Natural Language Processing with Deep Learning

CS224N/Ling284

Christopher Manning and Richard Socher

Lecture 1: Introduction

Lecture Plan

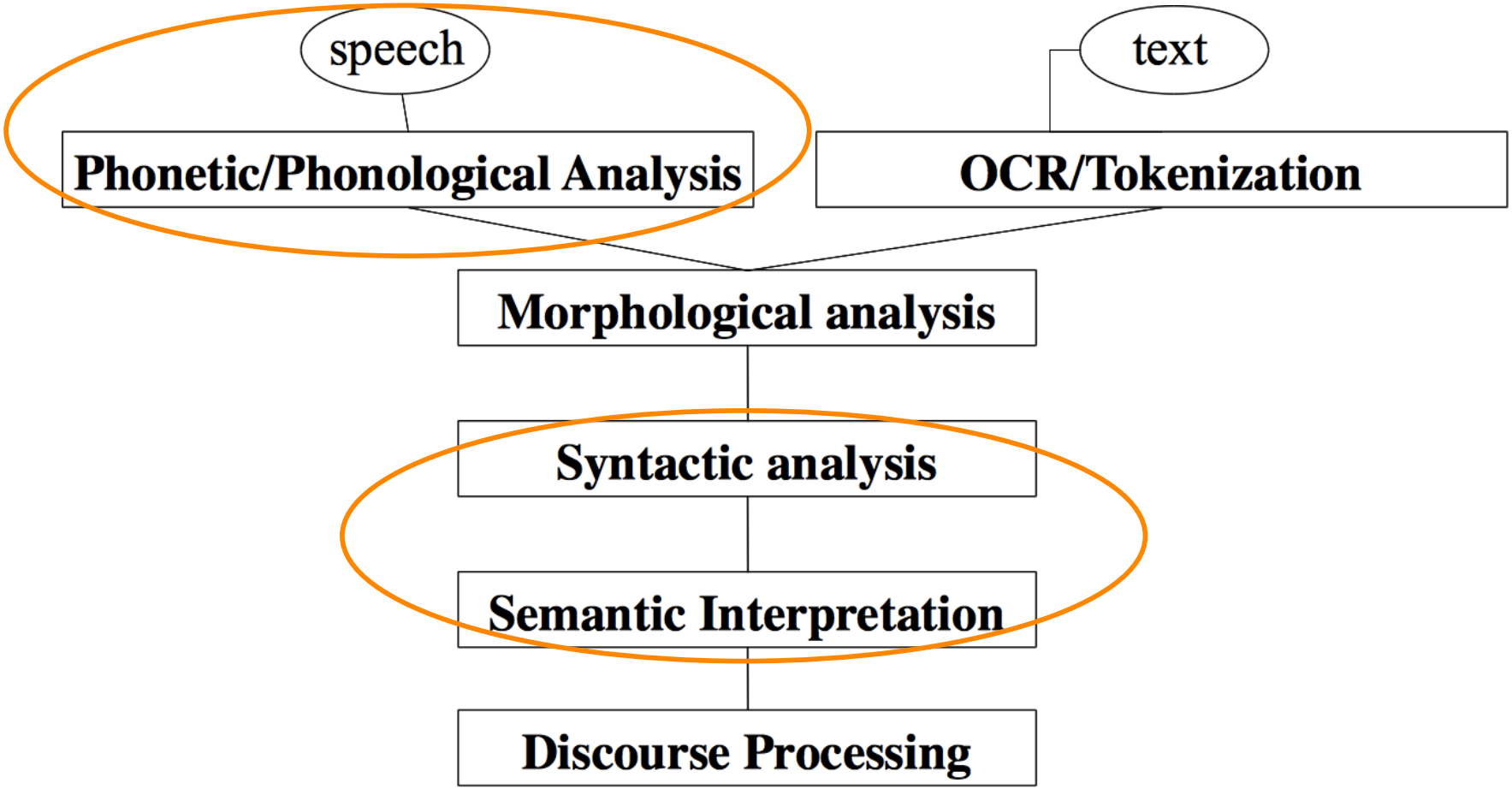
1. What is Natural Language Processing? The nature of human language (15 mins)
2. What is Deep Learning? (15 mins)
3. Course logistics (10 mins)
4. Why is language understanding difficult (10 mins)
5. Intro to the application of Deep Learning to NLP (25 mins)

Emergency time reserves: 5 mins

1. What is Natural Language Processing (NLP)?

- **Natural language processing** is a field at the intersection of
  - computer science
  - artificial intelligence
  - and linguistics.
- **Goal:** for computers to process or "understand" natural language in order to perform tasks that are useful, e.g.,
  - Performing Tasks, like making appointments, buying things
  - Question Answering
    - Siri, Google Assistant, Facebook M, Cortana … thank you, mobile!!!
- Fully **understanding and representing** the **meaning** of language (or even defining it) is a difficult goal.
  - Perfect language understanding is AI-complete

NLP Levels


speech
text
Phonetic/Phonological Analysis
OCR/Tokenization
Morphological analysis
Syntactic analysis
Semantic Interpretation
Discourse Processing

(A tiny sample of) NLP Applications

Applications range from simple to complex:

- Spell checking, keyword search, finding synonyms

- Extracting information from websites such as
  - product price, dates, location, people or company names
- Classifying: reading level of school texts, positive/negative sentiment of longer documents

- Machine translation
- Spoken dialog systems
- Complex question answering

NLP in industry … is taking off

- Search (written and spoken)
- Online advertisement matching
- Automated/assisted translation
- Sentiment analysis for marketing or finance/trading
- Speech recognition
- Chatbots / Dialog agents
  - Automating customer support
  - Controlling devices
  - Ordering goods

What's special about human language?

A human language is a system **specifically constructed to convey the speaker/writer's meaning**

- Not just an environmental signal, it's a deliberate communication
- Using an encoding which little kids can quickly learn (**amazingly!**)

A human language is a **discrete/symbolic/categorical signaling system**

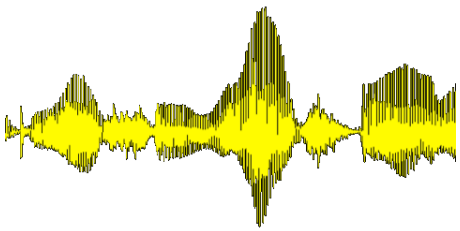
- rocket = 🚀; violin = 🎻
- With very minor exceptions for expressive signaling ("I loooove it." "Whoomppaaa")
- Presumably because of greater signaling reliability
- Symbols are not just an invention of logic / classical AI!

What's special about human language?

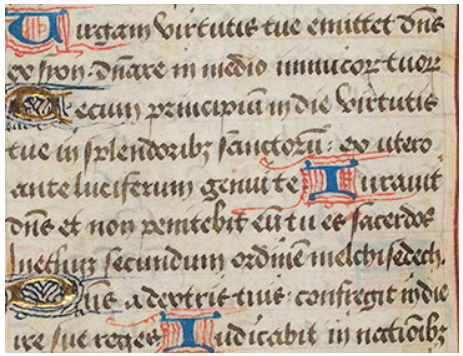
The categorical symbols of a language can be encoded as a signal for communication in several ways:

- Sound
- Gesture
- Images (writing)

The symbol is invariant across different encodings!



