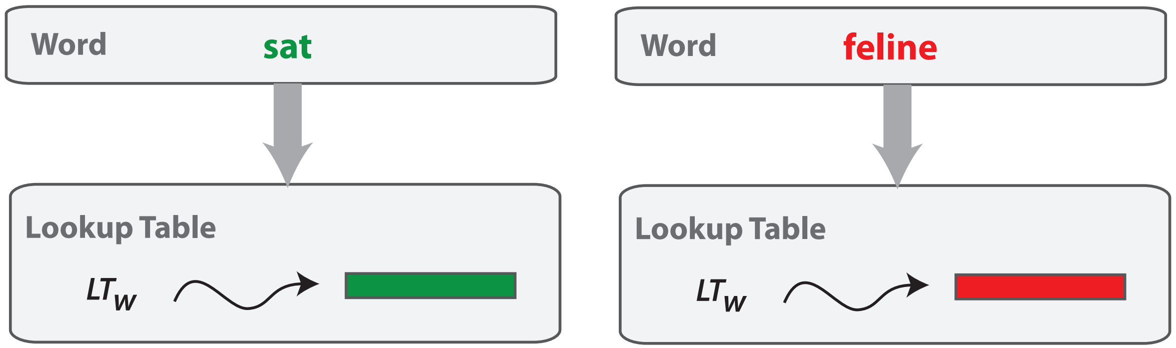
Multi-Task Learning



Good overview in Caruana (1997)

Improving Word Embedding

- Rare words are not trained properly
- Sentences with similar words should be tagged in the same way:
  - ⋆ The cat sat on the mat
  - ⋆ The feline sat on the mat



- Wordnet
  - ⋆ pull together linked words
  - ⋆ push apart other pair of words

Language Model: Think Massive

- Language Model: "*is a sentence actually english or not?*"
  Implicitly captures: $\star$ syntax $\star$ semantics

- Bengio & Ducharme (2001) Probability of next word given previous words. Overcomplicated – we do not need probabilities here

- English sentence windows: Wikipedia ($\sim 631M$ words)
  Non-english sentence windows: middle word randomly replaced

- Multi-class margin cost:

$$\sum_{s \in \mathcal{S}} \sum_{w \in \mathcal{D}} \max\left(0,\, 1 - f(s,\, w_s^\star) + f(s,\, w)\right)$$

$\mathcal{S}$: sentence windows $\mathcal{D}$: dictionary

$w_s^\star$: true middle word in s

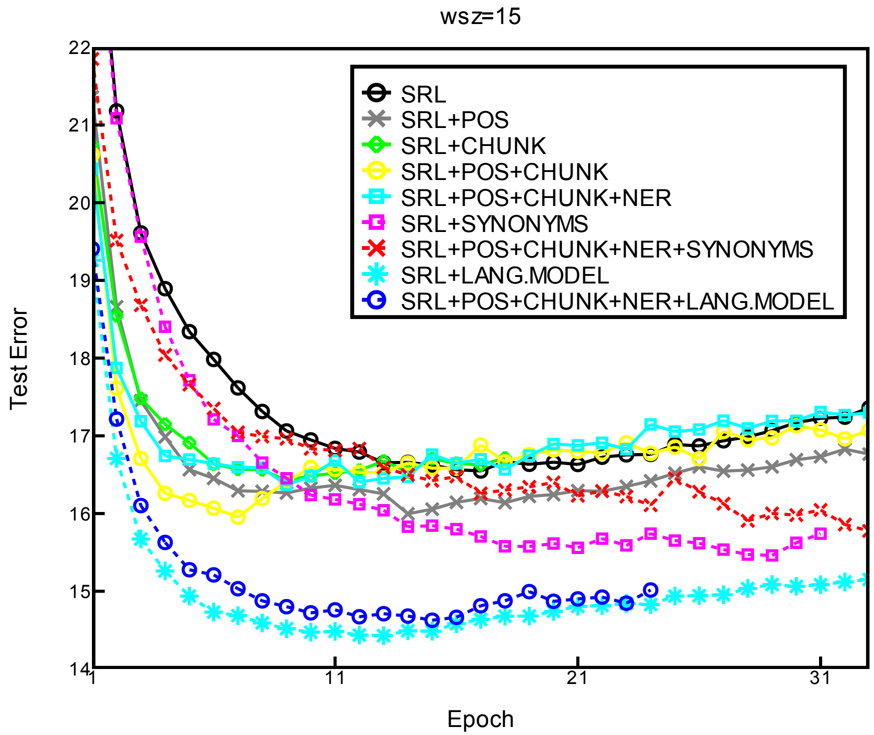
$f(s,\, w)$: network score for sentence s and middle word w