CC BY 2.0 David Fulmer 2008



National Library of NZ, no known restrictions

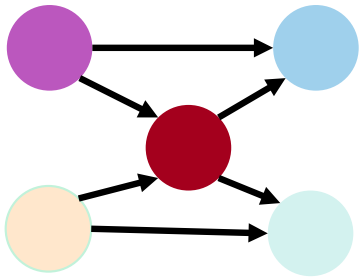
What's special about human language?

A human language is a **symbolic/categorical signaling system**

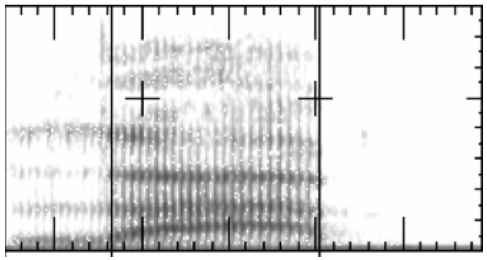
However, a brain encoding appears to be a **continuous pattern of activation**, and the symbols are transmitted via **continuous signals** of sound/vision

We will explore a continuous encoding pattern of thought

The large vocabulary, symbolic encoding of words creates a problem for machine learning – sparsity!



lab



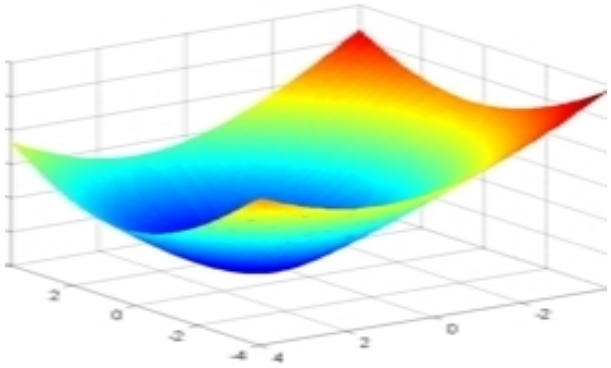
2. What's Deep Learning (DL)?

- **Deep learning** is a subfield of **machine learning**

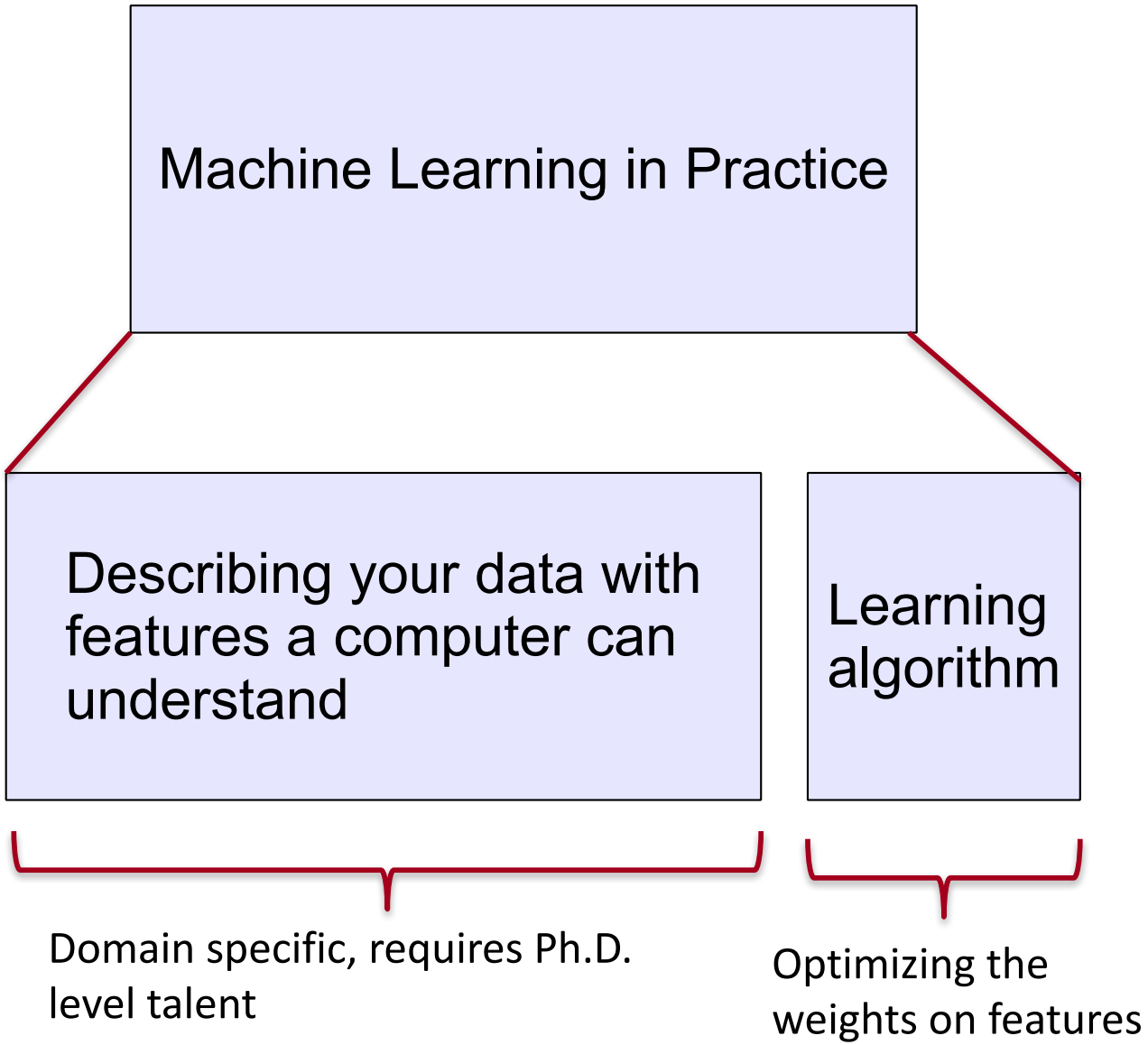
- Most machine learning methods work well because of **human-designed representations** and **input features**
  - For example: features for finding named entities like locations or organization names (Finkel et al., 2010):

Feature	NER
Current Word	✓
Previous Word	✓
Next Word	✓
Current Word Character n-gram	all
Current POS Tag	✓
Surrounding POS Tag Sequence	✓
Current Word Shape	✓
Surrounding Word Shape Sequence	✓
Presence of Word in Left Window	size 4
Presence of Word in Right Window	size 4

- Machine learning becomes just optimizing weights to best make a final prediction



Machine Learning vs. Deep Learning

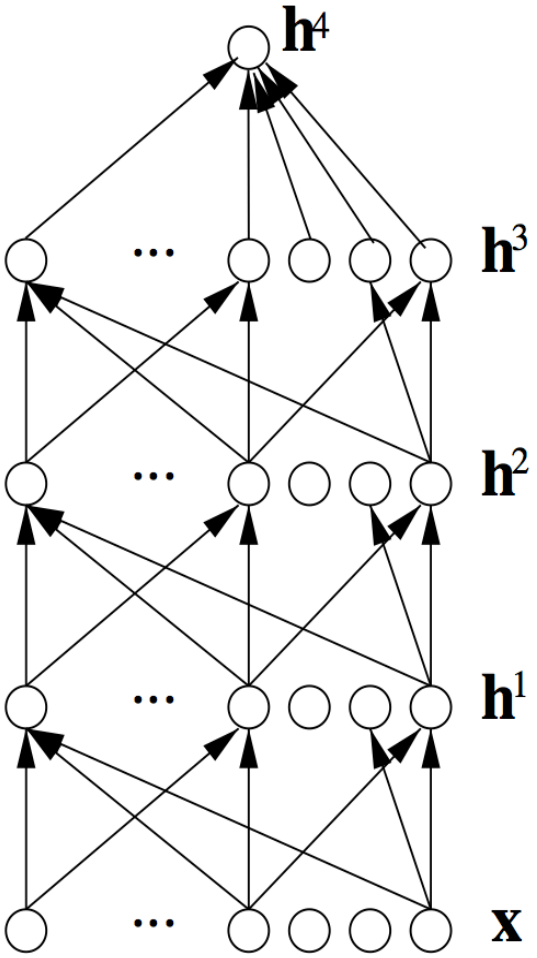


What's Deep Learning (DL)?

- Representation learning attempts to automatically learn good features or representations

- Deep learning algorithms attempt to learn (multiple levels of) representation and an output

- From "raw" inputs **x** (e.g., sound, characters, or words)



On the history of and term "Deep Learning"

- We will focus on different kinds of **neural networks**
- The dominant model family inside deep learning

- Only clever terminology for stacked logistic regression units?
  - Maybe, but interesting modeling principles (end-to-end) and actual connections to neuroscience in some cases

- We will not take a historical approach but instead focus on methods which work well on NLP problems now
- For a long (!) history of deep learning models (starting ~1960s), see: Deep Learning in Neural Networks: An Overview by Jürgen Schmidhuber

Reasons for Exploring Deep Learning

- Manually designed features are often over-specified, incomplete and take a long time to design and validate

- **Learned Features** are easy to adapt, fast to learn

- Deep learning provides a very flexible, (almost?) universal, learnable framework for **representing** world, visual and linguistic information.

- Deep learning can learn **unsupervised** (from raw text) and **supervised** (with specific labels like positive/negative)

Reasons for Exploring Deep Learning

- In ~2010 **deep** learning techniques started outperforming other machine learning techniques. Why this decade?

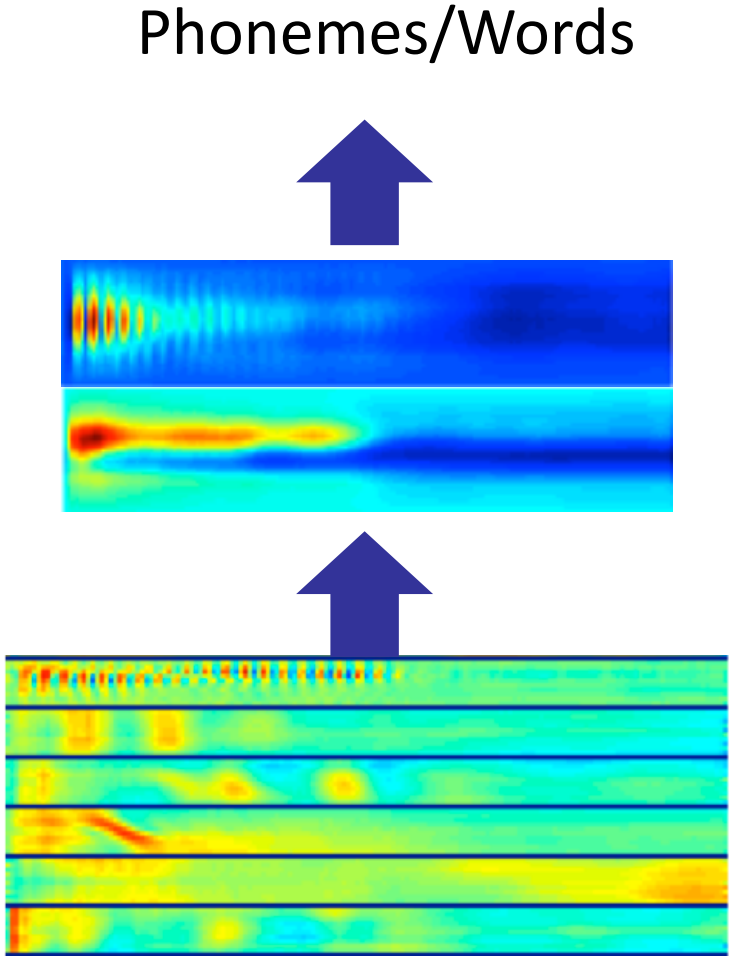
- Large amounts of training data favor deep learning
- Faster machines and multicore CPU/GPUs favor Deep Learning
- New models, algorithms, ideas
  - Better, more flexible learning of intermediate representations
  - Effective end-to-end joint system learning
  - Effective learning methods for using contexts and transferring between tasks

→ **Improved performance** (first in speech and vision, then NLP)

Deep Learning for Speech

- The first breakthrough results of "deep learning" on large datasets happened in speech recognition
- Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition Dahl et al. (2010)