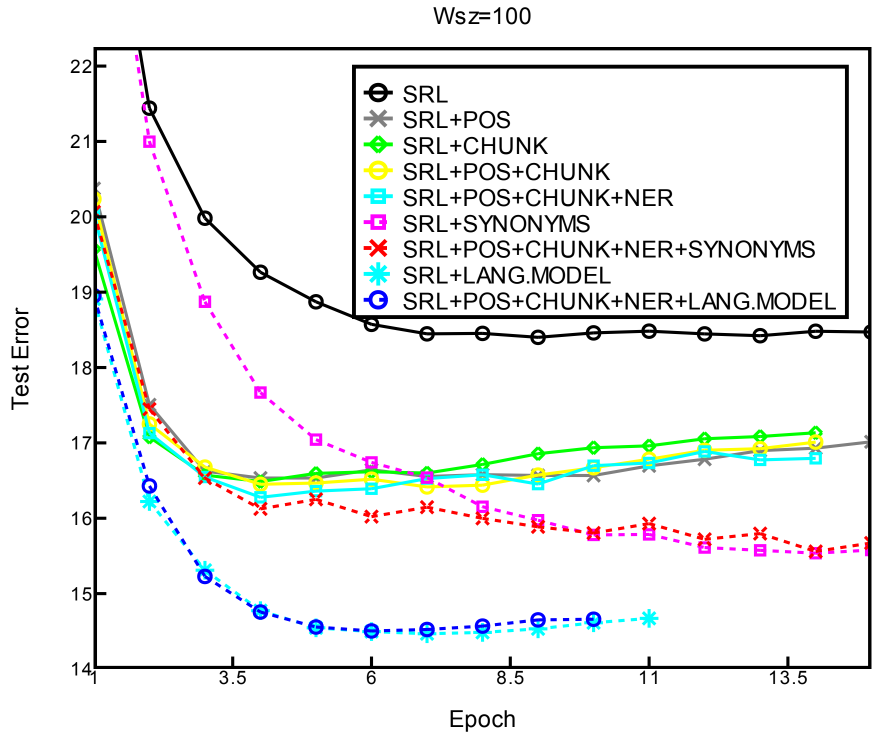
Language Model: Embedding

france	jesus	xbox	reddish	scratched
454	1973	6909	11724	29869
spain	christ	playstation	yellowish	smashed
italy	god	dreamcast	greenish	ripped
russia	resurrection	psNUMBER	brownish	brushed
poland	prayer	snes	bluish	hurled
england	yahweh	wii	creamy	grabbed
denmark	josephus	nes	whitish	tossed
germany	moses	nintendo	blackish	squeezed
portugal	sin	gamecube	silvery	blasted
sweden	heaven	psp	greyish	tangled
austria	salvation	amiga	paler	slashed

Dictionary size: 30,000 words. Even rare words are well embedded.

MTL: Semantic Role Labeling





- We get: 14.30%. State-of-the-art: 16.54% – Pradhan et al. (2004)
- $250\times$ faster than state-of-the-art. $\sim 0.01s$ to label a WSJ sentence.

MTL: Unified Network for NLP

Improved results with Multi-Task Learning (MTL)

Task	Alone	MTL
SRL	18.40%	14.30%
POS	2.95%	2.91%
Chunking – error rate	5.4%	4.9%
Chunking – F1-score	91.5%	93.6%

- POS: state-of-the-art $\sim 3\%$
- Chunking: Best system had 93.48% F1-score at CoNLL-2000 challenge `http://www.cnts.ua.ac.be/conll2000/chunking`. State-of-the art is 94.1%. We get 94.9% by using POS features.

Summary

- We developed a deep neural network architecture for NLP
- Advantages
  - ⋆ General to any NLP tagging task
  - ⋆ State-of-the-art performance
  - ⋆ No hand designed features
  - ⋆ Joint training
  - ⋆ Can exploit massive unlabeled data
  - ⋆ Extremely fast: 0.02s for all tags of a sentence
- Inconvenients
  - ⋆ Neural networks are a powerful tool: hard to handle
- Early Impacts
  - ⋆ Easy to apply to other tasks or languages: extending to Japanese
  - ⋆ Fast: developed a semantic search system