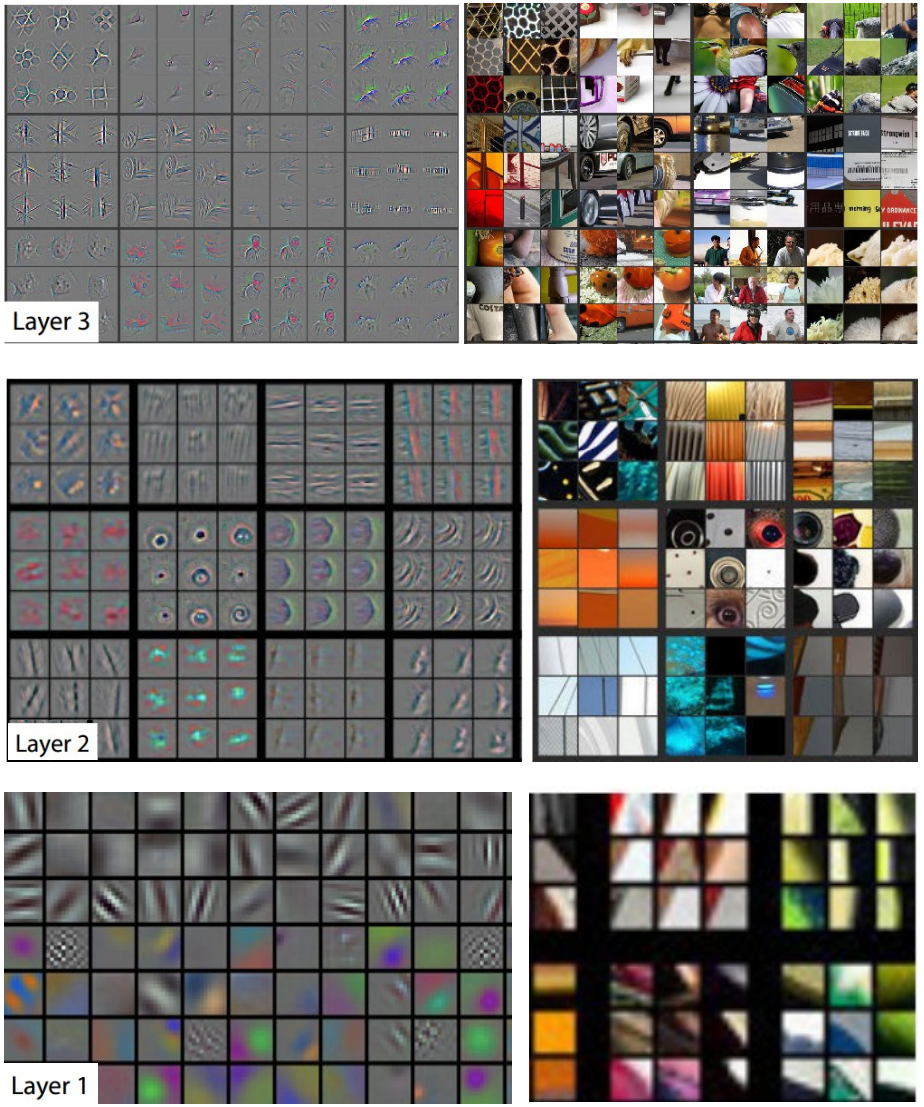
Acoustic model	Recog WER	RT03S FSH	Hub5 SWB
Traditional features	1-pass –adapt	**27.4**	**23.6**
Deep Learning	1-pass –adapt	**18.5** (–33%)	**16.1** (–32%)



Deep Learning for Computer Vision

Most deep learning groups have focused on computer vision (at least till 2 years ago)

The breakthrough DL paper: ImageNet Classification with Deep Convolutional Neural Networks by Krizhevsky, Sutskever, & Hinton, 2012, U. Toronto. 37% error red.



Zeiler and Fergus (2013)

3. Course logistics in brief

- Instructors: Christopher Manning & Richard Socher
- TAs: Many wonderful people!
- Time: TuTh 4:30–5:50, Nvidia Aud
- Apologies about the room capacity! (Success catastrophe!)

- Other information: see the class webpage
  - http://cs224n.stanford.edu/
    a.k.a., http://www.stanford.edu/class/cs224n/
  - Syllabus, office hours, "handouts", TAs, Piazza
  - Slides uploaded before each lecture

Prerequisites

- Proficiency in Python
  - All class assignments will be in Python. (See tutorial on cs224n WWW)
- Multivariate Calculus, Linear Algebra (e.g., MATH 51, CME 100)
- Basic Probability and Statistics (e.g. CS 109 or other stats course)
- Fundamentals of Machine Learning (e.g., from CS229 or CS221)
  - loss functions,
  - taking simple derivatives
  - performing optimization with gradient descent.

What do we hope to teach?

1. An understanding of and ability to use the effective modern methods for deep learning
   - Covering all the basics, but thereafter with a bias to the key methods used in NLP: Recurrent networks, attention, etc.
2. Some big picture understanding of human languages and the difficulties in understanding and producing them
3. An understanding of and ability to build systems for some of the major problems in NLP:
   - Word similarities, parsing, machine translation, entity recognition, question answering, sentence comprehension

Grading Policy

- 3 Assignments: 17% x 3 = 51%
- Midterm Exam: 17%
- Final Course Project or Assignment 4 (1–3 people): 30%
  - Including for final project doing: project proposal, milestone, interacting with **mentor**
- Final poster session (**must** be there: Mar 21: 12:15–3:15)**:** 2%
- Late policy
  - 5 free late days – use as you please
  - Afterwards, 10% off per day late
  - Assignments not accepted after 3 late days per assignment
- Collaboration policy: Read the website and the Honor Code! Understand allowed ‘collaboration’ and how to document it

High Level Plan for Problem Sets

- The first half of the course and Ass 1 & 2 will be hard

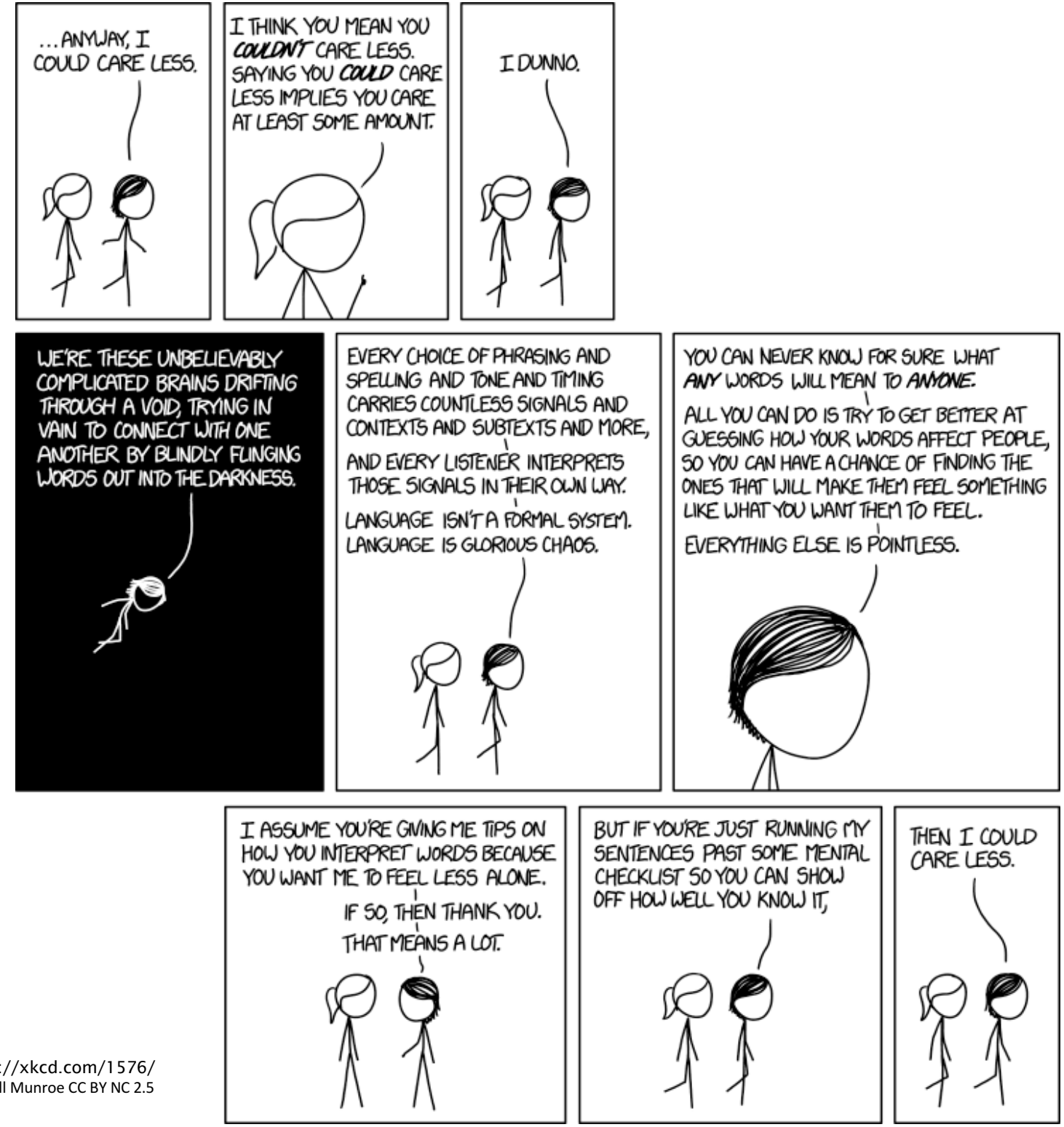
- Ass 1 is written work and pure python code (numpy etc.) to really understand the basics
- Released on January 12 (this Thursday!)

- Ass 2 & 3 will be in TensorFlow, a library for putting together neural network models quickly (→ special lecture)
- Libraries like TensorFlow are becoming standard tools
  - Also: Theano, Torch, Chainer, CNTK, Paddle, MXNet, Keras, Caffe, …

- You choose an exciting final project or we give you one (Ass 4)
- Can use any language and/or deep learning framework

4. Why is NLP hard?

- Complexity in representing, learning and using linguistic/situational/world/visual knowledge
- Human languages are ambiguous (unlike programming and other formal languages)
- Human language interpretation depends on real world, common sense, and contextual knowledge



https://xkcd.com/1576/

Why NLP is difficult: Real newspaper headlines/tweets

1. The Pope's baby steps on gays
2. Boy paralyzed after tumor fights back to gain black belt
3. Scientists study whales from space
4. Juvenile Court to Try Shooting Defendant

5. Deep NLP = Deep Learning + NLP

Combine ideas and goals of NLP with using representation learning and deep learning methods to solve them

Several big improvements in recent years in NLP with different

- **Levels**: speech, words, syntax, semantics
- **Tools:** parts-of-speech, entities, parsing
- **Applications**: machine translation, sentiment analysis, dialogue agents, question answering

Word meaning as a neural word vector – visualization

expect =

$$\begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \\ 0.487 \end{pmatrix}$$

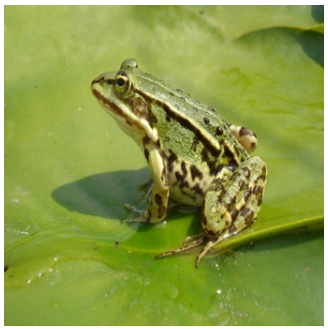
Word similarities

Nearest words to frog:

1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

litoria

leptodactylidae



rana

eleutherodactylus

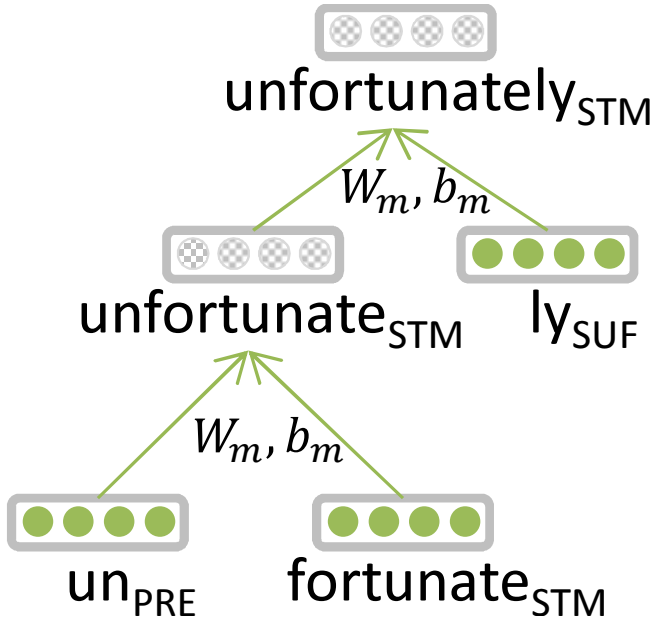
http://nlp.stanford.edu/projects/glove/

Representations of NLP Levels: Morphology

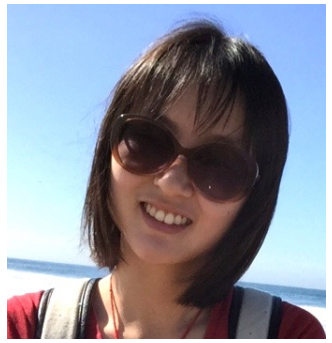
- Traditional: Words are made of morphemes

prefix	stem	suffix
un	interest	ed

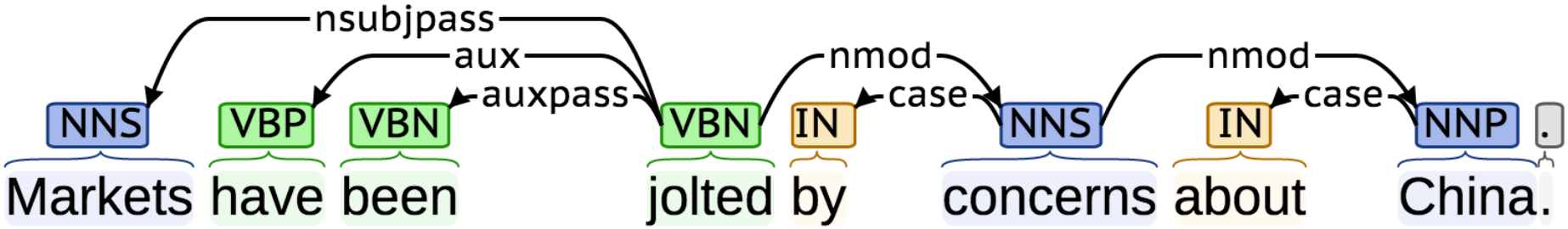
- DL:
  - every morpheme is a vector
  - a neural network combines two vectors into one vector
  - Luong et al. 2013



NLP Tools: Parsing for sentence structure



Neural networks can accurately determine the structure of sentences, supporting interpretation



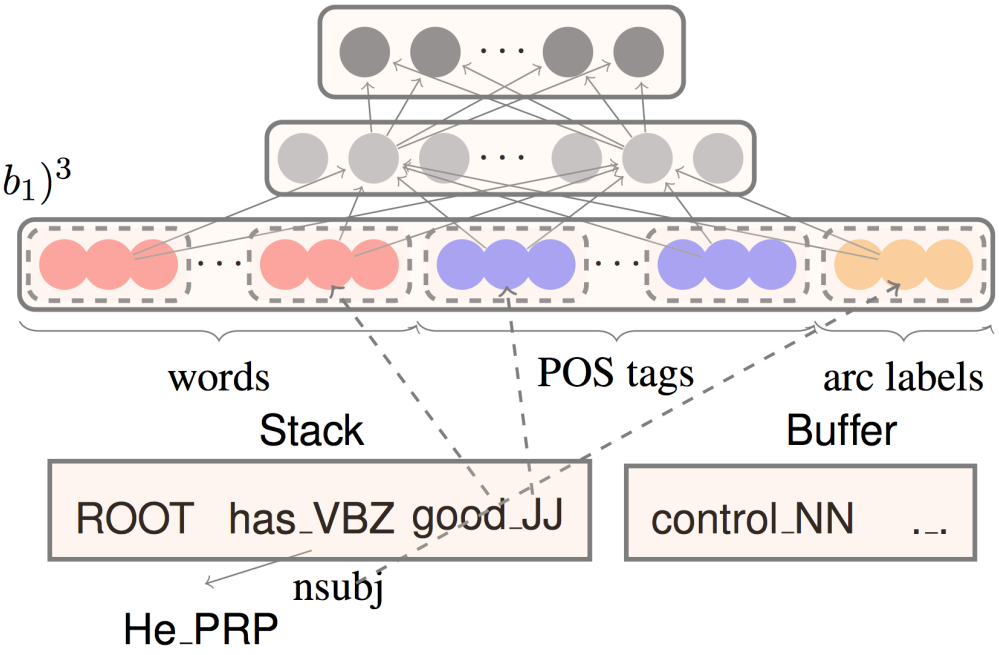
Softmax layer:

$$p = \texttt{softmax}(W_2 h)$$

Hidden layer:

$$h = (W_1^w x^w + W_1^t x^t + W_1^l x^l + b_1)^3$$

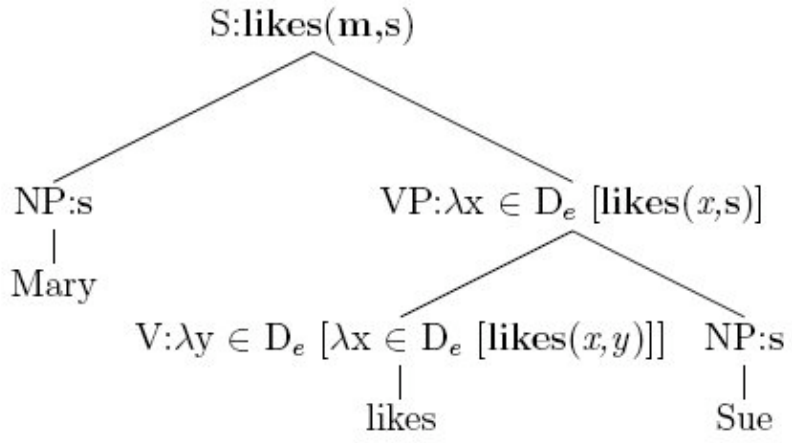
Input layer: $[x^w, x^t, x^l]$

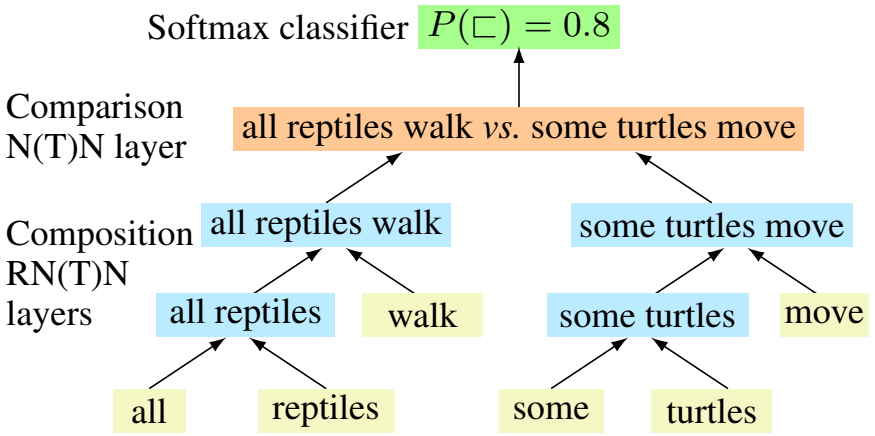


Configuration

Representations of NLP Levels: Semantics

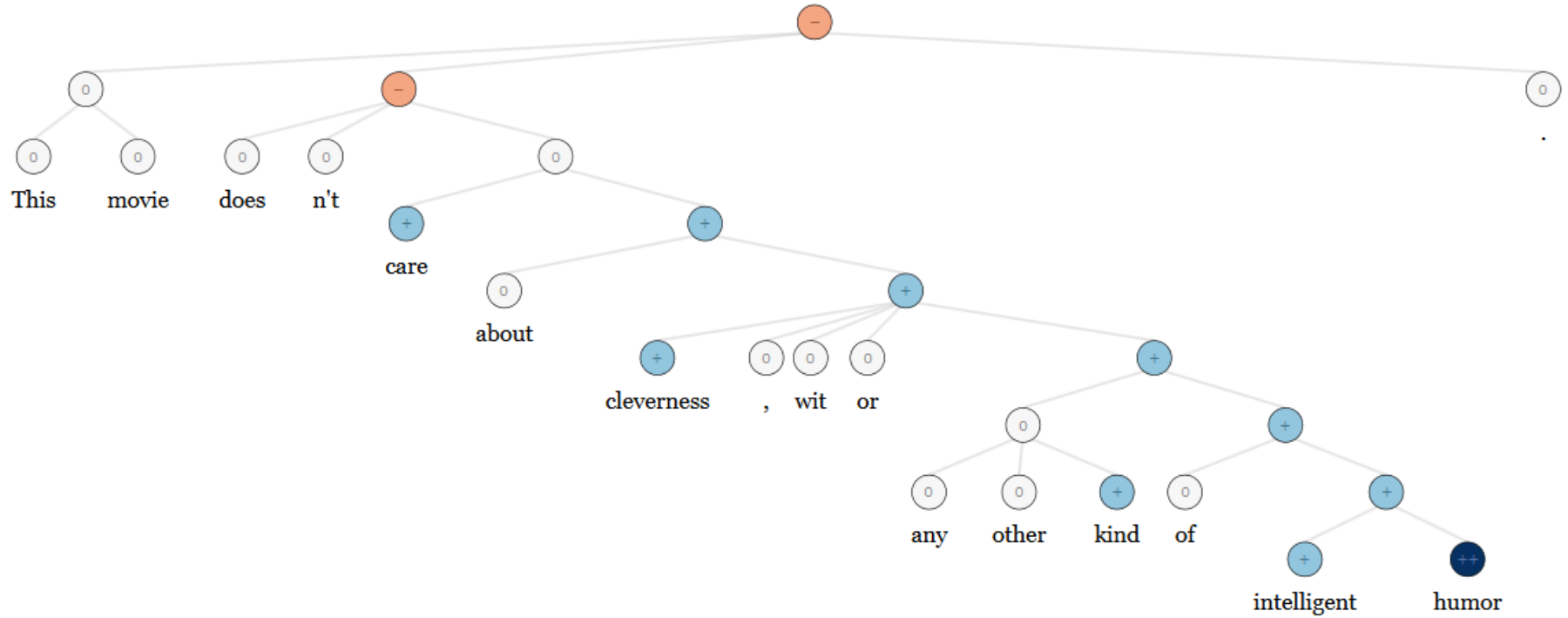
- Traditional: Lambda calculus
  - Carefully engineered functions
  - Take as inputs specific other functions
  - No notion of similarity or fuzziness of language
- DL:
  - Every word and every phrase and every logical expression is a vector
  - a neural network combines two vectors into one vector
  - Bowman et al. 2014





NLP Applications: Sentiment Analysis

- Traditional: Curated sentiment dictionaries combined with either bag-of-words representations (ignoring word order) or hand-designed negation features (ain't gonna capture everything)
- Same deep learning model that was used for morphology, syntax and logical semantics can be used! → RecursiveNN



Question Answering

- Traditional: A lot of feature engineering to capture world and other knowledge, e.g., regular expressions, Berant et al. (2014)

Is main verb trigger?

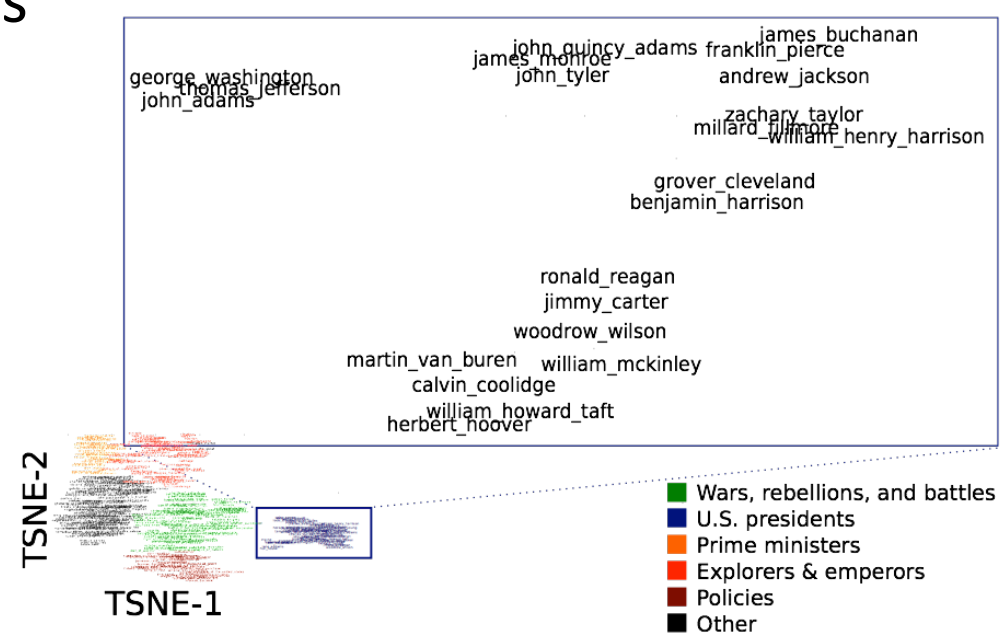
Yes

Condition	Regular Exp.
Wh- word subjective?	AGENT
Wh- word object?	THEME

No

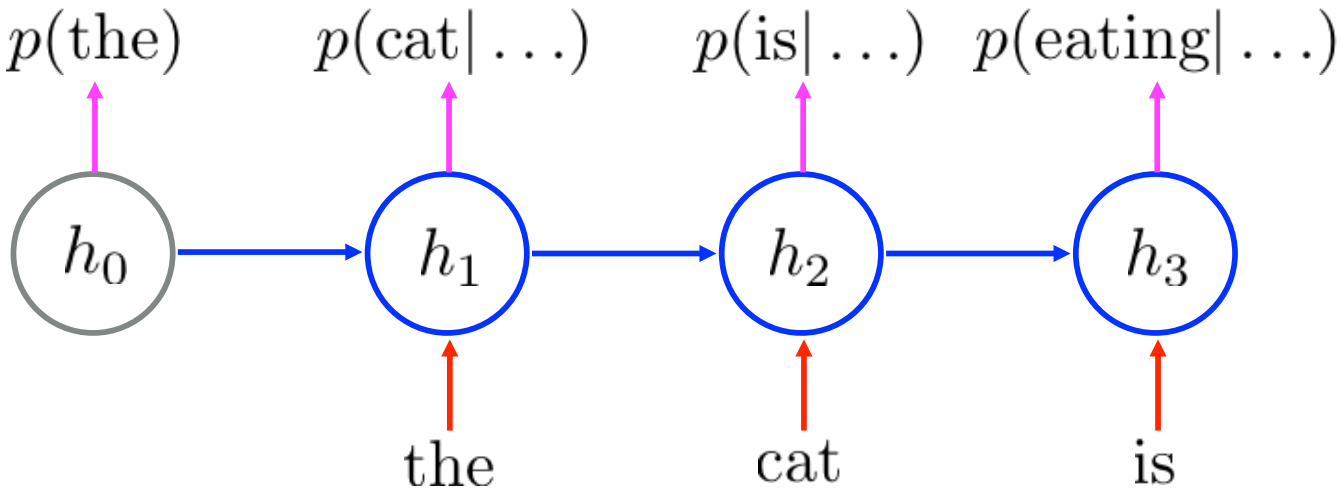
Condition	Regular Exp.
default	(ENABLE|SUPER)$^+$
DIRECT	(ENABLE|SUPER)
PREVENT	(ENABLE|SUPER)* PREVENT(ENABLE|SUPER)*

- DL: Again, a deep learning architecture can be used!
- Facts are stored in vectors



Dialogue agents / Response Generation

- A simple, successful example is the auto-replies available in the Google Inbox app
- An application of the powerful, general technique of **Neural Language Models**, which are an instance of Recurrent Neural Networks



Machine Translation

- Many levels of translation have been tried in the past:

- Traditional MT systems are very large complex systems

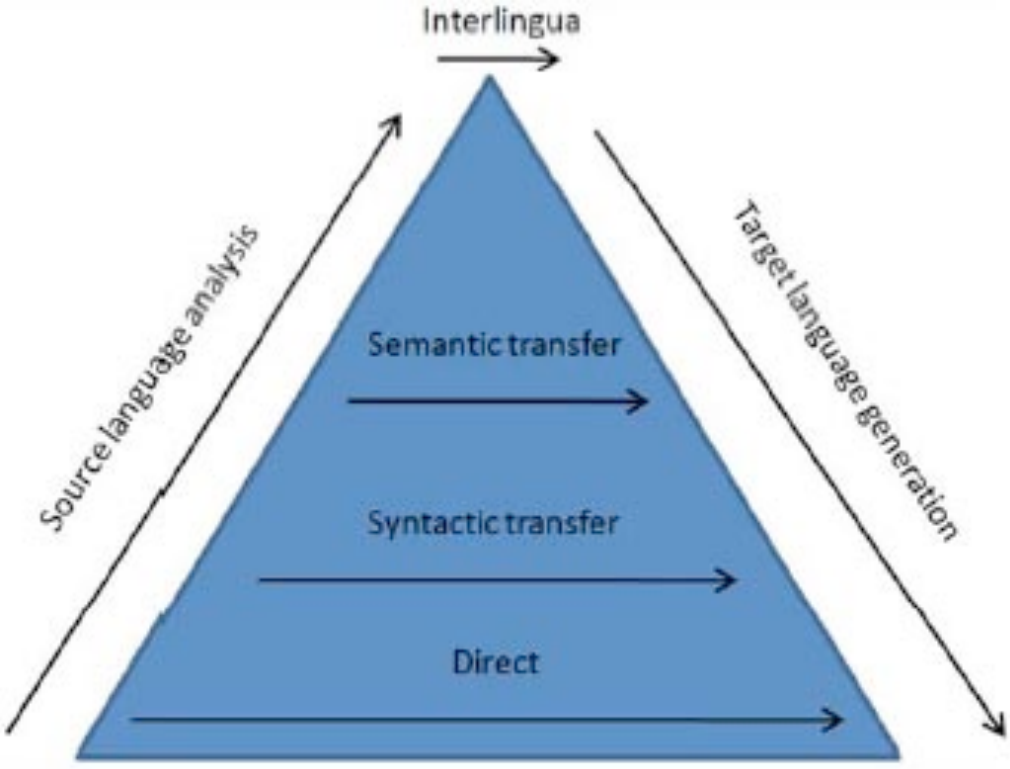
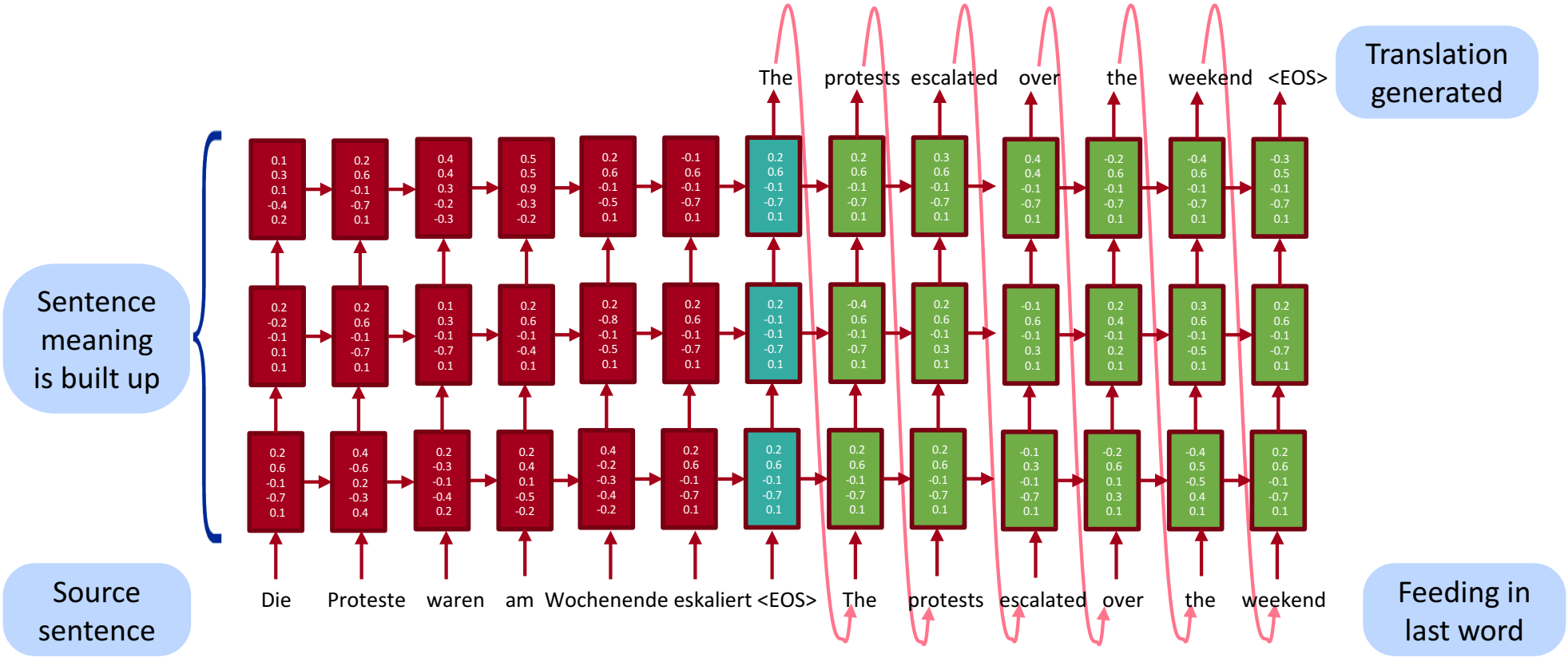


Figure 1: The Vauquois triangle

- What do you think is the interlingua for the DL approach to translation?

Neural Machine Translation

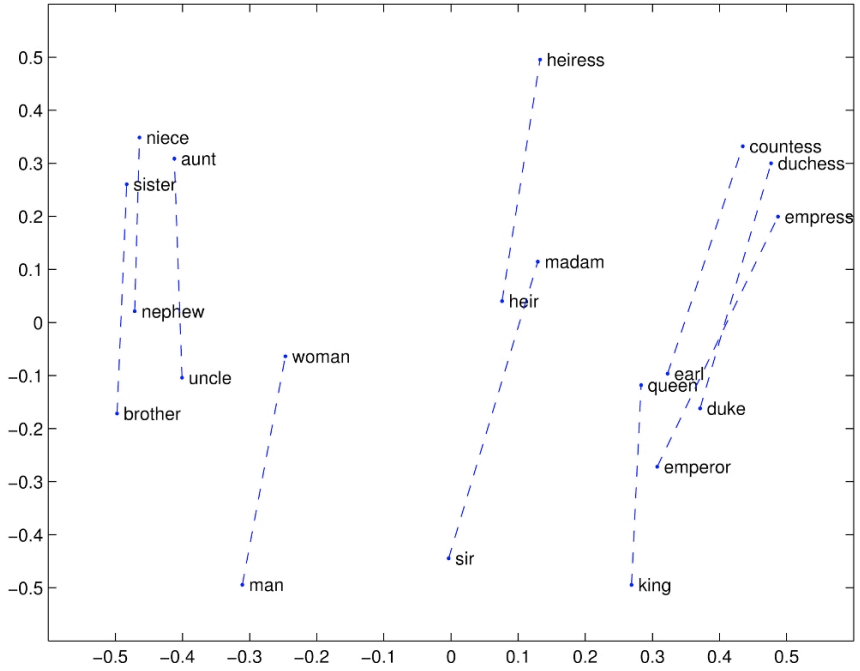
Source sentence is mapped to **vector**, then output sentence generated
[Sutskever et al. 2014, Bahdanau et al. 2014, Luong and Manning 2016]



Now live for some languages in Google Translate (etc.), with big error reductions!

Conclusion: Representation for all levels? Vectors

We will study in the next lecture how we can learn vector representations for words and what they actually **represent.**



Next week (**Richard**): how neural networks work and how they can use these vectors for all NLP levels and many different applications